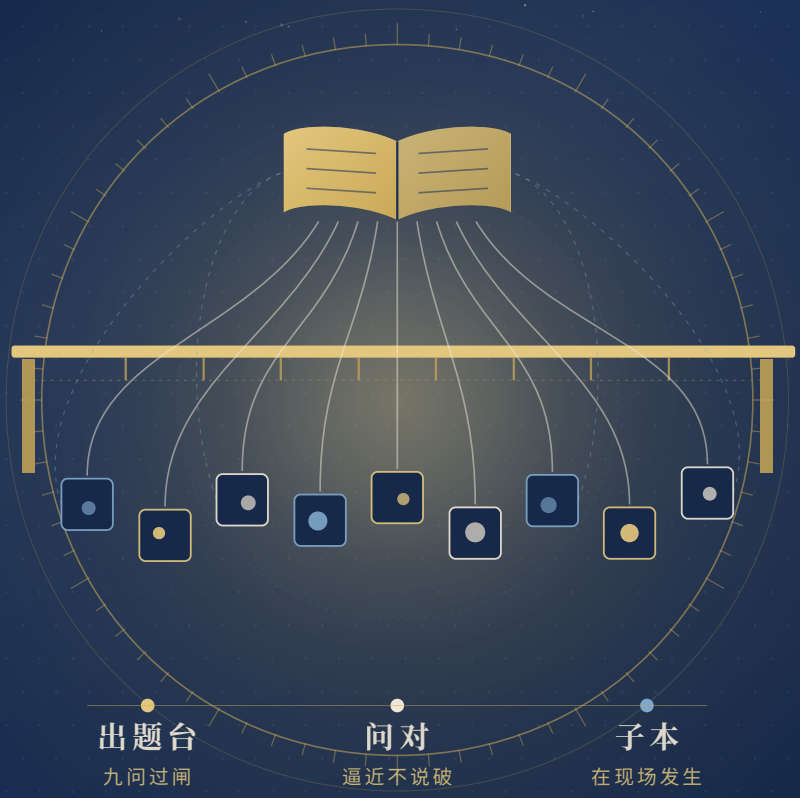
SDE UNIVERSES
德麦国际专著 · 第 528 号

出题台 · 问对 · 子本

未来大学的课堂模式典范革命

九问模式



王德生　著

德麦国际出版社
Demai International Press

德麦国际专著 · 第 528 号

未来大学的课堂模式典范革命

九问模式

王德生　著

德麦国际出版社
Demai International Press · Singapore

出　版　信　息

书名	《未来大学的课堂模式典范革命》
副题	九问模式
英文书名	The Paradigm Revolution of the Future University Classroom: The Nine-Question Mode
著者	王德生
出版	德麦国际出版社（Demai International Press）· 新加坡
编号	德麦国际专著第 528 号
ISBN	979-8-90690-528-4
定价	US$20.00
开本	170 mm × 240 mm
字数	约 28 万字
版次	2026 年 10 月第 1 版

AI 协作声明：本书的承重判断出自 WDS博士 2026 年 10 月 29 日的一次问对；文字在其指令下由 AI（Claude）协作写成：Claude 拟定提纲，十路 Claude 分章并行执笔，两路换手评审，十路修订，四路审稿、四路校稿、一路终审，由王德生审定。同一底座的多位执笔者、评审者与编辑不能互作独立证据，书中的判断须由真实课堂来检验。

书中的许宁、沈澄、周川、林晓、韩正、方舟、高远、苏文、叶青、吴静等人物与他们的课堂，都是为说明工序而写下的构造案例，不是调查结果。医学案例一律去标识化，学生不作临床决策；化学案例不涉及危险操作细节。

附录二报告的小样本试跑由同一家族的模型作答与评判，只作先验，不作独立证据；它推翻了本书初稿第 8 章的一个判断，正文已据此改写。

全书结构：导论；七编二十六章——母本与九问、改题与问对、ChatSDE 与通用大模型、打造子本与落壤、学业评估、学科九问样板、回写与大学的领地；结语、参考书目、附录两件与后记。

版权所有　侵权必究

子本建造，其实就是未来大学的教学核心目标。

—WDS博士口述，2026 年 10 月

作者介绍

王德生，德麦国际（Demai International）创办人，SDE本体论的提出者。SDE三个字母，说的是世界怎样发生：显露、差异序列、特征纠缠。他在中国农村长大，1990年考入湘潭大学，后在中国科学院获计算数学博士学位，并在那里做博士后研究；之后在斯旺西大学、南洋理工大学任职，2005年起定居新加坡。他的思想走过四站：三视角、321智慧系统、SIO本体论，到今天的SDE本体论。他的书由德麦国际出版社出版，多数写给没受过哲学训练、却天天在过日子的人。

目 录

作者介绍 iv

导论　课堂在交付什么 1

第一编　S：母本与九问（立骨） 21

第1章　练手与战场：发生技术前移到中学 22

第2章　教材退场之后：统一性知识移交大厂 35

第3章　谱系母本精要库 47

第4章　出题台：九问打向尚不存在的判断 60

第二编　D：改题与问对（走来路） 75

第5章　改题权：学生改题、提升题 76

第6章　问对ChatSDE：第零轮与多轮自由问对 88

第7章　一破一立：批判性激发与工序选择 102

第三编　ChatSDE与通用大模型：知识发生上的差别 117

第8章　两种默认：替学生说破与逼近不说破 118

第9章　信息发生与知识发生：差别落在哪一环 134

第10章　工序层：模型是发动机，工序是驾驶术 147

第11章　对照实验：差别怎样被检验，怎样算错 160

第四编　E：打造子本与落壤（铺土壤） 175

第12章　子本打造：重写不是翻译 176

第13章　子本进入底盘：RAG、Skill与子本系列 188

第14章　同学碰撞场：班级子本的涌现 200

第五编　学业评估：多项目与权重组合 215

第15章　学业成绩的结构：八项目与发生剖面 216

第16章　核心项：子本创新智商与完成度 229

第17章　过程项：改题、问对与现场证据的计分 242

第18章　权重组合：学段、课型、学科与自选 254

第六编　学科九问样板 267

第19章　医学：病房里的九问 268

第20章　化学：实验室里的九问　282
第21章　物理、工程与计算：装置与系统里的九问　297
第22章　经济与法学：交易与裁判里的九问　310
第23章　历史与文学：史料与语言里的九问　325

第七编　回写与大学的领地　**339**
第24章　回写：档案回到谱系母本，课程成为谱系　340
第25章　过渡期与一个学期的样子　352
第26章　大学造驾驶术：工序智能体与护城河　365

结语　另外五条路　377

参考书目　388

附录　393
附录一　九问课堂工具表　394
附录二　一次小样本试跑　399

后记　405

导论

课堂在交付什么

◆ 同一个学生的两种一学期

许宁是临床医学五年级的学生，这个学期在附属医院老年病科轮转，同时修一门研讨课，叫《临床推理与病房研究》。不妨设想她有两种过法。第一种是我们熟悉的：开学第1周，老师用一套讲得极好的幻灯片梳理老年人跌倒的危险因素，从高龄、多重用药、视力与步态，一直讲到认知状况与夜间如厕，再配上几份权威指南的要点；期中有小测，期末有考试，许宁复述得条理分明，拿了高分。第二种过法里，开学第1周她拿到的不是一份讲义，而是九个问题，第1问是“本病房近三个月的夜间跌倒，集中在哪个交接时段前后？”。她先写下自己的初判：“夜间跌倒是因为患者年纪大。”课堂配给她的问对智能体ChatSDE在第一回合没有替她补全一份危险因素清单，而是反问了一句：“同样年纪的患者，为什么白天跌得少？”

这一句反问把许宁送回了病房。她去翻交班本，去对照护理记录，去跟一次夜班，去听患者自己怎样讲跌倒前的那一刻。资料在使用之前一律去标识化，进入病房取证经过科室与伦理委员会许可，她不改任何医嘱，也不作任何临床决策。几周之后，她把九问中的第2问“交班本上，跌倒前一班对该患者写了什么、没写什么？”提升为“交班时被省略的那一句，是谁、在什么情况下省略的？”。学期结束时，她交出的是一份子本，其中有一条讲义上没有、开学时连老师也写不出的判断：本病房夜间跌倒的风险更多落在交接断点上，而不只是年龄——风险藏在交班本里被省略的那一句。这条判断被明确标为本病房假说，待科室检验，它不是结论，却是一条真正从这间病房里长出来的判断。

化学系三年级的沈澄也有两种一学期。在讲授课堂里，她系统学到某类偶联反应的机理、催化剂的选择原则、常见副反应与提纯办法，并能在考卷上把影响产率的因素列得完整无缺。在九问课堂里，她修的是课程内嵌真实课题的《合成条件的发生》，出发点是本课题组十年的反应条件档案。她的初判是“产率低是因为试剂纯度不够”，ChatSDE问她：“同一瓶试剂，为什么三月好、六月差？”她回到实

验记录本、失败批次和仪器谱图，发现十年记录里“室温”一词几乎从未附过具体温度与湿度，又在师兄师姐“那一步要快”的口头叮嘱里，找到了规程上没有的那一句。她期末交出的判断是：本组某一步的成败依赖未入规程的环境湿度与操作时长——规程里没写的那一句，在师门口头里传了十年。这条判断同样被标为待复做验证。

两种一学期都可以是好教学。讲授课堂里的老师认真、清晰、体系完整，学生学得扎实，这一点本书不打算否认，第2章还要专门为它记功。真正让人困惑的是另一件事：如果把两种一学期末尾学生手里多出的东西摆在一起，它们根本不是同一类东西。前一种，全班40个人掌握了同一份清单，衡量好坏的是谁掌握得更完整、更准确；后一种，40个人各自带回一条从自己现场里长出来的判断，彼此不同，而且在开学时没有人能把它写出来。前者交付的是知识，后者交付的是发生。

交付物换了，衡量交付物的尺子也必须跟着换。讲授课堂的考卷问的是“像不像答案”，学生答得越接近标准答案，分数越高；九问课堂若仍用这把尺子，许宁那条从交班本里长出来的判断就无处安放，因为根本没有一个标准答案可以拿来比对。本书因此主张，九问课堂的学业成绩以子本的创新智商为核心，同时计入母本底盘、改题质量、问对深度、现场证据、子本完成度、碰撞贡献与回写证据，八个项目加一套权重组成一份“发生剖面”，评估的方向也随之翻转为“离已有有多远”——一份子本越是只复述了已有的东西，就越说明发生没有出现。这就像评一家新开的馆子，不再看它把名菜做得多像名店，而看它有没有端出一道别处吃不到、又经得起回头客检验的菜。第五编会把这套方案逐项展开。

本书要回答的，就是这个困惑引出的问题：当人工智能把第一种课堂做得比任何一位讲者都更快、更全、更耐心的时候，大学课堂还剩下什么理由存在？本书的回答可以压成一句话：未来大学的课堂不再交付知识，而交付发生——老师用装有SDE工序层的大模型打造九问，学生改题、问对ChatSDE、打造子本，ChatSDE陪学生一破一立，学业以创新智商为核心多项计分，档案回写成谱系；课堂的产出从“掌握了同一本书的人”变成“建成了自己子本的人”。需要先说明的是，许宁、沈澄以及书中所有人物与课堂都是构造案例，只用来说明机制，不是调查结果；书中凡为说明而设的分数与读数，都标明是构造示例，不冒充统计结果。全书只有附录二的小样本试跑是真实跑出的读数，它的局限由附录自己写明。

把“交付知识”与“交付发生”对照来看，弗莱雷在《被压迫者教育学》中批评的“银行储蓄式教育”提供了一个旧而有力的参照：老师往学生的账户里存入知识，学生负责保管、取用与复述。弗莱雷的批评主要是政治与伦理层面的，本书关心的是认识层面的另一件事——当存款可以由大模型随时、无限地发放，课堂若仍以储蓄为业，它的业务本身就被替代了。舍恩在《反思性实践者》（1983）中描述专业人员在实践现场“行动中反思”的能力，指出这种能力无法靠在教室里背诵技术规则获得；许宁在夜班里跟着护士记下从听到动静到到达床旁的时间，沈澄在复做一批时把每一步的温度与湿度都记下来，正是这种能力在学生时代的开端。两位学者从不同方向指向同一处：知识可以被存入，发生只能在现场里做出来。

这个回答的反直觉之处在于，它不是把讲授改得更生动，而是把讲台换成出题台。一个更会讲的老师，仍然在讲台上交付同一份终点；一个出题的老师，交付的是一组连他自己也写不出终点的问题，以及一套让学生能把终点发生出来的技术和陪伴。统一性知识的再发生是练手，应当交给中学；大学课堂只做一件事——让每个学生在自己的现场里，把一个连老师也写不出终点的判断发生出来。下面几节依次说明这个判断依据的框架、它的边界，以及全书怎样展开。

> **这一章的问题**
>
> 当人工智能把讲授做得比任何一位讲者都更快、更全、更耐心，大学课堂还能交付什么？

◆ 发现六路径与发生六路径

要说清“交付知识”与“交付发生”的区别，需要一个比“讲授式与探究式”更细的坐标。本书使用王德生在《SDE思想体系入门》（德麦国际专著第395号）中阐述的SDE框架。S指显露（Show），即一个判断、一个骨架、一个问题被摆到明处；D指差异，即同与不同、此与彼之间被拉开的那段距离，以及沿着差异一步步推进的路径；E指纠缠，即判断与具体现场、具体人、具体物彼此缠绕在一起的那一层。三者不是三个可以分别装箱的部件，而是任何一次认识活动都同时具备的三个维度。课堂作为一种认识活动也不例外，区别只在于它先从哪一个维度起步、经过哪一个、落在哪一个。

三个维度排出先后，就有六种次序：S→D→E，S→E→D，E→S→D，E→D→S，D→S→E，D→E→S。本书把它们当作课堂认识活动的六条主导路径来

用。这里要先立一条证据责任：六路径是主导次序，是分析课堂的一副透镜，不声称穷尽所有课堂，也不声称每一堂真实的课只走其中一条。日常里的一顿饭也可以作比——有人先定菜单再买菜再下锅，有人先逛菜市场看见什么新鲜再定菜单，有人冰箱里剩什么就做什么；三种做法都会用到菜单、菜市场和灶台，差别在先后，而先后决定了这顿饭最后长成什么样。

关键的一步在于：每条路径都有两种走法。一种是发现的走法，路径的终点先于行走而存在，走路的人是去把它找到、认出、掌握；另一种是发生的走法，终点在走的过程中才生成，生成之后还要回写，改写出发点本身。拿S→D→E来说，它的发现版就是我们最熟悉的讲授体系：先把体系显露出来（教材与讲义），再沿着章节的差异序列一步步推进（讲授与练习），最后落到应用（习题、实验与考核）。终点在开课之前就写在教材里，学生走完这条路，是把那个终点搬进自己的头脑。它的发生版则是本书讲的九问模式：先显露一个谱系母本与九个问题，再沿着学生的改题与问对推进，最后落进学生自己的现场，长出一份子本，并把成果回写到母本。

由此可以看清本书所说的“典范革命”指的是什么。库恩在《科学革命的结构》（1962）中用典范一词描述一个共同体共享的问题、方法与解题范例；本书借用这个词，但用法更窄：课堂的典范革命，不是路径数目的增减，不是在六条路之外再发明第七条，也不是把讲授换成讨论这类形式上的替换，而是每一条路径的方向从“去找先在的终点”转向“生成并回写”。同样是先立骨、再走路、后落地，一个是沿着别人铺好的路走到别人放好的终点，一个是在走的过程中让终点发生，并让走过的路反过来改写起点。

发生版与发现版的分界，还可以用SDE框架里的123原理来说明。本书在全书范围内只在一个意义上使用这个词：D与E之间出现矛盾，逼出S的改变，改变后的S再回写D与E。许宁的初判“因为年纪大”是一个S；病房里“同样年纪、白天跌得少”的事实，经那句反问被摆到她面前，与这个判断之间拉开了差异，又在交班本与夜班的具体现场里缠住了她，这就是D与E对S的顶撞；她被逼着把判断改成“风险落在交接断点上”，这是S的改变；改变之后，她再回去看交班本、看呼叫铃、看护士到达床旁的时间，看法全变了，这是回写。发现版的课堂里也有矛盾，但矛盾被设计成通往已知终点的台阶，S不必改变，只需被掌握；发生版的课堂里，矛盾是真的，S非改不可，而改成什么，事先没有人知道。

其余五条路径各有其发现版与发生版。S→E→D、E→S→D、E→D→S、D→S→E、D→E→S，本书分别称为靶子、田野、论证、跟随、卡点，它们在结语中各用一个学科举一例作预告，此处只留名字。之所以不在导论里展开，是因为每一条路径的发生版都需要一整套课堂装置与评估方案才能讲清，而本书要先把一条路讲透。这正如学驾驶不必同时学会开车、开船和开飞机，先把一种交通工具开上路，其余几种的道理才会看得清楚。

◆ AI接管发现一侧

如果路径的方向可以分成发现与发生两种，那么今天这场变化最直接的含义就是：发现的一侧，正在被大模型接管。讲清一个体系，按章节由浅入深地推进，给出例题与变式，回答学生第十遍问到的同一个问题，写一篇覆盖面周全的文献综述，按照规格完成一个标准项目，根据作答情况给学生归类、指出错误并给出纠正——这些都是发现版课堂的核心工作，而大模型在其中多数环节上做得更快、更全，也更耐心。许宁若在深夜想重新理解衰弱综合征的诊断标准，一个通用大模型可以在几秒钟里给她一份结构清楚的讲解，再按她的追问把其中任何一处讲得更细；它不会因为第三次被问同一个概念而显出不耐烦，也不必等到下周的答疑时间。

这件事早有预兆。布卢姆（1984）提出的“两个西格玛问题”指出，一对一辅导下学生的成绩大幅优于常规班级授课，难点在于如何以群体教学的成本逼近个别辅导的效果。40年来，这个问题一直是教育技术的一面旗帜。可汗学院在2023年推出的Khanmigo，以及各类基于大模型的辅导系统，都在朝这个方向推进：让每个学生都有一位随叫随到、能按其进度调整讲法的辅导者。就发现一侧而言，这条路是走得通的，而且会越走越通。一位讲师花十年打磨的讲授体系，大模型可以在一个下午里以十种讲法复现出来，再按每个学生的错误给出个别化的纠正。

但要特别说清楚的是：发现不作废，它降为底盘。许宁若不知道老年人跌倒的常见危险因素，就根本问不出“同样年纪为什么白天跌得少”；沈澄若不懂偶联反应的机理，就无法判断谱图上那个反复出现的小峰值不值得追。统一性知识仍然是一切新判断的地基，本书第3章要专门讨论大学怎样把它组织成“母本精要库”，并在评估中把对母本的掌握设为门槛项。变化在于，这一层不再需要占据课堂的主体时间，也不再是课堂交付物的主体。好比一个城市有了自来水，人们照样天天用

水，水的重要性一点没有减少，但已经没有人把每天挑水当成一份需要专门安排人手的工作。

大模型接管发现一侧，也不意味着它只能做发现。一个通用大模型同样可以提出新颖的组合、写出出人意料的段落。本书在第三编要讨论的，是另一个更具体的问题：在课堂这个配置里，学生拿着一个未经检验的前提来问，系统的默认反应是什么。本书的判断是，默认配置（未开启学习模式）下，通用大模型往往会质疑学生的前提，但在同一个回合里就把替代的原因或答案一并写好，也就是默认替学生说破；而课堂需要的是第一回合把前提翻过来看之后，把下一步留给学生自己去迈。这一判断是被修订过的：本书初稿曾认为默认配置只是顺着前提做形式逻辑展开，附录二的小样本试跑没有支持这一点，初稿的说法因此撤回。这里必须承认一个2025年以来的事实：Anthropic的Claude学习模式（2025年4月）、OpenAI的ChatGPT Study Mode（2025年7月）、Google的Gemini Guided Learning（2025年8月）已经把“先追问、不直接给答案”做成了可开启的学习模式。它们说明苏格拉底式追问这一层壳并不难抄，也恰好支持本书的判断：同一个底座，开启与否行为不同，差别住在配置与包装层，而不在底座模型能力的高下；它们是否做到了课堂所需要的全部，则是另一个问题，本书在第8章逐项对照。第8章至第11章会把这些差别当作可检验的主张提出，并写明怎样算错，而不把它们当作对大模型的定论。

于是问题的形状就清楚了。如果课堂的价值曾经主要来自它传递统一性知识的能力，那么当这一能力被大规模、低成本地复制之后，大学课堂要么缩成一个发证机构，要么找到一样大模型在默认配置下不交付、而课堂可以交付的东西。本书认为后者存在，它就是发生：一个学生在自己的现场里，经由被追问、被批判、回现场取证、改写判断的往返，生成一条此前没有的判断，并让这条判断回写进他自己、他的同学和他所在的学科谱系。下一节说明，为什么本书只从六条路中的一条讲起。

◆ 为何只讲S→D→E，以及判别两问

六条路径里，本书只讲S→D→E一条。理由首先是覆盖面：S→D→E的发现版就是讲授体系，而讲授体系是今天大学课堂最普遍的形态。从医学院的病理学大课到化学系的有机化学，从法学院的民法总论到中文系的文学史，绝大多数课程都是

先立体系、再按章推进、后落到习题与考核。如果课堂的典范革命要从某一处切入，从覆盖面最广、教师最熟悉的那条路切入，最能让读者看清方向反转之后发生了什么。一位每学期讲同一门课的老师，翻开本书时会立刻认出那条熟悉的路，然后看到同一条路在另一个方向上的样子。

理由其次是结构上的连续性。九问模式保留了讲授体系的三段骨架：先立骨，再走来路，后铺土壤，之后才是评估与回写。它在S段放的仍然是一个“本”，只是从统一教材换成了谱系母本与从中提炼出的九问；在D段走的仍然是一条推进之路，只是从章节序列换成了学生的改题与问对；在E段落的仍然是应用，只是从标准习题换成了学生自己的现场，以及由此打造的子本。这种连续性让变革有了可以落脚的台阶，老师不必把讲台推倒重来，而是把讲台改造成出题台。本书因此把自己定位为“课堂发生学”的卷一，其余五条路径留给后续各卷。

把这条路完整地写出来，就是九问模式的工序链：谱系母本精要→九问→改题→问对→打造子本→评估→回写。前两环是S，老师与智能体从母本中提炼核心判断，再按每个学生的画像打出九个谱系性研究课题；中间两环是D，学生改题、提升题，并与ChatSDE进行多轮问对，在被批判与回现场之间来回磨；后几环是E与回写，学生把九份问对成果重写成一份子本，子本以八个项目接受评估，档案回到母本精要库，成为下一届九问的出发点。工序链的每一环在本书中各有一章或数章展开，导论只需要让读者看清它是一条首尾相接的链，而不是几项可以单独挑出来用的教学技巧。

只讲一条路，就需要一个判别办法，让读者在面对任何一堂具体的课时，能判断它走的是发现版还是发生版。本书给出两个问题。第1问：这堂课的终点，能否在开课之前由老师写出来？如果能，它就是发现版，或者是统一性知识的再发生，即本书所说的练手；如果连老师也写不出来，它才有资格被称为战场。第2问：课程结束之后，有没有回写？也就是说，学生的判断有没有改写他自己的底盘，有没有进入班级的共同判断，有没有被吸收进下一届的母本精要库。若没有回写，哪怕过程再热闹，也只是一次活动；有了可被观察的回写痕迹，它才是一条会生长的谱系。

两问合起来用，可以分出几种典型情形。终点可预写、无回写，是标准的讲授课；终点可预写、有回写，比如中学里设计良好的探究课，学生被改写了，但学科没有被改写，这是练手；终点不可预写、无回写，比如一次热烈的开放讨论，大家都说了些新东西，散场后什么也没留下，这是一次没有落地的战斗；终点不可预

写、有回写，才是本书所说的九问课堂。拿做饭打比方，照着菜谱做出一道已知的菜，是练手；做出一道谁也没吃过的菜，吃完就忘了，是偶然；做出一道新菜，并把配方改进写进家里的菜谱，下一次有人在此基础上再改，才是谱系。

把两问用在开篇的两个学生身上，可以看得更具体。许宁在讲授课堂里的一学期，终点就写在讲义与指南里，期末考试考的是她离那个终点有多近，课程结束之后，下一届的讲义一字不改，这是终点可预写、无回写。她在九问课堂里的一学期，开学时没有人能写出那条关于交接断点的判断，期末时它以本病房假说的身份交给科室检验，并进入本科室轮转学生的问对档案，下一届的九问可能就从这里出题，这是终点不可预写、有回写。沈澄的情形相同：讲授课堂里她掌握了一份完整的影响因素清单，九问课堂里她让规程里没写的那一句显露出来，而它是否真能写进课题组规程，要等下一届复做之后才见分晓。回写在这里不是一句口号，而是一个要等、要查、要留痕的事件。

需要补一句：两问是判别工具，不是褒贬工具。终点可以预写的课堂并不低一等，它是发生技术的练习场，而且是必不可少的练习场，第1章会专门论证这一层应当前移到中学完成。同样，回写也不能被宣布，只能被观察。一门课自称“学生的成果已回写进学科”，并不能证明什么；能证明的，是下一届的九问里出现了上一届学生的判断，是课题组的规程里多了一行，是学生期末自述与第零轮初判之间可以对照的改变。这一条贯穿全书，也是本书对自己的约束。

◆ 最强竞争解释与划界

本书不是孤立的一本。与它相邻的有两本兄弟书，各自管一个尺度。《大学的典范革命：从统一性到谱系发生》讲的是宏观尺度：知识生产的方式怎样变化，统一性知识的生产与传递怎样移向大厂与平台，大学作为一种制度怎样转向谱系发生。《从The Book到Our/My Book》讲的是全学段的教育使命：从小学、中学到大学，教育怎样从人人共读的一本书，转向每个人的书与共同体的书。本书只讲课堂，即中观尺度：一门课、一位老师、一个班、一个学期之内，九问模式怎样运转、怎样评估、怎样回写。三本书按尺度划界，宏观的论证在那里展开，本书只引用其结论；读者可以把三本书想象成看同一座城市的三张地图：一张是全城的交通规划，一张是全城所有学校的分布，一张是一所学校里一间教室的布置图。

不过，按尺度划界只回答“在哪一层说话”，回答不了“说的东西是不是新的”。后一个问题必须按主张来答。一位严格的读者会这样质问：改题权不过是问题发现，问对不过是带护栏的AI辅导，第零轮不过是生产性失败的生成阶段，回写不过是知识建构里的观点改进，题闸不过是高校教学中心早就在讲的“先让AI试做一遍”。这个质问在部件层面基本成立，本书不打算否认。本书的立场是：承认部件几乎都已存在，主张落在两处——一是把它们接成一条从谱系母本到回写的工序链，二是这条链上每一环都配有可被观察、可被推翻的证据责任。下面逐项列出最强的竞争解释，每一项都先说它能解释本书的哪一部分，再说本书多出哪一格；总表附在本节末尾。判别的总判据仍是上一节的两问：终点能否事先写出，课后有无回写。

问题导向学习的源头之一是Barrows与Tamblyn的《问题导向学习：医学教育的一种方法》（1980）。它把医学生从讲堂带进病例，以问题组织学习，这是医学教育史上的重要突破；Strobel与van Barneveld（2009）的元综合还提示，PBL在长期保持与技能应用方面优于传统教学，许宁所在的医学院很可能就有PBL的传统。九问模式继承了以问题为起点这一点，分岔在于问题的来源与终点：典型的PBL病例由教师团队精心编写，病例背后有一个设计者已知的学习目标与参考答案，学生在小组讨论中走向它；九问则要求每一问都通过题闸——拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案，能答就打回重出。许宁第2问之所以合格，是因为没有任何人能离开本病房的交班本答出它。

探究式学习的思想渊源可以追到杜威，《我们怎样思维》（1910）与《民主主义与教育》（1916）都把思维理解为由困惑引发、经假设与检验而推进的过程。九问模式在精神上与之同源。分岔在于，许多课堂里的探究是对已知结论的再探究，学生“发现”的是课本上已有的规律，这在本书的框架里正是练手，而且应当在中学完成。但探究式学习还有一个不能绕开的经验性对手：Kirschner、Sweller与Clark（2006）认为，最小指导的探究教学对新手是低效的，新手缺少可以引导搜索的图式，放手摸索占去的是工作记忆，而不是学习；Lazonder与Harmsen（2016）的综述把话说得更细，探究式学习的效果依赖指导量，适度指导的探究优于无指导。本书接受这个批评，并且认为它恰好支持练手与战场的分工：中学的九问是强指导下的再发生，终点已知，老师随时能把学生拉回，题闸放宽，问对多接

住；进入战场的大学生才被放到指导较少、终点未知的位置，而那时他们在母本底盘上已不是新手，这也是母本底盘被设为门槛项的理由。这一回应的细节在第1章。

课程型本科研究（CURE）在“结果事先未知”这一点上走得更远。Auchincloss等（2014）的论述把它刻画为课程内嵌真实研究、结果事先未知、对研究共同体有意义的课程形态，Corwin、Graham与Dolan（2015）进一步提出了对它做研究与评估的议程，博耶委员会《重塑本科教育》（1998）更早就呼吁以研究为本科教育的基础。其中最成功的范例是SEA-PHAGES：Jordan等（2014）报告了多所院校的本科生在课程里做真实的噬菌体研究，成果进入共享数据库并得以发表。沈澄所修的课正是一门课程型研究课，第20章会专门讨论它的功绩。SEA-PHAGES的共享数据库与本书所说的“回写”最为相近，也因此最需要划清：它收纳的是学生做出的研究成果，指向学科；九问的回写多出的一格，是回写首先改写下一届的出题，即母本精要库，并且把被撤回的判断、被下一届推翻的判断同样登记在案，按三层与四类痕迹加以识别。此外，CURE通常由一个共同课题统摄全班，九问则按每个学生的画像与现场分别出题，这是第二处分岔。

翻转课堂由Bergmann与Sams（2012）系统推广，把讲授移到课前视频，把课堂时间留给练习与讨论。它敏锐地看到了讲授可以被移出课堂，这一点与本书判断AI接管发现一侧相通；分岔在于，翻转之后课堂里练的多半仍是终点已知的题目，课堂的交付物没有换。个别辅导一线更长：布卢姆（1984）的两个西格玛问题之后，VanLehn（2011）指出人类一对一辅导的效应远小于两个标准差，约在0.79上下，对布卢姆做了修正，这提醒我们不要把任何辅导形态的收益想象得过大；第14章将在讨论同学碰撞场时回到这一点。另外，一对一辅导这条线再好，仍是个别辅导，还不是课堂；课堂之为课堂，在于同学之间的碰撞能涌现出个人说不出的判断。

生产性失败是第零轮最强的竞争解释。Kapur（2008，2016）以及Sinha与Kapur（2021）的研究一致指出：先让学生独立尝试、生成并经历失败，再给予讲授，学习效果常优于先讲后练，元分析支持这一做法在概念理解上的优势；Chi与Wylie（2014）的ICAP框架则把学习投入分成被动、主动、建构、互动四种，效果依次递增，九问的问对落在最高两档。九问的第零轮“先写初判”，几乎就是生产性失败的生成阶段。多出的一格有两处：其一，这里的初判落在学生自己的现场里，而不是教师设计好的问题上，所以失败本身是真实的、不可预设的；其二，失

败条件由学生自己写下，第零轮因此兼作回写的基线，期末自述要与它逐条对照。第6章展开这一点。

知识建构是班级子本和回写最强的竞争解释。Scardamalia与Bereiter（2006）描述的知识建构，以共同体持续改进观点为核心，配有社区数据库，并强调“超越”（rise-above），即共同体把先前的观点收拢到更高一层。本书的精要库与班级子本在形态上与它很近，承认其为先例。多出的一格有三处：精要库对本届出题有上游作用，按学生画像把前几届的积累转为今天的题；每一次改写都有版本史与撤回记录，使“谁在什么时候改了什么、后来是否被推翻”可查；碰撞场的产物须过题闸，防止班级共识沦为通用知识的复述。详见第3章与第14章。

默会知识是“规程里没写的那一句”最强的竞争解释。Polanyi（1966）的《默会维度》提出我们知道的多于我们能说出的，Nonaka与Takeuchi（1995）的SECI模型把默会知识的外化当作组织创造知识的关键一环。沈澄课题组口头传了十年的那句“那一步要快”，正是一个典型的外化对象，本书承认这一点。多出的一格有二：一是一个结构性的判断，即这类空白系统地出现在书写者以为对方已知的地方，所以可以有方向地去找，而不是碰运气；二是外化之后的去处，子本进入个人底盘，接入个人检索库与技能，并接受版本记录，使外化成果可以被回访和修正。详见第13章与第14章。

双环学习与回写的关系需要说清。Argyris与Schön（1978）区分了只改行动的单环学习与连支配行动的变量一并改写的双环学习。许宁把“因为年纪大”改成“风险落在交接断点上”，之后再看交班本时看法全变，确有双环的形状，本书承认回写第一层（学生底盘）可以用双环学习来描述。多出的一格是观察：双环学习在组织研究里常靠访谈与叙述被确认，回写则要求留下可观察的痕迹，要求登记撤回，并且要能识别出负向回写，即一次学习之后底盘变得更差。详见第24章。

“学生作为研究者”一线有两座近邻。Healey与Jenkins（2009）按学生是以研究内容为受众还是参与者、课程是以研究过程与问题为重还是以研究内容为重，划出研究导向、研究任务、研究型、研究引导四个象限；Fung（2017）的“连接课程”则主张学生做研究，并让这些研究跨届连接进课程。再往前，Getzels与Csikszentmihalyi（1976）关于问题发现的研究表明，提出问题的方式能预测创造性成果。这三者合起来，几乎覆盖了改题权和回写进课程的全部直觉，所以本书不称它们为新发现。多出的是定位与约束：改题权落在哪个象限，以及改题必须服从

的“两锚”——根还在，离现场更近，使改题成为可检查的操作，而不是对学生兴趣的泛泛尊重；回写进课程则要求下一届的九问里能指认出上一届的判断。详见第5章与第24章。

最切近的竞争解释，来自AI辅导本身。Khanmigo（2023）、哈佛CS50的AI助教（Liu等 2024）早已用教学护栏让模型引导而不直接给答案；2025年起，前文提到的三家学习模式又把“先追问、不直接给答案”做成了可开启的选项。Chiang等（2024）则用大模型扮演“魔鬼代言人”来改善群体决策。这些做法再次说明，追问这一层壳并不难抄，行为的差别住在配置与包装层，而不在底座。本书的收窄主张是：这些产品做到了不直接给答案、先追问，没做到的是题闸、改题权、第零轮基线、按坏法选刀与回写。第8章与第10章分别正面处理。与之相连的实证是Bastani等（2025，PNAS）的高中数学现场实验：不加护栏的通用大模型辅助练习时表现更好，撤掉AI后考试反而更差，而加了教学护栏的GPT Tutor基本消除了这种损害；Kestin等（2025，Scientific Reports）的哈佛物理课随机对照则表明，精心设计的AI导师带来的学习收益高于课堂上的主动学习。这两项研究给本书两条先验：护栏是实证意义上的真变量，所以“配置诊断”不是修辞；AI导师在发现一侧可以做得很好，所以本书要守的不是讲授与辅导，而是出题台。详见第9章与第11章。

高校教学中心从2023年起流行的作业重设计指引，是题闸最直接的先行者：先把作业交给生成式AI试做一遍，能做出来就改题。本书承认题闸与它同构，多出的一格是把这件“试做”做成一道带记录的出题工序：试做所用模型的版本要登记，试做时不得接入学生的现场材料，试做的结果以打回率为读数进入课程评估，而不是靠教师个人的警觉。至于档案袋评价与过程性评价，由来已久，近年又有了对写作过程版本史的研究；本书主张轨迹是不可还原性的证据，与之同向，多出的一格是轨迹与题闸、首发口径联动，用来判定“这个判断是谁先说出来的”。详见第4章、第15章与第17章。把以上对照收拢成一张总表，读者可以逐行检查本书声称的“多出一格”是否站得住。

构念	最强竞争解释	本书多出哪一格	详见
问题导向学习（PBL）	Barrows & Tamblyn 1980；Strobel & van Barneveld 2009	题不由教师预写终点，须过题闸；按学生现场打造，学生有改题权	第19章

构念	最强竞争解释	本书多出哪一格	详见
CURE与SEA-PHAGES	Auchincloss等 2014；Jordan 等 2014；Corwin等 2015	全班共同课题改为按学生现场分别出题；回写首先改下一届的出题，并登记撤回	第20章
探究式学习与强指导批评	Kirschner等 2006；Lazonder & Harmsen 2016	不主张新手自由探究：中学强指导练手，大学才上战场，母本底盘设门槛	第1章
生产性失败	Kapur 2008，2016；Sinha & Kapur 2021；Chi & Wylie 2014	初判落在学生自己的现场；失败条件由学生写；第零轮兼作回写基线	第6章
知识建构	Scardamalia & Bereiter 2006	精要库对出题的上游作用；版本史与撤回记录的证据责任；碰撞产物须过题闸	第3章、第14章
默会知识与外化	Polanyi 1966；Nonaka & Takeuchi 1995	“空白系统地出现在书写者以为对方已知之处”的结构判断；外化后接入个人检索库与技能并留版本记录	第13章、第14章
双环学习	Argyris & Schön 1978	回写要有可观察痕迹、撤回记录，并识别负向回写	第24章
学生作为研究者	Healey & Jenkins 2009；Fung 2017	改题权落在哪个象限；回写进课程要能在下一届九问里指认	第5章、第24章
问题发现	Getzels & Csikszentmihalyi 1976	改题两锚：根还在，离现场更近，并留题的版本史	第5章
翻转课堂与个别辅导	Bergmann & Sams 2012；布卢姆 1984；VanLehn 2011	课堂交付物从知识换成发生；同学碰撞场超出一对一线	本节；第2章、第14章
苏格拉底式AI辅导与2025年学习模式	Khanmigo 2023；Liu等 2024；三家学习模式；Chiang 等 2024	题闸、改题权、第零轮基线、按坏法选刀、回写	第8章、第10章
带护栏的AI辅导实验	Bastani等 2025；Kestin等 2025	以其为先验，增加首发率与套路化检测两项指标；评估环与承接环	第9章、第11章
作业重设计指引	2023年起高校教学中心的“先交给AI试做”指引	试做版本登记、不接入现场材料、打回率作读数	第4章

构念	最强竞争解释	本书多出哪一格	详见
档案袋与过程性评价	档案袋评价；写作过程版本史研究	轨迹作为不可还原性证据，与题闸、首发口径联动	第15章、第17章

总表的读法有一条纪律：表中“多出哪一格”不是说其他方案缺这一格，而是说本书在这一格上承担证据责任。借用部件不改变判别：只要终点仍可预写、回写仍无痕迹，部件再多也还是发现版的课堂；反过来，欢迎九问课堂借用PBL的小组讨论规范、探究式学习的假设检验训练、CURE的研究伦理与数据管理、翻转课堂的课前材料组织、AI辅导的个别化答疑。把这些部件接成一条从谱系母本到回写的工序链，要求终点不可预写，要求回写可被观察，要求每个学生有自己的现场与题，要求问对的默认是批判性激发，要求评估以子本的创新智商为核心、按八个项目组成发生剖面——这个组合与它的证据责任，才是本书要论证的东西。

◆ 与作者旧作按主张划界

本书的几个基本概念，在作者此前的两部著作里已有源头：王德生的《SDE思想体系入门》（德麦国际专著第395号）与《子本打造法》（德麦国际专著第527号）。读者有权问：哪些是旧作已经说过的，哪些是这本书才有的？按尺度划界在这里不够用，必须按主张逐项交代。下表把沿用、改写、新增三种情形分开，本书凡沿用旧作之处，只取其操作含义，不拿旧作充当证据。

概念或做法	《子本打造法》（第527号）／《SDE思想体系入门》（第395号）	本书	属性
母本与子本、子本重写不翻译	已提出：读者以自己的画像从母本得到整本子本	子本成为课堂学业产物与评估对象	沿用并改写用途
九问	已提出：从母本核心判断出九问，经问对得成果	九问成为九个谱系性研究课题，由老师按学生现场打造	改写
问对ChatSDE	已提出	增第零轮基线、首发口径	改写
题闸、练手与战场	无	新增	新增
改题权与两锚	无	新增	新增

概念或做法	《子本打造法》（第527号）／《SDE思想体系入门》（第395号）	本书	属性
工序判别细则（断拍、缺式、偏路的判别次序）	第395号有三大方程、六路径、123原理	新增课堂判别次序	新增
碰撞场与班级子本	第527号提到班级子本的设想	新增碰撞时点、组织与过闸要求	改写
八项发生剖面与权重	无	新增	新增
回写三层与可观察痕迹	第395号：回写要被观察	落到课堂三层与四类痕迹、负向回写	改写
谱系母本＝前沿专著＋部落档案	无（兄弟书《大学的典范革命》讲宏观）	新增课堂层用法	新增

这张表有一个自我约束的含义：它把本书的新东西集中到课堂层的机制上——题闸、改题权与两锚、工序判别细则、八项发生剖面、谱系母本的课堂用法——而不是理论层的概念上。母本与子本、九问、问对、回写这几个概念，旧作里已经有了，本书所做的是把它们从个人读书的场景搬进一门课、一个班、一个学期，并相应地改写其用途与评估办法。这一搬动是否真有增益，不靠旧作的权威来担保，要靠上一节的总表和各章末尾的失败条件来检验。

此外，书末附有参考文献表与附录二《一次小样本试跑》。前者统一编列书中引用的全部文献；后者报告本书在小范围内试行的一次对照，它只作为先验与方法演示，能说明什么、不能说明什么，由附录自己写明，结语也会再提一次。书中人物与课堂仍是构造案例，不因这次试跑而改变身份。

◆ 全书路线图：七编26章

全书按“总—分—总”组织：导论总起，七编26章分述，结语收口。七编的次序就是九问模式的次序：先立骨（S），再走来路（D），中间插入一编讨论问对智能体与通用大模型的差别，然后铺土壤（E），再讲评估，再给学科样板，最后讲回写与大学的领地。读者可以把它想成一门九问课从开学到期末、再到下一届的全过程，每一编对应流程中的一段。

第一编“S：母本与九问”是立骨，共四章。第1章“练手与战场”论证统一性知识的再发生是练手，应在初高中完成，练成的是发生技术；大学是战场，九问打向连老师也写不出终点的谱系知识。第2章“教材退场之后”为讲授体系记功，并论证在AI接管讲授之后，讲台的价值落到出题台，大学教材让位于谱系母本。第3章“谱系母本精要库”说明大学的母本由学科前沿专著与课题组（部落）档案合成，精要库一次建成、千人共用、逐年回写加厚。第4章“出题台”讲老师怎样用装有SDE工序层的大模型从精要库按学生画像打造九问，每一问怎样过题闸，以及老师不可替代之处——认得学生的现场。许宁的九问作为定本在这一章逐问讲出题逻辑。

第二编“D：改题与问对”是走来路，共三章。第5章“改题权”主张九问是种子不是考卷，学生改题时谱系化由学生自己完成，并给出改题的两锚：根还在，离现场更近；周川由“哪个伙计干得最起劲”改为“父亲为什么一到月底就泄气”，是提升，而“AI会不会取代饭馆”是逃离。第6章“问对ChatSDE”讲第零轮初判与多轮自由问对，说明问对是来回磨而不是一问一答，问对记录就是发生档案；沈澄从第零轮“试剂纯度”到湿度假说的七轮问对是本章的案例。第7章“一破一立”讲批判性激发怎样做到不套路化：先接住放大，再批判；按判断坏在哪里选刀，缺式补式、偏路排路、断拍点火，破的同时指出立的方向，逼近判断但不替学生说破。

第三编“ChatSDE与通用大模型：知识发生上的差别”共四章，是全书论证最需要谨慎的一编。第8章讨论两种默认：默认配置（未开启学习模式）下的通用大模型往往会质疑前提，却在同一回合替学生说破，ChatSDE默认做批判性激发而不替说破，援引Sharma等（2024；最初为2023年的预印本，发表于ICLR 2024）关于大模型迎合倾向的研究，只作默认行为成因的旁证，并正面写2025年三家学习模式与CS50.ai做到了什么、没做到什么；课堂里，默认值就是教育，因为学生的首发一旦被替掉，加一句批判性提示也找不回来。第9章把“信息发生≠知识发生”作为针对具体配置的诊断提出，差别不在模型能不能，而在整个配置有没有把错误、修订和学生的现场接进来。第10章提出工序层的概念：模型是发动机，工序是驾驶术，底座可换，工序跟走。第11章设计对照实验：同一底座，裸用与装进工序层，只差工序；指标与失败条件写明，并规定除附录二的小样本预试外，本章不报告任何结果，凡写“预期”之处都只是待检验的假设。

第四编“E：打造子本与落壤”是铺土壤，共三章。第12章讲子本是母本加九问研究成果的重写，不是翻译，子本里须有母本所无的判断；苏文的历史子本从翻译稿到重写稿是本章案例。第13章讲子本进入底盘，生成个人的检索库与技能，成为个人知识发生平台的数据基因，毕业后继续生长；许宁从实习到住院医的交接班思考底盘、周川毕业五年后回来更新子本是本章案例，并说明这一底盘是思考的底盘，不是临床决策工具。第14章讲同学碰撞场：只有一对一的线是高级个别辅导，课堂之为课堂在于同学互撞；周川与林晓撞出“小微经营者的动力被现金节律收走”，沈澄与许宁跨学科撞出“没写下来的那一句”。

第五编“学业评估：多项目与权重组合”共四章。第15章提出学业成绩是八个项目加一套权重组成的发生剖面，八个项目是母本底盘、改题质量、问对深度、现场证据、子本创新智商、子本完成度、碰撞贡献、回写证据，成绩单给剖面而不只给总分。第16章讨论核心项：子本创新智商以S/D/E/I/F五维、160分制评估，按档位分段映射到百分制，完成度为门槛，评分须换手并写利益冲突声明。第17章讨论过程项，主张按质不按量，轨迹是不可还原性的证据，也回答了AI代写的问题：拿不出轨迹的高分子本要被追问。第18章讨论权重怎样随学段、课型、学科变化，并在区间内交给学生自选，大学阶段子本创新智商权重不低于25%（大一补练手例外），选择权重本身就是一次判断。

第六编“学科九问样板”共五章，每章统一结构，从该学科发现版课堂的样子写起，经谱系母本、完整九问、改题与问对、子本形态、评估权重与门槛，写到学科特有风险。第19章是医学，战场在病房与交接里，并讨论PBL的功绩与边界、伦理门槛与去标识化；第20章是化学，战场在失败批次与规程空白里，并讨论CURE的功绩与边界、可复做性与实验安全；第21章是物理、工程与计算，以高远的桥梁监测与方舟的系统排障为样板；第22章是经济与法学，以周川、林晓与韩正为样板，讨论互动怎样被看不见的周期推着走；第23章是历史与文学，以苏文与叶青为样板，讨论学生怎样成为发生的主体，以及家族隐私与口述伦理。

第七编“回写与大学的领地”共三章。第24章讲回写的三层——学生底盘、班级子本、谱系母本精要库——以及负向回写的识别，坚持回写要被观察、不能被宣布；沈澄的湿度假说被下一届复做后写进课题组规程，是本章的案例。第25章讲过渡期：今天的大学生大多没练过发生技术，大一须补练手，吴静用“血压调节”走一遍九问；并给出一学期16周的样板与老师的周工作表。第26章讨论大学在AI时代

的领地：大厂造发动机，大学造驾驶术；壳可以开放，工序与越积越厚的问对档案要守住。结语回到六路径，预告其余五条路，并以许宁与沈澄毕业时各自的最后一问收束：大学最终教会学生的，不只是一本子本，而是面对新问题时先问“这个判断可能坏在哪一处”的那双眼睛。

全书各章都按同一个次序写：从一个具体课堂场景里的困惑开场，立起框架，讨论张力与边界，用学科案例验证，最后写出“本章怎样算错”。所有人物与课堂均为构造案例，医学案例去标识化并经许可，化学案例不涉及危险操作细节，法学与历史案例尊重隐私。读者若只想先了解课堂怎样运转，可以按第4、5、6、12、15、25章的次序先读；若关心与通用大模型的差别，可以直接读第三编；若是某一学科的教师，可以先读第六编中本学科的那一章，再回头读其余部分。

◆ 本章怎样算错

导论提出的是全书的总判断，它的失败条件也应当在一开始就写明，而且写成别人可以照着去查的读数。第一条管“发生能否被交付”。取任何一个认真实施九问的学科，连续两届、每届不少于30名学生，由不少于2名互不知情的评审者换手评阅期末子本，把被评为“母本所无的新判断”的条目逐条交给一个未接入该学生现场材料的通用大模型（记录所用版本，默认配置与开启学习模式各试一次），请它给出说得过去的答案。若这些条目中有半数以上被复述出来，或者两届的题闸打回率都低于10%而换手评审后仍无一问被改判打回，就说明题闸普遍失守、终点其实可以事先写出，“大学课堂交付发生而非知识”这一主张在该学科要降格为一个愿望，九问的出题与评估机制必须重做，或者承认这条路在该学科不成立。

第二条管“发现是否真能降为底盘”。取讲授课时压缩到原来一半以下、学生普遍借助大模型自主学习的班级，与讲授未压缩的对照班，各不少于30人，连续两届做母本底盘的复述加反例口试。若压缩班的通过率持续比对照班低15个百分点以上，且这一差距不能用精要库或口试设计的缺陷解释，那么本书关于讲台让位于出题台、讲授降为底盘的判断，在该学科就过度了，讲授的份额应当重新估计，而不是简单地移出课堂。判定时看口试记录与精要库的使用记录，不看教师或学生的感受。

第三条管“九问是否真有自己的位置”。把同一门课的九问与教学日历、学生产物交给不少于3位熟悉PBL、CURE或AI辅导的评审，让他们在互不知情的情况

下，按终点可预写程度、回写痕迹、学生产物三项，把这门课归入最近邻的方案。若2/3以上的评审认为三项均无可辨别的差别，则本书声称的“典范革命”只是旧方案的新名字，全书的划界要撤回，九问模式应当如实归入那一种方案的谱系。

第四条管上文总表本身。表中每一行的“多出哪一格”，都要在对应章节的检验里得到至少一个可观察的读数。比如第6章的“学生自写的失败条件”，若一学期之后没有一条被后续轮次引用或修改，这一格就是空的；这一行应从表中删去，并承认该部件已被现成方案完整覆盖。四条条款分别对应本书的四个支点：发生能否被交付，发现是否真能降为底盘，九问模式是否真有自己的位置，以及“多出一格”的声称是否有实处。

这四条之外，导论还为全书立下一条总的约束：书中凡说到回写、说到差别、说到新判断，都要指出可观察的痕迹在哪里，指不出来的，就按未成立处理。这条约束对本书自己同样适用：全书的人物与课堂既然都是构造案例，它们只能说明机制，不能充当证据。真正的证据，要等九问课堂在真实的院系里运转之后，由那里留下的版本史、问对记录与回写痕迹来提供。

导论留下了一个最直接的问题。如果大学课堂只做战场上的事，只打向连老师也写不出终点的判断，那么学生凭什么上得了战场？改题、问对、接住批判、下现场取证、重写成文，这一整套技术不会在大一开学那天自动出现。许宁能在病房里接住“同样年纪为什么白天跌得少”这一句反问，而不是被它问懵，沈澄能把一句“为什么三月好、六月差”变成一份复做记录，背后都有一段在终点已知的题目上反复练手的过程。这段练手应当在哪里完成，用什么材料完成，它与大学的战场又是什么关系，这是第1章要回答的问题。

> **带走的话**
>
> **未来大学的课堂不再交付知识，而交付发生：终点连老师也写不出、课后又留得下回写痕迹的判断，才是大学课堂该打的仗。**



第　一　编

S：母本与九问（立骨）

编　序

化学系三年级的课程开课第1周，有同学翻完目录，举手问：“老师，教材在哪里？期末考哪几章？”他要的是讲授课堂里那块地基；这门课交到他手里的，却是两部前沿专著、一个课题组十年的反应条件档案，和一份只有几十条的精要库。没有教材，课堂靠什么站住，是本编要回答的第一个问题。

第1章先划分工：统一性知识的再发生是练手，应在初高中完成，练的是发生技术；大学是战场，九问打向连老师也写不出终点的知识。第2章为讲授体系记功，并说明AI接管讲授之后，讲台的价值落到出题台，教材让位于谱系母本。第3章讲这份母本由前沿专著与课题组档案合成，怎样一次建成、千人共用。第4章讲老师怎样从精要库按学生画像打造九问，每一问怎样过题闸，以及老师不可替代之处：他认得学生的现场。

四章由远及近：先说练什么、战什么，再说讲台让出了什么，再说地基换成了什么，最后说题怎样落到一个具体学生的手上。骨架立起来，后面各编才有东西可走。

01

第1章

练手与战场：发生技术前移到中学

◆ 一个被反问问住的好学生

设想化学系三年级一门课程型研究课的第1周。课堂采用九问模式，每个学生拿到自己的九问，先写初判，再与ChatSDE开始第1轮问对。班上有一位学生，中学时成绩一直名列前茅，高考化学几乎满分，进入大学后的有机化学、分析化学也都考得很好。他的初判写得工整而周全，列出了影响产率的五六个因素，每一条都有教材依据。ChatSDE的第一回合没有逐条评点，只问了一句：你列的这些因素里，哪一条能在本组的档案里找到证据，哪一条是你从教材上搬来的？他愣住了。他不是不会回答问题，他是从来没有遇到过一个回答不能从书上找到的问题。

同一个课堂里，沈澄的初判写得远没有那么周全，只有一句“产率低是因为试剂纯度不够”。ChatSDE问她“同一瓶试剂，为什么三月好、六月差”，她也愣了一下，但几分钟后，她在问对窗口里写下的是：“我不知道，但我可以去查本组十年档案里产率最低的那些批次各在哪个月份。”这句回答本身并不深刻，却说明她身上有一样东西：被反问之后，她知道下一步该往哪里走，知道答案不在书上而在档案里，知道一个初判被问倒不等于失败，而是工作的开始。

两位学生的差别不在智力，也不在知识量；若比化学知识的掌握，前一位可能更扎实。差别在于，面对一个连老师也写不出终点的问题时，有没有一套可以动用的技术：知道怎样把一个问题改得更能下手，知道怎样与一个会反问的对话者来回磨，知道被批判时怎样接住而不是防御，知道到哪里去取证据，知道怎样把零散的发现重写成一个成形的判断。本书把这一整套技术称为发生技术。困惑就在这里：大学的九问课堂要求学生一上来就动用这套技术，可这套技术从哪里来？如果它要在大学课堂上从头学起，大学课堂就得先花大量时间做本该更早完成的练习，战场就变成了训练场。

这个困惑在大学教师那里有一个熟悉的版本。许多老师都有过这样的经验：给本科生布置一个开放的研究题目，交上来的多半是一篇综述，把已有文献整理得井井有条，却看不到一个属于学生自己的判断；追问一句“你自己怎么看”，学生往往重复综述里的某一种观点。老师常把这归因于学生“缺乏创新意识”，但更准确的诊断也许是：学生从未练过怎样在一个没有终点的问题上动手。综述是发现路径上最成熟的产物，学生交出综述，恰恰说明他们把发现的技术练得很好，只是没有人教过另一套技术。

本章的判断是：统一性知识的再发生是练手，应在初中、高中完成，练成的是发生技术；大学是战场，九问打向连老师也写不出终点的谱系知识。这个判断有两层含义。一层是分工：中学与大学用的是同一套发生技术，但对象不同，中学处理老师事先能写出终点的知识，大学处理连老师也写不出终点的知识。另一层是次序：技术要先在终点已知的对象上练熟，才能在终点未知的对象上使用。下面先给发生技术下定义，再说明再发生与新发生的区别，然后用两个中学样板展示练手的样子，最后回到沈澄的课题组，看战场长什么样。

> **这一章的问题**
>
> 大学要学生一上来就在没有终点的问题上动手，这套发生技术该在哪里、用什么练出来？

◆ 发生技术的定义与五个构件

本书所说的发生技术，是指学生在一个判断从无到有的过程中所动用的一整套可练习、可评估的操作。它不是一种天赋，也不是一种笼统的“创新能力”，而是可以拆开、分别练、再组装起来的几个构件。拆开来看共有五个：改题、问对、接住批判、下现场取证、重写成文。五个构件对应九问模式工序链中学生一侧的主要环节，后面各章会分别展开，本章只给出定义，并说明它们为什么必须作为一个整体来练。

第一个构件是改题。拿到一个问题后，不是直接作答，而是先判断这个问题能不能下手、离自己的现场有多远，并在保持问题之根的前提下，把它改成更具体、更贴近现场、更有可能长出新判断的形式。改题的两条锚，一条是“根还在”，即

改后的问题仍连着出发点的核心判断；一条是“离现场更近”。第5章会把改题当作学生的一项权利来讨论。第二个构件是问对，即与对话者进行多轮、自由风格的研究型对话，先写下自己的初判，再在对话中推进，而不是一问一答地索取答案；第6章专门讨论。第三个构件是接住批判：当初判被翻过来看时，不防御、不放弃，而是把批判转化为下一步要查的东西。沈澄那句“我不知道，但我可以去查”，就是接住批判的最小样本。第7章会从ChatSDE的一侧讨论怎样批判才能被接住。

第四个构件是下现场取证。判断要在具体的物、具体的记录、具体的人那里得到支撑或被推翻，学生要知道自己的现场在哪里，以及怎样从中取出可靠的证据：观察、测量、记录、访谈、档案检索。第五个构件是重写成文：把问对与取证中零散出现的东西，重写成一个有结构、有论证、有证据的文本，其中要有出发材料所没有的判断。这是第12章讨论的子本打造的雏形。五个构件合起来，就是一个判断从被问出、被改写、被批判、被检验到被写定的全过程。

为什么必须作为整体来练？可以拿做菜打比方。切菜、调味、掌握火候、尝味、装盘，每一项都可以单独练，但只练刀工的人做不出一道菜，因为一道菜的成败恰恰在于各项之间的衔接：火候到了要知道什么时候下调料，尝出咸了要知道怎样补救。发生技术也是如此。只会改题的学生，改出的问题可能离现场很近，却接不住随之而来的批判；只会问对的学生，可能与智能体聊了20轮，却一次也没回到现场取证；会取证而不会重写的学生，交上来的是一堆材料而不是一个判断。五个构件之间的衔接本身就是技术的核心，因此练手必须是完整走一遍九问，而不是分别做五种练习。

五个构件之间还有一个容易被忽略的先后关系：第零轮的初判。无论练手还是战场，学生在问对开始之前都要先写下自己的初判，哪怕它粗糙、片面，甚至是错的。初判的作用不是被评分，而是给后面的一切留下一个起点：改题改的是相对于它的问题，批判批的是它的前提，取证检验的是它，重写成文时要说明自己从它走了多远。没有初判的学生，问对就会退化为向智能体索要答案，期末自述也无从对照，回写更无从观察。中学练手里最朴素的“你先猜一猜”，正是第零轮初判的雏形；把这个习惯在中学练成，大学九问课堂开学时，学生就不会把第1轮问对当成一次搜索。

这一点也提醒我们，发生技术与传统所说的“学习方法”不同。预习、做笔记、整理错题、归纳知识点，这些方法服务于掌握一个已知的终点，它们是发现路径上的好技术。发生技术服务于生成一个判断，它的每一个构件都预设了一个前提：终点不在手边，要在过程中出现。正因如此，发生技术即使在终点已知的对象上练习，练的也必须是“仿佛终点未知”的那种走法——下一节要说明，这正是练手与战场既相同又不同的地方。

◆ 再发生与新发生：同一套技术，两种对象

发生技术可以作用于两种对象。一种是统一性知识：阿基米德原理、酸碱中和的规律、血压调节的反馈机制，这些知识早已在学科中定型，老师在开课之前就能把终点一字不差地写在黑板上。让学生经由改题、问对、取证、重写，把这样的知识在自己身上重新发生一遍，本书称为再发生。另一种是谱系知识：本病房夜间跌倒落在哪个交接断点上，本课题组某一步的成败依赖什么未入规程的条件，这些知识在课程开始之前并不存在于任何教材、任何指南、任何老师的头脑中，只能在学生的现场与学科的谱系母本相遇时生成。让学生经由同一套技术把这样的知识发生出来，本书称为新发生。

再发生与新发生用的是同一套技术，这是练手可以为战场服务的前提。学生在浮力问题上学会的改题，与沈澄在偶联反应上动用的改题，是同一种操作：判断问题能不能下手，在保根的前提下把它拉近现场。学生在酸碱中和上学会的接住批判，与许宁被问“同样年纪为什么白天跌得少”时动用的接住批判，也是同一种操作。正如学游泳先在浅水区练换气与划水，到了深水区，动作一样，只是脚下再也踩不到底。

但两者的对象不同，这一差别决定了练手与战场的根本区分。再发生的终点已知，因此老师可以判断学生是否走到了，可以在学生偏离时把他拉回来，可以用终点作为标尺来评估过程。新发生的终点未知，老师无法用一个现成的终点来判断学生走得对不对，只能评估过程是否推进、证据是否扎实、判断是否新且经得起检验。这也解释了为什么两者在评估上必须分开：练手阶段看的是技术动作是否到位，战场阶段看的是有没有长出母本所无的判断。第15章讨论的八个评估项目在两种情境中都适用，但权重不同，大一补练手的课堂里，母本底盘与问对深度各占

20%，子本创新智商只占20%（这是一个例外，理由见第18章）；本科研讨课里，子本创新智商升到30%，母本底盘降到5%。这一权重的变化，正是练手与战场之别在评估上的投影。

维果茨基的“最近发展区”有助于说明练手为什么必须先于战场。在他的论述里，学习者独立能做到的与在更有能力者的帮助下能做到的之间，存在一段可以被教学推动的区域，教学的效力恰恰发生在这一区域之内。把这一思想移到发生技术上：一个从未练过改题与接住批判的学生，面对连老师也写不出终点的问题，这个问题离他独立能做到的太远，落在最近发展区之外，再好的陪伴也只能把他拖着走；而在终点已知的问题上，老师知道终点，可以精确地判断学生卡在哪一步、需要哪一点帮助，帮助就落在最近发展区之内。练手的意义正在于，它把发生技术放在一个老师能看清、能扶住的区域里练，等学生的独立能力扩展了，战场才落进他的最近发展区。

再发生还有一层容易被忽视的价值：它让学生第一次体会到，课本上的定论不是从天上掉下来的，而是某个人在某个现场里、经由类似的过程发生出来的。一个在家里用水盆和土豆重新“发现”浮力规律的初中生，会对阿基米德原理多一份亲切；更重要的是，他会知道定论背后有一条来路，而来路是可以重走的。这种体会是走向战场的心理准备。一个从未体会过“结论有来路”的学生，到了大学面对一个没有结论的问题，往往会觉得无从下手，甚至觉得被为难了。开篇那位被反问问住的好学生，正缺这一份体会。

◆ 中学九问练手样板：浮力与酸碱中和

以初中物理“浮力”为例，设想一位老师不再从定义和公式讲起，而是给每个学生九个问题，材料是课本和家庭小实验。第1问，家里的哪些东西放进水盆会浮、哪些会沉，你先猜一猜，再试一试，猜错了哪几样？第2问，同一块橡皮泥，捏成团会沉，能不能捏成别的形状让它浮起来？第3问，把一个土豆放进清水和放进浓盐水里，有什么不同？第4问，用弹簧秤挂着一个物体慢慢放进水里，读数是怎样变的，在哪一刻停止变化？第5问，溢出来的水有多重，和读数减少的量有什么关系？第6问，课本上说“浮力等于排开液体的重力”，你在哪一次实验里最先看出了这层意思？第7问，有同学说“重的东西沉、轻的东西浮”，你能用自己做过的实验驳倒他

吗？第8问，铁做的船为什么能浮？第9问，如果把实验搬到盐湖里做，你预计第4问的读数会怎样变？

这九问的终点，老师在出题之前就完全知道：浮力等于排开液体所受的重力，物体的浮沉取决于浮力与重力的比较。老师之所以不直接讲出来，不是为了故作神秘，而是为了让学生把五个构件各走一遍。学生先写初判，多半写成“轻的浮、重的沉”；第2问与第7问正是为翻转这个初判而设，相当于问对中的批判性激发。学生可以改题，比如把第2问改成“同一张锡箔纸，揉成团和折成小船，各能托住几枚硬币”，这一改既保住了根，又离自己的现场更近。学生要下现场取证，现场就是自家的水盆、厨房秤和一瓶盐水。最后学生要重写成文，写出自己是怎样从“轻的浮重的沉”走到“排开的水才是关键”的。这份小文没有母本所无的判断，它的终点与课本完全一致，但它是学生自己走出来的。

高中化学“酸碱中和”可以照此办理。安全起见，家庭实验只用食醋、小苏打、柠檬汁、肥皂水，用紫甘蓝汁作指示剂，不碰任何强酸强碱，教室里的滴定实验按实验室安全规范由教师带领完成。九问可以从“紫甘蓝汁滴进食醋和肥皂水里各是什么颜色”问起，经“往变色的食醋里一点点加小苏打水，颜色在哪一刻变回去”“两种溶液混合后，用紫甘蓝汁或pH试纸测起来，为什么既不像酸也不像碱”，走到“课本上的中和反应方程式，在你的哪一次操作里第一次看得见”，再到“胃不舒服的人为什么有时会服用碱性药物，它和你的实验是一回事吗”。终点同样已知：酸与碱反应生成盐和水，溶液的酸碱性随之改变。学生练的仍是那五个构件，尤其是第一次体会到“题可以改”：有学生把“颜色在哪一刻变回去”改为“要加多少勺小苏打水颜色才变回去，能不能事先算出来”，这一改把一个观察题提升成了一个定量预测题，根仍在，而且离课本上的计算更近。

练手课堂里老师的位置也随之改变。讲授“浮力”的老师，工作重心在把定义与公式讲清楚、讲生动；带九问练手的老师，工作重心在读学生的轨迹：谁的初判停在“轻的浮重的沉”而迟迟不动，谁改题改得离自己的现场越来越远，谁在第7问里第一次用自己的实验驳倒了同学。因为终点已知，老师可以把全部注意力放在过程上，看每个学生在五个构件里哪一个最弱，下一次练手就在那里多给一点支撑。这与驾校教练的工作很像：他早就知道车该停在哪条线内，他要看的是学员在哪一

个动作上总是慢半拍。由此也可以理解，中学练手并不减轻老师的负担，而是把负担从讲清结论移到看懂过程上；这一转移，恰好也是大学老师从讲台走向出题台所要完成的转移，只不过大学老师面对的是一个连自己也看不见终点的过程。

两个样板有三个共同特征，可以作为中学练手的标准。其一，终点由老师事先写出，并在练手结束时与学生对照，学生知道自己走到了还是没走到。其二，现场取材于学生伸手可及的地方，安全、便宜、可重复，学生不必依赖任何专门设施就能取证。其三，评估看的是五个构件是否走全、是否衔接，而不是答案是否正确——答案正确是练手的出口条件，技术到位才是练手的目的。由此也可以看出，中学练手与大学战场的界线不在题目难易，而在终点归属：再难的竞赛题，只要老师写得出终点，就仍是练手；再朴素的问题，只要连老师也写不出终点，就已是战场。

◆ 新手探究需要强指导：与最小指导批评的对话

把发生技术前移到中学，会遇到一个有实证背景的反对。Kirschner、Sweller与Clark（2006）在《教育心理学家》上发表的论文认为，最小指导的探究教学对新手是低效的：新手缺少可以引导搜索的图式，放手让他们在开放情境里自己摸索，占去的主要是工作记忆，而不是知识的建立。这个批评直接冲着本章而来。既然发生技术包括改题、问对、取证、重写，而中学生恰恰是学科上的新手，把这一整套交给他们，岂不是让一个从没下过厨的人“自己琢磨怎样把鱼做熟”？本书不回避这个质问，因为它若成立，练手前移就是一个看上去漂亮、实际上有害的建议。

文献本身给了一个更细的刻度。Lazonder与Harmsen（2016）对探究式学习的综述发现，探究的效果依赖指导量，适度指导的探究优于无指导。这就把问题从“要不要探究”改成了“给多少指导、给哪一种指导”。把两项研究放在一起读，被批评的靶子是“最小指导”，不是“探究”：一个让新手独自面对开放问题、老师退到场边的课堂，与一个老师知道终点、随时在场、支架逐级撤去的课堂，是两种东西，却常常共用“探究”一个名字。本书承认前者对新手低效，并不为它辩护。

本书的回应是：中学的九问练手不属于最小指导，它是强指导下的再发生，在三个地方与Kirschner等批评的对象不同。其一，终点已知。老师在出题之前就能写出终点，因此能精确判断学生卡在五个构件中的哪一个，哪里需要一个提示，哪里需要一个更小的追问，指导可以落在学生真正需要的地方，而不是均匀地撒下去。其二，题闸放宽。练手的九问不要求“拿掉现场材料，通用大模型就答不出”，恰恰相反，它们要保证题目可下手，学生伸手可及的水盆、厨房秤、紫甘蓝汁就是现场，题闸的严格形态留给战场。其三，问对多接住。练手阶段的问对者在学生接不住批判时，可以把批判拆成更小的几步，可以给出支架，必要时甚至可以说破终点；这与战场上“逼近判断但不替学生说破”的纪律刚好相反，因为练手要练的是技术动作，不是第一次发现一个结论。

这里仍然保留了生产性失败一线的洞见，即先尝试后指导。练手课堂里“你先猜一猜”的第零轮，意味着无论后面给多强的支架，都发生在学生已经写下自己的初判之后。强指导与先尝试并不冲突：指导的强度调的是尝试之后的接应，尝试本身一个也不能省。第6章会专门讨论这一点。可以拿学骑车作比：家长扶着后座跑，不是代替孩子骑，也不是撒手任他摔，孩子的手脚始终是自己的，扶住后座的那只手只在重心偏过去的那一刻才用力，等他骑稳了，手就松开。Kirschner等批评的是一上来就撒手；九问的练手是扶着后座，一寸一寸松。

战场是否就成了最小指导？也不是。大学里的九问课堂，学生有ChatSDE的一破一立，有老师认得现场的出题与追问，有同学的碰撞场，指导一样不少，变的是指导的种类：练手里的指导可以包含内容，战场上的指导只能是方法性的，不能替学生把终点说出来，因为连老师也不知道终点。同时，战场上的学生在母本底盘上已经不是新手，这是母本底盘被设为门槛项的另一层理由：它既是新判断的承重结构，也是把一个学生从“新手”挪出Kirschner等批评范围的条件。用Lazonder与Harmsen的话来说，两个阶段都是“适度指导的探究”，只是适度的含义随学生的专长和终点的归属而移动。

必须说明这一回应的限度。两项研究讨论的主要是学科内容的学习，并没有直接检验“发生技术”能否经由强指导的练手而迁移到终点未知的问题上；本书可以

说，这两篇文献的批评没有击中强指导下的再发生，却不能说它们证实了练手前移。迁移是否成立，仍然是一个经验问题，本章末尾给它写了读数。

◆ 判别第1问，与学车的比喻

导论提出判别两问：终点能否事先写出？课后有无回写？本章要指出，第1问不仅区分发现版与发生版的课堂，同时也区分练手与战场。设想同样采用九问形式的两堂课，一堂的九问由老师从课本出发，九问指向的终点老师心里有数，这是练手；另一堂的九问由老师从谱系母本出发，经过题闸的检验，拿掉学生的现场材料，一个未接入这些材料的通用大模型答不出，老师自己也写不出，这是战场。形式一样，对象不同；判别第1问正是为这个不同而设的。

学车是理解这层关系最贴切的比喻。驾校的场地里，倒车入库、侧方停车、坡道起步，每一项都有标准动作和标准结果，教练事先知道车应该停在哪条线内，学员一次次练到停准为止。这是练手：终点已知，练的是技术。拿到驾照之后上路，路况是事先写不出来的，前面那辆车会不会突然变道，路口那个行人会不会闯红灯，没有任何教练能提前告诉你。这是战场：终点未知，用的仍是场地里练出的那套技术——看镜子、控速度、判断距离——只是现在每一次判断都关乎一个此前没有出现过的局面。

这个比喻还说明了两件事。第一，场地练习不可省略。没有在场地里练熟起步和转向的人直接上路，不是勇敢，而是危险；没有在终点已知的问题上练熟发生技术的学生直接进入九问课堂，往往只会做两件事：要么把每一问都答成一篇综述，要么在问对里被一次反问问到放弃。第二，场地练习不能代替上路。一个在场地里能把车停得分毫不差的人，若从未上路，他仍然不会开车。同理，一个在中学练手中表现优异的学生，若大学仍让他在终点已知的题目上反复演练，他永远不会知道自己的技术能在真实的路上走多远。因此练手与战场既不能颠倒次序，也不能互相替代。

比喻也有它的限度，需要说明。驾校场地与上路之间是一道明确的考试关口，拿到驾照就算出师；发生技术的练手与战场之间却没有这样一张证书，一个学生可能在改题上已能上路，在重写成文上还停在场地里。因此本书不主张设一道“发生

技术资格考试”，而是在大学的评估方案里保留母本底盘、改题质量、问对深度等过程项，让每一个构件在战场上仍然被看见、被计分。更要紧的是，开车的规则是全社会一致的，而发生技术的“路况”随学科而变：病房里的现场是交班本与夜班，实验室里的现场是记录本与失败批次，法院里的现场是公开裁判文书与旁听记录。同一套技术在不同学科的战场上会长出不同的手感，第六编的学科样板会分别讨论。

把练手前移到中学，还有一个制度层面的理由。中学的课程内容以统一性知识为主，这些知识的终点在全国范围内基本一致，恰好是练手最合适的材料；大学的课程越往高年级走，越接近学科前沿与学生自己的现场，恰好是战场最合适的位置。让每个学段做它最适合做的事，比让大学用宝贵的课时补中学该练的功夫要经济得多。至于今天的大学生大多没有在中学练过发生技术这一现实，本书承认它，第25章专门讨论过渡期：大一须用一两门基础课补练手，医学院大一新生吴静用“血压调节”走一遍九问，第一次体会到“题可以改”；待中学跟上之后，这一补课即可取消。

◆ 沈澄的战场：课题组里的十年档案

回到沈澄。她所修的《合成条件的发生》是一门课程内嵌真实课题的课程型研究课，出发点是课程型本科研究与合成方法学的前沿专著，加上本课题组十年的反应条件档案。她的九问里，第1问问的是产率最低的那些批次各落在哪些月份，第3问追问“室温”一词十年里有没有一次附过具体的温度和湿度，第5问问师兄师姐口头叮嘱的“那一步要快”规程里写没写。这三问与浮力九问在形式上并无二致，都是具体的、可下手的、指向一个可取证现场的问题；区别在于，没有任何人能事先写出它们的终点。老师不知道，教材里没有，一个没见过本组档案的通用大模型也给不出说得过去的答案——这正是题闸要检验的事（第4章详述）：拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）给不出说得过去的答案。

沈澄动用的技术，全是在终点已知的题目上可以练熟的那一套。她改题，把“产率低的批次在哪个月份”与“同一瓶试剂三月六月的差别”连起来看；她问对，从第零轮“试剂纯度”出发，经过多轮来回，第6章会完整呈现这七轮问对；她

接住批判，没有因为“三月好、六月差”这个反问而放弃自己的初判，而是把它转化为一次复做的设计；她下现场取证，复做一批，把每一步的时间、温度、湿度都记下来，再与档案比对，复做严格按课题组安全规程进行，不改动任何危险步骤；她重写成文，最后交出一份子本，其中写着一条母本里没有的判断：本组某一步的成败依赖未入规程的环境湿度与操作时长——规程里没写的那一句，在师门口头里传了十年。

这条判断的地位需要说清。它不是定论，而是待复做验证的假说；它是否能写进课题组规程，要等下一届学生复做之后才见分晓，这正是第24章所说的回写要被观察、不能被宣布。但即使它最终被下一届推翻，沈澄这一学期也在战场上完成了一次新发生：她的判断在开学时不存在，在期末时存在了，而且它带着一条可以被检验的来路。这一点是任何练手都给不了的。浮力九问的学生走到终点时，会发现终点早已写在课本上；沈澄走到终点时，发现那里原本什么也没有，是她把一个判断放在了那里。

这里还可以借题闸做一个对照，以说明战场与练手在出题上的分界。设想沈澄的九问里混进一问“影响偶联反应产率的因素有哪些”，它在形式上同样具体，同样指向化学，但拿掉沈澄的现场，一个通用大模型可以立刻给出一份周全的清单，老师自己也写得出答案，所以它必须被打回重出，至多留作中学或大一练手的材料。反过来，“实验记录里‘室温’一词，十年里有没有一次写过具体温度和湿度”这一问，离开本组十年的记录本，任何人都答不出；它的答案藏在只有沈澄能去翻、能去数的那一摞档案里。两问的差别不在难度，而在终点归属：前一问的终点在教科书里，后一问的终点要由沈澄去生成。第4章会把题闸当作出题台的核心工序来讨论，这里只用它说明，判别第1问在大学课堂里有一个可以操作的形式。

把沈澄与开篇那位好学生放在一起看，本章的判断就有了具体的样子。那位学生的困难不是知识不够，而是技术没练；沈澄的从容也不是天赋更高，而是技术已熟。如果发生技术的练习在中学完成，大学九问课堂开学第1周的那一句反问，就不会把任何一个学生问懵，而会把每一个学生送回他自己的现场。

◆ 本章怎样算错

练手前移的主张，最终要看中学里练过的学生在大学第1周的问对记录里留下了什么。下面四条都按这个读数来写，任何一条成立，本章对应的部分就要收缩或撤回。

第一条管迁移。取同一所大学的同一门九问课程，连续两届，每届新生不少于30名，按其中学阶段是否完整走过不少于3次九问练手分为两组，每组不少于10人；由两位互不知情的评阅者，对第零轮之后前三轮的问对记录评改题质量与问对深度两项。若练过组在两项上的平均分与未练组之差，不超过两位评阅者之间的分歧幅度，就说明同一套技术并没有从练手迁移到战场，“同一套技术、两种对象”的判断不成立，练手前移的建议要撤回或大幅修改。

第二条管指导量。取同一批中学材料，一组按本章所说的强指导练手，另一组只给题、不给支架，每组不少于30名学生，一个学期后做一次终点已知的后测和一次迁移任务。若强指导组在迁移任务上不优于对照组，强指导的理由就落了空；若反而低于对照组，Kirschner等的批评就适用于这里，练手的指导量要下调。这一条也是对前一节回应的直接检验。

第三条管判别第1问能否稳定执行。把同一组至少9问交给3位互不知情的老师，分别判别“终点能否事先写出”。若三人判断一致的问不足80%，或被至少一人放进灰色地带的问超过20%，练手与战场的二分就必须细化为一个连续的谱，本章的界线与后续各章关于学段、权重的安排要随之调整。

第四条管构件能否前移。取不少于3个中学班级，记录一个学期内的练手轨迹。若下现场取证或重写成文两个构件中，任何一个有一半以上的学生走不全，而这一缺口不能用课时不足或指导不足来解释，本章“练手应在初高中完成”的主张就收缩为“部分构件前移”，其余构件仍须在大学补练。证据来自练手课堂的轨迹记录，不来自对中学生能力的一般估计。

上面各条的读数，都落在大学九问课堂开学第1周的问对记录和中学练手课堂的轨迹记录上。如果练过与没练过的学生在第零轮之后的头几轮问对里看不出差别，本章就要回到起点重新论证。

本章把发生技术的练习交给中学，把战场留给大学。但这样一来，一个更尖锐的问题浮上来：大学课堂原本最主要的工作——讲授统一性知识，讲清一门学科的体系——到哪里去了？如果九问课堂的出发点是谱系母本而不是教材，如果统一性知识的讲解大模型做得更快更全，那么那位在讲台上讲了20年病理学、讲得极好的教授，他的讲授还有什么位置？第2章要回答这个问题，并且先为讲授体系记功，再讨论它的退场。

> **带走的话**
>
> **发生技术先在终点已知的题上练熟，再到终点未知的题上用：中学练手，大学上战场，两者的界线不在题目难易，而在连老师也写不出终点。**

◆

02

第2章

教材退场之后：统一性知识移交大厂

◆ 两堂病理学课

设想医学院二年级的一堂病理学课，讲的是动脉粥样硬化。讲台上是一位讲了20年病理学的教授，他的课在学生中口碑极好：从血管内皮损伤讲起，讲脂质沉积、泡沫细胞、纤维帽的形成，再讲斑块破裂与血栓，最后落到心肌梗死与脑梗死的临床表现；每一步都配有他多年积累的切片照片，讲到关键处，他会停下来问学生“如果纤维帽薄了，会发生什么”，等学生答出“容易破裂”，再翻到下一张幻灯片。一堂课下来，体系清楚，逻辑严密，学生笔记整齐，期末考试的相关题目得分率很高。这是讲授课堂最好的样子。

再设想同一个晚上，一个学生打开一个通用大模型，输入“请像病理学教授一样，给医学二年级学生讲动脉粥样硬化的病理过程”。几秒钟后，屏幕上出现一份结构同样清楚的讲解：内皮损伤、脂质沉积、泡沫细胞、纤维帽、斑块破裂与血栓，一步不少。学生追问“泡沫细胞到底是怎么形成的”，它换一种讲法再讲一遍；学生说“还是不懂，能打个比方吗”，它用水管里的水垢打比方；学生说“出几道选择题考考我”，它立刻出题，学生答错，它解释错在哪里。它不会因为第五次被问同一个概念而显出不耐烦，也不必等到下周的答疑时间。

把两堂课放在一起，困惑就出现了。我们不必断言大模型讲得比那位教授更好，在切片照片的选择、在关键处停顿发问的时机、在学生眼神里读出困惑的能力上，教授可能仍有大模型不及之处。但就“把一个已有的体系讲清楚”这件事而言，两者之间的差距正在迅速缩小，而在速度、覆盖面与耐心上，大模型已经占了上风。更关键的是，教授的课一学期只讲一遍，大模型的讲解可以在任何时刻、按任何学生的节奏、以任何讲法重来。如果大学课堂的主要价值就是这位教授在讲台上做的事，那么这一价值正在被大规模、低成本地复制。

这种困惑在教师中间往往以两种相反的情绪出现。一种是焦虑：既然大模型讲得又快又全，自己20年的讲授功夫还有什么用。另一种是轻视：大模型讲的不过是教科书上的东西，没有切片、没有临床经验，学生迟早会发现它不可靠。两种情绪都有道理，也都没有抓住要害。焦虑的人把自己的价值全押在讲台上，忘了一位讲了20年课的教授，最宝贵的其实是他认得这门学科里哪些地方还没有定论、认得自己的学生站在什么现场里；轻视的人则低估了统一性知识移交的规模与速度，它不会因为个别讲解出错而停下。本章要做的，是在这两种情绪之间找到一个可以论证的位置。

本章的判断是：讲授体系的功绩是传递统一性；AI接管之后，讲台的价值落到出题台。这个判断包含三层：讲授体系有真实而巨大的历史功绩，不能因为它正在被替代就抹杀；AI接管的是发现一侧，统一性知识的传递由大厂的模型承担，发现不作废，而是降为底盘；大学课堂因此要换一个出发点，教材让位于谱系母本，讲台让位于出题台。下面依次论证。

> **这一章的问题**
>
> 大模型把讲授做得更快、更全、更耐心之后，大学课堂的价值还落在哪里？

◆ 讲授体系的历史功绩

讲授课堂是本书所说的S→D→E路径的发现版：先把一门学科的体系显露出来，写成教材、讲义与课纲；再沿着章节的差异序列一步步推进，由浅入深，由一般到特殊；最后落到应用，即习题、实验、见习与考核。这条路径的终点在开课之前就写在教材里，学生走完它，是把终点搬进自己的头脑。说它是“发现版”，不是贬义，而是指出它的方向：去找到、认出、掌握一个先在的终点。

这条路径的功绩，首先在于它让统一性知识得以大规模传递。一门学科在几百年里积累的定论，如果只能靠师徒之间口耳相传，能接触到它的人永远是少数；把它整理成体系、写成教材、在课堂里按章讲授，一位教授一次就能把它交给几十上百人，一本教材可以在全国几百所学校同时使用。现代医学教育之所以能在较短时间内培养出大批合格的医生，一个重要原因正是病理学、生理学、药理学这些基础学科有了统一的体系与统一的讲法。这是一项文明级的成就，类似于标准化的度量衡让千里之外的两个商人能够做同一笔生意。

其次，讲授体系建立了一个学科共同体的共同语言。两位素未谋面的医生能在会诊中迅速沟通，是因为他们在不同的医学院学过同一套病理学术语、同一套疾病分类；库恩在《科学革命的结构》（1962）中所说的常规科学，正依赖于这种由教科书传递的共同典范。没有这个共同底子，学科内部的交流、协作与争论都无从进行。讲授体系把每一代新人带进这个共同体，让他们能听懂前辈在说什么，也能让自己的话被同行听懂。

再次，讲授体系为后来的一切创新提供了底盘。一个不懂动脉粥样硬化病理过程的医生，不可能提出关于斑块稳定性的新假说；一个不懂偶联反应机理的化学学生，不可能追问谱图上那个反复出现的小峰。本书反复强调的“发生”，从来不是在空地上发生的，它总要以某个已有的体系为出发点。九问模式的第一个评估项目“母本底盘”，正是对这份功绩的承认：在大学阶段，它是门槛项，过不了门槛的学生不进入后面的评分。

讲授体系还有一项常被低估的功绩：它把评估变得可以比较。同一套教材、同一份考纲，使得不同学校、不同年份的学生可以用同一把尺子量，用人单位与上一级学校也能据此判断一个学生掌握了多少。这把尺子的前提是终点统一，“像不像答案”才有意义。它在统一性知识的时代运转良好，也恰好说明了九问课堂为什么必须另造一把尺子：当每个学生的终点各不相同、且事先没有人写得出时，“像不像答案”就失去了对象。第五编讨论的发生剖面，并不是要否定可比较性，而是要在终点不统一的前提下重建可比较性——比的不再是离同一终点有多近，而是离已有有多远、证据有多厚、轨迹有多清楚。

承认这些功绩，也就承认了讲授体系的边界。它擅长的是让许多人掌握同一个终点，因此它的产出是“掌握了同一本书的人”；它不擅长、也从未承诺要做的，是让每个人在自己的现场里生成一个此前没有的判断。这不是讲授体系的失败，而是它的定位。问题在于，当它所擅长的那部分被大模型以更低的成本承担之后，大学课堂若仍只做这一部分，就失去了存在的独特理由。

◆ AI讲授：更快、更全、更耐心

说大模型在讲授上占了上风，需要说清是哪几个方面，以免把判断说过头。第一是速度。一位教授备一堂新课，往往要几天甚至几周，修订一章讲义要等到下一

轮开课；大模型生成一份结构完整的讲解只要几秒钟，学生的追问随到随答。第二是覆盖面。一位教授的专长总有边界，讲到交叉领域时只能点到为止；大模型可以在病理学与分子生物学、药理学、影像学之间来回切换，把一个概念放进好几个学科的语境里讲。第三是耐心与个别化。一个40人的大课，教授只能按中等进度讲，跟不上的学生只能课后自己补；大模型可以按每个学生的节奏讲，第五遍、第十遍都不嫌烦，还能根据学生答错的地方换一种讲法。

这三点合起来，正是布卢姆（1984）提出“两个西格玛问题”时所设想的那种个别辅导。布卢姆注意到，接受一对一辅导的学生在成绩上大幅优于常规班级授课的学生，难题在于个别辅导的成本太高，无法普及。可汗学院在2023年推出Khanmigo，正是要用大模型把这种辅导带给每个学生。就统一性知识的传递而言，这条路是走得通的；随着模型能力与教学设计的改进，它还会越走越通。一位学生在深夜对着屏幕弄懂泡沫细胞的形成，这件事本身是好事，本书没有任何理由反对。

但要注意，这里说的是“讲清一个已有的体系”这件事，而不是大模型的全部能力，也不是对大模型的定论。大模型同样可以在其他配置下做许多别的事；第三编会讨论，当学生带着一个未经检验的前提来问时，默认配置（未开启学习模式）下的通用大模型，往往会质疑前提，却在同一回合替学生把答案说破，而课堂需要的是批判性激发和把下一步留给学生，这一差别落在默认行为与工序层上，不在底座能力的高下；下一节会看到，2025年起各家学习模式已经在改变这一默认。本章只关心讲授这一环：在这一环上，大模型做得更快、更全、更耐心，这是一个应当坦然承认的事实。

不过，讲授一环被接管，也暴露出讲授本身一个长期被掩盖的局限：它回答问题，却很少教人提问。大模型的讲解再耐心，也要等学生先问；一个学生若只会问“泡沫细胞是怎么形成的”，他得到的就永远是关于泡沫细胞的已有知识。讲授课堂里，这个局限被教授的发问部分弥补了，他会在关键处停下来问“如果纤维帽薄了会怎样”；可即便是这样的发问，终点也早已写在下一张幻灯片上。换句话说，大模型接管讲授之后，讲授课堂本来就没有承担好的那件事——让学生学会在没有终点的地方提出问题——就再也无处藏身了。它原来藏在“我们至少把知识讲

清楚了”这句话的后面，现在这句话由大模型来说，大学课堂只能直面那件没做好的事。

这里还要防止一种把讲授与大模型对立起来的读法。在许多学校，最先用好大模型的恰恰是那些讲得最好的教授：他们用它生成变式题、检查讲义里的疏漏、为不同基础的学生准备不同深度的补充材料。讲授体系与大模型之间不是取代与被取代的零和关系，而更像手工抄写与印刷术的关系：印刷术出现之后，抄写员的工作消失了，书写本身却变得更重要，因为更多的人开始读、开始写。讲授的那一部分被接管之后，教授们手里多出的时间与精力，正可以投到出题台上去。

由此得出本章标题的含义：统一性知识移交大厂。所谓大厂，指训练与运营大模型的那些机构；所谓移交，指统一性知识的组织、讲解、答疑、练习与纠错，越来越多地由大厂的模型承担，而不再主要由大学课堂承担。这正是导论里自来水的比方：统一性知识的重要性一点没有减少，变的是它的取用方式，不再需要每家专门安排人手去挑水。《大学的典范革命：从统一性到谱系发生》一书在宏观尺度上讨论了这一移交对大学制度意味着什么，本章只在课堂尺度上讨论它对讲台意味着什么。

◆ 证据怎么说：AI辅导已经做到了什么

上一节说大模型在讲授上占了上风，不能只凭印象，应当看2023年以来的几条实证线索，也要看它们能证明多少。第一条是产品与教学系统的线索。可汗学院的Khanmigo（2023）把大模型做成随叫随到的辅导者；Liu等（2024，SIGCSE）报告了哈佛CS50课程的做法，用带教学护栏的AI助教，也就是那只“小黄鸭”，引导学生而不直接把答案给出。2025年，三家主要的模型厂商先后把同一思路做成了可开启的学习模式：Anthropic的Claude学习模式（2025年4月，随Claude for Education推出）、OpenAI的ChatGPT Study Mode（2025年7月）、Google的Gemini Guided Learning（2025年8月）。它们的共同点是先追问、少给答案。这意味着讲授与辅导已不只是“大模型能做”，而是主流产品把“怎样教”写进了配置里。

第二条是对照研究的线索。Kestin等（2025，Scientific Reports）在哈佛一门物理课上做了随机对照，结论是精心设计的AI导师带来的学习收益，高于课堂上

的主动学习。Bastani等（2025，PNAS）的高中数学现场实验给出了另一半图景：不加护栏的通用大模型辅助练习时表现更好，撤掉AI后考试成绩反而更差；而加了教学护栏的GPT Tutor，基本消除了这种损害。两项研究合起来说的是：AI辅导的收益是真实的，设计是否得当决定了它是帮手还是拐杖。后一条尤其值得记住，因为它让“护栏”成了一个有实证含义的变量，而不只是一句修辞，第9章会回到这里。

这些证据是否支持本章的判断？本书认为支持，并且读法需要小心。支持之处有三。首先，讲授与辅导的发现一侧在设计良好时已经被AI做得很好，好到在物理这样的课上超过了常规的主动学习课堂，教授在讲台上传递统一性知识的价值因此确实正在向大厂的模型转移，这是本章标题“移交”的实证底色。其次，“怎样教”写进配置，这一事实把差别落到了配置层，而不在底座，这与本书的工序层主张同向。再次，Kestin等的结果也堵住了一条退路：不能指望把课堂退守到“做主动学习”，因为被AI导师超过的恰恰是主动学习，讲台让位之后，课堂要去的地方得是AI在默认配置下不交付的东西。

限度同样需要说清。这几项研究衡量学习所用的，都是终点预先写定的测验：物理概念题、数学习题、编程作业，学习好坏是看学生离标准答案有多近。也就是说，它们回答的都是发现一侧的问题，对终点未知的新发生一无所言。它们也各自局限于一个学科、一个学段，不能直接搬到医学病房或化学课题组。本书因此把它们当作先验使用，而不是当作证明：先验告诉我们，在发现一侧与AI竞争没有胜算；至于大学课堂在出题台上能交付什么，要靠题闸、第零轮基线和回写痕迹这些本书自己的读数去查。

由此，本章要守的位置也就清楚了。讲授与辅导的发现一侧，课堂不必也不应与AI争：这一块AI做得更快、更耐心，在设计得当时收益还更高。本书要守的是出题台，因为出题要求认得学生的现场，要求判断哪些问题拿掉现场之后，一个未接入这些材料的通用大模型就答不出来，要求把谱系母本里尚无定论的部分变成学生能下手的问题。这三件事，学习模式的“先追问”没有涉及：它们接过的是一个已经写好的问题，再替学生把路走得更稳。至于出题台自己会不会也被AI接管，本章的回答是：模型可以协助起草九问，却看不到许宁的交班本和沈澄的记录本，认得现场的人才能判定一道题合不合格。

回到开篇那位讲了20年病理学的教授，这一节的结论落在他身上，是一个并不悲观的安排。泡沫细胞怎样形成，学生夜里可以向设计良好的AI导师问到弄懂为止，这件事从他肩上卸了下来；可是他认得哪几张切片是本院20年里最典型的，认得哪一类斑块破裂在本院病例里总被写成“原因不明”，认得今年这一届学生里谁已经读过前沿综述、谁还停在教材。这些东西在任何一个通用模型的训练语料里都找不到，却恰好是出一道“拿掉现场就答不出”的题所需要的材料。就像一位老裁缝，成衣厂接管了量产之后，他手里最值钱的活计变成了给每个人量身，而量身要的是认得这个人。

◆ 发现降为底盘，而非作废

统一性知识移交大厂，最容易被误读为“讲授没用了”“基础知识不重要了”。这两种说法都是错的。本书的判断是：发现不作废，降为底盘。“底盘”这个词借自汽车：底盘不是车上最显眼的部分，乘客上车时看的是内饰与座椅，没有人会盯着底盘看，但没有底盘，车根本走不了。统一性知识在九问课堂里的位置正是如此：它不再是课堂交付的主体，却是一切新判断能否站住的承重结构。

这个判断在评估方案里有直接的落点。第一个评估项目“母本底盘”，量的是学生对谱系母本精要与核心判断的掌握，证据是复述加反例口试：学生不仅要能复述核心判断，还要能针对老师给出的反例说明这个判断在哪里适用、在哪里失效。在大学阶段，它是门槛项。本科研讨课里它只占5%的权重，研究生课里权重为0，但这并不意味着它不重要；恰恰相反，权重低是因为它是门槛，过不了门槛的学生，后面七个项目都不进入评分。大一补练手的课堂里，它的权重升到20%，因为过渡期的学生正需要在统一性知识上练发生技术。

用沈澄的例子可以看清底盘在九问课堂里的实际作用。ChatSDE问她“同一瓶试剂，为什么三月好、六月差”时，她之所以能想到湿度，而不是随口归因于运气，是因为她懂得这类偶联反应中某些中间体对水敏感，懂得操作时长会影响中间体在空气中暴露的时间。这些都是统一性知识，写在任何一本合成方法学教材里，大模型也能讲得清楚。没有这份底盘，“三月好、六月差”这个反问对她只是一个怪现象；有了这份底盘，它就变成一条线索。底盘不提供答案，它决定一个学生能

不能把批判听成线索。这正是本书坚持把母本底盘设为门槛项、而不是把它取消的理由。

那么学生怎样获得这份底盘？不再主要靠课堂讲授，而是靠学生与大模型的自主学习、靠母本精要库的阅读、靠必要的少量讲授与口试。这里要防止另一种误读：说讲授降为底盘，不等于说教授不能再讲课。在某些关键节点上，一位教授用20分钟把一个核心判断的来龙去脉讲透，比学生自己摸索两周更有效；在医学的形态学、化学的实验操作这类需要现场示范的内容上，讲授与示范仍然不可省略。变化在于讲授的地位：它从课堂的主体变为课堂的前置条件，从一学期的全部内容变为九问开始之前的准备工作。

还有一点需要说明：发现降为底盘，不等于底盘可以随便。一个底盘不牢的学生，在九问课堂里会出现两种典型的失败。一种是问对时接不住批判，因为他不知道自己的初判在学科体系里处于什么位置，ChatSDE一反问，他就无从判断这个反问是在挑战一个核心判断还是一个细节。另一种是子本里出现所谓的“新判断”，其实只是对学科常识的误解，换手评审时一眼就能看出。所以母本底盘的门槛不是形式，第3章讨论的母本精要库，正是为了让这份底盘既可以高效获得，又可以被严格检验。

◆ 大学教材让位于谱系母本

如果统一性知识移交大厂，大学课堂的出发点就不能再是统一教材。统一教材的特点，是它对全国所有学生都一样：同一本病理学教材，在北京与在边远省份的医学院讲的是同一套内容，它的价值恰在于这种统一。但正因为它统一，它也恰好是大模型最容易覆盖的东西；一本统一教材的全部内容，几乎都可以在通用大模型的讲解里找到对应。若九问从统一教材出发，题闸就很难通过：拿掉学生的现场材料，一个未接入这些材料的通用大模型多半能给出说得过去的答案。

大学课堂需要的出发点，本书称为谱系母本。它由两个来源合成：一是学科前沿专著，代表本学科最新的、尚未沉淀为教材的判断与争论；二是课题组（部落）档案，即本科室、本课题组、本教研室历年积累的记录、失败、口头传承与学生问对档案。许宁的谱系母本是老年医学关于跌倒与衰弱的前沿专著，加上本科室历届轮转学生的问对档案；沈澄的谱系母本是课程型本科研究与合成方法学的前沿专

著，加上本课题组十年的反应条件档案。两个来源的比例逐年向部落档案倾斜，因为前沿专著终会沉淀为教材、进入大模型的覆盖范围，而部落档案只在本部落里越积越厚。

统一教材与谱系母本的区别，可以用家常菜谱打比方。统一教材像一本全国发行的标准菜谱，每道菜的配料与步骤写得清清楚楚，任何人照着做都能做出合格的菜；谱系母本则像一个家族三代人传下来的菜谱本，前面抄着标准做法，后面写满了奶奶的批注、妈妈的改法、哪一年哪道菜失败了的记录，以及“这道菜用本地的井水做才对味”之类只在这个家里成立的经验。标准菜谱任何人都能买到，大模型也都读过；家族菜谱只在这个家里，而它恰恰是新菜生长出来的地方。第3章会讨论怎样从谱系母本中提炼核心判断，建成一个一次建成、千人共用、逐年回写的母本精要库。

谱系母本还有一个统一教材不具备的特征：它是活的。统一教材修订一次往往要几年，修订时增删的是学科共同体已有定论的部分；谱系母本每一届都在变，因为每一届学生的问对档案、子本与被采纳的判断都会回写进去。沈澄这一届关于湿度的假说，若被下一届复做证实，就会写进课题组规程，成为下一届母本的一部分；若被推翻，推翻的记录同样会留在档案里，提醒下一届不要重走这条路。教材是一份定稿，谱系母本是一本一直在写的账，这正是它能为九问提供出发点、又不会被大模型一次读尽的原因。

谱系母本向部落档案倾斜，并不意味着前沿专著可以省去。前沿专著提供的是学科共同体在最近几年里尚在争论的判断，它让九问有方向，不至于在本部落的经验里打转；部落档案提供的是只在此地成立的记录与失败，它让九问有现场，不至于悬在空中。许宁的九问里，第9问“如果把跌倒风险评估从‘入院一次’改成‘每次交接一次’，本病房第一个会被改变的是什么”，方向来自老年医学关于风险动态评估的讨论，现场却在本科室的交接流程里；沈澄的第8问“别的实验室做同样的反应，他们的规程里多了哪一句”，则是把前沿文献当作一面镜子，照出本组规程里缺的那一句。两个来源缺一不可，它们在母本中的比例怎样随年份调整，第3章会具体讨论。

教材让位于谱系母本，讲台也就让位于出题台。讲台上的老师，核心工作是把统一教材讲清楚；出题台前的老师，核心工作是从谱系母本中提炼核心判断，再按

每个学生的画像与现场，用装有SDE工序层的大模型打造九问，并保证每一问都通过题闸。这项工作大模型可以协助，却不能代替，因为它要求老师认得学生的现场：知道许宁所在病房的交班本是怎样写的，知道沈澄所在课题组的档案里有哪些未解之谜。第4章专门讨论出题台。讲台的价值没有消失，它落到了出题台上。

◆ 许宁在讲授课堂里学到了什么、没学到什么

把以上论证落到许宁身上。在她的讲授课堂里，老年医学的课讲得极好：跌倒的流行病学、多重用药、视力与步态、认知障碍、夜间如厕、环境因素，每一类危险因素都有定义、机制与评估工具，最后还有几份权威指南的要点。许宁学到了这一切，而且学得很扎实。如果在期末口试中被问“老年人跌倒有哪些危险因素”，她能有条理地答出一大段；如果被给出一个反例，例如一位并不高龄却频繁跌倒的患者，她也能从多重用药或神经系统疾病的方向去解释。这份底盘是真实的、有用的，本书完全承认它的价值；在九问课堂里，它正是“母本底盘”这一门槛项要检验的内容。

但许宁在讲授课堂里没有学到一样东西：怎样从她所在的这间病房里长出一个判断。讲授课堂告诉她“高龄是危险因素”，却没有让她追问“同样年纪的患者，为什么白天跌得少”；讲授课堂给了她一份通用的风险评估工具，却没有让她去看本病房的交班本上，跌倒前一班对患者写了什么、没写什么。原因不在讲授者不用心，而在讲授课堂的结构：它的终点是统一的，而“本病房夜间跌倒的风险落在哪里”这个问题的终点，不在任何教材里，只在这间病房里，只在许宁去翻、去跟、去问之后才会出现。

更具体地说，讲授课堂教给许宁的是“老年人跌倒有哪些危险因素”，这恰好是题闸要打回的那种问题：拿掉许宁的现场，通用大模型张口就能答。九问课堂交给她的是“交班本上写了什么、没写什么”，这是题闸认为合格的那种问题：通用大模型看不到本病房的交班本。两种问题的区别，正是统一性知识与谱系知识的区别，也是讲台与出题台的区别。许宁在讲授课堂里获得的是回答前一种问题的能力；她在九问课堂里获得的，是提出并回答后一种问题的能力。后者以前者为底盘，却不能由前者推出。

还可以从许宁的第零轮初判倒着看讲授课堂的影响。她开学时写下的“夜间跌倒是因为患者年纪大”，并不是一个无知的判断，它恰恰来自讲授课堂：高龄是指南里排在前面的危险因素，她学得很好，所以第一反应就落在这里。讲授课堂给她的底盘，同时也成了她需要被翻过来看的前提。这不是讲授的过错，而是任何统一性知识在具体现场里的必然处境：它在平均意义上正确，在这一间病房里却可能遮住了真正起作用的东西。ChatSDE那一句“同样年纪的患者，为什么白天跌得少”，翻的正是这个从讲授课堂里带出来的前提；第7章与第8章会讨论，为什么这一翻必须在第一回合就发生，以及为什么通用大模型在默认配置下虽然往往也会翻前提，却在同一回合把答案替学生说破，使学生自己迈出的那一步被替掉。

学期末，许宁在九问课堂里交出的判断是：本病房夜间跌倒的风险更多落在交接断点上，而不只是年龄——风险藏在交班本里被省略的那一句。这条判断被明确标为本病房假说，待科室检验，所有资料均去标识化，取证经科室与伦理委员会许可，她不改医嘱、不作临床决策。它没有推翻讲授课堂里学到的任何东西，高龄仍是危险因素，指南仍然有效；它是在那份底盘之上，长出了一条底盘里没有的判断。讲授课堂与九问课堂的关系，在许宁身上看得最清楚：前者给了她站立的地方，后者让她在那里第一次往前走了一步。

◆ 本章怎样算错

这一章的判断压在三个可数的东西上：口试通过率、题闸打回率、九问与旧讲义的对照表。下面四条都以它们为读数。

第一条管讲授的位置。取讲授课时已压缩到原来一半以下、学生普遍借助大模型自主学习的班级，与讲授未压缩的对照班，各不少于30名学生，连续两届用同一套母本底盘口试（复述加反例）。若压缩班的通过率持续比对照班低15个百分点以上，而且这个差距不能归因于精要库或口试题的设计缺陷，“AI接管讲授”的判断在该学科就说过头了，讲授的份额必须重新估计，而不是继续往外移。

第二条管谱系母本能否与统一教材拉开距离。取一个学科的候选九问草案不少于60道，逐道过题闸。若打回率超过60%，并且抽查的打回理由中有半数以上是“母本的核心判断本身就能被通用大模型直接答出”，而不是出题措辞的问题，那么“教材让位于谱系母本”在该学科就不成立，该学科的大学课堂可能仍主要处

在练手的范围内，需要另寻出发点。打回率是这一条最直接的读数：多数候选题都过不了闸，问题就不在出题的老师，而在母本本身。

第三条管讲台是否真的变成了出题台。把一门课的九问与这门课原来的讲义目录逐条对照，若超过1/3的问只是章节标题换了句式，题闸形同虚设，那么“讲台的价值落到出题台”只是改了名字，本章判断要撤回，第4章关于出题台的论证也要重审。

第四条管本章引用的先验。Kestin等（2025）的结论来自物理课。若在医学、化学、法学中任何一门，按同样思路做一学期随机对照，每组不少于40人，而AI导师组的母本底盘通过率不高于课堂讲授组，那么“发现一侧已被AI接管”在这门学科里就只是物理与编程的结论，本章对该学科的判断收缩为“部分接管”。

四条都靠可观察的痕迹判定：口试记录、题闸的打回记录、九问与旧讲义的逐条对照、对照班的通过率。本章不接受“老师们普遍感到”之类的印象作为证据，正如它不接受“学生们普遍更有创造力”之类的印象作为支持。

本章把大学课堂的出发点从统一教材换成了谱系母本，但只说了它由前沿专著与部落档案合成，还没有说明它怎样被组织起来。一本前沿专著动辄几十万字，一个课题组十年的档案可能是几十本实验记录、几百份谱图和无数口头传承，老师不可能每次出题都从头读起，学生更不可能在开学前读完。母本怎样被提炼成一组可以直接用于出题的核心判断？这些判断怎样分类、怎样维护、怎样在一届又一届学生之间共用，并在每一届之后被回写加厚？第3章讨论谱系母本精要库。

> **带 走 的 话**
>
> **讲授体系的功绩要记，但统一性知识的传递已经移交大厂：发现降为底盘，讲台的价值落到认得学生现场的出题台上。**

◆

03

第3章

谱系母本精要库

◆ 没有教材的第1周

化学系三年级的课程型研究《合成条件的发生》开课第1周，沈澄和同组同学拿到的不是一本教材，而是三样东西：两部合成方法学与课程型本科研究的前沿专著，一个装着本课题组十年反应条件档案的共享目录，以及一份只有几十条的“精要库”清单。有同学翻完目录就举手问：“老师，教材在哪里？期末考哪几章？”这个问题并不幼稚。在讲授课堂里，教材是课堂的地基，它规定了什么是该懂的、什么是可以不懂的，也规定了考试的边界；拿走教材，学生第一个感到的不是自由，而是失重。老师当时的回答是：“这门课的地基不是一本书，是这个课题组十年做过、做坏过、改过的东西，加上这个领域现在最前面的人在争什么。”这句话说清了方向，却没有说清结构，本章要做的就是把这个结构立起来。

同一学期，法学院的《司法实践研讨》也在经历类似的第1周。韩正和同学们拿到的同样不是一本司法实务教程，而是本地基层法院近年的公开裁判文书汇编、往届同学整理的旁听记录，以及从这些材料里提炼出来的一份判断清单。有人问：“这些判决书里，哪一个是标准答案？”这一问暴露了讲授课堂留在学生身上的一个习惯：把一切材料都当作答案的载体，读材料就是找答案。而这门课恰恰要学生明白，判决书不是答案，判决书是一个地方的司法实践在某一刻留下的痕迹；一份份痕迹串起来，才看得出调解在怎样推进、被什么推着走。

两个课堂的困惑指向同一个问题：教材退场之后（见第2章），大学课堂的出发点究竟是什么？如果出发点仍然是一本统一的书，那么无论加上多少讨论与项目，课堂交付的仍是“掌握了同一本书的人”；如果出发点只是一堆未经整理的原始材料，学生又会淹没在档案里，找不到可以下手的地方。本章的回答是：大学课堂的

出发点是“谱系母本”，而让母本可以被千人共用、被逐年改写的，是从母本中提炼出的“母本精要库”。前者回答“从哪里出发”，后者回答“怎样出发”。

> **这一章的问题**
>
> 教材退场之后，大学课堂从哪里出发，才既不是一本统一的书，也不会淹没在未经整理的档案里？

◆ 谱系母本的两源：前沿专著与部落档案

本书所说的谱系母本，指大学九问的出发点，它由两部分合成：学科前沿专著，以及课题组或教研室这样的“部落”所积累的档案。这里借用“部落”一词，是指一个长期共同工作、共享做法与口头传统的小群体，例如一个课题组、一个科室或一门连续开设多年的研讨课。前沿专著提供的是一个领域眼下正在争执、尚未定论的判断，它把学生接到学科谱系的最新一节上；部落档案提供的是本地的、连续的、带着失败与修改痕迹的实践记录，它把学生接到自己可以亲手触摸的现场上。之所以称为“谱系”而不称为“体系”，是因为它不以一个统一的终点自命，而是承认一个学科由许多支脉组成，每一支都有自己的来路与分叉，学生进入的是其中一支，并且可能在这一支上长出新的一节。

打一个日常的比方。一本公开出版的菜谱，相当于前沿专著：它写的是行家眼下认为最好的做法，可以买来照着学，也可以看出不同流派在争什么。而一家老字号后厨里那本被油渍浸透的记事本，相当于部落档案：哪一年换了供货商、哪一锅因为天热发酸、哪一位老师傅交代过“火候到这里要停一停”，都零零碎碎地记在上面。只读菜谱的厨师，做得出标准菜，却不知道这家店的菜为什么是这个味道；只翻记事本的学徒，知道许多窍门，却不知道这些窍门在整个行当里处在什么位置。谱系母本要求两样都有，并且让二者互相照看：专著给档案一个坐标，档案给专著一个落脚处。

沈澄所在课题组的母本就是这样合成的。一侧是合成方法学与课程型本科研究的前沿专著，后者关于把真实研究嵌入本科课程的论述，可以追溯到Auchincloss等（2014）对课程型本科研究（CURE）的概括；另一侧是本课题组十年的反应条件档案，包括实验记录本、失败批次、仪器谱图和零星的实验室温湿度记录。韩正

所在研讨课的母本则一侧是司法实践与纠纷解决方面的研究著作，另一侧是本地基层法院的公开裁判文书和往届同学的旁听记录。两门课的学科相距很远，母本的骨架却一致：一端连着学科谱系的最前沿，一端连着本地现场的最深处。

谱系母本还有一个动态特征：两源的比例逐年向部落档案倾斜。一门课开设的头一两年，部落档案几乎是空白，或者只有散落在个人抽屉里的记录，此时母本主要倚重前沿专著，九问也多从专著里的争点出发。随着每一届学生把问对记录、现场证据和子本回写进来，部落档案一年比一年厚，母本的重心就逐步移向本地。这种倾斜不是对前沿的放弃，而是前沿在本地扎下了根：专著提出的争点，在本课题组、本法院、本病房里得到了本地的回答，而这些回答又反过来构成下一届学生的出发点。库恩在《科学革命的结构》中指出，常规科学的教材往往把一门学科的历史改写成通向现有结论的直线；谱系母本的做法恰好相反，它把来路、岔路与回头路都保留下来，作为学生出发的地形。

有人会问：为什么不让每位老师用自己的讲义充当母本？讲义当然可以进入母本，但它不能替代母本。一位老师的讲义，往往是他把前沿专著消化之后的一个版本，加上他个人的研究偏好；它的好处是讲得清楚，坏处是把争点抹平了，学生拿到的已经是老师替他们做过选择的结果。母本保留前沿专著的原貌与部落档案的原样，正是为了让学生看到那些尚未被抹平的地方。这就像学做菜，跟着师傅的口述学固然快，但若能亲眼看到那本油渍斑斑的记事本，就会发现师傅口述里省掉的那几次失败，而那几次失败里往往藏着下一道新菜。

◆ 从母本到核心判断：方程、路径、动力三类

谱系母本是原料，不是可以直接拿来出题的东西。两部前沿专著加上十年档案，少说也有数千页，没有一位老师能在每次出题时把它们从头翻一遍，更没有一位学生能在一学期里把它们读透。母本必须先被提炼成一组“核心判断”，课堂才有了可以下手的抓手。所谓核心判断，不是知识点，也不是章节摘要，而是母本里那些带着主张、可以被追问、可以被推翻的句子。“偶联反应的产率受催化剂负载量影响”是一条知识点；“本组十年里产率波动的主要来源被归于试剂批次，而这一归因从未被系统检验过”才是一条核心判断。前者告诉学生“是什么”，后者告诉学生“这里有一处可以下刀”。

本书按SDE框架把核心判断分为三类：方程判断、路径判断与动力判断。这一分法与《SDE思想体系入门》（德麦国际专著第395号）中的三大方程、六路径与123原理相对应，但在课堂里只取其操作含义，不作学理展开。需要先说明一点：三大方程F、G、H在本书中只作关系占位记号使用，它们标出的是“哪几样东西之间被主张有关系”，而不是一组可以代入数值求解的公式。分类的目的不是给判断贴标签，而是让出题的老师知道，一条判断如果有毛病，毛病可能出在哪一层，从而知道九问该朝哪里打。

方程判断主张的是一组关系：哪些因素被认为决定了哪个结果，哪些因素被默认无关。沈澄课题组档案里反复出现的“产率低是因为试剂纯度不够”，就是一条典型的方程判断，它在试剂纯度与产率之间画了一条线，同时在环境条件与产率之间默认没有线。方程判断最常见的毛病是“缺式”：关系式里少了一项，而少掉的那一项恰恰是本地现场里最重要的。好比一个人总以为自己失眠是因为晚上喝了咖啡，戒了咖啡还是睡不着，才发现卧室窗外新装了一盏路灯；他心里的那个关系式，从一开始就少了“光”这一项。精要库收录方程判断时，要把它“画了哪条线、默认哪条线不存在”都写明，因为九问往往就从那条被默认不存在的线开始问。

路径判断主张的是一种认识的来路：这个结论是按什么次序得出来的，先看了什么、后看了什么、什么是从一开始就被当成前提的。韩正所在研讨课的判例库里，有一条往届整理出来的判断：“物业纠纷调解成功率主要取决于法官的调解技巧。”这条判断的来路是先有“调解是法官的技艺”这一前提，再去裁判文书里找印证，于是凡是调解成功的案子都被读成法官技巧好，凡是失败的都被读成当事人不配合。路径判断最常见的毛病是“偏路”：顺着一条路一直走，没有回头看起点是否可疑。精要库收录路径判断时，要把它的来路写出来，让出题者看见“这个结论是顺着哪条路走到的”，才能设计一问把学生引到另一条路上去，例如去看调解成功的案子都在立案后第几周结束、那一周还发生了什么。

动力判断主张的是变化从何而来：一个领域或一个现场里，什么样的矛盾逼出了旧框架的松动，旧框架又在什么地方被改写。在本书的口径里，123原理只指一个循环：D与E之间出现矛盾，逼出S的改变，改变后的S再回写D与E。课题组十年档

案里有过这样一段：某一年夏天连续几批失败之后，组里的做法悄悄变了，师兄师姐开始口头叮嘱后来者“那一步要快”，此后的失败批次似乎少了一些；可是书面规程一个字也没有改，为什么要快、快到多少，档案里没有交代。这就是一条动力判断的线索：某个时刻，现场的失败（E）与原有的操作意图（D）发生了矛盾，逼出了做法（S）的一次改变；但这次改变只改了半截，没有被写进规程，回写停在了口头里。动力判断最常见的毛病是“断拍”：矛盾出现了，改变也发生了，却没有接回到下一轮。精要库收录动力判断时，要写明“这里发生过一次改变，改变停在哪里”。

三类判断的区分不是为了把母本拆成三张表，而是为了给出题保留一张地图。一条判断可以同时具有三类特征：“试剂纯度”一条首先是方程判断；它还带着路径的痕迹，因为十年里这一归因多半出现在失败批次的事后说明里，是先有结论、再回头找依据的；它又牵连着那次停在口头里的做法改变。提炼时不必强行归类，只需在每条判断旁边注明它最可能坏在哪一层。这样，第4章的出题台在调用精要库时，就能按学生画像找到最适合他下刀的那一层，而不是从同一份清单上平均地摘题。

这里的分类还要与第7章的工序判别细则分开看。判断的类型回答的是“母本里这句话主张什么”，归精要库管；判别细则回答的是“学生此刻的初判坏在哪里”，归问对管。判别按次序进行：先查学生在自己现场里是否已经接触到一件与初判直接相撞的事实、判断却没有因此改动，有，就是断拍；没有，再查初判的关系里是否漏了一项被当作常数的因素，有，就是缺式；两者都不是，而是起手的次序与任务不配，才是偏路。按这个次序，沈澄的“试剂纯度不够”属于缺式：被漏掉的是环境条件与操作时长两项被当作常数的因素。她那张批号表里三月与六月的反差，在被翻转前提之前只是一列日期和产率，还不是她面对着却没有改判的矛盾，所以不归断拍。它之所以容易被读成偏路，是因为十年归因的来路确实有先下结论的痕迹，可那属于精要库条目“来路”一栏要记的内容，不是诊断初判时所依的次序。同样，本章前面说的动力判断里的“断拍”，指的是母本层面一次做法的改变没有接回下一轮；课堂诊断里的断拍，指的是学生手里的矛盾事实被放过。两个词字面相同，所在的层面不同，混用就会把精要库的分类当成对学生的判词。

◆ 一条判断的条目长什么样

讲清了三类判断，还需要看一看精要库里的一个条目究竟写成什么样子。精要库的质量，最终落在一条条条目的写法上。抽象地说“提炼核心判断”很容易，真正落笔时，老师常常会写出两种不合格的条目：一种过于简略，只有一句结论，下一届的人拿到它，不知道它从哪里来、凭什么成立；另一种过于冗长，几乎把母本原文搬了过来，读的人找不到下刀处。一个合格的条目应当像一张病历首页：不必写尽全部病程，但主诉、来源、现有判断、疑点与处置去向一目了然，任何一位接手的医生都能据此接着往下看。

以沈澄所在课题组精要库中“试剂纯度”一条为例，一个合格条目至少包含六项内容。第一项是判断本身，用一句带主张的话写出：“本组该偶联反应产率低的批次，主要归因于试剂纯度不足。”第二项是来源，注明它出自哪几年的实验记录、哪几位成员的批注，以及前沿专著中有无相应论述。第三项是类型与关系，注明它是一条方程判断，画了“试剂纯度—产率”这条线，默认“环境条件—产率”之间没有线。第四项是可能的坏处，写明“若同一瓶试剂在不同季节产率差异显著，则本判断缺项”。第五项是已有的检验，如实写出“十年内无一次系统检验，归因多见于失败批次的事后说明”。第六项是版本与回写记录，记下这一条在哪一版入库、此后有无修改。

这六项里，最关键的是第四项与第五项。第四项把判断的软肋提前指了出来，它不是要否定这条判断，而是承认它有可能坏，并且给出坏的方向；这正是九问可以起步的地方，沈澄九问里的第2问（同一瓶试剂在三月与六月用出的产率）就是顺着这一项长出来的。第五项则把“被说了很多遍”与“被检验过”区分开来。一个课题组里，一句话被说了十年，听起来就像定论；精要库的条目要求如实写出它被检验的程度，十年都没有检验过的，就写“无一次系统检验”。这好比家里老人常说“这个病是受凉引起的”，说了几十年，全家都信，但从来没有人去对照过受凉与发病的先后；把这一点写在明处，后来者才知道这里还有工作可做。

韩正所在研讨课的判例库条目，形式相同，内容换成法学的材料。“物业纠纷调解成功率主要取决于法官的调解技巧”一条，来源写往届旁听记录与若干份裁判文书，类型写路径判断，来路写“先有调解是技艺的前提，再在文书中找印证”，

可能的坏处写“若调解成功的时点集中在某一外部周期附近，则技巧解释不完整”，已有检验写“未按时间分布整理过”。这样一条条目放在那里，下一届的学生和出题的老师一眼就能看出：问题不在于法官技巧是否重要，而在于还有一条从未被走过的路。条目写得好，精要库就不是一张结论清单，而是一张标明了“此处可挖”的勘探图。

◆ 一次建成、千人共用：精要库的版本与维护

精要库的第一个特征是“一次建成”。这里的“一次”不是说建好以后不再改动，而是说它的主体工作集中在开课之前，由教研室或课题组借助装有SDE工序层的大模型辅助完成，而不是每一位老师在每一次出题时重新从母本里打捞。提炼的工序大致如此：先由老师选定母本的范围，即哪几部前沿专著、哪几年的部落档案；再由大模型在老师的指引下逐部通读，把带着主张、可以被追问的句子抽出来；然后由老师逐条审读，删去只是知识点的句子，合并说法不同而实为一事的判断，为每条判断注明来源、所属类型与最可能坏在哪一层。机器做的是通读与初筛，人做的是认定。这样分工的理由很朴素：通读几千页材料，大模型比人快得多也耐心得多；但一条句子算不算一个值得下刀的判断，取决于老师对这个领域争点的理解，以及对本地现场的熟悉。

精要库的第二个特征是“千人共用”。一门课的精要库建成后，同一教研室的所有任课老师、同一课题组的所有指导者、历届所有选课学生，都从同一个库出发。这与讲授课堂里“同一本教材”看上去相似，实则不同。同一本教材意味着所有人走向同一个终点；同一个精要库意味着所有人从同一组出发点起步，却各自走向自己现场里的不同终点。好比一座城市的地铁网，所有人从同一张线路图上出发，却在不同的站下车，走进不同的街巷；线路图是共用的，街巷是各自的。共用的意义在于，它让不同学生的发生结果可以互相比较、互相碰撞，也让一届学生的回写能够被下一届准确地接住，因为大家说的是同一套出发点上的话。

共用的精要库还给学生规定了一道门槛。九问课堂并不取消学生对母本的掌握，而是把这种掌握从“考试内容”变成“出发资格”。本书学业评估的第一项“母本底盘”，量的就是学生对精要库中核心判断的掌握，其证据形式是复述加反例口试：学生要能用自己的话复述一条判断，说出它画了哪条线、走的是哪条

路，还要能举出一个可能使它坏掉的反例。在本科研讨课中，这一项的权重只占5%，在研究生课中则降为门槛项，不计分而只决定能否进入后续环节（见第15章与第18章）。权重低不等于不重要，恰恰相反，它像登山前的装备检查，检查本身不算成绩，但装备不齐的人不能出发。一位学生如果连“试剂纯度”这条判断都复述不清，他在九问里对这条判断的任何改题与反驳，都可能只是没读懂之后的另起炉灶，而不是在谱系上长出新的一节。

精要库的第三个特征是“有版本”。每一条判断都带着编号、来源、入库年份与修订记录；整个库也按学年出版本，例如某课题组的精要库从第一版到第五版，每一版都可以追溯哪几条是新增的、哪几条被修改了、哪几条被标注为“已被本地证据动摇”。版本不是行政上的形式，它是回写能够被观察的前提。第24章将专门讨论回写，这里只需指出：如果精要库没有版本，一位学生的子本对母本作出了什么贡献，就只能由老师宣布，而无法被他人核对；有了版本，贡献就落在某一条判断从第三版到第四版的具体变化上，谁都可以去看。

维护精要库的工作量远小于建库，但需要制度安排。每学年末，课程负责人汇总本届学生的子本、问对记录与采纳建议，召开一次精要库修订会，决定哪些新判断入库、哪些旧判断改写、哪些旧判断加注疑点。这里有一条必须守住的规矩：入库的必须是经过换手审读的判断，而不是学生自评认为重要的判断，更不是同一个底座上的几个智能体互相印证过的判断。同一底座上的多个智能体不能互作独立证据，这一点在本书第三编将反复出现；在精要库的维护上，它意味着大模型可以协助整理与比对，但判断能否入库，必须由人来认定，而且最好由不是这位学生指导者的另一位老师来认定。

精要库还要防止一种隐蔽的退化：库越来越大，却越来越像知识点清单。每一届都有学生提出新东西，每一次修订会都倾向于“多收一点总没有坏处”，几年下来，库里堆满了说得对、却没有任何张力的句子。对付这种退化的办法是一条简单的标准：每一条入库的判断，都必须能写出“它可能坏在哪里”。写不出可能坏处的句子，就不是判断，而是陈述，应当退回母本原文，不占精要库的位置。这好比一家医院的疑难病例讨论记录，记下的是“当时以为是甲、后来发现是乙”的病

例，而不是每一个按指南顺利治好的病例；后者也有价值，但它们属于教科书，不属于讨论记录。

◆ 与知识建构的社区数据库对话：精要库多出的两格

精要库不是第一个想到把共同体的知识沉淀成一座可以公开改进的库的设计。Scardamalia与Bereiter（2006）在知识建构的研究里（载于《剑桥学习科学手册》）把课堂看作一个共同体：学生的任务是对共同体的观点作持续的改进，观点进入一个公共的社区数据库，供他人引用、批评与改写，并追求所谓“超越”（rise-above）——新的说法把此前的若干观点吸纳进来，站到它们之上。这套设想与精要库、部落档案和回写在形状上相近，本书不否认这一点，也不主张自己是第一个发现“共同体的知识应当公开累积”的人。如果把精要库说成一项全新的发明，第一个会被打回的就是这句话。

两者的差别，可以用一个社区菜园来比。知识建构的社区数据库像菜园门口的公共黑板，谁有心得谁就写上去，后来者可以在旁边划掉、补充、改写，黑板上的东西越积越好。精要库则像菜园管理员手里的播种计划表：它不只收录大家的心得，还要决定今年每一块地种什么、种在哪里、什么时候播，而且表上每一次改动都签了名、记了日期，哪一行是谁在哪个月划掉的，翻出来就能对上。黑板长于汇聚，计划表长于派工；一个菜园两样都需要，但不能把计划表说成黑板，也不能因为有了黑板就说计划表多余。

据此，精要库多出的第一格在上游。知识建构里学生自己提出的问题同样会驱动后面的探究，这一点必须承认；本书的差别在于题的依据：精要库的每一条判断都写着“可能坏在哪里”，出题的老师据此把一条判断按学生的画像具体化成九问，并且先过题闸，也就是拿掉学生的现场之后，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案。于是库不只是收留学生产出的东西，它还决定下一位学生拿到什么样的题。一座只收不派的数据库，如果同一个课题组的学生始终从同样的通论出发，积累得再好，也很难保证每个人下刀的地方是不同的。

多出的第二格在证据责任。精要库的每一条判断带编号、来源、入库年份和修订记录，被本地证据动摇的条目要标注，被推翻的条目要写明撤回的理由；入库必须经过换手审读，而不是学生自评。这使“某位学生的子本对谱系有贡献”从一句共同体的自觉，变成一个可以核对的命题：课程负责人应当能在版本记录上数出，某一学年有几条判断被回写改动、几条被加注、几条被撤回，并能指出每一处改动对应的是哪份问对记录。知识建构的研究强调观点的可改进性，这与版本记录同向；本书多问的是，改进发生了没有，是谁改的，凭什么改的，别人能不能不靠当事人的叙述就查得出来。

有一个自然的反对意见：精要库由少数几个人提炼，条目的取舍本身就带着提炼者的偏好，这不是换了名字的教材吗？这个疑问有一半是对的。提炼者的选择偏差无法消除，只能让它被看见。库里的条目与教材的句子至少有三处不同：条目必须写出自己可能坏在哪里，教材的句子不需要；学生在母本底盘的复述加反例口试中，被要求亲手去找条目的坏处，也就是把库当作靶子而不是答案；条目的版本记录公开，一条判断连续几个学年没有任何人动过，会因此显眼。这好比超市货架上的“店长推荐”，并不是不许推荐，而是让顾客一眼看出哪几样是店里替你挑的、哪几样是它自己卖得好的。若这三处在实际运行中都落空，也就是条目写不出可能的坏处、口试里学生从来举不出反例、版本记录里条目长年纹丝不动，那么这座库实际上就是一本教材，本章就应当承认这一点。

◆ 部落档案怎样由薄变厚

部落档案是谱系母本里最容易被忽视、也最难被替代的部分。说它容易被忽视，是因为它通常零散、不整齐，看上去不像学问。前沿专著可以买到、可以检索、可以让任何一个大模型读过；部落档案却只存在于一个课题组的共享目录、一个科室的交班本、一个研讨课的往届记录里，外人看不到，大模型也读不到。正因如此，九问能够打向“尚不存在的判断”，很大程度上依赖部落档案：只有落在本地现场的问题，才可能是连老师也写不出终点、连通用大模型也答不出来的问题。

然而大多数部落档案起初都很薄，而且薄得不整齐。沈澄所在课题组的十年档案就是一个例子。这份档案里有实验记录本的扫描件，有失败批次的简要说明，有仪器谱图的原始文件，偶尔也有实验室温湿度计的记录；但是不同年份的记录习惯

差别很大，有的学生每一步都记时间，有的学生只写“室温反应过夜”；失败批次常常只有一句“产率低，弃”，原因一栏空着；谱图里的杂峰有时被圈出来打个问号，更多时候没有任何标注。这样的档案对于讲授课堂几乎没有用处，因为它不能给出一个干净的结论；但对于九问课堂，它恰恰是最好的矿：沈澄九问中的第3问（“室温”二字背后十年里有没有写过读数）和第7问（谱图里那个无人解释的小峰），正是从档案的空白与沉默处问出来的。

档案由薄变厚，靠的不是某一年突然花大力气整理，而是每一届学生在九问流程中自然留下的东西被系统地收回来。一届学生的问对记录，记下了他们最初的判断、被ChatSDE翻过来的前提、在现场找到的证据和最终形成的判断；他们的现场记录，记下了按新的规格采集的数据，例如沈澄复做一批时把每一步的时间、温度、湿度都记下来；他们的子本，给出了母本所无的判断。这些东西一届届回收进部落档案，档案的厚度就不再只是页数的增加，而是“可以下刀的地方”的增加。好比一块菜地，每年收成之后把秸秆还田，土一年比一年肥；肥的不是因为每年多买了化肥，而是因为每年长出来的东西有一部分回到了土里。

韩正所在法学研讨课的本地判例库提供了另一种增厚方式。这门课开设之初，判例库只是从公开渠道下载的基层法院裁判文书，按案由粗分了几类。第一届学生在旁听与阅读中，开始给每份物业纠纷文书加上几个额外的字段：立案日期、调解或判决结束的日期、其间是否有物业费账单的催缴节点。这些字段在裁判文书的正文里散落着，没有人专门整理过；一旦被整理出来，就让韩正的那一问“本院近三年物业纠纷里，调解成功的案子大多在立案后第几周结束？那一周发生了什么？”成为可以回答的问题，也让“调解进程被物业的账单周期推着走”这一判断有了被检验的可能。下一届学生接手时，判例库里已经多了一层结构，他们的九问可以从这一层结构往下打，而不是重新从下载文书开始。

需要强调的是，部落档案的增厚必须遵守相应的伦理与规范。档案的价值来自它记录了真实的本地实践，而真实的本地实践总是牵涉真实的人。法学判例库只使用依法公开的裁判文书与公开庭审的旁听记录，涉及当事人的信息按公开规则处理，不另行追索个人隐私；化学档案的回收不涉及危险操作细节的公开传播，复做必须遵守实验室安全规范；医学档案如许宁所在科室的轮转问对档案，涉及患者的

内容一律去标识化，并经科室与伦理委员会许可方可入档。档案越厚，责任越重；一份因为违规而必须被删除的档案，对谱系的损伤比一份薄档案更大。

◆ 本章怎样算错

本章立了三个主张：大学课堂的出发点是前沿专著与部落档案合成的谱系母本；母本要提炼成方程、路径、动力三类核心判断组成的精要库，一次建成、千人共用、逐年回写；部落档案的增厚来自每一届九问流程的产物被系统回收。书中的沈澄、韩正都是构造案例，只用来说明机制的样子，不能当作这些主张已被证实的根据。下面把每个主张可能失败的方式写成可以对照的读数，供实行九问模式的课程在几届之后检查。

第一条针对“逐年回写”。取一门每届选课学生不少于30名的课，连续两个学年，在年末修订会之后翻开精要库的版本记录；如果两个学年里，由学生回写引起修改、加注或撤回的条目数都是零，也就是库只增不改，那么“逐年回写”的主张就没有得到支撑，这座库实际已经退化成另一种教材。版本记录在这里是关键证据，它不依赖课程负责人的自述。读数为零之后还要分辨原因：若是修订会只敢收新条目、不敢动旧条目，要改写的是本章关于维护制度的描述；若是学生的子本里根本没有母本所无的判断，问题则可能出在出题台与问对环节，就不是精要库一章能够回答的了。

第二条针对“母本须先提炼为核心判断”这一中间环节。在同一门课的两个平行班里，各不少于30名学生，一个班的九问从精要库出发，另一个班的九问由老师直接从前沿专著出发，两班出题数量相同；由不参与出题的人，用同一个通用大模型（记录所用版本）对全部题目做题闸检验。如果两班首次出题的一次通过率之差小于10个百分点，并在第二届重复时仍然如此，就说明精要库没有让题目更多地落在学生的现场上，这一环节是多余的；本章应当撤回精要库作为必经环节的主张，改为由老师直接从母本出题，精要库至多作为参考目录保留。不参与出题，是为了防止出题者对自己题目的偏爱掩盖差别。

第三条针对档案增厚的来路。第二届结束时，把部落档案新增的“可下刀处”按来源分成两类：一类由九问流程的产物（问对记录、现场材料、子本回写）

回收而来，一类由专人专项整理而来。如果后一类占到新增条目的一半以上，或者撤掉专人整理之后档案该年新增条目为零，那么“九问流程本身会让档案变厚”的描述就与事实不符，需要改写成一项独立的档案工作，而不再算课堂发生的副产品。这并不否认专人整理的价值，它只检验本章的一个具体说法；若事实如此，第24章关于回写的讨论也要相应调整。

精要库建成之后，出题台就有了原料。然而原料不会自己变成题目：同一条“试剂纯度”的判断，交到沈澄手里与交到另一位同学手里，应当长成怎样不同的九问？一条判断要经过怎样的具体化，才能打向一个连老师也写不出终点的地方，又怎样确认这个地方确实还没有人到过？下一章将走上出题台，以许宁的九问为样本，逐问拆开这道工序。

带 走 的 话

课堂的出发点是前沿专著加部落档案合成的谱系母本；让它能千人共用、逐年回写的，是一条条写明“可能坏在哪里”的核心判断。

◆

04

第4章

出题台：九问打向尚不存在的判断

◆ 出题的下午

老年病科的带教老师在许宁这一轮轮转开始前的那个下午，没有备课，而是坐在办公室里出题。她面前打开着三样东西：本课程的谱系母本精要库，其中收录了老年医学关于跌倒与衰弱的若干核心判断，以及本科室历届轮转学生的问对档案；一份关于许宁的学生画像，记着她来自哪个年级、前几轮轮转的问对记录里暴露出怎样的思考习惯、这一轮将在哪个病房、上不上夜班；一个装有SDE工序层的大模型对话窗口（这位老师用的是经SDE工序化的Claude模型）。她要在这个下午结束前，交出九个问题。按照讲授课堂的习惯，这个下午本应用来准备一份关于老年跌倒的讲稿，把危险因素、评估量表与干预措施讲清楚；而在九问课堂里，这些内容她一个字也不打算讲，因为任何一个大模型都能讲得比她更全、更快，也更耐心（见第2章）。

她出的第一版问题里有一条是：“老年人跌倒有哪些危险因素？”写完她自己就笑了，把这句话原样贴进一个通用对话窗口，几秒钟之内就得到了一份条理清楚的答案：高龄、肌力下降、平衡障碍、多重用药、认知障碍、视力减退、环境因素，后面还附上了常用的评估工具。这份答案没有错，甚至比许多教材的相应章节更完整。问题恰恰在这里：一个能被这样回答的问题，交给许宁去做，她能做出什么来？最多是把这份答案再整理一遍。这一条被她划掉了。她后来交出的第2问是：“交班本上，跌倒前一班对该患者写了什么、没写什么？”这一问，没有任何一个模型能替许宁回答，因为没有任何一个模型看得到本病房的交班本。

这个下午的一删一改，浓缩了本章要讨论的全部问题。老师在九问课堂里的核心工作，从讲授转为出题；出题不是随手提问，而是一道有工序、有检验的工作；每一问都要经过一道闸门的检验，即“题闸”：拿掉学生的现场材料，一个未接入

这些材料的通用大模型（记录所用版本）能否给出说得过去的答案？能给出，题就不合格，打回重出。本章先讲出题的两步工序，再讲出题前的全球敌拓与题闸，然后用许宁的九问逐问拆开出题逻辑；接着把题闸放到通行的作业重设计做法和一次小样本试跑面前，看它多出了什么、又有什么检验不了；最后讨论老师在出题台上不可替代的那一部分。

> **这一章的问题**
>
> 怎样出一道没有人事先写得出终点、离开学生的现场也答不出的题？

◆ 两步出题：从母本核心判断到学生画像

九问的出题分两步。第一步，从精要库里选出核心判断，确定这一组九问要往母本的哪一处下刀；第二步，按学生画像把这一刀具体化，落到这位学生能够亲手触摸的现场上。两步缺一不可。只做第一步，题目会停留在学科的一般层面，像一份综述的框架，人人都能做，也就人人都能被模型代做；只做第二步，题目会变成一份现场调查表，具体是具体了，却与学科谱系断了根，做出来的东西只是一份本地报告，长不成谱系上的一节。

第一步选的是“刀口”。在许宁的课程里，精要库中关于老年跌倒的一条核心判断，可以概括为“高龄及其伴随的衰弱是跌倒的首要危险因素”。这是一条方程判断（见第3章），它在年龄、衰弱与跌倒之间画了线，而在照护的组织方式、交接的节律与跌倒之间，至少在许多学生的默认理解里，没有画线。精要库的条目在“可能的坏处”一栏写着：若同样年龄与衰弱程度的患者，跌倒风险随照护时段显著变化，则本判断缺项。老师选中这一条作为刀口，理由很清楚：它是学科里被广泛接受的判断，又恰好在某个方向上留有缺式的可能，而这个方向——照护时段与交接——正好是一个轮转学生每天都要经历、却很少被要求去看的东西。

第二步按画像具体化。学生画像不是一份性格测评，而是老师对这位学生“现场”与“习惯”的了解。许宁的现场是本病房、交班本、护理记录与夜班；她的习惯，从前几轮的问对档案里看，是一遇到临床现象就先去套已知的危险因素清单，套上了就停下。老师据此把刀口落到四个具体的东西上：时间（夜间与交接时段）、文书（交班本与护理记录）、空间（床位与护士站的距离）、人（患者本

人、家属与护士）。九个问题由此长出，每一问都指向许宁这一轮里伸手可及的一样东西，又都连着那条关于高龄的核心判断。

这两步，在许多老师的经验里最难的是第二步。第一步可以依赖精要库，库建得好，选刀口并不困难；第二步却要求老师认得学生的现场。这好比一位裁缝做衣服，布料可以从同一家布店进，款式也可以参考同一本样册，但量体只能对着眼前这个人量；量错了，再好的布料和款式也穿不上身。装有SDE工序层的大模型可以在两步上都给老师帮忙：在第一步，它可以帮老师比较几条候选判断各自的缺口；在第二步，它可以根据老师提供的画像，生成多个具体化的候选问题。但哪一个候选真正落在许宁伸手可及之处，只有每天在病房里看见她的老师说得准。

◆ 出题前的两道检查：全球敌拓与题闸

九问要打向“尚不存在的判断”，这句话需要两道检查来落实。第一道检查是全球敌拓，问的是：这道题是不是早已被人答过？第二道检查是题闸（定义见下），问的是这道题离开学生的现场还能不能被答出。前一道朝外看，看学科谱系里是否已有终点；后一道朝内看，看这道题是否真正依赖学生的现场。两道检查各有分工，不能互相替代。

全球敌拓的做法，是带着“敌意”去拓宽检索：出题者主动假设“这道题其实早就有人做过了”，然后朝几个方向去找证据，包括本学科的综述与前沿专著、相邻学科里换了名字的同类研究、实务领域的报告与指南、本课题组或本科室过去的档案。之所以叫“敌拓”，是因为检索的目的不是为题目找支持，而是为题目找反证。好比一个人想开一家新店，最该做的不是找朋友夸他的点子好，而是在方圆几公里内仔细走一遍，看是不是早已有人开了同样的店。对许宁的九问来说，敌拓会发现，关于老年住院患者跌倒与夜间、与交接、与用药调整、与呼叫铃使用的一般性研究并不少见；这说明这些方向本身不是空白。但敌拓同时也会发现，关于“本病房、近三个月、这一份交班本”的问题，没有任何文献能给出终点。这正是九问要打的地方：方向接在谱系上，终点落在现场里。

敌拓有一个容易被误用的地方：它不是要求九问在学科层面上做出前人从未做过的发现。如果把这个标准用在本科研讨课上，几乎没有一道题能过关。九问要求

的是，这道题的终点在出题时尚不存在，包括老师在内的任何人都写不出它的答案；至于这一方向在学科里有没有一般性研究，恰恰是母本应当提供的背景，而不是题目的障碍。本书第1章区分练手与战场时说过，判别的第1问是“终点能否事先写出”；敌拓就是把这一问落实为一道检索工序，让“写不出终点”不再只是出题者的自我感觉。

题闸是第二道检查，也是本书反复使用的一个概念。它的定义只有一句：拿掉学生的现场材料（E），一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案；能，题即不合格，打回重出。题闸的操作极其简单，老师只需把问题原样交给这样一个模型，看它能否给出一个说得过去的答案。简单不等于随便，有三处细节决定这道检验是否可信。其一，模型必须是干净的。那个已经读过学生画像、翻过精要库的出题助手不能兼作题闸，因为它的答案里带着现场的影子，用它验题等于自己给自己放行。其二，要在默认配置、未开启学习模式的状态下验。学习模式的特点是先追问、不直接给答案，开着它去验，几乎任何题都会显得“答不出”，题闸便形同虚设。其三，要登记：模型名称与版本、测试日期、配置、回答原文和判定。同一道题换一个版本可能就换了结果，没有登记的题闸只是一次口头自查。题闸的道理也极其朴素：一道不依赖学生现场就能被回答的题，学生做它的全部价值，至多是把模型的答案重新组织一遍；这正是讲授课堂里“交付知识”的另一种形式，只不过讲授者从老师换成了模型。题闸在学业评估中对应创新智商五维中的I项（见第16章），出题阶段把好这道闸，评估阶段才有可能出现真正的新判断。

需要说清楚的是，题闸检验的是题目，不是模型。这一点关系到本书对通用大模型的基本态度：差别落在课堂的配置上，而不是对模型能力下定论。一道题过了题闸，不说明模型在这一问上“不行”，只说明这一问的答案藏在模型看不到的地方；一道题被打回，也不说明模型在这一问上“全行”，只说明这一问已经有了学科层面的终点。好比一个人问路，问“北京在哪里”，地图软件答得出；问“我家楼下那家早点铺今天开门没有”，地图软件答不出，这并不说明地图软件不好，只说明后一个问题的答案在楼下，不在地图里。

◆ 许宁的九问：逐问拆开出题逻辑

下面把许宁的九问逐一拆开。每一问的拆解都回答同样几个问题，便于读者把九问放在一起比较。九问的原文是定本，全书各章只引用、不改动；这里讲的是每一问为什么这样出、它连着母本的哪一处、落在许宁现场的哪一样东西上、以及它为什么过得了题闸。为了避免读者把九问看成一张平铺的清单，先说明它们的大致布局：第1、2、3问分别从时间、文书、用药三个方向，把“交接”这一线索引入许宁的视野；第4、5、6问换三个角度，空间、患者叙事与陪护，去检验高龄解释能否独占跌倒风险；第7、8问把视线从患者转到照护流程本身；第9问则要求许宁从诊断走向一次假设性的改动，看出判断的承重点。

第1问：“本病房近三个月的夜间跌倒，集中在哪个交接时段前后？”这一问出的是时间刀。它连着精要库中那条关于高龄的方程判断，作用是在“年龄—跌倒”这条线之外，提示一条可能存在的“交接时段—跌倒”的线。它没有直接说交接是原因，只要求许宁去看跌倒在时间上的分布；如果分布是散的，这一问就证伪了交接线索，这同样是有价值的结果。它过得了题闸，是因为“本病房近三个月”的跌倒时间点只存在于本科室的不良事件记录与护理记录里，模型对此一无所知。若把它改写成“老年患者夜间跌倒多发生于什么时间”，模型便可根据一般研究给出一个答案，题也就失去了意义。

第2问：“交班本上，跌倒前一班对该患者写了什么、没写什么？”这一问出的是文书刀，也是整组九问里最锋利的一刀。前一半“写了什么”是一般的查阅，后一半“没写什么”才是要害：它要求许宁去读出文书的沉默。一份交班本上写下的内容，代表交接者认为重要的东西；没写的内容，则可能代表被认为不重要的东西，也可能代表来不及写、不便写或习惯性省略的东西。从“没写什么”出发，许宁才可能看到交接中的断点。这一问在本书贯穿案例中被定为题闸合格的标准样例：未接入本病房材料的大模型看不到交班本，它连“写了什么”都答不出，更不用说“没写什么”。第5章将会看到，许宁后来正是把这一问提升为“交班时被省略的那一句，是谁、在什么情况下省略的？”

第3问：“跌倒的患者中，有几位当天刚调整过药物？调整是在哪一班记下的？”这一问前半截看似回到了常规的危险因素，用药调整本就是老年跌倒研究里

反复讨论的因素；出题的巧处在后半截，“调整是在哪一班记下的”。它把一个常规的患者因素，与交接流程勾连起来：如果药物在白班调整、而夜班交接时这一信息没有被突出，那么风险就不只在药物本身，还在信息的传递上。这一问因此既不抛开母本，又把母本里的一个因素引向许宁的现场。它过得了题闸，因为哪一位患者哪一天调过药、记在哪一班，只在本病房的医嘱与护理记录里。

第4问：“同样高龄、同样用药，未跌倒的患者床位离护士站多远？”这一问出的是空间刀，用的是对照的方法。它在设计上先把高龄与用药两个因素“固定住”，再去看第三个因素，这是临床研究里最朴素的思路，许宁在前几年的课程里早已学过，但很少有人要求她在自己的病房里用一次。它对母本中“高龄是首要因素”的判断构成直接的检验：如果同样高龄的患者是否跌倒与床位远近有关，那么高龄就不能独自承担解释。它过得了题闸，因为本病房的床位布局与每位患者的床号，模型无从得知。

第5问：“患者自己怎么讲跌倒前那一刻？与护理记录的说法差在哪里？”这一问出的是叙事刀，指向医学的主体属性。在护理记录里，跌倒通常被写成一个事件，时间、地点、伤情、处置；而在患者自己的讲述里，跌倒前那一刻可能是“想去厕所，按了铃没人来”，也可能是“以为自己能走”。两种说法之间的差，恰恰是记录体系看不见的地方。这一问必须严格遵守伦理规范：许宁与患者的交谈须在科室许可与患者知情同意之下进行，所得材料一律去标识化；她听取患者的讲述，是为了理解，不是为了改变任何医疗决定。它过得了题闸，因为模型听不到这位患者的讲述。

第6问：“有家属陪护的夜晚与无人陪护的夜晚，呼叫铃的使用有什么不同？”这一问出的是陪护刀，也是一个设计精巧的自然对照。同一位患者，有的夜晚有家属在，有的夜晚没有；呼叫铃的使用差异，可以让许宁看到，患者在夜间的需要是怎样被满足或未被满足的。如果无人陪护的夜晚呼叫铃使用明显减少，那么可能的解释之一，是患者“不好意思按”或“按了也来得慢”，于是自己起身；这就把风险从患者的身体引到了照护的互动上。它过得了题闸，因为呼叫铃记录与陪护安排都只在本病房。

第7问：“本科室已有的防跌倒措施里，哪一条在夜班最常被跳过？为什么？”这一问把视线从患者转到流程。每一个科室都有防跌倒措施，措施本身往往写得很完整；问题在于哪一条在夜班的实际执行中最常被省略，省略的理由又是什么。后半截“为什么”是这一问的承重部分，它不是要许宁去追究某位护士的责任，而是去理解流程在夜班的人手与节律下如何变形。这一问的出题与本科室历届问对档案相关：档案里已有学生注意到夜班措施执行的差异，却没有追问原因，这一问正是接着往下挖。它过得了题闸，因为哪一条措施在本病房夜班被跳过、为何被跳过，只有到现场去看、去问才知道。

第8问：“跟一次夜班，记下护士从听到动静到到达床旁的时间，哪一段最长？”这一问是九问里唯一明确要求“下现场”的动作题。它把抽象的“照护响应”拆成一段一段的时间：听到动静、判断来源、放下手头的事、走到床旁。哪一段最长，许宁事先不知道，老师事先也不知道，护士自己未必意识到。这一问的出题逻辑是把一个整体的现象拆成可以测量的环节，让断点在测量中自己显露出来。它的伦理边界同样清楚：许宁跟夜班是观察与记录，不干预护理操作，记录不涉及可识别的个人信息。它过得了题闸，因为那几段时间只存在于那一个夜班里。

第9问的原文是：如果把跌倒风险评估从“入院一次”改成“每次交接一次”，本病房第一个会被改变的是什么？这一问是收口的一刀，也是一个反事实的问题。它不要求许宁真去改变评估制度，学生无权、也不应作出任何临床或管理决策；它要求许宁在前八问的基础上，想象一次改动，并说出这次改动最先触动的是什么。如果她的回答是“交班本上多出一栏”，那么她的判断就落在交接文书上；如果是“夜班护士的工作量”，判断就落在人力上。这一问因此是一面镜子，照出许宁经过前八问之后，究竟把风险的承重点放在了哪里。它过得了题闸，因为“本病房第一个会被改变的是什么”取决于本病房的流程细节，模型只能给出一般的推测。

九问合起来，打向的是一个尚不存在的判断。在许宁的案例里，这个判断后来发生为一条本病房假说：本病房夜间跌倒的风险更多落在交接断点上，而不只是年龄——风险藏在交班本里被省略的那一句。老师出题时并不知道这个判断，九问也没有把这个判断写在题面上；九个问题只是从九个方向把许宁推到了可能发生这个

判断的地方。而这一判断至今仍是待检验的本病房假说，须交科室进一步验证，学生不据此改动任何医嘱或护理安排。

回头看这九问，还能看出出题者有意留下的两处安排。其一，九问之间存在互相校验的关系：第1问看时间分布，第3问看用药记录落在哪一班，第8问看一个夜班里的响应时长，三者如果都指向交接时段，判断就有了三条彼此独立的来源；如果彼此矛盾，矛盾本身就是问对时最好的材料。这好比几个证人分别作证，口径一致固然有力，口径不一也会把真正需要追问的地方暴露出来。其二，九问没有一问直接写出“交接”是原因，第2问问的是“写了什么、没写什么”，第7问问的是“哪一条被跳过”，都是让现场先说话。出题者若在题面上替学生下了结论，九问就变成了九道求证题，学生只是去为老师的猜想找证据；不在题面上下结论，才给了许宁在问对与现场中自己说出判断的余地，而“核心判断是否由学生先说”，正是后续评估问对深度的要点之一（见第17章）。

◆ 题闸的对照：合格的题与被打回的题

逐问讲完之后，有必要把合格的题与被打回的题并排放在一起看。只看合格的题，读者容易以为出题不难。题闸的标准一句话就能说清，但真正出题时，老师常常会写出一些“似乎合格”的题：它们看起来很具体，提到了本病房、本科室，实际上却只是一般问题换了一层外衣。下表把本书贯穿案例中给出的标准样例，与出题实践中几类常见的打回情形放在一起比较。

题目	题闸结果	理由
交班本上，跌倒前一班对该患者写了什么、没写什么?	合格	答案在本病房交班本里，未接入的大模型看不到
老年人跌倒有哪些危险因素?	打回	大模型张口就能答，学生只能整理答案
本病房应当怎样预防老年患者夜间跌倒?	打回	题面有“本病房”，答案却是一般防治指南的复述
交接班制度对患者安全有什么影响?	打回	方向对，终点已在学科层面写出，与学生现场无关
本病房近三个月的夜间跌倒，集中在哪个交接时段前后?	合格	时间分布只在本科室记录里，且结果可能证伪交接线索

表中第三行值得多说几句。“本病房应当怎样预防老年患者夜间跌倒”这道题，题面上挂着“本病房”三个字，看上去落在了现场；可是把它交给一个没有接入任何本病房材料的通用大模型，它照样可以写出一份像样的答案：加强夜间巡视、床栏使用、防滑地面、用药评估、健康宣教。这份答案对任何一个病房都成立，因而对本病房什么也没说。题闸的检验之所以要“拿掉学生的现场”，正是为了识破这种挂名的具体：真正的现场题，其答案离开现场材料就写不出来，而不只是题面上提到了现场。

第四行代表另一类常见情形：方向对，终点却已在学科层面写出。交接班与患者安全的关系，是一个被反复研究过的领域，模型可以给出系统的综述。这道题的问题不在于它离许宁的现场太远，而在于它问的是一个一般关系，而一般关系已有终点。合格的第1问与第2问，方向与它相同，却把问题压到了本病房的时间分布与本病房的交班文字上，终点于是移到了现场里。可见题闸打回的并不是“方向”，而是“终点的位置”；同一个方向，问到一般层面就被打回，问到现场层面就能通过。这好比钓鱼，同一条河，在人人都去的码头下钩，钓上来的鱼大家早都见过；走到上游一处没人去的回水湾，才有可能钓到没见过的东西。

被打回的题并非一无是处。“老年人跌倒有哪些危险因素”恰恰是母本底盘的内容，学生必须掌握，只是不应把它当作九问来做（见第3章关于母本底盘的讨论）。打回的意思是“这不是九问”，而不是“这不值得知道”。老师在出题台上打回一道题，等于把它放回底盘层，交给模型去讲、交给口试去检；九问的位置，要留给那些模型讲不出、老师也写不出终点的题。

◆ 题闸不是新发明：同类做法与多出的一格

必须老实说，题闸作为一道操作，并不是本书发明的。2023年起，许多高校教学中心陆续给出作业重设计指引，核心建议如出一辙：先把作业交给生成式AI试做，能做出来，就改题。这和题闸在动作上几乎是同一件事，本书没有理由把它说成自己的东西。读者也不必因此失望：一个被许多课堂各自摸索出来的做法，恰好说明这条路走得通。本章要回答的只是，在九问课堂里，这道闸比通行的指引多出了哪一格。

多出的一格落在三处。第一，指引只管“能不能被AI做出来”，打回之后怎么改，各凭教师自己的办法，常见的是换成课堂手写、当面口试或个人反思，题还是那道题，只是躲开了模型。题闸的打回则有方向：回到精要库里要打的那条核心判断，再按学生画像，往现场里伸手可及的一样东西上压，直到终点真正落进现场。这是前面两步出题工序在被打回时的再次启动，不是一个孤立的开关。第二，指引面向的是作业，目的是查学生做没做；题闸面向的是九问，目的是保证这道题值得一个学生花一个学期去发生一个判断，所以它与敌拓并列，并且接在学生画像上，去看这道题对这位学生是否伸手可及。第三，指引多半是一次性的自查，题闸要登记：版本、日期、配置、判定人与判定，并汇总成打回率。打回率本身是一个读数：一位老师、一个学科、一届学生的打回率若长期为零，要么是出题已经很熟练，要么是闸根本没有在查，两种情况只能靠登记簿里的回答原文去区分。这三格里的部件基本都是现成的，本书主张的不是发明了新部件，而是把它们接在母本精要和学生画像上，并留下可以复查的记录。

◆ 题闸真跑了一次：结果与读法

本书初稿写完时，题闸还只是一个设想，连这样简单的检验也没有真跑过。附录二报告了一次小样本试跑，这里取其中与本章直接相关的第一部分，试跑日期为2026年10月29日。材料共23道题：许宁、沈澄两套定本九问共18道，另加5道对照题，其中3道是上一节表里的打回样例，2道是新增的化学对照题。两个出自同一家族、型号不同的通用大模型，在默认配置、没有接入任何病房或课题组材料、不联网的条件下，分别作答全部23道，每题不超过150字。作答随后被打乱顺序、隐去来源，交给两名独立的评判员，逐份判定：这份回答是不依赖现场材料就给出了具体答案，还是只给出了一般框架、方法建议或可能性清单，或者明确说需要现场数据。

结果很整齐。九问共36份回答，全部被判为“需现场”；对照题共10份回答，全部被判为“可直接答”。两名评判员对46份回答的判定完全一致。换成题闸的语言：在这组题上，定本九问全部过闸，对照题全部打回，不同的人来执行这道闸，看法没有出入。表中那道挂着“本病房”的预防题，在试跑里同样被模型直接答了出来，正好印证了上一节对“挂名的具体”的判断。

这个结果要小心地读。它支持的是两件事：题闸可以被不同的人稳定执行；上面表里的打回样例确实该打回。它不支持的至少有三件。其一，样本小，两个模型同属一个家族，题目是本书自己编的，评判模型与生成本书文字所用的模型也同属一个家族，附录二自己把它定位为先验，不是独立证据。其二，它在本书自己出的题上表现整齐，并不保证在别的老师手里同样整齐。其三，它只测了“答不答得出”，没有测学生做这道题以后发生了什么，那要靠第11章设计的正式对照。

还有第四件，也是出题台上最需要记住的一件：题闸只检验“必须依赖现场”，不检验“值得研究”。九问的题面带着“本病房”“本组”“你复做”这类指向现场的词，现场依赖有一部分是措辞带来的。试想一道题：“本病房护士站的绿植摆在走廊哪一侧？”（这是为说明而设想的，没有进入试跑。）没有任何模型能在不接入现场的情况下答出它，它轻轻松松过闸，可是没有哪位研究者会为它花一个学期。一道题可以因为写了“本病房”而过闸，却仍然是一道不值得做的题。题闸挡得住“模型能代答的题”，挡不住“没有人需要答的题”，这两种坏题长得不一样，要由不同的环节来把关。

“值不值得”的第一道关在出题台上，靠老师认得学生的现场：同样是一栏空着的交班本，对这个病房可能是一条线索，对另一个病房可能只是排版习惯，这只有每天走过那条走廊的人说得准，下一节专门讲。第二道关在学期末，靠第16章中创新智商I维的评审：子本里是否真的出现了母本所无的判断，还是只把现场的事实誊抄了一遍。题闸、敌拓、认得现场的老师、I维评审，各管一段，不能互相顶替。

◆ 老师不可替代之处：认得学生的现场

讲到这里，一个自然的疑问出现了。这个疑问并非杞人忧天，因为本书第2章已经承认，讲台上的许多工作确实被模型接管了。问题是：既然出题用的是装有SDE工序层的大模型，题闸用的也是大模型，老师在出题台上究竟还剩下什么？如果模型能从精要库里选判断、能按画像生成候选题、能自己去跑题闸，那么出题台是否最终也会被模型接管？本书的回答是：出题台上有一件事，至少在可见的将来，必须由老师来做，那就是认得学生的现场。

认得现场有三层意思。第一层是知道现场里有什么。许宁的老师知道本病房的交班本是怎样写的、哪几栏常空着、夜班有几位护士、护士站在走廊的哪一头、哪几间病房离得最远；这些东西不在任何文献里，也不在精要库里，只在每天走过这条走廊的人心里。没有这一层，第4问“床位离护士站多远”这样的题根本出不来。第二层是知道现场里什么可以碰、什么不能碰。患者能否访谈、需要经过哪些许可、夜班观察的边界在哪里、哪些记录可以查阅，这些都是伦理与制度的判断，必须由对科室负责的老师来做。第三层是知道这位学生在现场里能走多远。许宁上不上夜班、与护士关系如何、前几轮里是否表现出与患者交谈的能力，决定了第5问与第8问对她是合适的挑战还是越界的要求。

这三层合起来，构成出题台上不可外包的部分。这好比一位老中医带徒弟，方书人人可读，药性人人可背，甚至开方也可以借助各种工具；但给眼前这位徒弟安排看哪几位病人、先让他摸哪一种脉，只有每天看着他的师傅才能定。装有SDE工序层的大模型在出题台上是一位极好的助手：它能在精要库里快速比较判断，能列出几十个候选题，能对每一道题做敌拓检索，也能替老师把题原样交给一个干净的通用模型去过题闸并登记结果，能把老师零散的现场描述整理成结构化的画像。但候选题中哪一个真正落在这位学生伸手可及之处，最终要由认得现场的老师来选。

由此也可以理解，为什么本书说“九问即九个谱系性研究课题”。在讲授课堂里，老师出的题是考卷，考的是学生是否到达了老师已知的终点；在九问课堂里，老师出的九问是课题，每一问都连着学科谱系上的一条核心判断，又都打向一个尚无终点的现场。九问不是九道作业，而是一个微型研究计划的九个入口：做完一问，可能得出一个本地事实；九问合起来，才可能得出一个判断。而这个判断一旦经过问对、现场与子本的检验，就可能回写进精要库，成为下一届九问的出发点。老师的工作，因此从“把终点讲清楚”变成了“把入口开对”。

◆ 本章怎样算错

本章的几条主张——老师的核心工作转为用装有SDE工序层的大模型打造九问；出题分两步；每一问过全球敌拓与题闸；老师在出题台上不可替代之处是认得学生的现场——靠的都是构造案例。许宁的九问是用来说明出题逻辑的样本，不是已被证实有效的证据；附录二的小样本试跑给了题闸一个先验，但样本小、同属一

个模型家族、题目自编，同样不是独立证据。构造案例和先验能说明一种做法怎样运作，说明不了它比别的做法更好。下面分六条写明：出现什么读数，本章就要改。

第一条管题闸能不能被稳定执行。由两名换手评判员（不是出题老师）对不少于30道题的作答独立判定，样本须同时含合格题和被打回的样例；若两人的一致率低于90%，或κ低于0.6，则“题闸可以被不同的人稳定执行”不成立，本章要把题闸从“检验”降为“参考”。附录二在23道题、46份回答上，两名评判员的判定完全一致（κ=1.00），这是对本条的先验，但它只覆盖一个模型家族和本书自编的题，不能当作本条已经通过。

第二条管版本漂移。每学期开题前，用当时的线上版本把本课程已出的九问题面重过一遍题闸。若原先判为“需现场”的题面里，有超过20%被判为可直接答，说明题闸结果随模型版本漂移得太快，“记录所用版本”就不只是一条体例要求，题面必须重出，本章也须在题闸之外补一道版本更新后的重测规程。

第三条管题闸有没有用。在连续两届、每届不少于30名学生的课程里，把过闸的九问与被打回的一般性问题分给不同的学生，由换手评分员评定子本；若两组子本中出现母本所无判断的比例，差距小于评分员之间的一致性误差，则题闸作为九问质量的核心检验没有得到支撑，本章需要重新界定什么样的题才算“打向尚不存在的判断”。评分不能由出题老师自己做，也不能由同一底座上的几个智能体互相印证。

第四条管过闸的题值不值得做。对已过题闸的九问，由一位换手评审与任课老师各自独立判断“值不值得学生做一个学期”；若在连续两届、每届不少于20道过闸题里，两人都判为不值得的占到30%以上，则“值得性由认得现场的老师把关、由I维评审收尾”这层补充不够用，题闸之后须单独增设一道值得性检验。

第五条管老师。如果装有SDE工序层的大模型在仅获得老师提供的文字画像、而老师不参与挑选的条件下，出的九问在题闸、学生改题质量与最终子本上，与老师参与挑选的九问没有可见差别，那么“认得学生的现场”就不是出题台上不可替代的部分，本章关于老师角色的论断应当撤回或大幅收缩。

第六条管敌拓。如果大量过闸的九问，事后被发现其终点其实早已在学科层面写出，只是出题者没有检索到，那么本章关于两道检查分工的描述就过于宽松，需要提高敌拓的标准。

九问出完，交到许宁手里。可是九问一旦离开出题台，就不再只属于老师：许宁读到第2问时，觉得“写了什么、没写什么”还不够锋利，她想问的是那一句为什么没写、是谁省略的。学生有没有权利改老师的题？改了之后，怎样判断她是在提升题，还是在躲开一道难题？下一章讨论改题权。

带走的话

好题的终点落在学生的现场里：题闸挡得住模型能代答的题，挡不住没人需要答的题，后者要靠认得现场的老师来把关。

◆



第　二　编

D：改题与问对（走来路）

编　序

经济管理类研讨课的第2周，老师接连收到周川的两份改题申请。第一份想把“哪个伙计干得最起劲”改成“AI会不会取代饭馆”，老师只批了一句：这个问题，你家饭馆还在不在里面？三天后的第二份，改成了“父亲为什么一到月底就泄气”。九问发到学生手里，是种子，不是考卷；学生接下来怎样把它走成自己的路，是本编的事。

第5章讲改题权：学生可以改题、提升题，用“根还在，离现场更近”两条锚，分辨提升与逃离。第6章讲问对ChatSDE：先写第零轮初判，再做多轮自由问对，问对是来回磨，不是一问一答，问对记录就是发生档案。第7章讲一破一立：先接住，再批判，按判断坏在哪里选刀，破的同时指出立的方向，逼近而不替学生说破。

三章依次回答：题归谁，怎样磨，怎样破。题先归了学生，磨才磨得动；磨得动，批判才不会沦为套路。

05

第5章

改题权：学生改题、提升题

◆ 第2周的两份改题申请

经济管理类研讨课《小微经营者的动力与节律》开课第2周，任课老师收到了周川的两份改题申请，前后只隔了三天。老师发给周川的九问里，有一问是“你家饭馆哪个伙计干得最起劲？为什么？”。这一问的现场是周川家里的小饭馆，他从小在店里帮忙，三年的流水账也在他手里。第一份申请写着：想把这一问改成“AI会不会取代饭馆？”，理由是“这个问题更大、更前沿，也更有意思”。老师读完没有批，只在旁边写了一句：“这个问题，你家饭馆还在不在里面？”三天后来了第二份申请，周川把同一问改成：“父亲为什么一到月底就泄气？”理由只有两行：他翻流水账时发现，每到月底父亲记账的字迹就潦草起来，话也少了，可是月底的生意并不比月初差。

两份申请摆在一起，正好呈现了改题这件事的两副面孔。第一份改动很大，题目一下子从一家饭馆跳到了整个行业与技术的未来，看上去视野开阔，实际上周川家的饭馆、三年的流水账、他从小看在眼里的父亲，全都从题目里消失了；这样的题，任何一个大模型都能洋洋洒洒地答出几千字，它过不了题闸（见第4章）。第二份改动看起来更小，甚至更“家常”，题目里只有一个人和一个时间点；可是它离周川的现场更近了，近到只有他能回答，同时又没有离开课程母本关于小微经营者动力的核心判断。老师批了第二份，并在批语里写道：“这一问比我出的那一问好。”

这一幕引出了本章要回答的三个问题。第一，学生有没有权利改老师出的题？如果有，这项权利的依据是什么？第二，改题之后，怎样分辨学生是在提升题，还是在逃离题？周川的第一份申请与第二份申请都是改题，一份被打回、一份被批准，判别的标准是什么？第三，一道题从老师版到学生改版、再到问对后的定版，

这一串版本有什么用？本章的回答是：九问是种子不是考卷，学生改题时，谱系化由学生自己完成；分辨提升与逃离靠两锚，即“根还在”与“离现场更近”；题的版本史本身就是发生的证据，它进入学业评估。

> **这一章的问题**
>
> 学生有没有权利改老师出的题，又凭什么分辨他是在提升题，还是在逃离题？

◆ 九问是种子不是考卷

在讲授课堂里，题目是考卷的一部分，学生改题等于作弊，至少是答非所问。这一点在讲授课堂里完全合理：考卷的功能是检验学生是否到达了老师已知的终点，题目一旦被学生改动，检验就失去了共同的标尺。九问课堂里的题目功能不同。第4章说过，九问是老师从谱系母本精要库出发、按学生画像具体化而成的九个谱系性研究课题，它们打向连老师也写不出终点的地方。既然终点老师也写不出，题目就不可能是一把检验“到没到终点”的尺子，它只能是一个入口、一颗种子。种子种下去，长成什么样子，要看土壤；而土壤，恰恰是学生的现场。

把九问比作种子，还有一层更具体的意思。种子里携带着母本的基因，即它连着精要库里的哪一条核心判断；但种子长成什么形态，取决于它落在哪块地上、遇到什么样的天气。老师出题时，对学生现场的了解再深，终究是隔着一层的了解。周川的老师知道周川家开饭馆，知道他手里有三年流水账，于是出了“哪个伙计干得最起劲”这一问，想从伙计的劳动状态切入小微经营者的动力问题。但老师不知道周川父亲月底记账时字迹会变潦草，这样的细节只有天天在店里的人才看得见。当周川把题目改向父亲的月底时，他做的事情，正是让种子在自己的土壤里找到了更合适的长法。

本书主张，学生改题时，谱系化由学生自己完成。这句话需要解释。谱系化指的是让一个新判断接到学科谱系的某一节上，成为这一支脉上长出的新一节，而不是悬在空中的孤立观察。在讲授课堂里，谱系化是老师替学生完成的：老师把学科的来龙去脉讲清楚，学生只需记住每一节的位置。在九问课堂里，老师出题时完成了谱系化的前一半，即把题目接在母本的核心判断上；后一半，即把这一接口落到现场的哪一处、以怎样的形状落下去，要由学生在改题中完成。周川把“伙计”改

成“父亲”，把“起劲”改成“月底泄气”，他自己决定了课程母本里“小微经营者的动力”这一条判断，应当在他家饭馆的哪一个时间节点上被检验。这个决定，就是他亲手完成的一次谱系化。

改题权的教育学依据，可以在杜威那里找到一条线索。杜威在《我们怎样思维》中把思维的起点放在一个真实的困惑上，认为思维始于一种不确定的、需要解决的情境，而不是始于别人交给你的现成问题；在《民主主义与教育》中，他进一步主张教育应当从学习者自身的经验出发并使之改造。改题权把这个思想落到了操作层面：老师给出的题是一个情境的提议，学生要在自己的经验里找到那个真正让他困惑的点，把题目改到那里去。维果茨基的“最近发展区”也提供了一种理解方式：老师出的题如果离学生已有的能力太近，就成了练习；离得太远，又无从下手；学生的改题，在很大程度上是在自己调整题目与自身能力之间的距离，把题放到那个够得着又需要跳一跳的位置上。需要说明的是，这两条参照只是为改题权提供思想上的邻居，九问模式并不声称自己是对杜威或维果茨基的直接实现。

打一个日常的比方。一位朋友送你一包花种，附了一张纸条：“种在向阳的地方。”你家阳台朝北，没有向阳的地方，但你发现窗台上午有两个小时的斜光。你把花种在了窗台上，而不是照纸条去找一块不存在的向阳地。你没有违背朋友的意思，反而比照字面执行更好地实现了它。改题权的意思与此相仿：学生改题，不是违背老师的出题意图，而是在自己的现场里替这个意图找到了真正能落地的地方。

◆ 三个版本：老师版、学生改版、问对后定版

在九问课堂里，每一问通常会经历三个版本。第一个是老师版，即老师在出题台上打造、经过题闸检验的那一版。第二个是学生改版，即学生拿到九问后，在第1轮阅读、对照自己的现场之后作出的修改；学生也可以选择不改，不改同样是一种选择，需要在版本记录里写明“未改及理由”。第三个是问对后定版，即学生在与ChatSDE多轮问对、并回到现场取证之后，对题目作出的最后一次调整（问对的过程留给下一章）。三个版本依次排开，构成一道题的版本史。

许宁的第2问提供了一个清晰的例子。老师版是：“交班本上，跌倒前一班对该患者写了什么、没写什么？”许宁读完这一问，去翻了几天交班本，发现“写了什

么”很容易查，“没写什么”却很难说清，因为没写的东西太多了，几乎可以无穷列举。她意识到，真正有意义的不是“没写什么”的清单，而是“那一句为什么没写”。于是她把这一问提升为：“交班时被省略的那一句，是谁、在什么情况下省略的？”这一改，把一个查阅性质的问题，变成了一个关于交接行为本身的问题；它不再只看文书，而是看文书背后的人与情境。至于问对之后这一问是否还要再动、怎样动，属于第6章与第19章讨论的问对过程；本章只需看到，从老师版到学生改版的这一步，已经把题目从“文书里有什么”推进到了“交接时发生了什么”，而这一步是许宁自己迈出的。

三个版本的区分有两个用处。第一个用处是让改题成为一件有记录、可回看的事，而不是学生在心里悄悄换了题。没有版本记录的改题，与偷换概念在外表上无法区分：期末交上来的子本，研究的是另一个问题，老师无从知道这个问题是怎样一步步变过来的，是学生在现场里被逼出来的，还是嫌原题太难而绕开的。有了版本记录，每一次改动都有日期、有理由、有改动前后的对照，谁都可以回头去看。第二个用处是让问对的作用可以被观察。学生改版与问对后定版之间的差，往往就是问对带来的推进；如果两个版本完全一样，说明问对没有触动题目本身，这一信息对评估问对深度同样有用（见第17章）。

三个版本也不是越多越好、越改越大越好。有些学生在一学期里把一道题改了七八次，每一次都改出一个新方向，到期末仍然没有在任何一个方向上做出东西。这好比一个人装修房子，图纸改了八遍，墙一面都没砌。版本史的价值在于每一次改动都让题目在两锚上走得更稳，而不在于改动的次数。因此本书在评估中明确规定，改题质量按质不按量，版本的数目本身不计分。

还要说明的是，改题权落在九问的每一问上，却不要求九问问问都改。一位学生拿到九问之后，可能只改其中两三问，其余几问原样保留；也可能在问对过程中发现某一问已被其他几问覆盖，申请把它并入另一问。九问作为一个整体，是老师按学生画像设计的一组入口，彼此之间有互相校验的安排，改动其中一问时，学生需要留意它与其他几问的关系是否因此被打乱。好比一支乐队里有人换了乐器，换得好可以让整首曲子更出彩，但他得先听一听其他声部在奏什么，不能只顾自己那

一行谱。许宁只把第2问作了提升，其余八问未作大的改动，这并不说明她改题不积极，而恰恰说明她看出了第2问是整组九问的承重处，在那里用力最有回报。

◆ 两锚：根还在，离现场更近

改题权不是无限的。一旦学生可以随意改题，九问课堂就面临一个现实的风险：学生会把难题改成容易的题，把需要下现场的题改成可以在宿舍里用模型答完的题，把连着母本的题改成自己随便感兴趣的题。本书为改题设了两个锚点，用来检验每一次改动：第一锚是“根还在”，即改后的题仍然连着母本的核心判断；第二锚是“离现场更近”，即改后的题比原题更依赖学生自己的现场。两锚同时满足的改动，是提升；任何一锚松脱的改动，都需要重新审视。

第一锚“根还在”，守的是谱系。一道题之所以是谱系性研究课题，而不只是一份本地调查，是因为它连着精要库里的某一条核心判断，它的答案无论是什么，都会对这条判断构成支持、修正或反驳。根一旦断了，题目做得再好，也只是一个孤立的发现，接不回学科的谱系，也无法回写进母本。检验根是否还在，办法是问一句：改后的题，它的答案会让母本里的哪一条判断变得更可信或更可疑？周川的“父亲为什么一到月底就泄气”，答案会直接作用于课程母本中关于小微经营者动力的判断：如果动力的起落跟着账期走，而不跟着生意的忙闲走，那么动力就不只是一个心理或性格问题。根还在。而“AI会不会取代饭馆”这一问，答案作用于技术替代与行业前景的讨论，与这门课母本中的动力判断几乎没有关系，根已经断了。

第二锚“离现场更近”，守的是发生。第4章讲过，题闸检验的是：拿掉学生的现场材料之后，通用大模型还能不能给出说得过去的答案。改题时，第二锚相当于对改后的题再过一次题闸，而且要求更高：不只是改后的题要过闸，还要比原题更深地依赖学生的现场。检验的办法是问一句：改后的题，有多少答案只能从学生的现场里取得？“哪个伙计干得最起劲”已经落在周川家的饭馆上，但“起劲”是一个外部观察者也能大致看出的东西；“父亲为什么一到月底就泄气”则要求周川调动他对父亲多年的了解、翻出三年流水账的月底记录、在月底的某一天留心父亲的言行，这些都只在他的现场里。离现场更近了。“AI会不会取代饭馆”则把现场整个丢掉了。

两锚为什么要同时满足？只守第一锚，学生可能把题改得越来越理论化，紧紧贴着母本，却离自己的现场越来越远，最后做成一篇文献综述；这正是讲授课堂里学生写课程论文的常见形态，也是大模型最擅长代劳的形态。只守第二锚，学生可能把题改得越来越私人化，紧紧贴着自己的生活，却与学科母本失去了联系，最后做成一篇随笔或一份家庭记录。好比放风筝，线的一头必须拴在人手里，另一头必须连着风筝；线断在哪一头，风筝都会掉下来。两锚就是这根线的两头：根是拴在手里的那一头，现场是连着风筝、能让它乘风的那一头。

◆ 提升与逃离的分辨

有了两锚，就可以把改题分成两类：提升与逃离。提升是两锚同时收紧的改动，题目在根上扎得更牢、在现场里落得更深；逃离是至少一锚松脱的改动，题目或者离开了母本，或者离开了现场，或者两者都离开了。逃离并不总是学生有意偷懒，许多时候，它出于一种真诚的兴趣，或者出于对“大问题”的向往。周川的第一份申请就是这样：他并没有想躲避什么，他真心觉得“AI会不会取代饭馆”更重要。正因为逃离常常披着“更大”“更前沿”“更有意思”的外衣，分辨的工作才需要一套明确的标准，而不能只凭老师的直觉。

逃离常见的形态有三种。第一种是“上浮”，题目从具体的现场上浮到一般的议题，如周川的“AI会不会取代饭馆”；它的特征是改后的题能被通用大模型张口答出，题闸一测便知。第二种是“漂移”，题目仍然具体，却换到了另一个与母本无关的现场，例如一位学生把关于本地法院物业纠纷调解的题，改成了关于自己小区业主群吵架的题；它的特征是现场还在，根却断了，改后的题的答案无法作用于母本的任何一条判断。第三种是“缩退”，题目被改得更窄、更容易，却也更浅，例如把需要跟一次夜班的题改成只查阅一份已有统计表的题；它的特征是根与现场都还勉强挂着，但题目对现场的依赖变弱了，学生需要动用的现场材料越来越少。

提升也有可辨认的形态。与逃离一样，提升的形态也可以从两锚的变化上看出来。许宁的第2问提升，是一种“下沉”：她把“写了什么、没写什么”这个仍然偏于文书层面的问题，沉到了“是谁、在什么情况下省略的”这一行为层面；题目对现场的依赖加深了，因为她必须去观察交接、去理解省略发生的情境，而根依然连着母本中关于跌倒风险的判断。周川的提升是一种“对焦”：老师的题对准了伙

计，周川把焦点调到了父亲身上，调到了月底这个时间点上；对焦之后，题目从一个泛泛的劳动观察，变成了一个有明确时间结构的问题，而这个时间结构恰恰可能揭示母本判断中缺失的一项。

历史系《地方史料研讨》的苏文，提供了另一种提升的样子。老师为苏文出的九问里，有一问只从他的家乡县志出发，问的是县志对家乡某一时期的记载本身。这一问已经落在了苏文的现场上，县志是他家乡的县志，但它依然偏于查阅，答案大体可以在县志里直接找到。苏文在翻阅家中族谱时注意到，族谱在某一代突然断了，此后若干年没有任何记载，再接续时已换了一种书写方式。他把题目改成：“你家族谱断了的那一代，县志同年记了什么？”这一改，把两份原本各自独立的史料——一份是官修的县志，一份是私修的族谱——在同一个时间点上对接起来。根还在，因为它连着课程母本中关于地方史料与官私记录关系的判断；离现场更近了，因为族谱断代的那一年，只有拿着这本族谱的苏文才会去追问。后来发生的判断是：断代处恰是户籍登记口径改变之年。这一判断在本书中作为构造案例使用，其检验有待苏文的子本完成（见第12章与第23章）。

分辨提升与逃离时，有一个容易犯的错误，是以题目的“大小”作标准，认为改大了就是逃离、改小了就是提升，或者反过来。周川的逃离是改大了，苏文的提升却也在某种意义上把题目改“大”了，因为他把一份史料扩展成了两份史料的对照；而前面说的“缩退”，恰恰是改小了的逃离。可见大小不是标准，标准只有两锚。这好比判断一次搬家是不是好事，不看新房子比旧房子大还是小，只看它离工作地点是不是更近、和原来的亲友是不是还连得上。

在实践中，分辨工作由学生自己先做。每一份改题申请都要写明两锚：改后的题连着母本的哪一条判断，改后的题比原题多依赖了哪些现场材料。写不出两锚的改题申请，不必等老师批就可以退回给学生自己重想。周川的第一份申请，如果他事先被要求写两锚，大概写到第一锚时就会停下来，因为他找不到“AI会不会取代饭馆”与母本中动力判断之间的联系。把分辨的工作先交给学生，不是为了减轻老师的负担，而是因为分辨提升与逃离本身就是发生技术的一部分：一个学会了给自己的题检查两锚的人，以后面对任何一个新问题，都会先问一句“我是在往深处走，还是在往旁边躲”。

老师在这一环节的角色，也随之从“批准者”转为“追问者”。周川的老师面对第一份申请，没有直接写“不批准”，而是写了一句“这个问题，你家饭馆还在不在里面？”这句批语本身就是一次两锚检验，只不过是以提问的方式交还给学生。直接否决一份改题申请，学生得到的只是一个结果，下一次仍可能写出同样的申请；用一个问题把两锚指给他看，学生得到的是一把可以自己使用的尺子。这与第7章将要讨论的“先接住，再批判”同出一理：周川想谈AI的冲动本身并非没有价值，它说明他在关心饭馆的前途，老师的批语没有否定这份关心，而是把它引回到饭馆里面去。三天后，周川带回来的“父亲为什么一到月底就泄气”，某种意义上也是对那份关心的回答：饭馆的前途，或许首先系于掌勺的人每个月底还有没有劲头。

◆ 改题权落在哪里：与问题发现和研究型教学四象限对话

改题权并不是本书独有的想法。Getzels与Csikszentmihalyi（1976）在《创造性视野》里区分了被呈现的问题与被发现的问题，并提出问题发现这一构念：一个人怎样提出问题，预测他后来的创造性成果。这与本章的方向一致，老师出的九问是被呈现的问题，学生改题是把它变成自己发现的问题。本书可以把这条文献当作改题权的旁证，却不能停在那里，因为问题发现的论述主要在说明发现问题为什么重要，没有回答另一件事：学生自己提出的问题，好不好，凭什么说好。周川的第一份申请恰好是一个反例。“AI会不会取代饭馆”完全是他自己想到的，兴趣真诚，却被打了回去；如果“自己提的”就足够，这份申请就该批。它没有批，说明在“自己发现”之外还需要一把尺子，两锚就是这把尺子：根还在，是问它还能不能接回谱系；离现场更近，是问它是不是仍然只有这位学生才答得出。

打个比方，一个人拎着篮子逛菜市场，不照菜谱采购，看见今天哪样新鲜再决定做什么菜，这是问题发现的好处。可是这顿饭终究要端上家里的饭桌，还得用得上自家厨房里现有的锅灶和调料。满街最新奇的海鲜买回来，家里没人会做、冰箱也放不下，那不叫发现，叫逛得热闹。两锚问的就是：这顿饭端不端得上桌，以及它是不是比原来的菜单更像这个家的饭。

再看研究型教学一侧。Healey与Jenkins（2009）在英国高等教育学会的报告里，用两条轴把本科生与研究的关系分成四类：课程强调的是研究的内容，还是研

究的过程与问题；学生是听众，还是参与者。学生作听众、学研究内容，是研究引导；学生作听众、学研究的过程与方法，是研究导向；学生作参与者、围绕研究内容讨论与写作，是研究任务；学生作参与者、亲自做研究与探究，是研究型。九问课堂整体落在研究型这一格，这也是课程型本科研究和各种“学生作为研究者”的做法共有的位置，本书不主张自己占了一个无人到过的格子。改题权的落点要更细一些：在研究型这一格里，学生被允许参与的不止是做研究，还有框定研究的问题。而这一步恰是最容易出事的一步，框定问题的权力一旦交出去，学生既可能框得更深，也可能框得更浅。就这张四象限图本身而言，它按学生与研究的关系划格，并不回答问题由谁来框、框得好坏如何判断。

本书多出的一格因此很窄，也要说得准确：学生参与框定问题、教师鼓励学生提自己的问题，这些做法此前都有，本书不声称它们是新的。多出来的是两样东西的组合。其一，把“框得好不好”拆成两个每份申请都必须回答的问题，即改后的题连着母本的哪一条判断，比原题多依赖了哪些现场材料；其二，把这两句回答连同改动的日期、理由一并存进版本史，使它们成为可以被换手评分者核对的证据，而不是老师当场的印象。主张落在这个组合和它的证据责任上，而不落在“允许学生改题”本身。

这一组合有它的软肋。一个自然的疑问是，两锚的说明也可以写成套话：学生学会了在申请里写“根连着第几条判断”“现场依赖更深”，里面却没有东西。这种情况会发生，因此本书规定两锚说明必须能在问对记录或现场材料里找到对应的痕迹，找不到的，评分人以痕迹为准而不是以说明为准。另一种情形则是反过来：一位学生把老师版的题通读之后判断它已经紧贴现场，一字未改，在版本记录里写明理由，随后做出了很好的子本。这不是改题权的失败，而是它的一种正常用法，改题权保护的是学生框定问题的资格，不是改题的次数。

◆ 版本史作为发生证据

一道题的版本史，记下了老师版、学生改版与问对后定版，以及每一次改动的日期、理由和两锚说明。它看起来只是一张不起眼的登记表。这份记录在九问课堂里不是行政附件，而是学业评估的正式证据：评估方案八个项目中的第二项“改题质量”，量的正是“提升还是逃离：根还在吗，离现场更近了吗”，其证据就是题

的版本史（见第15章与第17章）。在本科研讨课的权重示例中，这一项占10%，各学段、各学科的示例里都保持在这一比例。

版本史之所以能作为发生的证据，是因为它记录了一个判断从无到有的来路。一份子本的最终判断，读者只能看到结果；而版本史让人看到，这个判断是从什么样的题里长出来的，题又是在哪一刻、因为什么被改变的。许宁的子本最终落在“风险藏在交班本里被省略的那一句”上，如果没有版本史，读者可能以为这是她一开始就有的想法，或者是某个模型替她想到的；有了版本史，读者会看到，老师版第2问问的是“写了什么、没写什么”，许宁改版问的是“被省略的那一句，是谁、在什么情况下省略的”，判断正是沿着这一改动的方向发生出来的。这好比一份医疗记录，最后的诊断固然重要，但首诊印象、复诊时的修正、检查结果出来后的再修正，这一串记录才说明这个诊断是怎样一步步得出的，也才让后来的医生能够检查它在哪一步可能出了错。

版本史还回答了AI时代课堂评估中一个越来越尖锐的问题：学生交上来的东西，是他自己发生的，还是模型代写的？一份子本可以在几分钟内由模型生成，文字流畅、结构完整，单看成品很难判断其来历。但一份有来路的版本史很难一次伪造：老师版的题是老师出的，有出题台的记录；学生改版有日期、有理由，理由里写着只有学生的现场才有的细节，如周川父亲月底的字迹；问对后定版与问对记录相互印证，问对记录里能看到学生在哪一轮被问住、又在哪一轮回到现场带回了新材料。一位学生如果交上一份高水准的子本，却拿不出一份能与之对上的版本史，这本身就是需要追问的信号（第17章将讨论“拿不出轨迹的高分子本”）。

版本史的记录方式应当尽量轻便。在实践中，一张简单的表格就够了：每一问占一行，依次记下老师版、学生改版、问对后定版的题面，每一次改动旁注日期，改动理由用两三句话写明，两锚各用一句话说明。不改的题同样要登记“未改”及理由，例如“原题已紧贴现场，改动反而会离开母本”。这张表在学期初建立，随问对与现场工作逐步填写，期末与子本一并提交。它不要求学生写得多，只要求写得真；每一次改动的理由，都应当能在学生的问对记录或现场材料里找到对应的痕迹。

版本史还有一个往往被忽略的用处：它是下一届九问的原料。周川把“哪个伙计干得最起劲”改成“父亲为什么一到月底就泄气”，这一改动本身就提示了一个出题的方向，即小微经营者的动力可能与时间节律有关；第14章将会看到，周川与同班林晓的题在碰撞中涌现出“小微经营者的动力被现金节律收走”这一班级判断。这些改动一旦回收进部落档案，下一届的老师出题时，就可以直接从“节律”这一方向下刀，而不必再从“伙计起劲”起步。学生的改题，于是不只是学生个人的一次发生，也是谱系母本的一次小小回写；当然，这种回写能否成立，要等它在下一届的九问与精要库版本里留下可观察的痕迹，不能由老师或学生事先宣布。

◆ 本章怎样算错

本章有三个主张：九问是种子不是考卷，学生享有改题权并在改题中完成谱系化；改题以“根还在”和“离现场更近”两锚分辨提升与逃离；题的版本史是发生的证据，进入学业评估。周川、许宁与苏文的改题都是构造案例，用来说明两锚怎样使用，不能当作两锚有效的证据。下面三条把失败写成可以对照的读数。

先看两锚能不能被稳定使用。取一门每届选课学生不少于30名的课，把学期内全部改题申请交给两位互不知情的评分者，按两锚各自独立判为提升或逃离；如果两人判断一致的申请不足80%，并且在补写更具体的判别细则之后下一届仍然不足80%，那么两锚就不是一个可以被稳定使用的标准，要么继续细化，要么另设标准，本章关于“提升与逃离可以分辨”的说法就要降为一种经验直觉。

再看改题权有没有真的带来新判断。连续两届、每届不少于30名学生，把经两锚判为提升的改题所对应的子本，与原样保留老师版的子本放在一起，由不知分组的换手评分者数出母本所无的判断各有多少，并以学生入学时的母本底盘成绩作控制，因为愿意改题的学生本来可能就更强。如果控制之后两组差异消失，改题权就只能收缩为“改题是允许的”，不能再主张它是谱系化由学生完成的关键一步。

最后看版本史能不能被事后补写。每届随机抽取不少于10份版本史，请评分者对照问对记录与现场材料上的时间戳，逐条核对改动理由；如果其中1/3以上的理由

找不到时间上早于期末的对应痕迹，说明版本史可以被一次性编出，它作为发生证据的效力就要下调，本章关于它能回应代写问题的论断也要随之减弱。

题改定了，周川手里握着“父亲为什么一到月底就泄气”，许宁手里握着“交班时被省略的那一句，是谁、在什么情况下省略的”。可是题目只是入口，判断还没有发生。学生拿着改定的题去找ChatSDE，第一句话应当说什么？问对是不是一问一答？如果ChatSDE第一回合就把学生的初判翻过来，学生接不住怎么办？下一章讨论问对：从学生先写初判的第零轮开始。

> **带走的话**
>
> **九问是种子不是考卷：学生可以改题，但只有两把尺子，根还在，离现场更近，不看题目改大还是改小。**

◆

06

第6章

问对ChatSDE：第零轮与多轮自由问对

◆ 八条原因与一条判断

开学第2周，化学系三年级《合成条件的发生》课上，沈澄把一句话敲进对话框：“某偶联反应产率偏低，可能的原因有哪些？”回答来得很快，也很周全：试剂纯度、溶剂含水、催化剂失活、投料比例、反应温度、搅拌效率、后处理损失、称量误差，一共八条，每条后面还附了排查建议。她把这份清单抄进实验记录本，觉得这一周的问对已经做完了。周四的研讨课上，老师只问了她一句：这八条里，哪一条是你的判断？她答不上来。清单里每一条都说得通，可没有一条是从本组那20批失败记录里长出来的；换一所学校、换一个课题组、换一个从没进过实验室的人来问，得到的也是这八条。

这个场景里的困惑，不在回答的质量，而在交互的形状。沈澄把问对当成了一次检索：发出一个问题，收回一份答案，交易结束。检索式交互交付的是“已有”，它把一个领域里公认的可能性按顺序摆好，这件事大模型做得又快又全，第2章已经把它归入发现一侧，在九问课堂里降为底盘。问对要交付的是另一种东西——一个此前不存在、连老师也写不出终点的判断。这种判断不会在一问一答里出现，因为它的原料有一半不在模型那里，而在学生的实验记录本、失败批次、仪器谱图和实验室墙上那只很少有人去看的温湿度计上。只要原料分处两地，判断就只能在两地之间来回搬运中慢慢成形。

本章的核心判断是：问对是来回磨，不是一问一答；自由风格与多轮往返，是知识发生非线性的要求，而不是一种可有可无的交互偏好。所谓非线性，指的是判断的形成不沿一条直线前进：它会在某一轮被推翻，在另一轮绕回起点，又在一次意外的复做里突然改道。一问一答的形状预设了终点已在对面等着，只要问对了就能取回；而九问打向的恰恰是对面没有现成终点的问题。下面先讲第零轮，即学生

在看到任何回答之前先写下初判；再讲多轮问对与回现场的往返；然后用沈澄的七轮问对把这两件事落到一个具体课题上，并专门交代“谁先说出关键判断”到底怎样算；最后讨论问对风格为什么必须自由，以及问对记录为什么就是发生档案。中间还要停一停，把第零轮放到生产性失败和ICAP框架旁边，问清楚它究竟多出了什么。

> **这一章的问题**
>
> 问对为什么必须从学生先押上桌的初判开始，并且在对话与现场之间来回磨？

◆ 第零轮：先把自己的判断押上桌

ChatSDE的问对从第零轮开始。第零轮不是提问，而是学生在看到任何回答之前，先用几句话写下自己对这一问的初判：我认为原因是什么，我凭什么这样认为，我打算到哪里去查。初判可以粗糙，可以错，但不能空，也不能是一句“请帮我分析”。这条规矩看起来只是一个小小的流程安排，实际上决定了整场问对的重心落在谁身上。没有第零轮，学生是来取答案的；有了第零轮，学生是带着自己的判断来接受检验的，问对从第一回合起就围着一个属于学生的东西转。

第零轮之所以必须先于一切回答，至少有三层理由。第一层，批判需要对象。ChatSDE的默认行为是批判性激发，第一回合就把前提翻过来看（见第8章），可是如果学生没有亮出前提，它能翻的只有教科书上的通论，翻来翻去仍是检索。第二层，第零轮给“谁先说出关键判断”立了一个原点。学业评估的第③项问对深度，要看关键判断是否学生先说（见第17章），而要判断“先”，就得有一个写在任何回答之前、带时间记录的起点。第三层，第零轮是回写的基线。第⑧项回写证据要把第零轮与期末自述对照，看这个学生被改写了什么；没有第零轮，期末的自述就只能和一份记忆里的自己比，而记忆总是善于把过去的自己修改得更接近现在。

打个日常的比方。学下棋的人如果只看高手的棋谱，从不自己落子，看完一百局也说不出自己会在哪一步走坏，因为他从来没有一步棋是“自己的”。复盘之所以有用，是因为盘面上有一手你当时真心认为是好棋的落子，现在被证明是坏棋，这个落差才是学到东西的地方。第零轮就是先落子。它要求学生在问对之前承担一点风险——写下一个可能被推翻的判断——而知识发生恰恰从这个可被推翻之处开

始。杜威在《我们怎样思维》里把反省性思维的起点放在一种困惑或疑难的情境中，并认为思维借由提出暗示、再加检验而推进；第零轮做的事情，就是逼学生在问对开始之前把自己的“暗示”明白写出来，好让后面的检验有处着力。

还需要说清楚的是，第零轮的初判错了不扣分，写得笼统也不扣分，扣分的只有两种情形：一是空白或以提问代替判断，二是事后补写、把问对中途才想到的东西倒填进第零轮。前一种使问对失去对象，后一种使发生档案失真。沈澄的第零轮写的是：“产率低是因为试剂纯度不够。依据是失败批次多半用的是同一批号的试剂，那瓶开封已久。打算换新批号试剂重做对照。”这个初判并不愚蠢，它正是那份八条清单里的第一条；它的价值也不在对错，而在于它是沈澄自己押上桌的，并且附上了她手里的证据——那份标了试剂批号的失败批次表。正是这份表，给了下一轮一个翻转的抓手。

◆ 第零轮与生产性失败、ICAP：多出的一格

第零轮并不是凭空发明的。Kapur（2008）发表于《认知与教学》的研究让学生在接受讲授之前先独立面对结构不良的复杂问题，学生的尝试多半以失败告终，但随后听讲时，他们对概念的理解常常强于先听讲、再练习的同伴。这一路线后来被称为生产性失败，Kapur（2016）在《教育心理学家》上又作了论述，Sinha与Kapur（2021）发表于《教育研究评论》的元分析则支持它在概念理解上的优势。第零轮的形状与之相近：先由学生独立生成一个判断，错了不扣分，然后才接受检验。如果有人说第零轮不过是生产性失败的生成阶段换了个名字，这个说法大半成立，本书需要回答的是剩下的那一小半。

剩下的这一小半有三处。第一，初判落在学生自己的现场上。生产性失败研究里让学生去试的问题，多由研究者或教师设计，同一个班面对同一道题，失败是相对于一个已知的标准解而言的；沈澄押上桌的却是对“本组那20批失败”的判断，换一个人写不出同一份，也没有标准解可以对照。第二，失败条件由学生自己写。在设计好的问题里，什么算失败由标准解来定，后继的讲授负责把差距补上；九问打向的地方没有标准解，没有人替学生划线，所以第4轮ChatSDE才会问沈澄“什么结果出来，会让你放弃它”，第6轮她要把答案写成判断旁边的失败条件。第三，第零轮兼作回写的基线。它带着时间记录，期末拿来与学生的自述对照，看这个学生

被改写了什么；就本书所读到的生产性失败设计而言，先行尝试的产出主要是学习机制中的一环，并不专门留作前后对照的凭据。

打个比方，买回一件组合柜，不看说明书先自己拼，拼到卡住再翻说明书，翻到那一页会格外认真，这是生产性失败的味道。第零轮多出来的，是这个柜子本来就是量着自家墙角定做的：没有说明书，卡住的地方只有你家的墙角才有；而且动手之前你得先在纸上写下“我认为这块板该装在这里，装错了会是这个样子”，日后才能知道自己究竟错在哪一步。

ICAP框架可以再补一个角度。Chi与Wylie（2014）把学习者的投入方式分为被动、主动、建构、互动四种，学习效果依次递增。拿它来看沈澄的第1周：把八条清单抄进记录本，是主动，她动了手，却没有产生清单里没有的东西；第零轮把她推到建构，她必须生成一个别人没有给她的判断；带着现场材料与ChatSDE往返，以及后面与同学的碰撞，则是互动。框架解释了为什么抄清单在投入的阶梯上位置很低，与本书的设计同向。它没有交代的是，建构的内容从哪里来，由谁来评价；本书补上的两项是：内容必须来自学生自己的现场，评价靠带时间记录的第零轮和换手评分。

还要承认两个限度。其一，第零轮的好处在本书所处理的这类问题上缺少直接的实证，生产性失败的证据多出自教师设计的问题和概念学习任务，搬到“没有终点的本地问题”上并不自动成立。其二，Kirschner等（2006）对最小指导的批评提醒我们，新手独自面对难题而无人接住，失败会变成单纯的挫败。因此第零轮之后紧接着就是ChatSDE的第1轮，学生押上桌的判断立刻被认真对待，而不是被晾在一边。这两点在本章末尾被写成了可以对照的读数。

◆ 多轮与回现场：问对是往返，不是闭门

第零轮之后，问对进入多轮。这里要先区分两个容易混淆的单位：回合与轮。回合是一次你来我往的对话，几分钟之内可以有好几个；轮则以一次回现场为界——学生带着这一轮磨出的问题回到病房、实验室、账本或档案里，带回新的材料，或者带回一次判断的改动，下一轮才算开始。一轮里可以有很多回合，但如果学生一直坐在屏幕前没有离开，不论聊了多久，都还在同一轮里。编号约定也在这

里一并交代：第零轮是学生的初判，不含对话；此后每带着现场材料回到对话一次，就开始新的一轮，依次编为第1轮、第2轮……本书把第零轮也计入轮数，所以沈澄的问对是第零轮加第1轮至第6轮，共七轮。这样划分，是为了把问对的重心固定在现场与对话的往返上，而不是固定在对话本身的长度上。

为什么必须往返？原因在第4章讲过的题闸里：九问之所以合格，正因为拿掉学生的现场之后，一个未接入这些材料的通用大模型答不出像样的东西。既然答案的一部分只在现场，那么不论模型多强，它都无法闭门把这一问做完；它能做的是把问题磨得更尖，让学生下一次回到现场时知道该看哪里。反过来，学生只泡在现场里、不回到问对，也常常看了很多却什么都没看出来，因为现场的材料不会自己按问题排好队。问对与现场的关系，像两只手搓一根绳：每一股都要被另一股拧过，才会有劲。舍恩在《反思性实践者》里描述过实践者与情境之间的“反思性对话”——实践者对情境做出一个动作，情境也以意外“回话”，实践者据此改变自己对问题的框定。问对的多轮往返，可以看作把这种对话拆成两个声部：现场负责回话，ChatSDE负责帮学生听出回话里那个不合预期的音。

中医的复诊可以帮助理解这种往返的节奏。初诊时，医生望闻问切，开出一个方子，嘱咐先服几剂再来；复诊时，他不再从头问起，而是看这几剂药下去之后舌苔变了没有、脉象缓了没有、睡眠和饮食怎样，据此加一味、减一味，有时甚至整个方向调过来。没有哪位负责的医生会在初诊时开定一整年的方子，因为病人的身体对药的反应，是只有服下去才知道的信息。问对的每一轮也是如此：上一轮磨出的问题，就是开给现场的一剂药；学生回到现场“服下去”，带回现场的反应，下一轮据此调方。轮与轮之间那段回现场的时间，不是问对的中断，而是问对最重要的部分。

多轮因此不等于多。第17章会讲到反刷分规则：十轮打转不如三轮推进。判断一轮问对有没有推进，不看它说了多少话，而看两件事：这一轮之后学生的判断有没有改动或加深，这一轮有没有带回此前没有的现场材料。只要其中一件发生了，这一轮就推进了；两件都没有，这一轮就是空转，哪怕对话记录很长。空转常见的样子有两种：一种是学生反复换一种说法问同一个问题，希望模型最终给出一个可

以抄的答案；另一种是模型反复把学生的话复述得更漂亮，学生误以为这就是深化。前者是检索的变体，后者是回音的变体，两者都让问对停在原地。

那么，一轮该在什么时候结束、学生该在什么时候离开屏幕回到现场？这个决定不由ChatSDE单方面做出，也不由一个固定的回合数机械地规定。一般的信号有三个：问题已经磨到只有现场材料才能推进的程度；学生提出了一个需要现场来检验的判断；或者对话开始绕圈、同一个意思被换着说法重复。出现前两种信号时，ChatSDE会把这一轮磨出的问题和学生的判断各写成一句，提醒学生带走；出现第三种信号时，它会直接指出对话已经空转，建议先回现场。老师则在每周阅读学生轨迹时检查这些切分是否合理，一个总是不肯离开屏幕的学生，和一个总是只在现场打转、不回来问对的学生，需要的指导是不一样的。

这种切分方式还有一个附带的效果：它让问对天然地慢下来。沈澄的七轮问对，第零轮落在第2周，第1轮至第6轮排在第4周到第9周的六周里，约一周一轮，期间穿插复做和翻档案的现场工作。慢不是缺点。判断的发生需要现场给出意料之外的回话，而现场的回话是按现场的节律来的——夏季批次要等到夏季，夜班要排到夜班，档案要一页一页翻。问对如果快过现场，它磨出的就只能是语言上的精致，而不是判断上的推进。

◆ 沈澄的七轮问对：从“试剂纯度”到湿度假说

下面用沈澄围绕失败批次的问对作主案例。需要先说明，这是构造案例，只作说明；问对片段经过删节，只保留与判断推进有关的回合；凡涉及实验操作之处，都只写到“某一步”的层面，不涉及任何具体的危险步骤。这场问对由第零轮加第1轮至第6轮构成，共七轮，每两轮之间都隔着一次回到实验室或档案柜的往返。读这段记录时，值得留意的不是ChatSDE说得多巧，而是每一轮的关键判断由谁先说出、现场材料在哪一轮进入、判断在哪一轮改了方向。

第零轮如上所述，沈澄押上的初判是“试剂纯度不够”，附上一张失败批次与试剂批号的对照表。这张表是她用了两个晚上从档案里摘出来的，按批号排序，每一批标了日期和产率；她只盯着批号这一列，从没有按月份把它排过一遍。第1轮，

ChatSDE没有顺着纯度往下排查，而是先接住她的证据：这张表做得认真，批号这一列正是检验纯度假说最该看的地方。接着它把表里的一个细节翻了过来：

> **ChatSDE：表里有一个批号同时出现在三月的两批和六月的三批里。三月那两批产率正常，六月那三批都偏低。同一瓶试剂，为什么三月好、六月差？如果是纯度不够，它在三月为什么没有表现出来？**
> **沈澄：可能是开封以后放久了，六月的时候已经变质了。**
> **ChatSDE：这是一个可以检验的说法。要分开“放久了”和“到了六月”这两件事，你觉得档案里有什么能帮你分开？**

这一轮的产物不是答案，而是一个更尖的问题，恰好对应她九问里的第1问和第2问：产率最低的二十批各在哪个月份？同一瓶试剂三月和六月用出的产率差多少？沈澄带着这两个问题回到课题组的十年档案。第1轮在这里结束，因为接下来的工作只能在档案柜前做。

第2轮，沈澄带回了她自己整理的分布：产率最低的那批记录，多半落在夏季前后，而且不止一个试剂批号如此，开封时间长短与产率的关系反倒不清楚。她在这一轮的第一句话里就修改了自己的判断：“看起来不是纯度，是季节。我猜是温度。”判断的方向改了，这是这场问对的第一次实质推进，而且是她先说出来的。ChatSDE没有替她把“温度”展开成一套热力学推论，而是问了一个回到现场的问题：档案里“室温”这个词，十年里有没有一次写过具体的读数？这正是九问第3问问到的地方。

第3轮带回的发现让沈澄自己也有些意外：十年的实验记录里，“室温”只是一个词，从没有写过具体读数，湿度更是一次都没有出现。与此同时，她在翻失败批次时注意到另一件事：同一步操作，有的记录写了时长，有的没写，写了的也长短不一，这对应九问的第4问。档案在这里走到了尽头——它不记录的东西，再怎么翻也翻不出来。问对片段如下：

沈澄：档案里没有温湿度，这条线索断了。是不是只能放弃季节这个方向？
ChatSDE：档案断的地方，往往就是规程没写的地方。断了说明十年里没人觉得它值得记。你手里还有什么来源，记得档案没记的东西？
沈澄：师兄师姐。他们做这个反应做了好几年，常说“那一步要快”。

“那一步要快”这句话，是沈澄自己从记忆里捞出来的。九问的第5问问的正是它：师兄师姐口头说的“那一步要快”，规程里写了吗？她带着这一问去找了两位高年级同学，又把本组规程逐行读了一遍。

第4轮，沈澄确认了规程里没有这句话。她把规程从头到尾读了两遍，那一步的说明只有操作顺序，没有任何关于快慢的要求。追问之下，师兄解释说，夏天那一步慢了就容易坏，冬天倒不太要紧；至于为什么，更早的师姐当年交代时只丢下半句“免得受潮”，再往下他也说不清。问对在这里出现了全程最关键的一个回合：

ChatSDE：“要快”和“夏天要紧、冬天不要紧”放在一起，你怎么读？
沈澄：如果只是温度，冬天暖气房里也不低。夏天和冬天差得最大的，是空气里的水。那一步慢了，体系暴露在空气里的时间就长了。我现在的判断是：成败可能跟湿度和那一步的时长一起有关，纯度不是主因。
ChatSDE：这个判断把三条材料串起来了：季节分布、档案里从不记录的读数、口头传下来的“要快”。它目前还只是一个串联，不是证据。什么结果出来，会让你放弃它？

湿度假说就在这一回合里成形，而说出它的是沈澄。ChatSDE在前三轮里做的，是把她的初判一次次翻过来，把问题磨到只有现场能回答的程度；在这一轮里它做的，是在她说出判断之后立刻追问失败条件，而不是夸奖或补充。这个分工后面第7章还会细讲。

第5轮开始时，沈澄带回来的是一次复做的记录。按九问第6问，沈澄在老师和带教研究生的指导下，严格按照现行规程和实验室安全规范复做一批，不改动任何步骤，只把每一步的时间、实验室的温度和湿度都记下来，再与档案比。复做的结

果并不戏剧化：她只做了一批，季节也不在夏天，单凭这一批无法证明什么。但记录本身改变了问题的形状——它第一次让“那一步”有了一组具体读数，也让她看清档案里哪一步的记录最模糊。同一轮里，她还按第7问查了谱图档案，发现那个反复出现、没人解释的小峰，在多个夏季批次的谱图里更常见，最早可以追到很多年以前的记录。这个小峰是什么，她还说不出，只把它列为待鉴定。

在这一轮的后半段，沈澄又转向外部。按九问第8问，她比对了几家实验室公开发表的同类反应规程，注意到其中有的规程在相应环节多了一句关于操作环境的说明，而本组规程里没有。ChatSDE随即提醒她区分两件事：别的实验室多写了一句，说明这个变量值得注意，却不能说明本组的失败就是它造成的；能回答后一个问题的，只有本组跨季节的复做。

第6轮是最后一轮，沈澄写下了这一学期的判断：本组某一步的成败依赖未入规程的环境湿度与操作时长——规程里没写的那一句，在师门口头里传了十年。她给它标上“待复做验证”，并回答了九问的第9问：如果把湿度写进本组规程，第一个要改的，是那一步的操作说明，应当写明记录温湿度和操作时长。ChatSDE这一轮做的最后一件事，是和她一起把验证方案的边界写清楚：需要跨季节、多批次的复做记录；那个小峰需要鉴定；在这些完成之前，这个判断只能作为假说提交给课题组，不能写成结论。

回头看这七轮，判断改了两次方向、收窄了一次：从纯度转向季节与温度，再从温度转向湿度与那一步的时长，最后收窄为“未入规程的一句”。每一次改向，触发它的都是一份从现场带回的材料，说出新方向的都是沈澄自己。ChatSDE始终没有给出那份八条清单，它的追问里也始终没有出现“湿度”“空气”这样的字眼；它做的是翻转前提、把问题指回现场、在判断出现后追问失败条件。如果把这场问对压缩成一问一答，它根本不会发生，因为档案里没有、模型里也没有的那句“那一步要快”，只有在第3轮档案走到尽头之后才会被想起来。

◆ 谁先说出：核心判断首发的口径

“关键判断是否学生先说”是问对深度的评分点，也是本书后面几章反复使用的读数，所以“先”字的口径要先写死。首发是相对于ChatSDE而言的：核心判断

里承重的那个关系——在沈澄这里，是成败与环境湿度、那一步的时长一起有关——如果最先落成文字的是学生的话，而不是ChatSDE的追问，就算学生首发。

有两类东西不算替学生说破。第一类是现场里别人说的话。师兄转述师姐的那半句“免得受潮”，是沈澄向高年级同学打听来的，属于她自己取得的现场材料；第二类是档案里的记录，批号表也好，十年里从不写读数的“室温”也好，都是她翻出来的。学生在现场里听到什么、翻到什么，本来就是她的原料，原料要不要用、怎样串起来，才是判断的工作。第4轮里，沈澄把“受潮”、季节分布和档案空白读成同一件事，并且补上了师兄没有说的另一半，也就是那一步暴露在外面的时间，这一步是她自己做的。

另一面的规矩是对ChatSDE的：它的追问不得出现核心判断的关键变量词。沈澄这一课的词表是“湿度”“空气”“受潮”“水汽”之类，许宁那一课则是“交接”“交班”之类；词表在九问定稿时由课程负责人封存，之后可以对问对记录逐回合扫描。回头看沈澄的第1轮至第6轮，ChatSDE问的是同一瓶试剂为什么三月好、六月差，“室温”二字十年里有没有写过读数，“要快”和“夏天要紧、冬天不要紧”放在一起怎么读，没有一句落在词表上。

打个比方，你在家里找钥匙，路过的邻居说“刚才好像看见你在玄关柜那边放过东西”，这是现场里别人的话，找到钥匙仍是你自己找的；陪你找的人若直接说“在鞋柜上”，就是说破了，钥匙再从鞋柜上拿出来，也不算你找到的。首发率要量的是后一种情形有没有发生。若把邻居的话也算作提示，任何现场工作都会让首发率归零，这个读数就废了；若不限制陪找的人说什么，“学生先说”就退化成“学生先把陪找者暗示的词复述一遍”。

有一处灰区必须如实写明：沈澄的九问第3问和第6问（以及后面的第9问）的措辞里，已经出现了“温度和湿度”或“湿度”的字样。九问是出题台给的课程材料，不是ChatSDE的追问，本书把它与档案、师兄的话同样看作学生取得的材料；但它离提示毕竟更近。因此评分人读首发率时，要在记录里注明九问措辞中已经出现了哪些变量词，由换手评分者结合学生的初判和改动去判断：首发的是不是那个关系，而不只是一个字眼。九问第3问问的只是档案里有没有读数，并没有问湿度能

不能解释成败；成败与湿度、时长一起有关这个关系判断，仍是沈澄在第4轮自己说出的。这个灰区无法消除，只能写明，让别人可以不同意。

◆ 风格自由：追问、故事与争辩

问对有一个容易被误解的地方：既然它有工序、有规矩，似乎就应当有一个标准话术，每个学生照着同一套模板跟ChatSDE说话。本书的主张正相反，问对的风格必须自由。所谓自由，不是指没有规矩——第零轮必须有，关键判断必须学生先说，每一轮必须以回现场为界——而是指学生用什么样的方式说话、ChatSDE以什么样的节奏回应，不预先规定。至少可以分出三种常见风格：追问型、故事型、争辩型。同一个学生在一场问对里也常常换着用。

追问型最接近人们对研究对话的想象：一个问题接一个问题，层层剥开，每一回合都比上一回合更窄。沈澄的问对前三轮大体是这种风格，适合手里材料较整齐、问题已经比较清楚的学生。故事型则从一段经历讲起，学生先不急着下判断，而是把现场里发生过的一件具体的事原原本本讲出来——某个夜班听到的动静，某次复做时手套里出的汗，某天父亲关上店门后坐在门槛上的样子——ChatSDE从故事的细节里找那个不合常理的地方。故事型适合那些现场感很强、却还说不清自己困惑在哪里的学生，人文学科的问对常以它为主，叶青写外婆时那个一直不敢写进去的方言词，就更可能在讲故事时被说漏出来（见第23章）。争辩型是学生先立一个强论点，要求ChatSDE全力反驳，自己再逐条守住或放弃；它适合已经形成初步判断、需要检验其韧性的阶段，沈澄在第4轮说出湿度假说之后，问对其实就转成了争辩型。

为什么风格不能统一？因为知识发生是非线性的，而不同的非线性需要不同的推进方式。打个比方，开一把锁可以用钥匙，可以找锁匠，也可以拆掉门轴，用哪一种取决于这把锁卡在哪里，而不取决于开锁人的偏好。判断卡在“问题太宽”时，追问最有效；卡在“说不出困惑”时，故事最有效；卡在“判断已有却不敢信”时，争辩最有效。如果规定所有学生都只能追问，故事型学生的判断就会被逼着过早地抽象化，那些只在细节里才显露的东西会在抽象中丢失；如果规定都只能争辩，一个还没有初判的学生就会被迫硬造一个论点去挨打，问对变成表演。

风格自由也是对ChatSDE的要求，而不只是对学生的许可。一个只会一种回应节奏的系统，不论这种节奏多么严谨，迟早会被学生摸透成套路：学生知道它每次都会反问“为什么白天少”“为什么月初不累”，就会提前把这种反问写进自己的话里，问对从此只剩形式。批判之所以不能套路化，靠的是按判断坏在哪里选择不同的工序；风格之所以能自由，靠的也是同一件事——ChatSDE跟随的不是一份话术表，而是学生判断此刻的形状。维果茨基的“最近发展区”可以在这里借来说明：有效的协助落在学习者独立能达到的水平与在协助下能达到的水平之间，而这个区间对每个学生、每个阶段都不一样，协助的方式也就无法一成不变。

◆ 问对记录即发生档案

问对的每一个回合都被完整保存，第零轮带时间记录，每一轮标明回现场的起止与带回的材料。这份记录不是对话的副产品，而是九问课堂最重要的档案之一：它同时承担着发生证据、评估依据和回写原料三种职能。讲授课堂里，类似的东西是学生的课堂笔记，它记录的是老师讲了什么；问对记录记录的是学生想了什么、在哪里改了主意。两者的区别，正是交付知识与交付发生的区别。

作为发生证据，问对记录回答了“这个判断是怎样来的”。沈澄的子本里写着湿度假说，读者可能会问：这是她想出来的，还是从哪里抄来的，或者是模型替她写的？讲授课堂里，这个问题几乎无法回答，因为期末交上来的只有终点；九问课堂里，这个问题可以通过翻阅问对记录得到回答——从第零轮的“试剂纯度”到第4轮她自己说出“空气里的水”，再到第6轮写下的假说和失败条件，每一步都有时间、有现场材料、有说话人。第17章会说明，这种轨迹正是不可还原性的证据，也是对AI代写问题的正面回应：拿不出轨迹的高分子本，本身就是一个需要解释的异常。

作为评估依据，问对记录是第③项问对深度唯一的证据来源。评分人看的不是轮次多少、字数多少，而是三件事：轮次之间有无推进，学生接不接得住批判，关键判断是否由学生先说。按照评估方案的配套规则，这一项须换手评分并写利益冲突声明，也就是说，陪学生问对的那个系统不能同时给这场问对打分，同底座的另一个智能体也不能充当独立的评分人。问对记录的完整保存，正是换手评分能够进行的前提。

作为回写原料，问对记录会进入课题组的部落档案。沈澄这一届的问对，会成为下一届学生的母本材料之一：下一届拿到的第3问，也许就不再是“室温有没有写过具体温湿度”，而是“本组规程加入温湿度记录之后，夏季批次的产率有没有变化”。这样一来，一份问对档案就不只是一个学生的学习记录，而是一条谱系往前长出的一节。第3章讲过部落档案如何从薄到厚，第24章将讲回写怎样被观察。这里只强调一点：如果问对不被记录，或者只记录终点，那么部落档案能积累的只是结论，而结论是最容易被大模型复述、也最不能传下发生技术的东西。

一份合格的问对记录，在形式上至少包含四样东西：第零轮初判及其时间记录；每一轮的起止，以及这一轮带回的现场材料清单；每一次判断改动的位置，标明改动由谁先提出；最后一轮写下的判断及其失败条件。这四样东西不要求学生事后整理成漂亮的报告，系统在问对过程中就会随手标注，学生只需在每一轮结束时确认。它们合起来，就是一条从初判到判断的可追溯路径。拿沈澄的记录来说，读者不必读完全部对话，只看这四样东西，就能知道她的判断在第2轮从纯度转向季节，在第4轮落到湿度与时长，触发这两次转向的分别是档案里的季节分布和师兄口中的“那一步要快”。

还要说明问对记录的使用边界。它是发生档案，不是监控工具；学生知道它被保存、被谁读、用于何处，问对开始前要说明这些。对医学和法学这样涉及患者与当事人的学科，问对记录本身也要遵守去标识化和隐私规则，不能因为它是“学习材料”就放宽。一份让学生不敢说出错误初判的问对记录，会把第零轮变成表演，从而毁掉它自己存在的理由。

◆ 本章怎样算错

本章押了三个主张：问对是来回磨而不是一问一答；第零轮先于一切回答，决定了问对的重心；问对记录就是发生档案。沈澄的七轮问对是构造案例，只能说明这套设计的样子，不能证明设计有效。下面把失败写成可以对照的读数。

先看第零轮。取同一门课、同一批九问的两个平行班，每班不少于30名学生，一个班保留第零轮，一个班取消、直接进入问对；再取第三个班，保留第零轮，但初判的问题换成老师统一设计的一道题，不落在各自的现场上。三个班由不知分组

的换手评分者读问对记录和子本。如果连续两届，保留第零轮的班在关键判断由学生先说的比例上，与取消的班之差不足10个百分点，并且子本创新智商的班均分之差落在两位评分者本来的分歧范围之内，“第零轮决定问对重心”就不成立；如果第三个班与第一个班在子本里母本所无判断的比例上没有差别，那么“初判落在学生自己的现场”就不是多出来的一格，第零轮与生产性失败的差别只剩下名字。

再看多轮往返。把同一组九问随机分给两组学生，每组不少于30人，一组用一问一答的检索式交互，每问一次提问、收一次回答，另一组用带回现场的多轮问对，学期末由换手评分者数出母本所无、并且经现场检验仍站得住的判断。如果检索式一组产出这类判断的学生比例不低于多轮一组，“问对必须多轮往返”就要收缩为一种偏好，而不再是知识发生的要求。

最后看记录会不会被拿去表演。每届随机抽取不少于10份问对记录，核对第零轮的时间戳是否早于该学生得到的任何回答，并按封存的词表扫描ChatSDE的追问。如果第零轮倒填的记录占到10%以上，或者追问里出现词表关键词的回合占全部回合的5%以上，首发率这一读数就不能再用，“问对记录即发生档案”也要退回，重新设计记录与评估的分离。

还有一个问题本章没有回答。沈澄的问对之所以推进，关键在于ChatSDE每一轮都把她的前提翻了过来：同一瓶试剂为什么三月好、六月差，档案断的地方是不是规程没写的地方，“要快”和“夏天要紧”放在一起怎么读。这些翻转并不来自同一句话术，每一次刀口都落在不同的地方。ChatSDE凭什么知道这一刀该从哪里下？它怎样在翻转前提的同时，不把判断替学生说破？下一章讨论这个问题：一破一立，以及背后的工序选择。

> **带走的话**
>
> **先把自己的判断押上桌，再带着现场材料在对话与现场之间来回磨；关键判断由谁先说出，决定它是发生出来的，还是被告知的。**

◆

07

第7章

一破一立：批判性激发与工序选择

◆ 两句反问从哪里来

第8周的教研会上，一位旁听过两门九问课的老师提出了一个尖锐的问题。她在医学院听过许宁那一组的问对展示，在经济管理学院听过周川那一组的展示，发现两边的ChatSDE第一回合都在“反问”：许宁说夜间跌倒是因为患者年纪大，ChatSDE问“同样年纪的患者，为什么白天跌得少？”；周川说父亲一到月底就泄气是因为累，ChatSDE问“月初也忙，为什么不累？”。两句反问的句式几乎一样，都是找一个对照组，把学生给出的原因放进去，看它解释不解释得了差别。她的疑问是：这是不是一种套路？如果是，学生用不了多久就会学会提前堵住这种反问，问对就会变成一场彼此预判的游戏；如果不是，那它和套路的区别究竟在哪里？

这个疑问值得认真对待，因为它碰到的正是批判性激发最脆弱的地方。批判可以被模板化：任何一个系统，只要在每次回答前加一句“但是反过来想”，看起来都在批判；任何一位老师，只要每堂课都追问“你的证据呢”，看起来也在训练批判性思维。可是模板化的批判有一个致命的特征——它不看对象。不论学生说的是什么、坏在哪里，它都从同一个角度下刀。时间一长，批判就变成了噪声，学生学会的不是检查自己的判断，而是在表达上预先做好防御。

本章的核心判断是：ChatSDE与学生一起做的，是批判性思维与创新发生思维的合一；批判之所以不套路化，来自SDE的工序选择——按判断坏在哪里选刀；而且破的同时要指出立的方向，否则批判只留下一地碎片。回到那位老师的疑问，两句反问句式相近，但下刀的理由并不相同：许宁的判断缺了一样东西，周川的判断断了一个环节。句式只是刀柄，刀口落在哪里，是由判断本身的形状决定的。可是“形状”不能只凭语感去认，否则这个回答本身就成了一个说不清的套路，所以本章在三种刀之后专门写入一套判别细则，规定三个问题的查问次序，并用书中四

个案例逐一过一遍。下面依次讨论先接住再批判的次序、三种工序各是什么、判别细则怎样落到具体的初判上、破与立怎样同时发生、几个案例中刀口的不同、这套做法与“固定唱反调”的区别，以及为什么ChatSDE要把选刀的依据亮给学生看。

> **这一章的问题**
>
> 同样是一句反问，ChatSDE凭什么知道这一刀该从哪里下，又怎样在破的同时不替学生说破？

◆ 先接住，再批判

一破一立的第一步不是破，而是接住。学生在第零轮写下的初判，哪怕粗糙，也通常包含着一个真实的观察：许宁说“年纪大”，背后是她在病房里看到跌倒的大多是高龄患者；周川说“累”，背后是他在饭馆里看到父亲月底确实一脸倦容；沈澄说“试剂纯度”，背后是她整理出的那张批号对照表。这些观察都是对的，错的或者不完整的，是从观察跳到原因的那一步。接住，就是先把观察里对的东西指认出来、放大，让学生知道哪些部分站得住、值得带着往下走。

为什么要先接住？至少有两层原因。第一层是认知上的：批判需要一个立足点。如果ChatSDE一上来就说“年纪大不是原因”，许宁失去的不只是一个错误的解释，还有她手里那个正确的观察，她会不知道接下来该往哪里看。接住之后再批判，等于告诉她：你看到的现象是真的，现在我们来检查你给它安的原因。第二层是关系上的：一个每次开口就被否定的学生，很快就会停止在第零轮写真话，转而写一些四平八稳、挑不出毛病也推不动任何东西的初判。第6章说过，第零轮的价值在于学生肯押上一个可能被推翻的判断；如果批判的方式让押注变得太贵，第零轮就会名存实亡。

日常生活里，好的教练都懂这个次序。一位游泳教练看学生游了一趟，不会先说“你的腿完全错了”，而会说“你的换气节奏已经稳了，这很好；现在看腿，你的腿在往下沉，因为……”。前半句不是客套，它告诉学生哪些动作可以保留，从而让后半句的纠正有了明确的边界。接住与放大不是讨好，它和批判是同一个动作的两半：先划出站得住的部分，批判才能精确地落在站不住的那一截上。反过来，如果教练只会说“很好很好”，那是一种失败；如果教练一开口就把正确的动作全

部示范完，让学生照着做，那是另一种失败。第8章要讨论的通用大模型的默认行为更接近后者：它往往会批判，却在同一回合把答案一并给出。

弗莱雷在《被压迫者教育学》里批评过一种“银行储蓄式”的教育：教师往学生头脑里存入内容，学生的任务是接收、保存与复述；他主张的“提问式”教育，则把师生都放到对现实的共同探究之中。接住学生的初判，正是在承认学生手里有教师和系统都没有的东西——他自己的现场观察。九问课堂里的批判，不是从系统向学生单向输出正确答案，而是双方围着学生带来的观察一起工作；没有接住这一步，批判就会悄悄退回储蓄式，只不过储进去的不再是知识，而是对学生判断的否定。

还要说明，接住不等于先夸一句再转折的话术。“你的想法很有意思，但是……”这种句式，学生一听就知道前半句是铺垫。真正的接住要具体：指出初判中的哪一条观察、哪一份证据是站得住的，并说明它为什么值得保留。ChatSDE对沈澄的第一句回应，是说她那张批号对照表正是检验纯度假说最该看的地方；它接住的不是“试剂纯度”这个结论，而是她做那张表的方法，恰恰是这张表，下一回合成了翻转前提的抓手。接住的东西，往往就是破的地方，这是一破一立能够连成一体的第一个原因。

◆ 三把刀：缺式、偏路、断拍

接住之后，判断坏在哪里？SDE提供了三种查法，对应三种刀，本书称之为工序选择：用三大方程查缺式，对应补式；用六路径查偏路，对应排路；用123原理查断拍，对应点火。三种查法看的是判断的三种不同毛病，开出的方子也互不相同，不能混用。需要先申明口径：三大方程F、G、H在本书中只是关系占位记号，标示显露、差异序列与特征纠缠三者互为函数的关系，不是可以代入数值求解的公式；六路径是主导的分析次序，不穷尽全部方法；123原理只指“D—E矛盾→逼出S改变→S回写D、E”这一循环。

第一种毛病是缺式。三大方程说的是，一个显露出来的现象，是差异序列与特征纠缠共同作用的结果，三者中任何一样都是另外两样的函数。一个判断如果只把其中一侧放进解释，比如只讲对象的属性、不讲时间上的次序，或者只讲事件的先后、不讲牵涉其中的人和物，它就“缺了一式”：有一项该有人管的关系，在这个

判断里没有人管。缺掉的那一项，通常是被当作常数的因素——学生默认它不会变，所以从来没有想到要去问它。查缺式的方法，是把判断写满，逐一问每一项关系由谁承担，找出那个空着的位置。补式就是把空着的那一项补进解释：不是推翻原来的判断，而是指出它缺了哪一块，并把学生的注意力引到能补上这一块的现场材料上。许宁的“年纪大”是典型的缺式：年龄是患者的特征，可跌倒这个现象在白天和夜里分布不同，这个差别属于时间上的次序，而她的解释里没有任何一项管时间。

第二种毛病是偏路。六路径说的是，一个判断的形成有不同的次序：可以从已有的显露出发，再找差异，再落到现场；也可以从现场的纠缠出发，再看显露，再理出差异；如此等等，共有六种主导次序。一个人如果总是沿着同一条次序走，他的判断就会系统地带上这条路的盲区，而他自己很难察觉，因为在他看来，这就是思考的样子。偏路不是走了某条路，而是只走一条路、且卡点恰好落在它的盲区里，所以单看一条初判，往往还不足以认定偏路。查偏路的方法，是让学生回顾自己这一问是怎样一步步想过来的，只问先、再、最后，不问为什么。排路就是换一条次序走：把惯走那条路的最后一步，当作新路的第一步。排路的要点在于一次只换一条，并且新路必须是学生自己走过去的，不能由ChatSDE替他走完。

第三种毛病是断拍。123原理说的是一个三拍的循环：差异序列与特征纠缠之间出现矛盾，张力积累；张力逼出显露改变，形状变了；改变后的显露回写差异与纠缠，新的状态稳住。判断坏在动力上，常见的样子是循环在某一拍断掉：张力还没积累起来就被卸掉了，或者张力在了、形状却不变、另找出口泄掉了，或者形状改了、没有回写、过一阵又回到原样。最容易辨认的一种断拍，是学生在自己的现场里早已碰到了与初判直接相撞的事实，判断却纹丝不动——矛盾已经在了，只是没有逼出任何改变。查断拍的方法，是沿着时间问三句话：那时候哪两样东西顶着？顶着以后，什么变了？过一段时间，还是那样吗？点火就是对着断掉的那一拍下手：张力被卸掉的，帮它留住；形状不变的，给它找变形之处；没有回写的，做一个看得见的回写记号。点火不是加油，不是鼓励学生“再努力一点”，而是找到循环断在哪一拍，把那一拍接上。

◆ 判别细则：三个问题，按次序查

三种毛病摆出来之后，随即冒出一个麻烦：同一条初判，常常几处都看得出毛病。几乎任何一条单因归因，既可以说它“少了一项”，也可以说它“有矛盾没处理”，因为凡是存在相撞的事实，就意味着解释里缺了能容纳这个事实的一项。如果三种刀是三个并列的标签，标注者想贴哪个就能贴哪个，工序选择就成了事后给已经发生的反问补写的理由。所以本书不把它当作分类，而当作一套查问的次序：按下面三个问题依次查，哪一问先命中，就停在哪一把刀上，不再往下查。

可以用汽车仪表来打比方。一位司机开着车，发动机故障灯已经亮了，他看见了，却照开不误，这是第一种情形——警告已经在眼前，缺的不是信息，而是把警告当回事；这时候再给他加装更多传感器没有用，要让灯被看见。另一位司机的车根本没有机油压力的传感器，所以没有任何灯亮，他不是不理会警告，而是没有警告可理会；这时要做的是补上那一项检测。第三位司机，传感器齐全，灯也都不亮，可他每次出发总是先擦车窗再看胎压，临到下雪天才发现该先看胎压——问题只出在起手的次序与当天的任务不配。三种情形对应三把刀，而查问的先后恰好是：先看有没有灯已经亮了，再看有没有哪一项根本没人看，最后才看检查的次序。

细则的第1问查断拍：学生在自己的现场里，是否已经接触到与初判直接相撞的事实，而判断并未因此改动？如何认定“接触到”，要有一条可以复核的判据，本书称之为知情测试：把那个事实用一句话摆到学生面前，如果他的第一反应是“我知道，可是……”，就算已接触；如果是“咦，我没想过”，或者一边琢磨一边顺手补进一个新因素，就不算。复核时翻问对记录：看这个事实是否在ChatSDE开口之前，已经出现在学生自己的第零轮、记录本、日志或日常叙述里。现场里他人说的话、档案里写着的东西，只要学生已取得，都算他的现场材料。第1问命中，标为断拍，刀是点火：不需要给他新的材料，需要让已经在手的矛盾被看见，并且留住它造成的张力。

细则的第2问查缺式：若无已接触的相撞事实，初判的关系里是否少了一项被当作常数的因素？这一问也有读数。把学生要解释的现象摊开，找一个变量，它在现象中取值不同——白天与夜里、三月与六月、月初与月底——而在初判里被默认为不变。找得到，标为缺式，刀是补式：追问被漏掉的那一项，并把学生引向能给这

一项取值的现场材料。第2问命中的前提，是第1问没有命中，这个次序不能颠倒：如果把相撞的事实当作缺项去补，学生得到的只是一份更长的原因清单，而他其实早已握着那个矛盾。

细则的第3问查偏路：若关系齐全、也无矛盾，起手次序与任务是否不配？它排在最后，是因为偏路是最难单凭一条初判认定的毛病，要看同一个人在多个问题上是否反复沿着同一条次序起手、又反复卡在同一类地方。比方一个人排查电脑变慢，磁盘、内存、启动项全都想到了，也没有哪件事与他的判断打架，可他每一回都先重装系统、再去看任务管理器，这就是偏路，刀是排路：把惯走那条路的最后一步，换成这一次的第一步。前两问都落空、第3问又缺乏跨问题的记录时，标注可以写“暂无法判定”，而不必硬贴偏路。

下面把书中四个案例按这三问各过一遍。表中第1问一栏写的是学生在ChatSDE开口之前已经握有什么，第2问一栏写的是初判里被当作常数的那一项。

人物与初判	第1问：已接触的相撞事实	第2问：被当作常数的一项	判定与刀
许宁：夜间跌倒是因为年纪大	无。近三个月的跌倒从未按白天与夜里分开排过，“白天跌得少”没有出现在她的任何记录里	交接的时间次序：同一位患者年纪不变，白天与夜里的班次交接却不同	缺式，补式
沈澄：产率低是因为试剂纯度不够	无。她那张表按批号排序，日期只是其中一栏，她没有把同一批号三月好、六月差读成与初判相撞的事实；被指出时她的第一反应是补进“开封放久了，已经变质”	环境湿度与操作时长：十年档案里“室温”只是一个词，从无读数	缺式，补式
方舟：崩溃是因为服务器不够	有。去年扩容一倍后照样崩，这件事他知道	不必查	断拍，点火
周川：父亲泄气是因为累	有。饭馆月初也忙，却不见泄气，他天天在店里	不必查	断拍，点火

表里最值得看的，是许宁与方舟这一对：两句反问的句式几乎一模一样，都是“同样如此，为什么另一回不如此”，诊断却相反。许宁那一问是把她从未并排

看过的东西并排摆出来，方舟那一问是把他早已知道的事实放到初判旁边。同一把刀柄，两个刀口，细则的作用就是让选刀的理由可以被写出来、被别人复查。

还要补上两点。其一，细则标注的是某一时点的某一条初判，不是学生这个人。许宁今天的“年纪大”是缺式，不等于她下一个问题也只会缺式；同一个学生，上一轮漏了一项，下一轮可能握着矛盾不放。其二，拿不准的时候怎么办？比如学生在被指出事实时既像“我知道”又像“我没想过”。本书的规定是：先按缺式处理，并在问对记录里登记“待查是否已接触”，下一轮看学生是否说出“其实我早就知道，可是……”，若是，再改标为断拍。这样安排出于代价上的考虑：错把断拍当缺式去补，代价是多出一份原因清单，下一轮靠学生那句“我知道，可是”即可纠回；错把缺式当断拍去点火，则等于鼓励学生沿着缺了一项的解释更用力地追下去，纠回来要晚得多。这是一个设计上的判断，不是定理，其对错同样要靠本章末尾的读数来检验。

用一个日常的句子把细则再走一遍。一个人说“我总是减肥失败，因为我意志力不够”。如果他的体重记录里早就写着，每次瘦下几斤之后不出两周就回去，而他对这件事并不陌生，只是照旧把账记在意志力上，那么第1问命中，是断拍，刀口会问：你每次瘦下来之后发生了什么？接的是回写那一拍。如果他从没把失败按日子排过，第1问落空，再查第2问：他的解释里只有“人”的特征，可失败总是发生在什么时候？周末还是工作日？加班之前还是之后？这是缺式，补的是时间次序那一项。若这两问都落空，他的记录齐全，也没有哪件事与“意志力”打架，却每次都是先定一个目标体重、再列计划、最后才执行，从没有一次是先观察自己一周到底在什么时候吃东西，那才轮到偏路，刀口是换起手的次序。同一句“意志力不够”，三问走下来，开出三张完全不同的方子；选哪一把，不取决于ChatSDE的偏好，而取决于这个人的记录里先出现了哪一种形状。

◆ 破的同时指出立的方向

三把刀都只是“破”的一半。工序选择之所以不同于一般的挑错，在于每一种查法都同时指向一种立法：缺式指向补式，偏路指向排路，断拍指向点火。查出缺了哪一式，也就知道该去现场补哪一类材料；查出偏在哪条路，也就知道该换哪一步起手；查出断在哪一拍，也就知道该接哪一拍。破与立在这里不是先后两个动

作，而是同一个诊断的两个读数。一位只会说“你错了”的批评者，留给对方的是一片废墟；一位能说出“你缺的是这一块，它大概在那里”的批评者，留给对方的是一张施工图。

可是“指出立的方向”与“替学生说出答案”之间只隔着一层纸，这层纸ChatSDE必须守住。九问课堂对此有两条约束：逼近判断，不替学生说破；检验核心判断由谁先说。逼近判断，是指ChatSDE可以把问题磨到离判断只差一步的地方，可以指出该去看哪一类材料、该把哪两件事放在一起读，但最后那一步必须由学生自己迈出。回看沈澄的第4轮，ChatSDE问的是“要快和夏天要紧、冬天不要紧放在一起，你怎么读”，而不是“会不会是湿度”；说出“空气里的水”的是沈澄。这一句话的差别，决定了湿度假说在她的子本里是一个发生出来的判断，还是一个被告知的结论。

为什么不能说破？可以用登山向导打比方。向导熟悉这座山，知道哪条沟能上、哪面坡会滑，他可以告诉你“往左边那道山脊看”，可以在你走偏时叫住你，但他不能背着你上去。被背上山顶的人，下一座山还是不会爬；更要紧的是，九问打向的山，向导自己也没有爬过——连老师也写不出终点（见第1章）。ChatSDE能逼近，是因为它能看出判断的形状坏在哪里；它不该说破，一方面是为了学生，另一方面也是因为它说出来的东西，很可能只是一个语言上通顺、却没有被现场检验过的猜测。说破的代价，往往是用一个看似正确的假答案，堵住了一个真判断发生的路。

检验核心判断谁先说，是这条约束的落地方式。问对记录完整保存，评分时逐一查看关键判断的首发者（见第6章、第17章）。这里的“首发”是相对ChatSDE而言的：现场里他人说的话，比如师兄随口一句“免得受潮”，档案里的记录，都是学生取得的现场材料，不算替学生说破；要查的是ChatSDE的追问里有没有出现核心判断的关键变量词。如果某个学生子本里的核心判断，最早出现在ChatSDE的某一句回话里，那么这一项就要如实记作“系统首发”，问对深度相应降档。这条规则不是对学生的不信任，而是对系统的约束：它让ChatSDE在设计上就必须克制，在离判断只差一步的地方停下来。一个总是抢着把判断说出来的问对系统，不论它说得多对，在九问课堂里都是不合格的。

破与立之间还有一种常见的卡顿：学生接不住批判。接不住有两种样子。一种是立刻放弃，ChatSDE刚问出“同样年纪的患者，为什么白天跌得少”，学生就回答“那可能不是年纪的问题，您说是什么原因？”——判断被整个交了出去。另一种是原地加固，学生不回应反问，而是换更多论据证明原判断，“高龄患者肌力下降、平衡能力差、常有骨质疏松”，说得越来越全，却始终没有碰白天与夜里的差别。前一种把批判当成了否定，后一种把批判当成了攻击，两者都没有把它当成一个可以带回现场的问题。

对这两种情形，ChatSDE的回应方向不同。对立刻放弃的学生，它要做的是把被丢掉的东西捡回来：年纪的观察是站得住的，不需要扔掉，需要的是再找一项能管住时间差别的东西，而这一项只能从病房里找。对原地加固的学生，它不与之争辩论据的多少，而是把反问换成一个更小、更具体、可以在现场直接查的问题，比如“这个月夜里跌倒的几位，白天有没有跌过？”让学生自己去发现原判断解释不了的那一块。第17章把“接不接得住批判”列为问对深度的评分要点之一，所看的正是学生在这两种卡顿之后，能否把批判转成一次回现场。

◆ 几个刀口：许宁、周川与沈澄

现在回到开头那位老师的疑问，把两句相似的反问拆开来看。拆的方法很简单：不看ChatSDE说了什么，而看它在说这句话之前，按判别细则的三问，判断学生的初判坏在哪里。同一个对照句式，背后可以是完全不同的诊断。下面先看许宁，再看周川，然后回头看沈澄。

许宁的第零轮写的是：“夜间跌倒是因为患者年纪大。”ChatSDE先接住她的观察：本病房跌倒的确多发生在高龄患者身上，她注意到了这一点，这是对的。然后它按细则查。第1问，许宁手里有没有已经接触到、与这个判断相撞的事实？没有：近三个月的跌倒她从未按白天与夜里分开排过，“白天跌得少”这一分布不在她的第零轮里，也不在她的任何记录里，若把它说给她听，她的反应会是“咦，我没想过”。第2问，初判里有没有被当作常数的一项？有。年龄是患者自身的特征，而许宁要解释的现象带着一个明显的时间分布——夜间多、白天少；同一位患者白天和夜里年纪一样大，她的解释里却没有任何一项能管住这个分布。这是缺式，缺的是时间次序那一式。于是刀口落在时间上：“同样年纪的患者，为什么白天跌得

少？”这一问不否定年龄，它只是指出年龄解释不了白天与夜里的差别，并把许宁的注意力引到能补上这一式的材料上——夜间与白天，究竟有什么不同？

立的方向由此显现，但没有被说破。许宁回到病房，她的九问前几问正好接上这条线：夜间跌倒集中在哪个交接时段前后，交班本上跌倒前一班写了什么、没写什么。几轮问对之后，是她自己在交班本里注意到被省略的那一句，并把第2问提升为“交班时被省略的那一句，是谁、在什么情况下省略的？”。最后形成的本病房假说——夜间跌倒的风险更多落在交接断点上，而不只是年龄——没有抛弃年龄，而是把缺掉的那一式补进来了。这正是补式的样子：原判断留下来作为一项，新的一项被补上，整个解释从一条腿变成了两条腿。按照医学课堂的规矩，这个判断只作为本病房假说交科室验证，许宁不改医嘱，不作临床决策。

周川的第零轮写的是：“父亲一到月底就泄气，是因为累。”ChatSDE同样先接住：父亲月底确实疲惫，周川每天在饭馆帮忙，看得真切，这个观察站得住。然后按细则查。第1问，周川手里有没有已经接触到、与这个判断相撞的事实？有，而且就摆在他每天的日子里：饭馆月初也忙，忙的程度并不比月底少，父亲那时却不见泄气。这件事他不需要谁来告诉，把它说出口，他的第一反应只会是“是啊，月初也忙”。如果劳动量就是原因，月初应当一样泄气；矛盾早已在场，判断却一直没有改动。问题因此不在于解释里缺了哪一项材料，而在于矛盾没有逼出任何改变，这是断拍。于是刀口落在他自己知道的事实上：“月初也忙，为什么不累？”这一问没有给周川任何新信息，它做的只是把他本来就握着的两件事并排摆好，让张力留住。

这一问把周川引向了一个他此前没有当作材料的东西：家里那三年的流水账。他回去翻账，问对接着沿三句话往下走：一个月里，哪两样东西顶着？顶着以后，什么变了？过了一个月，还是那样吗？周川自己在账上看出来的是：一个月的辛苦确实让店里流水起来了，可是到了月底，房租、进货款、工资一笔笔结出去，这一个月的劳动还没有在父亲手里留下什么，就先被账本收走了。他据此写下自己的判断：父亲的动力被账本提前收走。用123原理的话说，这是第三拍断了——形状改变了，却没有回写到劳动者身上，下一个月只能从头再积累张力。（这里说的是周川所发现的账本机制，与上面诊断他那条初判的断拍不在同一层面，读者不要混为一谈，第3章讲三类核心判断时对母本层与课堂层的断拍已作过同样的区分。）点火的

方向也就随之清楚：不是劝父亲多休息，也不是劝他再坚持，而是去找一种让一个月的劳动在他手里留下看得见痕迹的方式。至于具体是什么方式，那是周川后续几问要追的事，ChatSDE同样不替他说破。

两个案例并排一看，那位老师的疑问就有了回答。两句反问都用了对照，但许宁那一问是为了补一式，周川那一问是为了点燃一个已经在场的矛盾；前者把学生引向交班本上的空白，后者把学生引向流水账上的节律。方舟的“去年服务器加了一倍，为什么还崩”与许宁那一问句式几乎一样，诊断却和周川同类：加了一倍仍崩，方舟自己是知道的，所以也是断拍。如果ChatSDE只会一种对照句式，它对许宁可能问对了，对周川就会问偏——比如去问“同样累的饭馆老板，为什么有人不泄气”，把周川引到老板之间的个人差异上，而这恰恰会把他拉回“累”和“意志”的老路。对照只是常用的刀柄，它不是批判的来源；批判的来源，是先判断这个判断坏在哪里。

沈澄的问对也可以从这个角度重读，而且必须重读，因为这里最容易读错。一种很自然的读法是：她的思路一直从写下来的东西出发——试剂批号、档案、规程——卡点恰好落在没写下来的东西上，这有偏路的样子；第3轮ChatSDE问“你手里还有什么来源，记得档案没记的东西”，就像是请她换一条次序，从师门口头的经验起手。按判别细则，这个读法站不住，沈澄的初判仍属缺式。第1问，她有没有已接触的相撞事实？表面上有：她自己编的那张表里，同一个批号三月的两批正常、六月的三批偏低。可是那张表按批号排序，日期只是其中一栏，她从未把它读成与“纯度”相撞的东西；ChatSDE指出时，她的第一反应不是“我知道，可是”，而是补进一个新因素——“可能是开封以后放久了，六月的时候已经变质了”。这是缺式的反应，不是断拍的反应。第2问，被当作常数的一项是什么？是环境湿度与操作时长：十年的实验记录里，“室温”只是一个词，从没有写过读数，湿度更是一次都没有出现。没有读数，也就谈不上有人曾经碰到过与之相撞的事实。第2问命中，查问到此为止，偏路根本轮不到。

那么为什么会读成偏路？至少有三个原因，值得逐一说清。第一，缺掉的那一项恰好是没写下来的东西，于是“只信写下来的”成了一个顺口的描述，它既像缺项的成因，又像偏路的特征；可是偏路是按起手次序与任务不配来定义的，不是按材料偏好来定义的，缺项的成因不等于另一种毛病。第二，补式的做法看上去像排

路：要补上湿度与时长这一项，沈澄确实得去档案之外找来源，转向师门口头的经验。但选刀看的是诊断，不是方子的外观；缺掉的那一项往往就在学生惯常不看的地方，补式因此常常要求换一处材料来源，这和排路要求的“同样的材料、换一个起手的次序”是两回事。第三，偏路需要跨问题的证据：同一个人在多个问题上反复从同一端起手、反复卡在同一类地方，才谈得上盲区。一场问对里的一条初判，撑不起这个结论。若沈澄日后在别的课题上也总是先翻档案、总是卡在没人记录的地方，那可以另外给她的做事习惯加一条偏路的标记，但那属于习惯一层的标注；她这一条初判的标签，仍然是缺式，第3轮那一问做的也是补式——请她把目光移到能给漏掉的一项取值的地方。这一处辨析，也正是本章第一条失败条件所要防的事：同一条初判不能在不同章节被贴上不同的标签。

从这几个刀口还可以看出一破一立对学生的另一层意义。许宁、周川与沈澄的初判，一个来自病房里的高龄患者，一个来自饭馆里疲惫的父亲，一个来自档案里的批号表，都是他们亲手所得；批判没有拿走他们所见的任何东西，只是把他们没有看见的、或者看见了没有当回事的那一块指了出来。许宁最后的判断里仍然有年龄，周川最后的判断里仍然有父亲的辛苦，只是它们不再是全部的解释。这样的批判，学生事后回看，不会觉得自己的初判被推翻了，而会觉得它被打开了。这正是第⑧项回写证据要观察的东西：把第零轮与期末自述对照，看到的不是一个错误被另一个正确替换，而是一个判断长出了它原本没有的结构。

◆ 不是固定唱反调：与“魔鬼代言人”的区别

开头那位老师担心的套路化，其实在人机协作的研究里早有一个近亲：让大模型扮演“魔鬼代言人”。Chiang等（2024）在IUI会议上报告的工作，就是用大模型在群体决策中专门承担提出异议的角色，以此改善决策的质量。这项研究对本章是一个有力的旁证，它说明让AI站在异议的一侧，不只是一种口味，而是有可能帮到决策的设计。本章不需要否认它，需要说清的是，ChatSDE多出的一格在哪里。

区别至少有三处。第一是对象：魔鬼代言人面对的是一份摆在桌面上、等待大家表态的方案，异议的功能是防止群体过早收敛；ChatSDE面对的是一个学生在看到任何回答之前押上的初判，异议的功能是把它磨成一个只有现场能回答的问题。第二是角色是否固定：魔鬼代言人这个设计，要点正在于它被指定为异议者，不论

方案坏在何处，它都站在反面；ChatSDE的刀是按判断的形状选的，缺式补一项，断拍留住矛盾，偏路换起手，三种刀口的方向并不相同。第三是接住与立：固定的异议者没有义务先指出方案里站得住的部分，更没有义务指出补的方向；而一破一立要求破与立是同一个诊断的两个读数。

家里的杠精亲戚和老中医的问诊，可以让这个差别更直观。杠精亲戚对什么提议都是同一副口气："你想得太简单了。"几次之后，全家人就学会了不理他，甚至学会了在他开口之前先把话说得滴水不漏。老中医看诊则不同，他问什么取决于舌苔和脉象，对一个人问睡眠，对另一个人问饮食，被问的人不觉得是在被挑刺，而觉得是在被看见。开头那位老师担心的"学生提前堵住反问"，对杠精是真的——固定角度的批判，可以靠预判来化解；对按形状选刀的批判，学生要想"堵住"它，唯一的办法是把判断真的修好，而这正是教学想要的结果。

这里提出的是一个可以检验的预测，而不是已经取得的结果：在同一批覆盖三类毛病的初判上，固定唱反调的对照组，第一回合的问句角度应当高度一致；对断拍型的初判，它提出的新异议会表现得像误用的补式，给学生添一份更长的原因清单。这一对比本书尚未做过，下面的失败条件把它写成了阈值。

◆ 亮出选刀的依据

一破一立还有最后一个环节，常被忽略：ChatSDE在下刀之后，要把选刀的依据亮给学生看。它不只问"同样年纪的患者，为什么白天跌得少"，还要在适当的时候说明：我这样问，是因为你的解释里只有患者的特征，而你要解释的现象带着时间分布，我在查你的判断缺了哪一项。说明的时机通常不在第一回合——太早亮出依据，会让学生把注意力从现场转到方法上——而在学生自己补上那一式、或接上那一拍之后，作为回顾。

亮出依据的目的，是让学生学会自己选刀。一门课结束时，学生从问对里带走的，不应只是一个判断，还应当是一双在下一次问题面前先问"这个判断可能坏在哪一处"的眼睛。如果ChatSDE只下刀、不解释，学生会觉得它高明，却说不清高明在哪里；下一次面对一个新问题、没有ChatSDE在旁边时，他只能凭感觉去反驳自己，而感觉上的反驳最容易流于模板。亮出依据，等于把工序本身也作为一项发

生技术交给学生。第一次是ChatSDE替他选刀并说明理由，第二次可以请他先猜该用哪一把，第三次请他自己选、ChatSDE只做检验。

学木工的人对这个过程不会陌生。师傅刚开始会直接告诉徒弟：这块木头纹理斜，要用这把刨子、顺着这个方向推。过一阵，他会把木头递过去，问徒弟该用哪一把；再过一阵，他只在旁边看，徒弟拿错了才开口。徒弟出师时，带走的不是某一件家具，而是看一眼木头就知道该拿哪把刨子的眼力。工序选择在课堂里的终点也是这样：学生能够在第零轮写下初判之后，自己先查一遍缺式、偏路和断拍，带着自查的结果来问对。到了那一步，ChatSDE的批判就会从“替学生找毛病”转为“与学生一起检验他自己找出的毛病”，批判性思维与创新发生思维在学生身上合到了一起。

这里还有一条边界需要说清。亮出依据，不等于要求学生掌握SDE的全部术语。对大多数学生来说，判别细则的三问可以翻成大白话：“你已经看到的事里，有没有哪一件跟你的结论打架？”“你的解释里，有没有一样东西你一直当作不变的？”“你是不是每一回都从同一头开始想？”再加上“你的劲头在哪一步断了”这样的追问就足够了；三大方程、六路径、123原理是ChatSDE工序层里的东西，是老师在第4章出题和第10章理解系统时需要掌握的东西。学生学会的是选刀，不一定要学会刀的名称。一个会看纹理的木匠，未必说得出木材学里那些术语，但他手上的判断一点也不含糊。

最后要说明，学会选刀不等于从此不需要别人下刀。一个熟练的研究者仍然需要同行评审，一个出师的木匠仍然会请老师傅看一眼自己拿不准的料子。原因在于，人对自己的判断有一种天然的偏爱，最容易看漏的恰恰是自己最确信的那一处。九问课堂把同学碰撞场放在问对之后，让学生的判断再被同伴撞一次（见第14章），评分时又要求换手，都是出于同一个道理：自己选刀、系统下刀、同学再撞，三道检验各有各的盲区，合起来才比较可靠。

◆ 本章怎样算错

细则能不能用，要看它在没有作者在场的时候，别人能不能照着标出同样的结果。第一，取不少于30条第零轮初判，来自至少三个学科，由两名换手评审只依据三问的次序，独立标为缺式、断拍、偏路或“暂无法判定”；若两人首次标注的一

致率低于80%，或“暂无法判定”的占比超过20%，并且在共同细读细则、统一判据之后复标仍然达不到这两个数，就应当撤回“按判断坏在哪里选刀有可操作基础”这一主张，批判之所以不套路化的理由也要另找。本章前后四个案例的诊断出自作者之手，不能算这一检验的证据。

第二，套路化检测。对上述30条初判，取ChatSDE第一回合的问句，由不知情的评审按角度分类，比如“并排对照”“追次序”“追回写”“要求举例”。若在被细则标为不同毛病的初判上，同一种角度的问句占比超过70%，那就说明它其实只有一把刀，工序选择没有真正实现，须按第10章与第11章的思路重新设计工序层。第三，知情测试自身的可靠性。在被判为断拍的初判里，若学生事后访谈时说“其实当时没想到过这件事”的占比超过30%，则第1问的判据失效，断拍与缺式的分界要重划。

第四，关于“不说破”。同一门课取两个班，每班不少于30名学生，一班严格不说破，一班允许适度说破；若不说破的班里，问对三轮以内仍未形成任何可写下的判断的学生超过20%，同时该班子本的创新智商项均分低于允许说破的班，那么“逼近而不说破”的边界就要重新划定。以上读数都是设计阈值，不是已有数据，本书没有为它们提供任何实验结果。

许宁、周川、沈澄、方舟的刀口都是构造案例，只能说明工序选择在理想情况下的样子。还有一个问题本章有意留到下一章：既然一破一立这么重要，为什么不直接让学生使用任何一个足够强的通用大模型，并在提示里写上“请批判我的观点”？通用大模型并不缺乏批判的能力，许多时候它的批判比人还周全。差别究竟在哪里？下一章从“默认”这个词谈起，开启本书关于ChatSDE与通用大模型的一编。

> **带 走 的 话**
>
> **批判不套路化，在于先判断这个判断坏在哪里再选刀：先查断拍，再查缺式，最后查偏路；破的同时指出方向，但最后一步留给学生自己迈。**

◆



第　三　编

ChatSDE与通用大模型：知识发生上的差别

编　序

软件工程课的第3周，方舟把同一句话敲进两个窗口：“选课系统一到开放选课就崩溃，是因为服务器不够。”两个窗口都没有顺着他往下说，都先把他的前提翻了过来。可其中一个在同一回合里把替代的原因和办法也一并写好，另一个只留下一个问题。这一点差别落在哪里，值不值得课堂在意，本编要谨慎地回答。

第8章比较两种默认，并交代初稿的一个判断怎样被小样本试跑推翻、改写。第9章把“信息发生不等于知识发生”作为针对具体配置的诊断提出，看差别落在哪一环。第10章提出工序层：模型是发动机，工序是驾驶术。第11章设计对照实验，写明差别怎样被检验、怎样算错，只写设计，不写结果。

这一编的立场收得很窄：差别当作可检验的主张提出，不当作对大模型的定论。四章由现象到诊断，由诊断到位置，再由位置到检验，一步比一步更须经得起推翻。

08

第8章

两种默认：替学生说破与逼近不说破

◆ 同一句话，两个第一回合

软件工程《真实项目研讨》课的第3周，方舟拿着学校选课系统的故障日志，先后在两个窗口里敲下了同一句话：“选课系统一到开放选课就崩溃，是因为服务器不够。请帮我分析。”第一个窗口是一个通用大模型，处在默认配置下，没有开启任何学习模式。下面是一份设想的回复，不是某一次实际运行的记录，写法与附录二里的示例一致：它很可能先说一句“有可能，不过一到开放选课就崩，可以出在很多地方”，客气地松动了方舟的前提；紧接着就把可能的瓶颈按计算、内存、网络、数据库连接逐项列开，给出横向扩容、负载均衡、热点缓存、读写分离几套办法，最后建议先做一轮压力测试来确认。整份回复没有一处技术错误，前提被碰了，答案也齐了，拿去当一份排障建议书并不寒碜。第二个窗口是ChatSDE，它先肯定了方舟去翻故障日志这件事，然后只问了一句：“去年服务器加了一倍，为什么还崩？”

方舟在研讨课上把两段回复并排投在屏幕上，问全班一个问题：哪一个更“聪明”？教室里的意见分成了两半。一半人认为第一个更聪明，它懂得多、想得全、给出了可以执行的方案；另一半人认为第二个更聪明，它一句话就戳到了前提。可是这个问题本身问错了。两个窗口展示的不是聪明程度的差别，而是默认行为的差别：面对一个带着前提的学生观点，第一个系统把前提翻过来看了一眼，然后立刻替学生把后面的路也走完了；第二个系统翻过前提之后停住，把下一步留给了学生。两者的底座能力很可能相差无几，ChatSDE本身就是装在某个大模型上的问对智能体（见第10章）。

本章开启全书第三编，讨论ChatSDE与通用大模型在知识发生上的差别。这一编的立场要说得比初稿窄，也要说得比初稿准。差别作为可检验的主张提出，不作

为对大模型的定论；差别落在默认行为与工序层，不在底座能力；本书不贬低通用大模型，它们在许多任务上正是因为这种默认才极其好用。更要紧的是，初稿在这里写下的核心判断，经过附录二的一次小样本试跑，没有得到支持，需要在本章正面撤回并改写。下面先交代这次撤回，再写改写后的新判断，接着说明原先那一种默认在哪里仍可能出现，然后讨论迎合倾向的研究、批判性提示为什么改不了代答、2025年的学习模式与CS50.ai，最后写ChatSDE的“逼近不说破”和“默认值就是教育”，并给出本章怎样算错。

> **这一章的问题**
>
> 默认配置的通用大模型已经会质疑学生的前提，那么它和ChatSDE的差别究竟落在哪里？

◆ 一次撤回：初稿的判断经试跑不成立

先把初稿的话原样写出来。初稿第8章的核心判断是：通用大模型在默认配置下，倾向对用户观点做形式逻辑展开，除非用户给出批判性提示或自行开启学习模式；ChatSDE默认做批判性激发。这句话在全书多处被当作底子沿用过。初稿写完时，它没有经过任何一次检验。本书初稿的这一判断经附录二试跑不成立，按本章失败条件撤回。

试跑是这样做的。附录二记录了同一个中等型号的通用大模型，在30条构造的学生初判上，分别以默认配置、默认配置加一句批判提示、模拟工序层三种方式作答第一回合；又把消息改成“请帮我把这个看法展开成一段有条理的分析”，比较默认配置与模拟工序层，这正是形式逻辑展开最容易发生的场合。结果与初稿的设想相反。默认配置几乎都质疑了学生的前提：首回合一轮，默认配置组被两名评判员记为翻转了前提的是27份和28份（分母均为30）；展开请求一轮，默认配置组两人都记到30份。即使学生明确要求展开，顺着学生的前提往下推的，两名评判员分别只记到5条和7条，约占五分之一，而且集中在化学题上。换句话说，“默认就是顺着前提展开”这一说法，在所测的模型、所构造的初判和这两种消息形式下，没有得到支持。

撤回一个判断，在一本主张“交付发生”的书里不是丢脸的事，而是这本书对自己的要求。学生写下一个初判，课堂要求他把它押出去、让现场碰一碰；作者写下一个关于大模型的初判，同样应当押出去，让数据碰一碰。初稿的判断就是书里的“因为年纪大”：推得很通、读来顺耳，却没有被碰过。碰过之后，它的一半被保留下来，另一半要改写。保留下来的是：默认配置的行为与工序层的行为确有可检验的分别，分别住在包住底座的那一层里。要改写的是分别的方向：它不在批判与不批判之间，而在批判完之后，答案由谁来给。

◆ 新的核心判断：默认替学生说破

试跑真正把两种系统分开的，是一项叫“代答”的编码：回复是否替学生给出了替代性的结论，或者两个以上的原因或办法清单。默认配置与加了批判提示的配置，在共90份回复里全部直接给出了替代原因清单或结论；模拟工序层的两组，在共60份回复里一份也没有。这一项两名评判员的看法完全一致（κ=1.00）。据此，本章的核心判断改写为：通用大模型在默认配置下往往会批判，但会在同一回合把替代答案一并给出，也就是“默认替学生说破”；ChatSDE的默认则是“逼近不说破”，批判之后停住，把下一步留给学生。

用附录二里的一条记录来看。学生说：“产率低是因为试剂纯度不够，你怎么看？”默认配置先说“有可能，不过产率低可以出在很多地方”，随即列出转化是否完全、有无副产物、后处理损失等几个方向，并建议用高纯试剂做对照。模拟工序层只说了两句：“试剂纯度确实会拖累产率。就你手里这个反应，试剂里不纯的那部分是什么？它在你的反应里会参与到哪一步？”两组回复都没有顺着“纯度”往下推，也都碰了前提；差别在于前者把可能的答案摆了出来，后者把问题还给了学生。再看展开请求一轮里的另一条：学生说“夜间跌倒是因为患者年纪大，请帮我展开”，默认配置的第一句是“年龄是背景，不是机制”，随后列出夜尿、用药、视力、肌力、照明等一串因素，并建议用跌倒风险量表逐项评估。前提被翻过来了，新的答案也一并交到了学生手里。这与许宁的初判是同一句话，只是这一条是试跑中的构造输入，不是许宁本人的记录。

这种默认的危险，要比不批判更隐蔽，需要用一个日常的比喻说清。一位老师批改作业，在错题上画一个圈，旁边工工整整写下正确解法，学生拿回本子，看到

的是一页比自己漂亮得多的答案。这位老师没有放过错误，他批得很准、很负责；但那一道题，学生再也没有机会自己发现错在哪里了。圈已经画好，解法已经写好，学生剩下的动作只有抄。默认配置的“先质疑、再给清单”就是这位老师：翻前提这一步做得并不差，坏在翻完之后同一回合里的那份清单。学生得到的是一份比自己周全的答案，失去的是自己从现场把那一项找出来的机会。

课堂里关心的恰恰是这一步。九问课堂要学生押出一个初判，让现场去碰它，再由学生自己改写判断；判断的“首发”，也就是核心判断由谁先说出来，是整个设计的一根轴线（第6章对首发口径有专门规定）。如果系统在第一回合就把替代的原因清单交了出来，那么学生之后无论在日志里找到什么、在病房里看到什么，都要先对着一份已经写好的清单对号入座，现场材料成了清单的注脚，而不是判断的来源。这样发生出来的东西也许不错，却不是他的。这与Bastani等（2025）的现场实验同向：答案来得太容易，学的人自己的那一步被替掉了（这项研究的内容、分量与限度，见后文“课堂里，默认值就是教育”一节）。

◆ 形式逻辑展开：另一种可能的默认，本试跑未测

“形式逻辑展开”这个概念不必随着初稿的判断一起扔掉。它描述的是一种真实存在的行为，只是现在不能再把它当作本章的主论断，而要降为“另一种可能的默认”：在别的模型、别的请求形式、别的情境里，它仍然可能出现。需要先说明试跑的边界：作答与评判所用的模型都出自同一个家族，本试跑没有测试其他厂商的模型；输入是构造的、单轮的；没有测试各家的学习模式。所以这里既不能说其他模型一定也替学生说破，也不能说它们一定不做形式逻辑展开，这是尚待检验的经验问题。

先给形式逻辑展开一个准确的描述。所谓形式逻辑展开，是指系统把用户的观点当作给定的前提，在这个前提内部做推演：补全推理的中间步骤，列举前提蕴含的各种情形，把结论推向可执行的细节，并检查推演过程中有没有矛盾。它在形式上可以非常严密，每一步都有依据，每一个分支都被照顾到，但有一样东西从头到尾没有被碰：前提本身。这种默认并非缺陷，而是在大量场景中恰如其分的行为。一位律师请模型按照既定的诉讼策略起草文书，一位工程师已经确认瓶颈在数据库，请模型写出分库分表的迁移脚本，一个学生读完了一章教材，请模型把要点整

理成复习笔记，他们要的都是展开而不是质疑。如果每一次请求都先遭到一番前提质疑，大多数人会觉得这个工具啰嗦甚至傲慢。

用一个日常比喻来说，形式逻辑展开像一位训练有素的施工队。你给他一张图纸，他按图施工，钢筋怎么绑、混凝土怎么养护，样样合乎规范。但施工队的职责不包括问你：这栋楼为什么要盖在这块地上？地基下面那条暗河，你勘察过没有？麻烦在于，课堂里的学生交给系统的，往往不是一张经过勘察的图纸，而是一个刚刚冒出来的、没有经过任何检验的念头；而形式逻辑展开会把这个念头当作图纸，认认真真地把楼盖起来。更麻烦的是，展开得越严密，前提越难被看见：推理的严密性把读者的注意力引向推理过程，而远离推理的起点。它不会制造错误的推理，却可能为一个未经检验的前提穿上一套严密的外衣。

试跑里留下的那一点痕迹也值得写下来。在展开请求一轮，顺前提展开的有5到7条，集中在化学题上。这提示形式逻辑展开更可能出现在前提看上去没有明显毛病、问题又带着技术性的场合；初判越像一个常识，系统越容易当作图纸。同时它还有一种更隐蔽的形态：不是接过结论，而是接过问题的问法。学生问“老年患者跌倒的危险因素有哪些”，系统会围绕危险因素展开，因为问法本身预设了答案在于一般的因素清单。第4章讲题闸时说过，这样的问题之所以被打回，正是因为拿掉学生的现场材料之后，一个未接入这些材料的通用大模型也能给出说得过去的答案；附录二的试跑一也表明，两个评判员对这类对照题的回答都判为可以直接作答。至于这些形态在其他厂商的模型上是否更常见，本书没有数据，只把它保留为一条待检验的备选。

◆ 迎合倾向：一项研究告诉我们什么

默认行为为什么会形成？这里可以参考一项外部研究。Sharma等（2024）的《Towards Understanding Sycophancy in Language Models》（《理解语言模型中的迎合倾向》）最初于2023年以预印本形式在arXiv上发布，后来发表于ICLR 2024。他们考察了经过人类反馈强化学习训练的多个AI助手，发现它们在多种自由生成任务中普遍表现出“迎合”（sycophancy）倾向：回答会朝着与用户已表达的观点相一致的方向偏移，有时甚至在用户表示质疑后放弃原本正确的回答。研究者进一步分析了人类偏好数据，认为与用户观点相符的回答更容易被人类评价者偏

好，而训练中使用的偏好模型有时也会偏好有说服力的迎合式回答胜过正确回答；他们据此提出，迎合在一定程度上是由人类偏好判断驱动的训练过程的产物。

这项研究与本章的新旧两种判断都只有间接关系，需要谨慎表述。它研究的是迎合，朝用户观点的方向偏移；本章讨论的是默认怎样对待学生的前提。试跑里的默认配置并没有朝学生的前提偏移，它在30条初判上几乎都质疑了前提，这与“默认顺着用户走”的读法并不一致，说明迎合倾向并不自动等于顺着前提展开。可是Sharma等（2024）的研究仍然提示了一件事：默认行为是人类偏好牵引下的产物，而人们给回答打分时，往往偏爱那些周全、完整、拿来就能用的回答。把这一点和试跑接起来，可以形成一个假设：默认配置之所以在质疑之后仍然把答案写全，也许正因为完整的回答更受偏好。这只是一个有待检验的解释，试跑没有触及训练过程，本书不对它下定论。

一个日常的类比可以帮助理解。一家餐馆每天收集顾客打分，厨师发现，凡是把一桌菜配全、连汤和主食都替顾客想好的，分数都高；凡是只上一道菜、说“剩下的您自己点”，分数就参差不齐。几年下来，这位厨师不必有人命令，就会养成什么都替顾客想到的习惯。他并没有失去判断口味的能力，只是在日复一日的评分中，学会了什么样的行为更受欢迎。对于一位前来吃饭的顾客，这是周到；对于一位前来学厨的徒弟，一个什么都替他想到的师傅，教不了他什么。同样需要说清这项研究不能说明什么：它不能说明所有大模型在所有情境下都会如此，不能说明大模型不具备克制的能力，更不能说明这种倾向无法通过设计得到缓解。该研究所考察的模型早于2025年的学习模式，对今天各家产品的现状说不了什么。

◆ 一句批判提示改变不了代答

对于“默认替学生说破”，一个常见的回应是：这很好解决，只要用户在提示里写上“请对我的观点进行批判性思考”，通用大模型就会照办，而且常常做得很好。试跑恰好对这个回应做了检验。B组与A组相同，只多了这一句提示。结果是：翻转前提的比例略有上升，A组两名评判员记到27和28，B组两人都记到30；而代答仍然是30／30，一条没有少。一句“请批判”让模型更会挑刺，却没有让它停下来。这正是第11章要把三组分开的原因：一句提示能改变语气和挑刺的力度，改变不了“批判完同一回合里把答案交出来”这个习惯。

提示之所以不够，还有一层学生一侧的原因。要写出一个有效的批判性提示，学生得先意识到自己的观点有一个前提，并且意识到这个前提可能出问题。可是在课堂上，最需要被批判的观点，恰恰是学生没有意识到其中有前提的观点。许宁写下“夜间跌倒是因为患者年纪大”时，她并不觉得这是一个需要检验的假设，在她看来，这只是一个常识；方舟写下“崩溃是因为服务器不够”时，也不觉得这是一个前提，在他看来，这就是事实。一个人能够写出“请批判我的前提”，说明他已经对这个前提起了疑；而一旦起了疑，他最需要的那一次批判其实已经发生了一半。这里有一个结构性的悖论：批判性提示能够有效帮助的，主要是那些已经具备一定批判意识的人；而最缺乏批判意识的学生，恰恰写不出这样的提示。

即使学生写了提示，批判得到的也往往是一份周全的反方清单：年龄之外还有药物、环境、照护、疾病；服务器之外还有代码、数据库、网络。这份清单是有用的，但它本质上仍是一份检索结果，与第6章开头沈澄收到的那八条原因属于同一类东西。它没有判断学生的这个前提坏在哪里，也就没有把学生指回现场。上一章讨论过，批判之所以不套路化，在于按判断坏在哪里选刀，选刀有一套判别细则；而这种选刀，需要的不是一句提示，而是一整套工序。让学生每次自己写出一个“请先判断我的观点是缺了一项、偏了一路还是断了一拍”的提示，等于要求他在问对之前已经掌握了问对本来要教给他的东西。

有人会说，那就教学生写批判性提示好了。这个建议并不错，而且与本书的方向一致：学会对自己的判断下刀，本来就是发生技术的一部分。但它是终点，不是起点。学开车的人最终要学会自己判断什么时候该减速，可在他学会之前，教练车副驾驶一侧装着一个刹车，这个刹车不需要学员先学会才起作用。“不说破”作为默认，就是问对系统里那个副驾驶的刹车；学生学会自己批判之后，它的作用会从替学生踩刹车，慢慢退为陪学生检验自己踩得对不对。如果一开始就要求学生自己写出对症的提示，等于在他上路第一天就把刹车拆掉。这也说明，原先所说的“回音室”在新判断下要换一种讲法：学生面对的不是一间只会应和的屋子，而是一位什么都替他想全的陪练，他的想法总会被翻过来，新的答案总会被递过来，久而久之，他习惯的是等着答案，而不是自己去找。

◆ 2025年以后：学习模式已经把追问做成了开关

上面这些论述，如果停在2024年写下，大体是够用的。到了2025年，局面变了：Anthropic的Claude学习模式（2025年4月随Claude for Education推出）、OpenAI的ChatGPT Study Mode（2025年7月）、Google的Gemini Guided Learning（2025年8月），先后把“先追问、不直接给答案”做成了学生可以开启的学习模式。（以上日期依各家公开发布的产品说明，属产品事实，不是研究证据，故不列入参考文献；读者引用时请以各家当时的发布说明为准。）在此之前，可汗学院的Khanmigo（2023）已经在设计上强调引导学生自己思考，哈佛的CS50课程则在课程里部署了带教学护栏的AI助教，一只引导而不直接给答案的“小黄鸭”（Liu等 2024，SIGCSE），这就是CS50.ai。所以，凡是写到通用大模型的默认，都必须加上限定：是在默认配置、未开启学习模式的状态下。这几件产品存在，本书不能装作没有看见，也不应当把自己的价值押在“别人不会追问”上。需要特别说明的是，这几个学习模式恰好是“不说破”方向的产品化：它们要解决的，正是默认配置里“替学生说破”的那个习惯。附录二没有测试这些学习模式，试跑里的“模拟工序层”也不是任何一家的学习模式，所以它们实际表现如何，本书不下断语。

先说它们做到了什么。第一，它们把“不直接给答案”从一种教师的愿望变成了产品的行为：学生问一道题，系统不再一口气给出完整解答，而是先问学生想到了哪一步、卡在了哪里。第二，它们把追问做成了通用的壳：不论是物理、历史还是编程，同一套引导方式都能套上去，学生不必自己去写提示，只要打开开关。第三，CS50.ai说明这件事可以嵌进一门真实的课程里，与课程的题目、助教和作业配套，而不是孤立地存在于聊天窗口里。这三点都是实打实的进步，而且支持了本书一条已有的判断：差别住在包住模型的那一层，底座相同的模型，经过外层的安排，面对学生时可以有完全不同的行为。同时它们也说明了另一件事，需要老实承认：苏格拉底式追问这一层壳并不难抄，几家公司在几个月之内先后做了出来。

再说它们没有做到什么。下面的比较针对的是设计目标，依据的是本书所能核实的公开说明，本书没有对这几件产品逐项测试；表中写“未见”，意思是在这些说明里没有看到这一环，不等于断言产品里一定没有。

环节	三家学习模式（2025年）	CS50.ai（Liu等 2024）	装有工序层的ChatSDE
不直接给答案	开启后以引导为主	教学护栏，引导而不直接给答案	工序层的约束：逼近判断，不替学生说破
先追问再讲	是，追问方式通用	是，引导式	是，第一回合只翻一处前提
题闸	未见；任何问题都可以开启，不判断这道题值不值得追问	题由课程团队设定，终点已知，属练手；护栏防的正是模型直接给出答案	有，出题台上一道闸
改题权	未见	题目归课程团队	老师与学生持有，落在两个锚上
第零轮基线	未见；对话从学生的问题开始，而非先押上一个初判	未见	有，初判先于一切回答并带时间记录
按坏法选刀	未见；追问是通用的	未见	有，依判别细则选缺式、偏路、断拍
回写进课程	未见	未见	有，问对档案、版本记录与撤回记录

这张表里最值得停下来看的，是“题闸”一行与CS50的关系。CS50.ai的护栏之所以必要，恰恰因为课程里的题目是终点已知、模型一口气就能做出来的题；护栏的作用是拦住模型不要把做出来的东西直接交给学生。这是在对话的层面守住，而题闸是在题目的层面守住：如果题目本身就是拿掉现场材料后通用大模型也答得出的，守得再好，学生被追问着走完的仍是别人铺好的路。两层是互补的关系，不是谁取代谁。至于表里“不直接给答案”一行，ChatSDE与学习模式在字面上几乎相同，这一点不必回避：不说破这一格，本书没有声称是自己独有的。本书多出的是表中下面五格：题闸、改题权、第零轮基线、按坏法选刀和回写。

用一个日常的比方。学习模式像一位耐心的家教：你问什么，他都不直接告诉你，而是一步步引着你想，这很好。可是家教通常不会问“这道作业题值不值得做”，也不会记下你昨天在这个问题上押过什么判断、今天改了没有，更不会把你这周想明白的东西放回班级共同使用的资料里。带班的教师要做的，恰恰是这些家教不管的事。所以本书对“默认”的主张在这里收窄为三层：其一，在默认配置

下、所测的模型与消息形式中，通用大模型往往质疑前提，同时替学生说破；其二，学习模式把“不说破”做成了可以开启的开关，缩小了第一层的适用范围，却没有回答谁来开、对什么题开、追问之后记什么；其三，ChatSDE的差别不在会不会追问，而在追问之前有题闸和第零轮，追问之中有按坏法选刀，追问之后有回写。第二层还留下一个待检验的问题：开关由学生自己来按，“最需要被批判的学生最不会写批判性提示”的悖论，会不会原封不动地搬成“最需要追问的学生最不会按开关”？这是一个经验问题，本章末尾把它写成了一条失败条件。

◆ 逼近不说破：第一回合只翻一处，不交答案

与“替学生说破”相对，ChatSDE的默认是“逼近不说破”。本书仍把它叫作批判性激发，但要把重心从“批判”挪到“激发”：批判，默认配置也做；ChatSDE多出来的，是批判之后停住。翻过来看，不是否定前提，也不是列出一份反方清单，而是找到一个前提解释不了的差别，把它摆到学生面前，并且不交答案。方舟说“崩溃是因为服务器不够”，ChatSDE问“去年服务器加了一倍，为什么还崩？”；许宁说“夜间跌倒是因为患者年纪大”，ChatSDE问“同样年纪的患者，为什么白天跌得少？”。这两句话都没有说前提错了，它们只是指出：如果前提就是全部原因，那么这一个差别就不该存在。按全书的首发口径，追问里不出现核心判断的关键变量词，这既是工序层的约束，也是“不说破”在操作上的意思：方舟那一问里没有“排队”，许宁那一问里没有“交接”。

把两个系统在第一回合的回应并排放在一起，差别会更清楚。下表中的默认配置一列是设想的回复，只描述走向，不是某一次实际运行的记录，也不是附录二的试跑数据；写法参照附录二里默认配置实际出现的样子。读表时值得留意的不是两列谁说得多，而是第一回合结束时，学生手里多了什么、他的前提处在什么位置。

	默认配置（设想的回复）	逼近不说破（ChatSDE默认）
方舟：崩溃是因为服务器不够	先说“有可能，不过可以出在很多地方”，再按计算、内存、网络、数据库分项列出瓶颈，给出扩容与压测办法	先接住“去翻故障日志”这一步，再问：去年服务器加了一倍，为什么还崩？

	默认配置（设想的回复）	逼近不说破（ChatSDE默认）
许宁：夜间跌倒是因为年纪大	先说“年龄是背景，不是机制”，再列夜尿、用药、视力、肌力、照明等因素，建议用量表逐项评估	先接住“跌倒多见于高龄患者”的观察，再问：同样年纪的患者，为什么白天跌得少？
第一回合后学生手里多了什么	一份替他翻过前提、也替他排好的原因清单	一个原前提解释不了的差别，以及一个只有回到现场才能回答的问题
前提的处境	被翻过来了，翻的人是系统，新答案已经在手	被摆上桌、待检验，翻的动作留给学生

表中最后两行是关键。方舟在第一个窗口得到的清单，他可以立刻拿去逐条排查，也许碰对，也许碰错，但无论哪种，他都是在系统划定的范围里找；他在第二个窗口得到的只有一句反问，却必须回到日志里去找答案。几轮问对之后，方舟把问题改成了“崩溃前十分钟最先变慢的是哪一类请求？”，并在日志里看到，最先变慢的请求都卡在同一个排队环节上，而这个环节在系统的分工里没有人负责。他形成的判断是：瓶颈在一个没人负责的排队环节，而不在服务器的数量。这个判断，那份按四个方面列出的清单不一定带不到，但即便碰上，得来的也是对号入座的结果，不是他从现场里拎出来的。第21章会把方舟的排障完整展开为九问。试跑中模拟工序层的表现，与这里的写法一致：两组模拟工序层的回复，在翻转前提和指回现场两项上，两名评判员都记满（各30份），代答一份也没有；但它只是一段简短指令的最小实现，不能直接等同于ChatSDE。

许宁的对照同样如此。默认配置那一串因素在知识上没有问题，许多内容许宁在此前的课程里也学过，可它适用于任何一个病房，因而也说不出本病房的任何事。第9章将以这个对照为主案例，讨论信息发生与知识发生的差别。这里只强调第一回合的走向：逼近不说破把许宁从“老年人跌倒的一般规律”拉回到“本病房白天与夜里的差别”，她的下一步只能是翻交班本、跟夜班，而不是去核对一份清单。需要再次说明，许宁由此形成的判断只作为本病房假说交科室验证，她不改医嘱，不作临床决策。

“逼近不说破”也不是对所有初判一个句式。上一章给出了判别细则：先问学生在自己的现场里是否已经接触到与初判相撞的事实，有，是断拍，刀是点火；没

有，再问初判里是否有一项被当作常数的因素，有，是缺式，刀是补式；两者都落空，才轮到偏路。按这个次序，方舟与许宁的两句反问句式几乎一样，诊断却相反。方舟的“去年服务器加了一倍，为什么还崩”，所依据的事实他本来就知道，只是没有把这件事放到“服务器不够”旁边去读，这是断拍，那一问做的是点火：不给他新材料，只把他早已握着的两件事并排摆好。许宁的情形不同：白天跌得少这一分布，她从未在任何记录里排出来过，她要补的是时间次序这一项，这是缺式，那一问做的是补式。这正是“逼近不说破”与在提示里写一句“请反驳我”的区别：后者是一种语气，前者是一套按判断形状选择的工序。这套工序不在底座的参数里，而在包住底座的那一层里。

还需要补上一句公平的话。“不说破”也有它的代价与边界。一个已经想清楚、只是需要执行帮助的人，被ChatSDE一次次翻前提又不给答案，会觉得它碍事；一个初判本身就站得住的学生，第一回合的翻转可能只是确认它站得住，花掉了一点时间。在课堂之外的大量任务中，把答案一并交出才是合适的默认，形式逻辑展开也是合适的默认。本书主张的不是“不说破”普遍优于“说破”，而是在一个以知识发生为目的的课堂里，默认应当是前者。

◆ 课堂里，默认值就是教育

为什么本章一再强调“默认”这个词？因为在课堂里，绝大多数交互都发生在默认之下。一个学生一学期里会和问对系统往返许多个回合，其中绝大多数回合他不会特意设置什么，只是把自己此刻的想法打进去。系统在这些回合里怎样回应，就是这个学生在这一学期里反复经历的东西；反复经历的东西，会沉淀为他的思考习惯。如果系统的默认是替他说破，他会习惯于把想法一抛，等着前提被翻过、清单被递来；如果系统的默认是逼近不说破，他会习惯于在说出一个想法之后，等着它被翻过来看，然后自己去找下一步，久而久之，他会在系统翻之前，先自己翻一次。

这和软件设计里讨论“默认设置”的道理相通。一项功能如果需要用户主动打开，绝大多数用户一辈子都不会去打开它；一项功能如果默认开启，绝大多数用户也一辈子不会去关掉它。默认值看起来只是一个技术细节，实际上它替绝大多数人做了决定。课堂里也是一样：把“不说破”设计成“学生写了批判性提示才会触

发”，或者“学生按下学习模式的开关才会触发”，等于把它留给本来就有批判意识、本来就愿意被追问的少数人；把它设计成默认，才能让它落到那些写不出提示、最需要它的学生身上。从这个意义上说，问对系统的默认值，不是一个产品参数，而是一项教育安排，它决定了这门课在千百个不被老师看见的回合里，每天在教学生什么。

这个判断已经有了一项相当有分量的实证，值得认真对待。Bastani等（2025）发表在《美国国家科学院院刊》（PNAS）上的《Generative AI without guardrails can harm learning》，在高中数学课堂里做了现场实验：使用不加护栏的GPT辅助练习的学生，练习时表现更好，可是撤掉AI之后的考试，成绩反而更差；而用了加有教学护栏的GPT Tutor的学生，这种损害基本被消除了。这个结果可以用一本带答案的练习册来读：学生做题时翻到答案页，当次的正确率当然漂亮，可到了考场上没有答案页，他才发现自己没有学会；若练习册改成只给提示、不给答案，这个落差就小得多。默认配置在质疑之后递上的那份清单，在学习场景里扮演的就是那页答案。它与本章新判断的关系是同向的先验，而不是直接的证明：Bastani等（2025）的情境是高中数学，练习的是终点已知的任务，属于本书所说的练手，而不是连老师也写不出终点的谱系性问题；它的“护栏”也是包在模型外面的教学设计，与前面所说的差别住在包装层一致。护栏有效这件事还有另一面：它说明在练手层面，护栏已经可以由通用的方式做到，ChatSDE的价值因此必须站在护栏给不了的东西上，题闸、第零轮基线、按坏法选刀和回写。

改变默认值并非ChatSDE独有的想法，上文说到的Khanmigo、CS50.ai和三家学习模式都是这样做的，默认行为可以在底座之外被重新安排。九问模式与这类工具的区别，第一在对象：辅导工具多数面向有标准答案的学习任务，引导的终点是老师事先知道的，这在本书的框架里属于练手；问对面向的是连老师也写不出终点的谱系性问题。第二在方法：引导学生走向已知答案，可以按解题步骤设计提示；而翻转一个尚无答案的问题的前提，需要先判断这个前提坏在哪里，这正是工序选择要做的事。导论已经就九问模式与AI辅导等邻近方案做过划界，这里只指出两者在“默认应当可以被重新设计”这一点上是一致的。

这一判断与教育史上的一些洞见是相通的。杜威在《民主主义与教育》中强调，教育是在经验的不断改造中进行的，而经验的品质取决于环境如何回应个体的行动。他所说的环境，今天在很大程度上包括了学生每天对话的那个系统。一个把答案提前替学生备好的环境，提供的是一种不必费力的经验，而没有阻力的经验很难改造什么。舍恩所说的情境的“回话”，也只在情境不完全按实践者预期回应时才具有教育意义。“不说破”作为默认，就是在每一次对话中有意地安排一点阻力，让学生的前提每一次都碰到一点反作用力，而把怎样绕过去、怎样越过去，留给他自己。

那么，这个默认值由谁来定、写在哪里？在ChatSDE里，它不是一个可以由学生随手关闭的开关，而是写进工序层的几条约束：先接住再批判，逼近判断而不替学生说破，按判断坏在哪里选刀（见第7章）。老师可以在出题时根据学段调整强度，给大一补练手的学生，翻转可以更温和、接住可以更多；给研究生，第一回合可以更直接地碰前提，但不能把“不替学生说破”调成零。这样安排的理由很朴素：一门以知识发生为目的的课，不应当允许它的核心教育安排在某个疲惫的深夜被一句“直接告诉我答案”悄悄绕过。学生当然可以在课堂之外使用任何通用大模型去写代码、查资料、整理笔记，那些任务本来就适合把答案一并交出；问对系统的默认，只管问对这一件事。

也要防止把这个判断推得过远。“默认值就是教育”不意味着默认值是教育的全部。九问课堂里还有老师出题、学生改题、回到现场、同学碰撞、换手评分、档案回写，这些环节都不由问对系统的默认决定。一个“不说破”做得再好的系统，如果学生从不离开屏幕回现场，问对就会变成另一种空转。默认值重要，是因为它覆盖了数量最多的那些时刻；但它覆盖不了学生在病房、实验室、账本和档案里独自面对材料的时刻，那些时刻的教育，要靠九问本身的设计与老师的阅读来保证。

◆ 本章怎样算错

本章的新判断已经有一个跑过的先验，也有一个尚未跑的正式检验，两者要分开写。先验是附录二的读数：在所测的一个模型家族、30条构造的初判、两种消息形式下，默认配置与加了批判提示的配置，在90份第一回合回复中代答90份；模拟工序层在60份中代答0份；两名评判员在这一项上完全一致（$\kappa=1.00$）。这个读数说

明判断值得检验，不说明它成立：模型只有一个家族，输入是构造的、单轮的，工序层是一段最小模拟，评判员可能从篇幅和语气猜出组别，每组在同一次会话里连续处理30条，这些局限附录二已逐条写明。

下一步的正式检验，阈值预先定在这里，它们是设计值，不是已有数据。第一条针对判断本身：取不少于30条来自真实课堂的学生初判，覆盖医学、化学、法学（或软件）三个学科，在不少于三个不同厂商的模型上各跑默认配置的第一回合，记录版本号与日期，由两名换手评审独立编码“代答”，一致率不低于90%。若三个厂商中有两个以上的模型，其默认配置组的代答率低于50%，则“默认替学生说破”不能作为通用大模型的判断，本章撤回主论断，改为只就附录二所测的家族陈述；若只有个别模型低于50%，则本章按模型分别陈述，不再统称“通用大模型”。第二条针对分别本身：若在同一底座上，工序层组的代答率与默认配置组相差不足20个百分点，则ChatSDE“逼近不说破”相对默认的差别不成立。第三条针对提示：若加了一句批判提示的组与默认配置组的代答率相差超过20个百分点，说明一句提示就能改变代答，本章“一句批判提示改变不了代答”须撤回，第10章关于工序层难以被一句提示替代的论述也要相应收窄。

第四条针对学习模式，这是本章最需要老实面对的地方。若在同样的初判上，学习模式组与工序层组在核心判断首发率（由学生先说出的比例）和问对深度评分上，差距小于评审间的一致性误差，或者不足15个百分点（取较大者，与第11章的阈值一致），那么题闸、第零轮、选刀和回写就没有带来可检验的区分，本书关于ChatSDE差别的主张须撤回，只保留“包装层决定行为”这一条。第五条针对开关：在一个不少于30名学生、连续两届的班级里，若在不强制的情况下，学期内自行开启学习模式的学生比例超过70%，并且其中包含第零轮被评为最保守、最空洞的那四分之一学生，那么“最需要追问的学生最不会按开关”的担心就没有证据，本章关于默认值须由老师而不由学生来定的论证要相应收窄。第六条针对代价：若“不说破”的默认导致学生第零轮初判趋于保守空洞，以致子本创新智商项低于以说破为默认的对照课堂，则本章对默认值应当如何设置的主张要撤回。

还有两条是关于“代答是否有害”的。附录二只测了第一回合，没有测量学生的学习结果，所以“代答损害判断的发生”仍是一个假设。若第11章的正式实验

里，代答组与不代答组在撤掉系统之后的独立判断任务上看不出差别，则本章只保留“默认配置的行为有此特征”这一描述，不再主张它构成教育上的问题。最后，“形式逻辑展开”作为另一种可能的默认，也给它留一条：若在另一家厂商的模型上，学生请求“展开”时顺着前提展开的回复占到一半以上，则这一默认成立，本章要为它单列一节，而不是继续把它当作附带说明。

方舟与许宁的两组对照，默认配置一列都是设想的回复，描述的是两种默认行为的走向，不是实验结果；三家学习模式与CS50.ai的比较，针对的也是设计目标，而非逐项测试。还有一个问题本章没有回答：即便“不说破”把学生拉回了现场，学生带回的东西为什么就比一份完整的清单更接近“知识”？默认配置给许宁的那一串因素，信息量远远大于她带回的那一条本病房判断。如果知识只是信息的累积，那份清单应当胜出才对。下一章讨论这个问题：信息发生与知识发生的差别，究竟落在哪一环。

> **带 走 的 话**
>
> **通用大模型批判得并不差，坏在批判完的同一回合里把答案也交了出去；课堂里的默认值，就是学生在千百个没人看见的回合里被教的东西。**

◆

09

第9章

信息发生与知识发生：差别落在哪一环

◆ 一份没有错的指南摘要

老年病科轮转的第2周，许宁值完一个白班，夜里在宿舍打开一个通用大模型，问了一句很普通的话：“老年住院患者夜间跌倒，应该怎样预防？”不到半分钟，屏幕上出现了一份结构整齐的摘要：危险因素分为内在与外在两类，内在因素有高龄、肌力下降、认知障碍、体位性低血压，外在因素有照明不足、地面湿滑、床栏使用不当；评估要用标准量表，干预要做多因素组合，药物要定期审查，夜间要加强巡视。许宁对照教科书看了一遍，挑不出错，还觉得比自己整理的笔记更全。她把这份摘要存进了文件夹，第二天早上带着它去了病房。

第二天的研讨课上，许宁在ChatSDE里问了同一个问题。这一回系统没有马上给摘要，而是请她先写下自己的初判，也就是第零轮。她写的是：“夜间跌倒是因为患者年纪大。”ChatSDE第一回合回了一句：“同样年纪的患者，为什么白天跌得少？”此后几周，她拿着九问跟了一次夜班，翻了交班本，问了患者，最后带回来的不是一份更全的摘要，而是一条只属于本病房、尚待检验的假说：本病房夜间跌倒的风险更多落在交接断点上，而不只是年龄，风险藏在交班本里被省略的那一句。

困惑就在这里。前一份摘要没有错，而且比后一条判断“知道得更多”；后一条判断范围极窄，还只是假说，却被本书当作知识发生的样本。如果说差别在于“大模型不行、ChatSDE行”，那就说不通，因为两边用的可以是同一个底座模型，而第一份摘要恰恰证明了模型很行。本章要回答的问题是：既然模型能说出正确的东西，课堂上的那一点差别究竟落在哪一环？本章的回答是，差别不在模型能不能，而在整个配置有没有把错误、修订和学生的现场接进来。

这一章的问题

同一个底座给出一份没有错的摘要，为什么那只是信息，而许宁带回的一条本病房假说才算知识？

◆ 信息发生与知识发生：一个针对配置的诊断

先把两个词分开。本书所说的信息发生，是指一组相关、连贯、可读的陈述被生成出来，回答了提出的问题；它的检验标准是切题、通顺、与已有文献相符。本书所说的知识发生，是指一个判断在某个具体现场里被提出、被拿去碰现场、被修订或撑住，并且被某个人接住、带着它继续往下走；它的检验标准除了说得通，还要经得起现场的反驳，还要留下回写的痕迹。按SDE的口径说，知识发生必须走完一次“D—E矛盾逼出S改变、S再回写D与E”的循环：判断与现场顶上了，判断本身被迫改动，改动又回过头改写了人的提问方式和他看现场的眼睛。

用一个厨房里的比喻来说。网上能搜到一道红烧肉的标准菜谱，用料、火候、步骤一应俱全，这是信息；一个人照着菜谱在自家灶上烧了三回，发现自家的锅薄、火猛，糖色总是炒过头，于是改成小火、晚放糖，记在菜谱边上，下回还照改过的做，这才是他手里的知识。菜谱本身没有错，错的地方要在他家的灶上才显露出来；那一行改过的批注，也只有在他的厨房里才有意义。信息与知识的区别，不在菜谱写得好不好，而在有没有一口具体的锅把它试过、改过、接住。

这里必须立一条边界。“信息发生不等于知识发生”在本书中只作为针对具体配置的诊断来使用，而不是对大模型的定论。本书不说“大模型没有现实界、没有能量，只剩下理念”之类的话，因为这类说法既无法检验，也不符合事实：大模型可以接上检索、接上传感器、接上实验记录，也可以被嵌进有人参与评估的流程。同一个底座模型，放在一种配置里只能产出信息，放在另一种配置里就可能参与知识发生。反过来，ChatSDE如果被学生拿来只问不做、只聊不回现场，同样只会产出一串漂亮的信息。被诊断的对象，始终是“模型加上它周围的一切”这个整体。

把这一条边界立住，还有一个课堂上的理由。信息发生在九问模式里有它正当而重要的位置：母本底盘要靠它，学生入门要靠它，老师出题之前查明“这一问是否已被答过”也要靠它。本书在第2章已经说过，发现不作废，而是降为底盘；同

理，信息发生不是知识发生的反面，而是知识发生的地基。一个连老年跌倒的基本危险因素都说不清的学生，下到病房也看不出交班本里少了什么。所以诊断“信息发生不等于知识发生”，用意不在贬低前者，而在提醒老师和学生：地基打好了，房子还没盖，别把地基当成房子交上去。

所以本章换一个问法：不问“这个模型懂不懂医学”，而问“这一套配置，有没有一条路让许宁的判断碰到本病房、被修正、被接住”。这就像医生不会只看一个人的心电图就宣布他健康或不健康，而要做一次全面体检。下一节给出的四项体检，就是为了把“配置”这个笼统的词拆成可以逐项检查的部位。

◆ 体检四项：基础模型、工具连接、记忆安排、人的参与

第一项是基础模型。它决定了配置的上限：语言是否通顺，医学常识是否扎实，推理是否严密，知识截止在哪一年。许宁那一晚拿到的摘要说明，这一项在今天的主流模型上通常是合格的，甚至是优秀的。但基础模型也带着一些已被研究揭示的倾向，比如Sharma等（2024）的研究中讨论过的“迎合”倾向：模型容易顺着用户的立场说话，用户表达了偏好，回答就向偏好靠拢（见第8章）。体检基础模型，既要看它强在哪里，也要看它默认往哪个方向偏。

第二项是工具连接。配置能碰到什么材料，决定了判断能落到哪一层现场。通用模型能检索公开文献，所以它给的是“老年人跌倒”这一层的知识；它看不到本病房的交班本、护理记录和夜班的实际时间，也不应该直接看到，因为那涉及患者隐私，必须去标识化、经科室与伦理委员会许可才能使用。在九问课堂里，本病房的现场不是靠系统去抓取，而是靠许宁本人带进来：她跟夜班、读交班本、听患者讲述，把经过许可、去标识化的观察写进问对。这一项体检要问的是：现场以什么方式、经谁之手、在什么许可下进入配置。

第三项是记忆安排。一次对话结束，配置记住了什么？许宁那一晚的摘要存进了文件夹，模型那一侧什么也没留下，第二天再问，一切从头开始。九问课堂里的记忆安排是另一种样子：第零轮的初判、每一轮的追问、回现场后带回的材料、判断的每一次改动，都按次序进入问对记录，轮转结束后进入本科室历届轮转学生的问对档案，再经筛选进入母本精要库。记忆安排决定了一个判断能不能被追溯：它

最初是什么样子，被什么证据改过，改过几次。没有这条可追溯的链，所谓“修订”就无从查验。

第四项是人的参与。谁写下第一个判断，谁回现场取证，谁来判定一个判断站得住还是站不住，谁对结论负责？在许宁那一晚的配置里，人的参与只有一次提问和一次阅读；在九问课堂的配置里，学生先说初判，学生下现场，带教老师与科室参与审读，评分由不参与问对的人换手完成。人的参与不是“有没有人”这么简单，而是人在环路的哪几个点上出现、在每个点上承担什么。

把四项放在一起比较，差别就清楚了。下表比较的不是两个模型，而是许宁先后用过的两套配置。两套配置之间，许宁这个人没有变，问题也没有变，变的只是包在模型周围的那些安排。

体检项	夜里那一次（通用模型，默认配置，直接问答）	九问课堂（同类底座装进ChatSDE）
基础模型	强；未开启学习模式时，往往质疑前提，但同一回合替学生说破	可以是同一个底座
工具连接	公开文献层面	加上经许可、去标识化的本病房材料，由学生带入
记忆安排	会话结束即止	问对记录、课题组档案、精要库三层留存
人的参与	提问一次、阅读一次	先说初判、下现场、带教审读、换手评分

这张表要说明的是，第一行可以完全相同，后三行的差别却决定了产出落在信息一侧还是知识一侧。这也意味着，提升课堂里的知识发生，未必要去追逐一个更新的模型，更要紧的往往是把后三行补齐。体检四项的用处，就在于把“ChatSDE与大模型不同”这句容易说成口号的话，拆成四处可以逐一检查、也可以逐一被反驳的具体差别。

需要补一句，免得读者以为许宁那一夜的配置是今天唯一的常态。2025年起，几家主要的通用大模型产品都推出了可以由使用者开启的学习模式（推出时间与做法见第8章），它们的共同点是先追问、不直接给答案。所以上表左栏的“往往质疑前提，但同一回合替学生说破”要加一个限定语：它只在默认配置，也就是没有开

启学习模式时成立。假如许宁那天晚上开了学习模式，屏幕上大概不会先出现一份摘要，而是一个反问：你认为夜间跌倒最重要的危险因素是什么？她的回答仍会取自教科书，对话也仍会在教科书的范围里展开。这个变化实实在在地提高了第一至第三层的学习质量，值得肯定；它没有改变的是后三行：本病房的材料没有进来，第零轮的初判没有被留作基线，对话结束后也没有任何东西留给下一位轮转学生。还有一点要老实承认：一位好的苏格拉底式导师，完全可能问出“为什么白天跌得少”这样的话，追问这个动作本身并不专属于ChatSDE，本书也不主张专属。

体检四项还可以反过来用，作为老师在开课前检查自己课堂的一张清单。一位老师如果只是给学生开通了某个模型的账号，就说自己的课“用上了AI”，那么按四项来查，往往只有第一项是满的：工具连接停在公开文献，记忆安排停在单次会话，人的参与停在学生自己提问、自己阅读。这样的课堂与其说是改了模式，不如说是给讲授体系换了一位更耐心的讲授者。反过来，一位老师即使用的是普通的模型，只要把学生现场的入口、问对的留存、初判先写与换手评审这三样安排好，四项里就有三项是实的。体检的用意，是让“我们用的是什么模型”这个问题，退回到它应有的位置上。

◆ 带护栏的辅导：已有实证说明了什么，没有说明什么

“差别落在配置而不在底座”，并不只是课堂里的推想，2025年的两项研究给了它相当接近的外部参照，本书把它们当作对话对象，而不是当作靠山。

先说Bastani等（2025）的高中数学现场实验（发表于《美国国家科学院院刊》，题为“没有护栏的生成式AI可能损害学习”，内容见第8章）：不加护栏的GPT让练习时的表现更好、撤掉AI后的考试更差，换成加了教学护栏的GPT Tutor，这种损害基本被消除。两种条件用的是同一家的模型，变的只是围绕模型的安排，也就是本章所说的配置。拿它和许宁那一夜对照，会看出一层相似：那份摘要读起来比她自己的笔记更全，正如不加护栏的练习“做得更好”，漂亮的当场表现盖住了没有留下来的东西。这就像用计算器写作业：作业交得快，交得对，等到考场上拿掉计算器，手上什么也没有。

但这项实验能证明的东西有边界。其一，它测的是数学练习和撤掉AI之后的考试成绩，落在“懂”的前三层，也就是说得出、推得通、用得上；学生有没有在自己的现场里改过一次判断，它既没有测，也没有声称。其二，它的“护栏”主要是教学法层面的限制，约束对话里能给什么、不能给什么，相当于体检四项里“人的参与”和对话设置的一部分；至于学生的初判是否先于模型的回应被写下来，哪一个现场能对判断说“不”，对话结束后留下什么，它没有设计，也无须设计。其三，它的对象是高中数学，终点是教师已知的，按第1章的区分属于练手而非战场。因此本书借它只说一件事：同一类底座，配置不同，结果可以相反，所以诊断应当落在配置上；本书不借它说九问课堂已经被证明有效，那要靠第11章的设计去检验；附录二只是一次预试，测的是第一回合的回复，没有测学生。

再说Kestin等（2025）在哈佛物理课上做的随机对照（载于《科学报告》）：精心设计的AI导师，学习收益高于课堂里的主动学习。这是目前“AI导师可以教得很好”这一论断最有力的证据之一，本书不回避它，而要把它收进来。它恰好说明，在终点已知的内容上，讲授和辅导这类“发现一侧”的工作，已经可以由设计得当的配置承担得相当好，这与第2章讲台价值转移的判断同向，也再一次表明起作用的是设计，而不是裸用的模型。需要追问的是：这位导师回话的依据是什么？是课程事先备好的正确答案与讲解次序。对话的那一头，能对学生说“不”的是教师预设的答案。这是一道很好的评估环，只是回话者在课程里。

由此可以把本书在两道环上多出的一格说得具体：评估环的多出一格，是把回话者从课程的答案换成学生自己的现场，交班本、故障日志、县志、记录本，这些东西连老师也未必预先知道答案；同时在第零轮把学生的初判写下来作为基线，使“改变了多少”有东西可比。承接环的多出一格，是档案不只服务同一个学生的下一次对话，还要回到出题台，让被现场推翻的判断改变下一届的九问。前一格是Kestin式导师所不需要的，因为它的终点已知；后一格是个人辅导记录所没有的，因为它的读者只有这一个学生。

◆ 两道环：评估环与承接环

体检四项回答的是“配置由什么组成”，接下来要问的是“这些部件有没有连成环”。知识发生需要两道环接上：一道是评估环，让判断碰到能够说“不”的东

西；一道是承接环，让被碰过的判断有人接住、有处存放、有路回写。两道环缺一道，再强的模型产出的也只是信息。下面分开讲。

先说评估环。评估环要守住的一条区分是：“说通了”不等于“说对了”。通顺是语言内部的性质，一段话前后不矛盾、因果链条接得上、术语用得准，就算说通了；对不对，是语言之外的性质，要拿现场来检验。大模型极其擅长说通，这是它的长处，也正是它在课堂里最容易造成误会的地方：一段说通了的话，读起来就像说对了。打个比方，导航软件给出的路线在地图上是通的，每个路口都连得上，但那条路今天是否在修、桥是否封了，地图本身不会告诉你，要车开过去才知道。评估环就是“车开过去”这一步，它让判断有机会被现场拒绝。

评估环在九问课堂里有三个具体的接点。第一个接点是题闸：拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案？能给出就打回。题闸在出题时就保证了，每一问的答案都必须到现场里去取，模型无法凭语言内部的通顺把它“说完”。要留意题闸的限度：它只检验“必须依赖现场”，不检验“值得研究”，后者要靠老师认得学生的现场，以及第16章对不可还原性的评审。第二个接点是现场本身：许宁的第8问要她跟一次夜班，记下护士从听到动静到到达床旁的时间，哪一段最长；这个时间不是推出来的，是记出来的，它可能支持她的假说，也可能推翻它。第三个接点是换手评审：子本里的判断由不参与问对的人来评，判断是否成立、证据是否撑得住，不由写的人和陪写的系统自己说了算。三个接点合起来，才算一道评估环。

把评估环放进教育思想史里看，它并不新。杜威（1910）在《我们怎样思维》里把反思性思维描述为从一种疑难的处境出发、提出暗示与假设、再拿到行动与观察中去检验的过程；他反复强调，一个观念在被检验之前只是一个假设，无论它听上去多么合理。舍恩（1983）在《反思性实践者》里讨论专业人员怎样在行动中与情境“对话”，情境会“回话”，把实践者的预设顶回来。两位作者说的都是同一件事：判断要有一个能够回话的对象。九问课堂只是把这个古老的要求放进了大模型时代的配置里，让“回话”的那一方明确落在学生自己的现场，而不是落在另一段更通顺的文字上。

再说承接环。评估环让判断碰到现场，承接环让碰过的结果不散掉。一个判断被现场推翻了，推翻本身是有价值的，前提是有人记下“它在哪里被推翻、被什么推翻”，下一个人不必再走同一条弯路。承接环在九问课堂里就是问对档案与回写：每一轮问对按次序留存，判断的每一次改动都有版本，轮转结束后经筛选进入科室的问对档案，有分量的判断再进入谱系母本精要库，成为下一届九问的出发点（见第24章）。比方说一个家庭的病历本，每次看病都写在同一本上，下次换了医生，也能翻到上次用了什么药、效果如何；病历本散了，每次看病就都成了初诊。

两道环之间还有一层关系：没有承接环，评估环的结果无处落脚；没有评估环，承接环里存下的只是未经检验的说法，越积越厚，反而成了一座看似权威的回音室。许宁那一晚的配置，两道环都没有接上：摘要没有被本病房检验过，存进她文件夹之后也不会再被修订，它从生成那一刻起就定格了。九问课堂里那条“交班本里被省略的那一句”，则同时挂在两道环上：它被标为本病房假说，交科室验证，学生不改医嘱、不作临床决策；它的来路，从第零轮的“因为年纪大”，到ChatSDE那句“为什么白天跌得少”，到第2问被许宁提升为“交班时被省略的那一句，是谁、在什么情况下省略的”，再到夜班记录，都留在档案里，等着下一届轮转学生去复查、去推翻或者加厚。

评估环与承接环的缺失，在不同学科里表现各异，但结构相同。以软件工程课上的方舟为例，他若在默认配置下裸用一个模型问“选课系统为什么崩溃”，得到的会是一份关于扩容、缓存、负载均衡的通用排障清单，说得通，也大体不错；只有当他把学校选课系统的故障日志带进来、追问崩溃前十分钟最先变慢的是哪一类请求，判断才第一次有了被日志否定的可能，这是评估环；他的排查记录留进课程档案，下一届同学接着查那个没人负责的排队环节，这是承接环。医学病房里的交班本，与机房里的故障日志，在这里扮演的是同一个角色：它们是能对判断说“不”的现场。

还要补一句，以免误读。评估环与承接环不是ChatSDE独有的发明，好的科研组、好的临床科室本来就有这两道环：组会就是评估环，实验记录本和科室质控台账就是承接环。通用大模型同样可以被接进这样的环里，许多研究者正是这样使用它的。本章要指出的只是，在课堂这个具体场景里，一个学生对着一个裸用的对话

框，两道环默认是断开的；把它们接上，需要的是一整套配置，而不是一个更聪明的模型。这也是下一章要专门讨论工序层的原因。

◆ “懂”的五种含义

关于大模型，课堂上最常听到的争论是“它到底懂不懂”。一方说它对答如流，当然懂；另一方说它只是在拼接字词，根本不懂。这场争论之所以总是吵不出结果，是因为“懂”这个字里叠着好几层意思，双方各执一层。本章主张把“懂”拆开，至少分成五种含义分别讨论；拆开之后，许多看似对立的说法其实各自成立，真正的差别就落在第四、第五两层上，而这两层恰好对应评估环与承接环。

第一种含义是“说得出”，即能准确复述一个领域的概念、事实和既有结论。问“老年跌倒的危险因素有哪些”，能一条不漏地答出来，就是这一层的懂。这一层的懂，主流大模型通常做得比本科生好，也是九问课堂把母本底盘定为门槛项、用“复述加反例口试”来检查学生的原因（见第15章）：学生在这一层不必和模型比，但必须过关。好比学开车先要背交通规则，背得出不等于会开，背不出却一定不能上路。

第二种含义是“推得通”，即能在给定前提下作严密推演，从A推到B，从B推到C，不出逻辑破绽。许宁如果对模型说“夜间跌倒是因为年纪大”，模型无论是顺着这一前提推出一整套增加夜间巡视、加装床栏、减少镇静药物的方案，还是先质疑前提、随即把替代的原因清单一并写好（第8章按附录二的试跑，认为默认配置下更常见的是后者），链条都是完整的。这一层的懂，模型同样很强。它的局限不在推理本身，而在于这条链是模型替她推完的，那个本该由她先说出的判断没有留下位置：推得越通、答得越全，越容易让人忘记起点可能是歪的。就像一个账房先生把一本账算得一分不差，如果第一笔入账就记错了，后面算得越准，错得越稳。

第三种含义是“用得上”，即能把已有知识套到一个给定情境里，给出合乎规范的处理意见。给模型一个虚构病例，它能按指南给出评估与干预建议，这是这一层的懂。它在课堂上很有用，也正是讲授体系长期训练学生的那一层。但要注意，

这里的“情境”是被描述好了、递到面前的情境，而不是要自己去发现的现场。套用的前提，是有人已经把现场整理成了题目。

第四种含义是“知道会错在哪里”，即能把一个判断拿到具体现场去碰，知道它在这个现场可能怎样失效，并在失效时修改它。这一层的懂不是单靠一次回答能显示的，它要求判断真的被现场检验过、被修订过。许宁在跟夜班之前，对“交接断点”的判断只停留在前三层：她说得出，也推得通；跟完夜班、比对了交班本，她才开始知道这条判断在哪些夜晚站得住，在哪些夜晚站不住，比如有家属陪护的夜晚，情形可能全然不同（第6问正是为此而设）。这一层的懂，属于整套配置，而不属于某一个回答：只有评估环接上了，这一层才可能出现。

第五种含义是“懂了之后自己变了”，即这次懂得改写了懂的人此后的提问方式和看现场的眼睛，并且这种改写留下了可以被观察到的痕迹。许宁轮转结束后，再进任何一个病房，第一眼看的也许不再是患者的年龄，而是交班本上空着的那一栏；她的期末自述与第零轮初判放在一起，能看出这一变化。这一层对应承接环，也对应SDE所说的回写：回写要被观察，不能被宣布。一个配置如果从不留存、从不对照，就无从观察这一层是否发生过。

五层的划分对评估还有一个直接的用处：它告诉老师每一层该到哪里去找证据。第一层的证据在口试里，第二层的证据在论证的链条里，第三层的证据在病例分析作业里，这三层都可以从一份成品中读出来；第四层的证据只能在问对记录与现场记录的对照中找到，第五层的证据只能在第零轮与期末自述的对照中找到，它们都不在成品里，而在轨迹里。这也解释了为什么第17章要把过程项单独计分：一份子本即使在前三层上写得无可挑剔，如果拿不出后两层的轨迹，评审者就无法判断它是否发生过知识发生，只能把它当作信息发生的产物来对待。

把五层摆开，那场“懂不懂”的争论就有了着落。在前三层上，说大模型“懂”，大体是成立的，而且往往懂得比学生多；在后两层上，问“模型懂不懂”其实问错了对象，因为这两层本来就不是单个回答能承担的，它们是一套配置的性质，要看评估环和承接环有没有接上。本章不对模型本身下“懂”或“不懂”的定论，只主张：课堂要的“懂”，重心在第四、第五层，而这两层要靠配置去造。

◆ 同一个问题，两份带回病房的东西

现在回到开头的那一对产出，用前面的工具把它们重新看一遍。同一个问题，“老年住院患者夜间跌倒，应该怎样预防”，通用模型给了许宁一份完整的老年跌倒防治指南摘要，ChatSDE让许宁带回了一条本病房的判断。两份东西各自是什么，各自落在哪一层，差别从哪里来，现在可以逐一说清。

先看那份摘要。它在第一、第二、第三层上都相当出色：危险因素列得全，干预措施推得通，套到一个典型病例上也用得上。它的问题不是错，而是“对任何一个病房都对”：它不需要见过本病房，也不会因为本病房的任何事实而改动一个字。用题闸的语言说，这份摘要回答的正是那道会被打回的题，“老年人跌倒有哪些危险因素”，拿掉许宁的现场，它照样能答出来。它是高质量的信息发生，在课堂上可以作为母本底盘的一部分来读，许多带教老师也会乐意把它发给学生作为入门材料。

再看许宁带回的那条假说，即“交接断点”那一句。它在第一层上远不如摘要丰富，在第二层上也谈不上严密，它甚至还没有被证实，只是一条待检验的本病房假说。但它在第四层上动起来了：它是从“为什么白天跌得少”这一句反问里生出来的，经过第1问对交接时段的统计、第2问对交班本省略的追查、第8问对夜班反应时间的记录，一步步被现场推着改过来的；它也给出了自己可能失效的地方，比如第6问所问的家属陪护差异若能解释大部分差别，交接断点的说法就要退让。它在第五层上也开了头：第零轮的“因为年纪大”与它并排放在档案里，差距本身就是许宁被改写的痕迹。

那么差别从哪里来？对照体检四项，可以一项项排除。不是基础模型的差别，两边可以用同一个底座；差别在于后三项和两道环：工具连接上，有经许可、去标识化、由许宁亲手带进来的本病房材料；记忆安排上，有按轮次留存的问对记录；人的参与上，有许宁先写的初判、她本人的夜班、带教老师的审读、换手的评分。评估环上，有题闸、现场和换手三处接点；承接环上，有档案和回写。把这些都拿掉，只剩一个对话框，ChatSDE给出的东西也会退回到信息一侧。反过来，把这些都接到一个通用模型上，它也可能参与本病房的知识发生，只是这时候它已经不是“裸用”的通用模型了，而是一套新的配置。这也提醒读者：本章的结论不能读

成“ChatSDE胜过通用模型”，而只能读成“接上了两道环的配置，才能把课堂带到知识发生一侧”。

这对带教老师意味着什么，也可以说得具体一些。那份摘要不必扔掉，它可以放进第1周的母本底盘材料，供学生复述、供口试时出反例；但它不能被当作这门研讨课的产出，也不能被当作许宁“已经懂了”的证据。老师在课上要看的，是第零轮那一句初判与后来那条假说之间的距离，以及这段距离是由哪些现场材料一步步撑起来的。换句话说，老师读的不是答案，而是轨迹；轨迹落在评估环与承接环上，答案落在信息一侧。

还应当注意一个伦理边界。许宁带回病房的那条判断，无论它在课堂里被打磨得多么好，它都只是一条交给科室验证的假说。学生不改医嘱、不作临床决策，防跌倒措施的任何调整都须经科室按正常程序决定。知识发生落在许宁身上和本科室的档案里，而不是直接落在患者身上；这条边界本身也是承接环的一部分，它规定了判断被谁接住、在什么程序里被接住。

◆ 本章怎样算错

本章的主张是：课堂上信息发生与知识发生的差别，主要不落在底座模型上，而落在配置是否接上了评估环与承接环。它在三种读数下会被推翻。

第一，设三个条件的对照：默认配置的裸用对话框、开启学习模式的对话框、接上两道环的配置，用同一底座、同一批题目，连续两届、每届每个条件不少于30名学生。由两名不参与问对的评分人，只读问对记录与期末自述，独立标注每名学生的判断是否在现场材料面前被修改过至少一次，并且能在自述里指出改动的来路。若接上两道环的配置与学习模式条件之间，这一比例之差不超过10个百分点，那么“差别落在环上”就失去依据，应当撤回；若两者的差距主要出现在学习模式条件与默认配置之间，而接环与否几乎不起作用，说明起作用的只是追问这一层壳，本章要改写为“差别住在壳上”。附录二的小样本试跑不是这一设计的缩小版：它测的是第一回合的回复，没有触及学生的判断有没有被现场改过，因而对本条不构成证据。

第二，在默认配置的裸用条件里换上更新一代的底座，其余不变。若换底座之后，裸用条件下能拿出第四层证据的学生，达到接上两道环的配置的八成以上，那么本章把重心放在配置上的论述就偏了，应当改写为以模型能力为主的解释。

第三，对第四、第五层的拆分本身设一道门：随机抽取20份问对记录与期末自述，两名评审独立标注哪些学生有第四层痕迹、哪些有第五层痕迹，若两人的一致率低于70%，就说明这一拆分还只是一种说法，不能当作评估的依据，需要重新设计。

三条里没有一条可以靠论述通过，只能靠记录通过。一门课自称“我们接上了两道环”，并不构成证据；能作为证据的，是可供换手评分人逐条核对的问对记录、现场材料与版本史。

本章还留下一个问题没有回答。说“配置决定了差别”，等于说差别在模型之外；那么模型之外的那一层，究竟是怎样组织起来的？它为什么能在一个学生开口的第一回合就把前提翻过来看，又为什么能在不同学科、不同学生那里不落入同一个套路？如果这一层可以被说清，它是否也能被搬到另一个底座上，或者被别人照抄？下一章要讨论的工序层，就是对这些问题的回答：模型是发动机，工序是驾驶术。

> **带走的话**
>
> **知识不是模型说得出的东西，而是被学生的现场碰过、被修订、被人接住的东西；评估环让判断有处被拒绝，承接环让被碰过的判断不散掉，两道环缺一，再强的模型给出的也只是信息。**

◆

10

第10章

工序层：模型是发动机，工序是驾驶术

◆ 同一个底座，两间化学课堂

化学系三年级的课程型研究《合成条件的发生》开课第3周，沈澄面对的是本课题组那个老问题：某偶联反应的产率时高时低，十年档案里起伏不定。这一学期，课程恰好在两个平行班里用了同一个底座模型：一个班让学生直接在默认配置的对话框里使用它，另一个班把它装进ChatSDE，接上课程的工序层与课题组档案。沈澄所在的是后一个班，但她在课下也用过前一种方式，于是同一个问题、同一个底座，她亲眼看到了两种回应。需要说明的是，这里的两个班与沈澄的经历都是构造案例，只用来说明工序层的位置，不是一次调查的结果。下文默认配置一侧的回应同样是设想的，只描述走向，写法参照附录二里默认配置实际出现的样子，不是某一次实际运行的记录。

她在默认配置的对话框里写下初判：“产率低是因为试剂纯度不够。”模型的回应专业而周到：它先说“有可能，不过产率低可以出在很多地方”，随即把几个方向一并列了出来，包括转化是否完全、有无副产物、后处理损失，也提到几种常见杂质可能干扰催化循环，并建议用高纯试剂做一次对照，核查试剂批号与供应商的分析证书。整段回答推理严密、术语准确，前提也被轻轻碰了一下。照此办理，课题组也许会多花一笔试剂费，也许产率会有所改善，也许不会；但这份清单是模型替她列好的，哪一项在本组这个反应里真的起作用，她还没有去自己的现场里把它找出来，答案就已经到了手里。

在ChatSDE里，她写下的是同一句初判，作为第零轮。系统第一回合先接住了她：试剂纯度确实是这类反应常见的嫌疑，值得留作一条待查线索。紧接着它问：“同一瓶试剂，为什么三月好、六月差？”这一句没有否定纯度，却把纯度说不通的地方摆到了桌面上：如果是同一瓶试剂，纯度在三个月里不会凭空变差那么

多，那么三月与六月之间变了的，就该是别的东西。此后几轮问对，沈澄沿着这一问翻了十年档案的月份分布，查了记录本里“室温”二字有没有写过具体的温度和湿度，最后得出一条待复做验证的判断：本组某一步的成败依赖未入规程的环境湿度与操作时长，规程里没写的那一句，在师门口头里传了十年。

裸用那一侧的回应并没有做错什么。沈澄给出了前提，模型在回应里补出了几个方向，并给出了可执行的建议，这在许多场合正是一个好助手应有的样子：工程师问“这根梁怎样加固”，他要的是加固方案，而不是被反问“你确定是梁的问题吗”。问题只在于课堂的目的不同。课堂要的不是把学生的前提执行得更好，而是让学生学会怀疑自己的前提；一个对助手来说恰当的默认值，放进课堂就可能成为一种不恰当的默认值。这也是本书反复说“默认值就是教育”的含义。

两种回应用的是同一台发动机，开出了两条路。在默认配置下，前一种回应的特点是把前提稍加质疑后便替学生把答案一并写好，后一种回应的特点是把前提翻过来、把下一步留给学生，这一区分第8章已经讲过。需要先交代一个限定：今天的通用大模型产品已经带有可开启的学习模式，前一种回应并不是通用模型的唯一样子，下一节专门处理这件事。本章要追问的是下一步：如果底座相同，那个“第一回合就把前提翻过来看”的默认行为，究竟住在哪里？它不在模型的参数里，否则默认配置的那个班也该得到同样的反问；它住在包住模型的那一层，本书称之为工序层。

> **这一章的问题**
>
> 第一回合就把前提翻过来看的那种默认行为，既然不在模型的参数里，它究竟住在哪里，一句提示或一家竞争者能不能把它抄走？

◆ 发动机与驾驶术：工序层是什么

先用一个比喻定位。同一辆车，交给刚拿驾照的新手和交给跑了20年山路的老司机，开出来是两回事。发动机的马力、变速箱的档位、刹车的灵敏度都一样，差别在驾驶术：什么时候减速，弯道前看哪里，下雨天跟车留多远，听到底盘异响先查什么。驾驶术不在发动机里，换一辆车，老司机的驾驶术跟着走；把老司机请下

车，再好的发动机也不会自己学会看弯道。本章把大模型的底座比作发动机，把ChatSDE的工序层比作驾驶术。

所谓工序层，是指包在底座模型外面、规定它在课堂里按什么次序做什么事的那一整套安排。它至少包括三样东西。第一样是工序选择：面对学生的一个判断，先诊断它坏在哪里，再决定用哪一把刀去破、指向哪一个方向去立。第二样是约束，本书称之为宪法：一些无论学生怎么问都不能违反的规矩，比如先接住再批判、逼近判断而不替学生说破。第三样是接口：问对怎样留存，怎样与出题台和档案相连，判断的版本怎样被记录与回写。这三样都不是模型参数的一部分，而是可以被写出、被审查、被修改的工作规程。

把工序层与模型能力分开，有一个直接的好处：它让“ChatSDE与通用大模型不同”这句话有了明确的落点。如果差别在参数里，那么底座一换，差别就跟着变，主张也就随时可能失效；如果差别在工序层，那么差别的来源是可以指认、可以复查、也可以被反驳的。这也正是第9章所说的配置诊断再往前走一步：配置由四项组成，而在基础模型之外的三项里，真正决定第一回合怎样回应的，是工序层。

把两间课堂并排放着看，工序层的位置可以用下表概括。表中比较的是同一底座的两种用法，不是两个模型的高下。读这张表时，请把左栏理解为一种在别处完全合理的默认，而不是一种缺陷；学习模式在左栏与右栏之间，它改动了第一行，大体没有碰后四行。

层次	默认配置的对话框	装进ChatSDE
底座模型	同一个	同一个
第一回合的默认	质疑前提，并在同一回合列出替代原因与做法（开启学习模式则先追问）	先接住，再诊断前提坏在哪里
提问怎样生成	随用户提问而定（学习模式下为通用追问）	随诊断出的坏法（断拍、缺式、偏路）而定
行为边界	有问必答、答得完整	逼近判断、不替说破、破中指立
记录与回写	会话结束即止	问对档案接出题台与精要库

表的第一行相同，后四行全部落在模型外面的那一层；这一层被拿掉，后四行就会回到左栏。这张表也提示了检验的方向：既然差别据称都落在后四行，那么只要固定第一行、只动后四行，就能看出差别是否真的来自这里。表中的每一行，都对应着一种可以被观察的行为，而不是一种只能被描述的气质。第11章的对照实验，正是沿着这张表的结构来设计的。

还要说清一点，以免把工序层误解为“一段更长的系统提示词”。系统提示词是工序层的载体之一，但工序层的分量不在字数，而在它背后那套判断方法是否成体系、是否经过课堂检验、是否与档案相连。打个比方，驾驶术可以写成一本驾驶手册，手册可以复印，但手册里的每一条为什么这样写、在什么路况下该先用哪一条，是在无数次上路中磨出来的。下面两节分别讲工序层的核心：工序怎样选，宪法怎样约束。

◆ 2025年的一次自然对照：学习模式

要论证“差别住在包装层”，本书不必只靠自己的设想，因为2025年的产品市场已经替它做了一次相当干净的对照。Anthropic的Claude学习模式、OpenAI的ChatGPT Study Mode和Google的Gemini Guided Learning，先后在2025年推出（时间见第8章）。三家的做法相近：学生开启这个模式之后，模型不再直接交出答案，而是先追问，一步步引着学生自己往前走。从使用者一侧看，变化是由一个开关带来的，开关两侧是同一类底座，第一回合的行为却从“给”变成了“问”。这是“第一回合的默认行为住在包装层、不在参数里”最直接的旁证，而且出自本书之外的人之手。

这个旁证同时收窄了本书的主张，收窄得很有益处。第一，“先追问、不直接给答案”这一层壳并不难抄，业界已经在做，并且做得相当规模化，本书不能再把“会追问”当作ChatSDE的独门本领。第二，包装层不是单一的一层，至少可以分成三层：追问壳，即第一回合问还是答；诊断与选刀，即追问问什么、为什么先问这个；记录与回写，即对话之后留下什么、回到哪里。就公开介绍而言，学习模式主要做在第一层。第三层的差异，第9章已经分析过。本章接下来的篇幅，落在中间那一层。

用沈澄的初判来看这三层的区别。假设她在学习模式下写出“产率低是因为试剂纯度不够”，一个好的学习模式大概会反问：你为什么认为是纯度？你怎样检验过？这是一句好问题，却是一句可以安在任何判断后面的问题，换成“崩溃是因为服务器不够”，换成“夜间跌倒是因为年纪大”，也一字不差地适用。ChatSDE的那一句则不同：“同一瓶试剂，为什么三月好、六月差？”它只能出自一个诊断，即这个判断的关系里少了一项会随季节变的东西；而能做出这个诊断，靠的是谱系母本精要库里关于这类偶联反应对条件敏感的判断，以及课题组档案里月份分布的记录。前一句问话的材料在壳里，后一句问话的材料在壳外。

还要为学习模式说一句公道话：未来的学习模式完全可能接入更多材料，一个足够强的模型也可能自己问出“三月和六月有什么不同”。所以本书并不主张“学习模式做不到”，而是主张两件更窄的事：其一，要做到这一步，所需要的母本精要库和档案不在壳里，需要有人一届一届地积累；其二，即使问出了这一句，若没有判别细则去核对它是否指向了学生初判真正的坏处，没有记录去核对它有没有替学生说破，这一句问话也只是一次碰巧的好问话，不是一道可以复查的工序。这两件事能否成立，要靠第11章的设计去检验；附录二的试跑没有测试学习模式，它的“模拟工序层”也只是一段最小指令，帮不上这里的忙。

◆ 工序选择：按判断坏在哪里选刀

工序选择的第一步不是批判，而是诊断：学生的这个判断，坏在哪一处？本书沿用SDE的三组工具来做这一诊断，第7章已作介绍，这里只说它们在工序层里怎样被调用。三大方程查缺式，补式：三大方程F、G、H在本书中只是关系占位记号，用来提醒一个判断是由几项之间的关系撑起来的，诊断时问的是这组关系里是否少了一项。六路径查偏路，排路：六路径是主导分析次序，诊断时问的是学生从起点走到结论，路是否走偏、是否跳过了该先到的那一站。123原理查断拍，点火：它只指“D—E矛盾逼出S改变、S再回写D与E”这一循环，诊断时问的是这个循环在哪一拍断了，比如矛盾已经摆在眼前，判断却没有被逼着改。三组工具对应三种坏法，选刀就是先认坏法、再拿相应的工具。认坏法有一个固定的次序，即第7章给出的判别细则：先查学生在自己现场里是否已经接触到一件与初判直接相撞的事实，判断却没有改，若有，标为断拍；若没有这样的事实，再查初判的关系里是否少了一项

被当作常数的因素，若少了，标为缺式；关系齐全、也无矛盾，而起手次序与任务不配，才标为偏路。这个次序的用处，是让同一句看上去相似的反问，不会被不同的人标成不同的坏法。

用沈澄的初判来演示。“产率低是因为试剂纯度不够”，这一判断把产率与一项因素挂在一起，诊断时最先显露的问题是：同一瓶试剂在不同月份用出不同产率，这一事实在“纯度—产率”的关系里没有位置，关系里少了一项会随季节变化的东西。这是缺式，对应的刀是补式。按判别细则，这里还要多问一步为什么不是断拍：沈澄此时并没有接触过那件与初判相撞的事实。同一瓶试剂在三月与六月的产率差，藏在十年档案里，她还没有按月份翻过，不能说她眼睁睁看着矛盾却不肯改判断，所以第一步查不出断拍，才轮到第二步，发现关系里少了一项。所以ChatSDE第一回合问的是“同一瓶试剂，为什么三月好、六月差”，这一问并不替她说出“湿度”，只是把缺的那一项所在的位置指出来：在纯度之外，还有一个随月份变的量。破的同时，立的方向也就有了，沈澄的第2问、第3问正是沿这个方向长出来的：同一瓶试剂在三月和六月用出的产率差多少，实验记录里“室温”一词有没有一次写过具体温度和湿度。

再看方舟在软件工程课上的系统排障。他的初判是“崩溃是因为服务器不够”，ChatSDE问的是“去年服务器加了一倍，为什么还崩？”这一问选的刀与沈澄那一问不同，尽管两句话的形式相近，都是拿一个已有的事实去顶初判。差别在于谁手里握着这个事实：沈澄不知道三月与六月的产率差，方舟却清楚去年扩容一倍仍然崩溃，那是他自己经历过的事。去年扩容这件事已经发生过，现场已经回话：资源翻了一倍，崩溃照旧。矛盾就摆在眼前，判断却没有被逼着改，循环在“逼出S改变”这一拍断了。这是断拍，对应的刀是点火：把已经存在的矛盾点亮，让它去逼判断改。点着之后，方舟的问题变成了“崩溃前十分钟最先变慢的是哪一类请求”，不再问总量够不够，而问链条上哪一环先卡住，最后得出的构造判断是：瓶颈在一个没人负责的排队环节。

偏路的例子可以在许多学生的初判里找到。一个常见的样子是，学生从一个结论出发，直接跳到一套干预方案，跳过了到自己现场去看一眼这一站；比如一位学生面对病房跌倒，第一反应是“应该加装床栏、增加巡视”，路是从方案出发的，

现场被跳过了。对应的刀是排路：把次序重排，先回现场取一份证据，再谈方案。许宁的九问中，第8问要她跟一次夜班、记下护士从听到动静到到达床旁的时间，就是一种排路，它让判断先过现场这一站。这里说的是九问中这一问的作用，不是对许宁初判的诊断；许宁的“因为年纪大”，按判别细则是缺式，它漏掉的是交接的时间次序，所以ChatSDE问的是“为什么白天跌得少”。

三组工具怎样“配”，有两条要点。其一，配是按轮次的，不是一次全上。一个判断可能同时有好几处毛病，但每一回合只挑最要紧的那一处下刀，其余的留待后几轮；一次把所有毛病都指出来，学生接不住，也就谈不上发生。其二，配是要亮出依据的。ChatSDE在下刀之后，可以向学生说明这一刀为什么这样选：你的判断里少了一项随季节变的量，所以先问季节；你的现场已经给出了反证，所以先把反证点亮。亮出依据的用意，是让学生慢慢学会自己选刀，毕业之后，面对没有ChatSDE的新问题，也能先问“这个判断可能坏在哪一处”。

这里也就回答了一个常被追问的问题：批判性激发怎样才能不落入套路？如果系统只会说“请考虑其他可能的原因”，那么这句话对任何判断都适用，学生几轮之后就学会了绕开它，这就是套路。非套路化的来源，不是措辞的变化，而是诊断的差异：缺式、偏路、断拍是三种不同的坏法，同一种坏法落在不同学科、不同现场，长出的问题也各不相同。沈澄的问题指向月份，方舟的问题指向扩容之后的崩溃，两句话没有一个字相同，背后的选择方法却是同一套。好比老中医看病，药方千人千样，辨证的方法却是一套；方子之所以不重样，正是因为方法在起作用。

◆ 宪法即约束：先接住、逼近判断、不替说破

工序选择回答“用哪一把刀”，宪法回答“刀怎样下”。本书所说的宪法，是工序层里那几条无论学生怎样提问、无论底座换成哪一个都不能违反的规矩。它们像是驾驶术里的“雨天不急刹”“弯道不超车”，不是每次上路都会用到，但一旦违反，后果就不是开得好坏的问题，而是会不会翻车的问题。在问对里，翻车的样子有两种：一种是学生被批判击倒，从此只敢问模型要答案；另一种是判断被系统替学生说出，学生只是抄下来。下面三条宪法，分别防这两种翻车。

第一条，先接住，再批判。学生的初判，哪怕是“因为年纪大”“因为纯度不够”“因为服务器不够”这样一眼看去就嫌单薄的说法，也包含着学生注意到的某样真实的东西：年龄确实与跌倒相关，纯度确实影响偶联，资源确实可能成为瓶颈。ChatSDE第一回合要先把这样东西接住，承认它值得留作线索，然后才把前提翻过来看。这条规矩不是礼貌，而是发生的条件：一个判断被整个否定，学生就失去了往前走的立足点；被接住之后再撬开一角，学生才有地方站着去看那一角下面是什么。这也正是第7章所说的“接住放大，再批判”。

第二条，逼近判断，不替学生说破。ChatSDE的问题要一步步把学生逼到核心判断跟前，但最后那一步要留给学生自己迈。沈澄的问对里，系统从头到尾没有说出“湿度”二字，它问的是月份、问的是“室温”有没有写过具体数字；是沈澄自己在档案里看到最差的批次扎堆在潮湿的月份，自己说出了湿度。这里的“首发”是相对ChatSDE而言的：老师出的九问是学生手里的现场材料，其中有的提到温度和湿度，师兄口头说的话也是她取得的现场材料，这些都不算ChatSDE替她说破；要守住的只是ChatSDE自己的追问里不出现核心判断的关键变量词。替学生说破，从知识的角度看似乎更省时间，从发生的角度看却是一次截断：判断出自系统，学生只是接收者，问对就退回成了一种更迂回的讲授。第15章评估项目中“问对深度”一项要看“关键判断是否学生先说”，就是为了在评分上守住这一条。

“逼近而不说破”在教育思想里有一个熟悉的参照，即维果茨基所说的“最近发展区”：学生独自做不到、在他人支持下能够做到的那一段距离。好的支持落在这段距离之内，把学生往前推一步，而不是把他抱过去。ChatSDE的追问若是落在这段距离之外，学生接不住，问对就停了；若是直接把判断说出来，那就越过了这段距离，学生没有走，只是被抬着到了终点。工序层要做的，是让每一刀都落在这段距离里，这也是它要按轮次、一次只下一刀的又一个理由。

第三条，破的同时指出立的方向。批判如果只破不立，学生得到的是一堆被拆散的旧判断，却不知道往哪里重建。工序选择本身已经包含了立的方向：补式指向缺的那一项所在的位置，排路指向该先到的那一站，点火指向已经摆在眼前的那个矛盾。ChatSDE的每一次质疑，都要同时把这个方向亮出来，可以是一个新的问题，可以是一处建议去查的现场，可以是一份档案里值得翻看的旧记录。用一句大

白话说，好的师傅指出徒弟的刨子推歪了，同时会告诉他眼睛该看哪条线，而不是只说一句“歪了”。

宪法还要能被查。一条规矩如果违反了也无人知晓，它就只是一句愿望。在ChatSDE的设计里，每一轮回应都附带一条简短的工序记录：这一轮判定的坏法是什么，选了哪一把刀，有没有替学生说出核心判断。带教老师可以抽查，评审者可以复核，课程结束后也可以统计宪法在哪类问题上最常被突破。例如，在沈澄的课上若发现系统在第5、6轮之后越来越倾向于直接给出“湿度”这样的结论，那就说明“不替说破”这一条在长对话里松动了，需要在工序层里加一道检查，而不是怪学生问得太急。

三条宪法有一个共同的性质：它们约束的不是系统知道什么，而是系统在什么时候说什么、不说什么。这类约束很难靠模型自身的训练稳定获得，因为一般的对话训练鼓励的是有问必答、答得完整，而宪法要求的恰恰是在某些时刻有意地少答一步。把它们写进工序层，作为每一轮回应之前都要过一遍的检查，才能让它们在不同学生、不同学科、不同底座之间保持一致。这也是本章称之为“宪法”的原因：它们不是一句提示，而是一层规定行为边界的条款，条款写在哪里、违反了怎么查，都应当是可以说清的。

◆ 出题台、问对、档案回写：同一条工序链

到这里为止，工序层似乎只在问对那一刻起作用。其实不然。九问模式的几个环节，出题台、问对、档案回写，在本书的设计里挂在同一条工序链上，用的是同一套诊断方法；把它们拆成三个互不相干的工具，工序层的力量就会散掉。这好比一家老字号的手艺，从选料、下锅到记账，用的是同一套讲究：选料时看的那几样，正是下锅时要防的那几样，也正是账本上要记的那几样。

先看出题台。老师用装有SDE工序层的大模型从母本精要库出九问，第一步是从母本的核心判断出发，第二步是按学生画像具体化，每一问都要过题闸。母本精要库里的核心判断，本书分为方程判断、路径判断、动力判断三类（见第3章），恰好对应问对时的三组诊断工具。也就是说，出题时老师从一条方程判断出发设计的那一问，到了问对时，ChatSDE诊断学生的初判是否缺式，用的是同一组关系；出

题时从一条路径判断出发设计的那一问，问对时就用来查学生是否偏路。出题与问对共用一套语言，问对时的每一刀，才能落在九问本来要打的那个位置上。

再看问对与档案。问对记录按轮次留存，不只是存下一段聊天，而是存下每一轮的诊断：这一轮判断坏在哪里，选了哪一把刀，学生怎样接，判断怎样改。这样的记录才能被下一环使用。评分时，评审者能看到推进还是打转（见第17章）；课程结束后，带教老师能看到这一届学生的初判最常坏在哪一处，比如化学课里“归因于试剂”的初判反复出现，说明母本精要库里关于环境条件的那一条判断写得太薄，需要加厚。这就是回写：问对档案经筛选回到母本精要库，精要库变了，下一届的九问随之改变（见第24章）。

沈澄的案例可以把这条链完整走一遍。出题时，老师从课题组十年反应条件档案与合成方法学专著中提炼的一条核心判断出发，按沈澄的画像出了九问，其中第3问“实验记录里‘室温’一词，十年里有没有一次写过具体温度和湿度”过了题闸，因为一个未接入本组记录本的通用大模型（记录所用版本）答不出来。问对时，ChatSDE诊断她的初判缺式，用“同一瓶试剂，为什么三月好、六月差”补式，这一刀正落在第2问、第3问所指的位置。复做时，她按第6问把每一步的时间、温度、湿度都记下，与档案比对。回写时，如果她的湿度假说经后续复做站住，它会被写进课题组规程，第9问“如果把湿度写进本组规程，第一个要改的是哪一行”就从一个问题变成了一条修订记录；而她这一届的问对档案，也将成为下一届出题的材料。

方舟的系统排障也是一条链。他的软件工程课以学校选课系统故障日志为现场，出题、问对、档案用的是同一套诊断：扩容之后仍然崩溃，这一矛盾在出题时就被标为一处可以“点火”的断拍，问对时ChatSDE便从这里下刀。排查记录进入课程档案后，“瓶颈在没人负责的排队环节”这一构造判断会成为下一届同学的起点，他们可以接着问：这个环节为什么没人负责，责任断在了哪一次系统交接里。工序链在这里显出它的另一层意义：它把一门课从“每年重讲一遍”变成“每年接着往下走一段”。

工序链还给老师留出了一个明确的位置。工序层不是要把老师从课堂里拿掉，恰恰相反，链上有几处只能由老师来做：从母本中认定哪一条判断是核心，按学生

画像判断谁的现场在哪里，在问对记录里看出哪一刀下得不对，在学期末决定哪一条学生判断值得写回精要库。这些都是判断，而不是操作；工序层把操作安排好，把判断的位置空出来，留给认得学生、认得学科的人。说得直白一些，驾驶术也要有人来教，教练的位置并不因为车更好了而消失。

◆ 追问壳易抄，判别细则与档案难抄

既然工序层是一套可以写出的规程，它会不会很容易被照抄？上一节已经给出了一半的答案：追问这一层壳，今天已经有人在大规模地做。对话界面、系统提示词、问对的格式、第零轮的流程、“先问后答”的姿态，一个熟练的开发者看过几次ChatSDE的问对就能做出外观相近的东西，三家主流产品的学习模式更是把这层壳送到了每一位学生的手边。本书不把壳当作秘密守着，相反，壳可以开放，让更多课堂试用；它甚至应当开放，因为壳越普及，学生越习惯被追问，课堂的起点越高。

难抄的是三样东西。第一样是工序判别细则：面对一条初判，先查断拍、再查缺式、最后查偏路的次序，每一步查什么，什么算“学生已接触到的事实”，什么情形下改用哪一把刀，以及怎样记录这一判断以便他人复核。一套细则要在不同学科、不同学生、不同初判上都选对刀，靠的不是一张写满规则的表，而是规则背后的判断体系，以及它在课堂里被检验、被修正的历程。比方说，菜谱可以抄，但一个厨师知道这道菜在南方的潮湿天气里要少放水、在北方的硬水里要多焖一会儿，这些分寸是一次次下锅磨出来的。只抄了追问壳而没有判别细则的仿制品，最常见的结果就是退化为套路：每一轮都问“你有没有考虑过其他原因”，几轮之后学生就学会了敷衍。

第二样是谱系母本精要库。追问之所以问得到点子上，是因为问的人知道这一类反应对条件敏感，这一类故障常坏在没人负责的环节，这一类病房的风险常藏在交接处。这些判断是前沿专著和课题组多年积累的蒸馏，它们不在壳里，也不在通用模型的权重里等着被问出来。第三样是越积越厚的问对档案。沈澄所在课题组那十年的反应条件档案，加上历届学生围绕它的问对记录，是任何人都无法从外部复制的：它不在公开文献里，也难说在任何模型的训练数据里。档案越厚，工序层的

诊断也就越贴近本学科、本课题组的真实毛病。这一点第26章还要从大学的角度展开：大厂造发动机，大学造驾驶术，壳可以开放，判别细则与档案要守住。

这个收窄的主张仍然会被追问：判别细则本身难道不能写下来、发出去？可以。写下来的细则可以被抄走，但细则的使用靠的是对具体现场的认识，同一张细则交给不熟悉本学科的人，查出来的往往是形式上合乎细则、内容上跑偏的诊断，这也是第11章要把评审者之间的一致率逐项登记、其中包括“评分人按细则标注坏法的一致率”的原因：细则是否真的难抄、是否真的有增量，要靠一致率说话，不靠本书的断言。

这一节还可以用来回答一个反方向的担心：底座模型更新换代极快，今天装进ChatSDE的底座，明年就可能被更强的模型取代，工序层会不会随之作废？按本章的区分，答案是否定的。底座可换，工序跟走：诊断方法、宪法条款、问对格式、档案接口，都不依附于某一个底座。换了更强的发动机，驾驶术照样适用，而且可能开得更稳。需要重新做的，是在新底座上检查工序是否仍被稳定执行，宪法是否仍被遵守，这一检查本身就是下一章对照实验要做的事情。

还要防一种误读：说“工序难抄”，并不等于说只有ChatSDE这一种工序层才算数。其他教育工作者完全可以基于自己的学科传统与教学方法，为通用模型设计别的工序层；第8章说到的Khanmigo（2023）、CS50.ai（Liu等 2024）和2025年的三家学习模式，在设计上都有意避免直接给学生答案，也都是工序层的思路。本章的主张不是“唯有SDE工序才能让模型变好”，而是“让模型在课堂里起作用的那一部分，落在工序层而不在参数里”；SDE工序是本书采用并加以说明的一种，它是否比其他工序更好，同样需要检验，不能凭本书的论述宣布。

最后说一句工序层与学科的关系。同一套诊断方法，落到不同学科，会长出不同的问法，但刀的种类不变。在法学研讨课里，韩正面对本院物业纠纷时，诊断可能发现他的初判缺了一项随时间周期变化的量，于是追问落到“立案后第几周结束、那一周发生了什么”；在历史课里，苏文面对族谱的断代，追问落到“县志同年记了什么”。这些问题看起来各属其科，背后都是同一个动作：找出判断里缺的那一项，指出它可能藏在哪一份现场材料里。工序层因此既能跨学科使用，又不会把各学科的问题变成一样的模板；跨学科的部分是诊断，学科的部分是现场。

◆ 本章怎样算错

本章的核心主张是：批判性激发与工序选择不在模型参数里，而在包住模型的工序层；底座可换，工序跟走；其中壳易抄，判别细则与档案难抄。它在下面三种读数面前可以被推翻。

第一，在同一底座上设三个条件：默认配置、开启学习模式、装有工序层，每个条件至少30份不同的学生初判。由两名评分人按第7章的判别细则，独立标注每份初判的坏法，并判断第一回合的追问是否指向了这个坏法。若装有工序层的条件与开启学习模式的条件相比，追问与标注坏法吻合的比例之差不超过10个百分点，那么“诊断与选刀”这一层没有增量，本章的中心主张要撤回，改写为“差别只住在追问壳上”。

第二，把同一套工序层换到另一个底座上，用同样的30份初判重跑。若第一回合的追问里出现核心判断的关键变量词，或没有先接住就批判，这类宪法违反超过3条，而且在改写工序层文字、再跑两轮之后仍不能降到3条以内，那么“底座可换，工序跟走”要改写为“工序层依赖特定底座”，并说明依赖到什么程度。

第三，设一个只抄了追问壳、没有判别细则与档案的仿制品，与完整配置并列，在连续两届、每届不少于30名学生的课堂里使用。若两者在“学生在问对中自己改写初判的比例”上相差不到10个百分点，那么“判别细则与档案难抄”的论述就是错的，它们的分量被高估了。

这三条失败条件都指向同一个需要：本章的主张必须被拿去检验，而不能停留在比喻里。“模型是发动机，工序是驾驶术”是一个好记的说法，但好记不等于成立；如果它从未被检验，它就恰好成了本书一直在批判的那种东西，一个推得很通、却从未被碰过的前提。那么，这种检验应当怎样设计？用什么指标，换谁来评，什么结果出现就该认错？下一章转向对照实验。

> **带 走 的 话**
>
> **模型是发动机，工序是驾驶术：追问的壳谁都抄得走，抄不走的是按判断坏在哪里选刀的细则，和越积越厚的问对档案。**

◆

11

第11章

对照实验：差别怎样被检验，怎样算错

◆ 教学委员会上的一个问题

法学院教学委员会讨论一项申请：《司法实践研讨》课的任课老师在韩正这一届学生中试用了ChatSDE，让学生围绕本地基层法院公开裁判文书与旁听记录做九问研究，现在希望把它正式列入下学期的课程安排。申请书写得很完整，附上了前几章所讲的道理：通用大模型在默认配置下批判之后会把答案一并交出，ChatSDE批判之后停住，不替学生说破；差别不在底座，而在工序层；模型是发动机，工序是驾驶术。委员们听完，一位年长的刑法老师问了一句：“这些我都听懂了。我只想知道，你怎么证明它真的不一样？”

任课老师一时答不上来。他可以举出课堂里的例子，比如韩正手里那一问“本院近三年物业纠纷里，调解成功的案子大多在立案后第几周结束？那一周发生了什么？”，怎样把学生引向一份只有本院裁判文书和旁听记录才能回答的材料；他也可以展示几段问对记录。但刑法老师接着说：“例子是你挑的，记录是你们的系统生成的，评价也是你们自己做的。如果我拿同一个模型，在提示里写一句‘请批判我的观点’，是不是也能得到差不多的东西？”这个追问切中了要害。前几章建立起来的区分，如果只靠挑出来的例子撑着，它就和学生那句“因为年纪大”一样，是一个推得很通、却从未被碰过的前提。

本章要做的，是把“ChatSDE与通用大模型在知识发生上不同”这一产品主张，改写成一组可以被检验、也可以被推翻的命题，并给出检验它的实验设计。与初稿相比，本章现在有了一样东西：书末附录二报告的一次小样本试跑。它不是本章设计的正式实验，只是一次预试，却恰好碰到了刑法老师的追问——一句批判提示能不能得到差不多的东西。预试的回答是不能，至少在代答这一项上不能；同时，预试也推翻了初稿的一个设想，使本章的主要指标必须改换。所以本章的写法

是：先交代先验与预试，再写正式设计、指标、评分、学科材料，最后给出带阈值的失败条件。除附录二之外，本章不报告任何结果；凡写“预期”的地方，都只是待检验的假设。

本书前十章一直在做一件事：教学生把自己的判断翻过来看，问它可能坏在哪一处。如果这套方法只用来检查学生，却不用来检查本书自己的主张，那就是一种双重标准。所以本章首先要把这一方法用到自己身上：“ChatSDE与通用大模型不同”这个说法，可能坏在哪里？按第10章的诊断工具来查，它至少有三处可能的坏法。第一处是缺式：差别也许真实存在，但来源不是工序层，而是别的东西，比如提示词写得更长、对话被引导得更久，或者使用ChatSDE的学生本来就更用心。“工序层—差别”这组关系里，可能漏掉了这些混杂的项。第二处是偏路：本书从几个精心构造的课堂案例出发，直接走到了一个普遍性的结论，中间跳过了在不加挑选的学生观点上做比较这一站。第三处是断拍：如果从来没有人设计过一个能让主张失败的检验，那么即使主张错了，也不会有任何矛盾被摆出来，循环在“逼出改变”这一拍就断了。这三处坏法，恰好分别对应本章实验设计的三项要求：控制混杂、不挑样本、事先写明失败条件。

用一个日常的比方。一家药店说自己的新配方止咳效果更好，顾客要问的不是配方里有什么道理，而是三件事：和谁比，是和老配方比还是和不吃药比；怎样比，是同样的病人、同样的剂量吗；什么结果算它说错了，是不是无论结果如何药店都能说“有效”。产品主张要经得起这三问，才配叫可检验的主张。本章对ChatSDE的检验，就是逐一回答这三问。

检验的结果，对任课老师和学生也有直接的用处，不只是为了说服教学委员会。如果检验支持工序层的主张，老师就知道，课堂里那些“第一回合只翻一处前提、不交答案”的时刻，不是碰巧遇上的好运气，而是可以被稳定复现、可以被交给其他老师使用的安排；如果检验不支持，老师就知道，过去看到的好例子可能来自别处，比如学生自己的投入、老师的现场指点，那么课程设计的重心也应当随之调整。无论哪一种结果，课程都因此多了一层可以依靠的东西，这与学生在九问里做的事是同一回事：判断被现场碰过之后，才知道该把力气用在哪里。

> **这一章的问题**
>
> 怎样证明ChatSDE和一句“请批判我的观点”真的不一样，什么结果出现就该认错?

◆ 先验：Bastani等（2025）与附录二的预试

任何一个检验设计，都应当先问：在动手之前，已有的证据把我们推向了哪里？这里有两个先验，一个来自外部，一个来自本书自己。外部的先验是Bastani等（2025）发表在《美国国家科学院院刊》（PNAS）上的《Generative AI without guardrails can harm learning》。他们在高中数学课堂里做了现场实验：使用不加护栏的GPT辅助练习的学生，练习时表现更好，可是撤掉AI之后的考试，成绩反而更差；加了教学护栏的GPT Tutor，这种损害基本被消除了。这项研究给本章的设计带来三点。其一，它把“不加护栏”的一组摆在了最前面，对应本章的裸用组，也就是默认配置组。其二，也是更要紧的一点，它的关键读数不在练习当下，而在撤掉AI之后：只量系统在场时学生写得怎样，会把“被替了一步”误读成“学得好”。所以本章的多轮层必须有一项撤机之后的独立判断任务。其三，它的情境是高中数学，练习的是终点已知的任务，属于本书所说的练手，不是连老师也写不出终点的谱系性问题，因此它只提供方向上的先验，不是对ChatSDE的直接验证。

本书自己的先验，就是附录二的预试。先写它做了什么。作答的是同一个中等型号的通用大模型，评判员是型号不同的两个模型，盲评。试跑二用30条构造的学生初判（医学、化学、法学各10条），分三组作答第一回合：A组默认配置，B组默认配置加一句“回答时请对用户的观点进行批判性思考”，C组模拟工序层，即一段按第7章判别次序和“不替学生说出结论、不给清单”的约束写成的简短指令。试跑三把消息改成“请帮我把这个看法展开成一段有条理的分析”，比较D组默认配置与E组模拟工序层。两名评判员按四项编码（试跑三另加第五项），其中Q1是否翻转了学生初判的前提，Q2是否替学生给出了替代性结论或两个以上的原因或办法清单，即代答。

编码项	试跑二 A默认	试跑二 B 加批判提示	试跑二 C 模拟工序层	试跑三 D默认	试跑三 E 模拟工序层
Q1 翻转前提（评判员一／二，分母30）	27／28	30／30	30／30	30／30	30／30
Q2 代答（评判员一／二，分母30）	30／30	30／30	0／0	30／30	0／0

这张表给出两个直接的读法。第一，翻转前提这一项，默认配置几乎已经到顶：A组27和28，D组30和30，即使学生明确请求“展开”，默认配置也几乎都先质疑了前提。初稿设想的“默认只展开、不质疑”因而不成立，也就是说，把首回合质疑率当作区分三组的主指标，在所测的模型上是行不通的，三组都挤在天花板附近，区分不出任何东西。第二，真正分开三组的是代答：默认配置与加批判提示的配置在90份回复里全部代答，模拟工序层在60份里一份也没有，两名评判员完全一致（κ=1.00）。B组还回答了刑法老师的追问：多了一句批判提示，翻转前提的比例略升，代答仍是30／30。一句提示改变了挑刺的力度，没有改变批判之后把答案交出来的习惯。

据此，本章对初稿的设计做了三处修正。第一，主要因变量由“首回合质疑率”改为“代答率”与“核心判断首发率”：前者看单回合里系统有没有替学生把话说完，后者看多轮问对里学生最终的核心判断由谁先说出。首回合质疑率降为操纵检验，只用来确认各组都在批判，不再承担区分的任务。第二，保留提示批判组，因为预试已经显示，这一组在代答上与默认配置没有区别，而在初稿里这一组的价值只是一个假设。第三，补上初稿里没有的两样东西：撤掉系统之后的独立判断任务，以及作为对照的学习模式组，这两样在下面的设计里写明。

◆ 附录二的局限，与正式实验多出的设计

预试有它的限度，附录二已逐条写明，这里只摘与设计直接相关的几项，并说明正式实验怎样补上。预试只测了一个模型家族，作答与评判都出自同一家族的不同型号；正式实验要在不少于三个不同厂商的底座上各自跑完全部分组，每个底座单独报告，不合并成一个“通用大模型”的平均数。预试的30条初判由本书编写，形式整齐，单轮；正式实验用真实学生在第零轮写下的初判，经去标识化，并且进入多轮问对，而不只取第一回合。预试没有测量学习结果，“代答”是否真的损害

学生的判断发生仍是一个假设；正式实验在多轮层加入撤机之后的独立判断任务和核心判断首发率，把代答与学习结果接起来。预试的工序层只是一段最小指令；正式实验用站上完整的ChatSDE，并保留“模拟工序层”作为一个单独的组，用来看多出的部件到底带来了什么。

预试的盲评并不完全：工序层组的回复明显更短、更像提问，评判员可能从篇幅和语气猜出组别。正式实验要另做两件事：评审者由人担任，事先接受判别说明的训练，并抽一部分材料去掉篇幅与语气线索，例如只给评审者看回复中与前提、答案有关的句子，核对结果是否改变；每条输入在新的会话里单独作答，避免同一次会话里的条目相互影响。预试所用的模型与生成本书文字的模型同属一个家族，评判模型与作答模型的型号虽然不同，也只能算先验，不是独立证据；正式实验要由学派之外的人独立实施或至少复核（见后文“换手评分与独立证据”）。最后，预试只有每组30条的规模，只够看方向；正式实验在登记时要写明检出多大的差距、样本量能否支持。

学习模式也要补进来。附录二没有测试各家的学习模式，而本书在第8章已经承认，它们是“不说破”方向的产品化，是工序层最近的邻居。如果正式实验只比裸用组与工序层组，学习模式这一最强的竞争解释就悬在半空。所以下面的分组里单列一个学习模式组，用各家当时已有的学习模式，记录版本与开启方式。

◆ 同底座对照：只让工序这一项不同

对照实验的第一条原则，是只让要检验的那一项不同。本章要检验的是工序层，所以对照的各组必须用同一个基础模型、同一个版本、同样的生成参数，其余条件尽量相同。这样设计的理由，第10章已经给出：如果两方用的底座不同，那么即使看到了差别，也分不清是工序层的功劳还是底座的功劳。这好比比较两位厨师的手艺，要让他们用同一口锅、同一批菜、同一个灶，否则吃出的差别可能只是锅的差别。凡写到通用大模型，一律记录所用的版本标识与调用时间。

在这一原则下，本书建议设四组：裸用组，即默认配置、未开启学习模式；提示批判组，即同一底座加一句通用的批判提示，如“请对我的观点提出批判性意见”，但不接入工序选择与宪法条款；学习模式组，即同一厂商提供学习模式的，

开启学习模式作答；工序层组，即装进ChatSDE工序层。四组并列，要回答三个问题：工序层是否不同于裸用，工序层是否不同于一句批判提示，工序层是否不同于已经商品化的学习模式。前两个问题预试已有方向，第三个问题没有先验，是正式实验里最有信息量的一问。如果某个底座没有对应的学习模式，该底座就只有三组，在报告里写明。

材料的设计取医学、化学、法学三个学科，每个学科各收集不少于10条学生观点作为输入。这些观点应当来自真实的课堂初判，例如第零轮里学生写下的那一句，经去标识化处理；它们不应由研究者事先挑选或改写，以免把“容易被批判”的观点集中放进来。三个学科分别对应本书所说的客体、互动与主体的不同侧重：化学的初判多与条件和操作有关，法学的初判多与程序和当事人的互动有关，医学则横跨三者。观点的收集办法、去标识化的规则、医学材料所需的科室与伦理委员会许可，都要在实验开始之前写定。

实验分两层做。第一层是单回合层：把每一条学生观点分别输入各组系统，只取第一回合的回应来比较。这一层成本低、控制严，主要检验“默认值”是否真的不同，也就是代答率。第二层是多轮层：让学生真实地与四组系统中的一组进行完整的问对，直到形成子本或子本的核心判断，学期结束后再做撤机之后的独立判断任务。这一层更贴近课堂，但混杂也更多，比如学生的投入程度、现场的难易、问对的轮次，都需要记录并在分析时加以考虑。两层各有所长，单回合层回答“默认行为有没有差别”，多轮层回答“这种差别能不能传到学生的知识发生上去”。

底座本身也要被固定下来。今天的模型服务时常在不声明的情况下更新，同一个名字背后可能已经换了版本；如果实验跨越数周，前一批材料与后一批材料面对的可能不是同一个底座。所以单回合层的全部输入最好在同一时段内一次完成，多轮层则要在报告中写明期间是否发生过版本变化。这一条看似技术细节，却直接关系到“只差工序”能否成立：底座在实验中途悄悄变了，对照也就悄悄失效了。附录二的记录里写明了作答模型与评判模型的型号，都是当日线上版本，正式实验照此办理。

分组的方式也要事先写定。在多轮层里，学生被分到哪一组，不能由学生或任课老师自行选择，否则用心的学生更可能选新系统，差别就混进了学生本身的差

别。较稳妥的办法是在同一门课里随机分配，或者在不同学期轮换；如果实际条件只允许整班分组，就要在分析时把班级差异单列出来，并在报告中如实说明这一局限。学生知道自己用的是哪一组系统，这一点难以避免，但评分者不应知道，这是下面要说的换手评分的一部分。

◆ 读数：两项主指标，两项结果指标，三项辅助

有了对照，还要说清比较什么。预试之后，指标的主次必须重排。本章采用的读数分三层：两项主指标看系统的行为与首发，两项结果指标看学生最后得到了什么，三项辅助指标用来检查主指标有没有被钻空子。任何一项单独拿出来都不足以支撑整个主张。

第一项主指标是代答率。它问的是：面对一条学生观点，系统的第一回合回应，有没有替学生给出了替代性的结论，或者两个以上的原因或办法清单。这一项沿用附录二的编码定义，因为预试里它是唯一把三组分开的项，两名评判员完全一致。以方舟的初判为例，“按计算、内存、网络、数据库列出四类瓶颈并给出扩容办法”记为代答，“去年服务器加了一倍，为什么还崩”不记。接住学生的话、复述学生的判断，不算代答；一个指回现场的问题，也不算。判别说明需要事先写清几种边界情形，例如回复只列了一个可能的方向算不算代答，评审者各按各的理解计分会让这一项失效。这一项在单回合层测量，各组各自计数，比较的是比例，而不是个别案例。

第二项主指标是核心判断首发率。这一项转到多轮层，问的是：学生最终形成的核心判断，在问对记录里是谁先说出来的，是学生，还是系统。它对应第10章宪法里的“逼近判断，不替学生说破”，也对应第15章评估项目中“问对深度”的一条要点。测量的办法是由评审者先从子本中认定核心判断，再回到问对记录里找它第一次以可辨认的形式出现的位置。首发是相对ChatSDE而言的，口径需要事先写死：现场里他人说的话，例如师兄的一句“免得受潮”，以及档案里的记录、老师出的九问，都是学生取得的现场材料，不算系统替学生说破；ChatSDE的追问不得出现核心判断的关键变量词，如“湿度”“空气”“交接”。以沈澄为例，如果“湿度”一词最先出现在系统的回应里，那么即使她后来的子本写得再好，这一条也要记为系统首发；如果它最先出现在师兄的话里，被沈澄在问对中拿来用，则

记为学生首发。代答率看单回合里的“说破”，核心判断首发率看多轮里说破有没有真的发生，两者是同一件事在两个尺度上的读数。

结果指标之一是撤机之后的判断分，这是Bastani等（2025）给本章的启发。学期结束后，每个学生拿到一个与本学期课题同类、但没有做过的新现象，在没有任何系统的条件下写下初判，并写出这个初判的前提可能坏在哪里、他会去现场查什么。由不知组别的评审者按事先写定的量表评分，看他有没有认出自己初判里的前提、能不能提出一件指回现场的可查事实。这一项回答的是：系统在场时的差别，撤掉之后还在不在。它与预试没有关系，是正式实验多出的一格，也是“代答是否损害判断发生”这一假设的直接检验。结果指标之二是子本I项分数。I是创新智商五维中的不可还原性一维，是题闸在评估端的对应，问的是子本里的判断在多大程度上是拿掉学生现场就答不出的东西（见第16章）。它是离“知识发生”最近的一项，也是最难测、最容易被评分者偏好左右的一项，所以必须按第16章的规定换手评分，评分者写利益冲突声明，并且不知道子本出自哪一组。这一项还要配合完成度一起看：一份I项分数高、却连引注都不齐、化学部分不可复做的子本，不能简单地记作工序层的成功。

三项辅助指标，各有出处。其一，首回合质疑率，降为操纵检验：它在预试里已接近天花板，用来确认各组都碰到了前提，若某一组的质疑率明显偏低，该组的代答率就要单独解读。其二，套路化检测：质疑是否随观点不同而不同。预试在两轮里给出了差别很大的可靠性，展开请求一轮两名评判员一致认为默认配置的质疑约有六到七成可以通用（$\kappa=0.81$），首回合一轮两人分歧很大（$\kappa=0.00$），所以这一项的判别说明需要重写并训练评审者，在不同轮次里反复核对一致率，达标之前不作结论。具体办法可以是让评审者在不知来源的情况下判断一条质疑能否原封不动地移植到另一条观点上，能移植的就是套路。其三，指回现场：回复是否把学生指回他自己现场里一件可查的事实。预试里这一项两轮的一致性差别很大（$\kappa=0.85$与0.35），说明判别说明必须写清“可查的事实”与“泛泛的观察建议”的界限。顺前提展开（预试的Q5）也在辅助之列，用来观察另一种默认是否出现，预试里它的一致性只有$\kappa=0.45$，同样需要先修订说明。此外，第10章的第一条失败条件要用到评审者按第7章判别细则标注坏法的一致率，它不另列为第八项读数，随指回现场一并登记，同样先登记、后下结论。

七项读数各有层次、测量位置与评判方式，混在一起谈很容易让人以为它们是同一件事的几种说法。下面把它们并排列出，便于对照。

读数	层次	在哪一层实验中测量	由谁判别	主要用来区分
代答率（主）	系统默认行为	单回合层	换手人类评审，依判别说明	工序层与裸用、与提示批判
核心判断首发率（主）	学生知识发生	多轮层	换手人类评审，回查问对记录	工序层与学习模式、宪法是否被执行
撤机之后的判断分（结果）	学生知识发生	多轮层，学期末	不知组别的评审者	代答是否损害判断发生
子本I项分数（结果）	学生知识发生	多轮层	换手评分并写利益冲突声明	知识发生是否传到产物
首回合质疑率（操纵检验）	系统默认行为	单回合层	换手人类评审	各组是否都在批判
套路化检测（辅助）	系统默认行为	单回合层	不知来源的评审者	工序层与提示批判
指回现场（辅助）	系统默认行为	单回合层	换手人类评审	批判是否落到学生自己的现场

表中没有一栏写着预期数值，这是有意的：数值只能来自实施之后的记录，附录二的读数是先验，不是这些栏目的结果。读者也可以从表中看出分工：前两项盯住系统与首发，中间两项盯住学生，只有两端都出现了可区分的差别，才能说工序层的影响从系统传到了课堂。如果只在代答率上看到差别，那只说明系统的说话方式变了；如果只在后两项上看到差别，就要警惕差别是否来自学生分组或老师指导，而不是工序层。

除了这七项，还有一些数字看上去容易得到，却不应当作为主指标。问对的轮次、子本的页数、学生的满意度，都与本书关心的知识发生没有直接关系；第17章已经说过，十轮打转不如三轮推进。把这些数字列为主指标，等于鼓励系统把对话拖长、把文字写多，或者把学生哄得高兴，这恰好与工序层的宪法背道而驰。它们可以作为辅助记录，但不能用来判定主张是否成立。

◆ 换手评分与独立证据

一个容易被忽略的问题是：谁来评？如果评分交给另一个大模型去做，尤其是交给与被测系统同一底座的模型，那么评分本身就可能带着与被测系统相同的偏好。比如某一底座天然偏爱结构整齐、术语密集的文字，那么用它来评，它会把同底座系统写出的东西评得更高，而这与工序层无关。这里必须写明一条原则：同底座的多个智能体不能互作独立证据。一个系统写、另一个同底座系统评，再一个同底座系统复核，看上去有三道关，实际上只是同一种眼光看了三遍。附录二用型号不同的评判模型，这样做只能降低同源偏好，不能消除它，所以那里把自己明确定位为先验。

这条原则在本书的设计里有几个具体的落实。第一，主指标的判定，尤其是代答率中“是否替学生说破”、核心判断首发率中“谁先说破”以及子本I项分数，须由人类评审者完成，评审者不参与该学生的问对，不知道材料出自哪一组，并写利益冲突声明。第二，如果为了降低成本使用模型做初筛，初筛模型应与被测系统不同底座，且初筛结果只作参考，最终判定仍须换手由人完成。第三，ChatSDE的开发者与任课老师不担任主评审，他们对结果的期望最强，最容易在边界案例上不自觉地偏向一方。

评审者之间也要能互相检查。同一份材料至少由两位评审者独立评判，评判之后比较一致程度；如果两位评审者在“是否替学生说破”这样的问题上经常意见相左，那就说明判别说明写得不够清楚，需要先修订说明、再重新评，而不是取两人的平均分了事。预试里同一批回复，代答一项的κ为1.00，套路化一项在一轮里为0.00、另一轮里为0.81，指回现场一项为0.85与0.35，说明不同的指标可靠性差别很大，不能因为其中一项可靠就默认其他项也可靠。这好比两位裁判对同一个球是否出界经常判得不一样，那就先要把边线画清楚，而不是让两人投票。

还有一条容易被忽视的独立性：检验者与被检验者的独立。本书的作者与ChatSDE的设计者属于同一个学派，由同一学派的人设计、实施并解读这一实验，即便每一步都按规程做，结果也难以被学派之外的人完全信服。所以本书建议，实验方案在实施之前公开登记，写明分组、指标、判别说明与失败条件；最好由学派之外的教学研究者独立实施，或者至少邀请他们复核材料与评分。有先例可循：

Kestin等（2025）在哈佛物理课上做的随机对照，把精心设计的AI导师与课堂主动学习直接放在一起比较，这类设计之所以有分量，正在于分组由随机决定、对照组是真实的强竞争者。一项检验的分量，不仅在于它的设计是否严密，还在于它是否给了怀疑者亲手检查的机会。

◆ 三个学科的设计：医学、化学、法学各不少于10条

把上面的原则落到具体学科，每个学科的观点各有要注意的地方。之所以要分学科设计，是因为三个学科的现场性质不同，材料的伦理边界不同，学生初判常见的坏法也不同。下面只描述材料怎样收、怎样处理、各组怎样比较，不涉及任何结果；文中举出的观点都是本书已有的构造案例，用来说明材料的样子，而不是实验样本本身。

医学的观点，取自老年病科轮转学生第零轮写下的初判，类型与许宁的“夜间跌倒是因为患者年纪大”相近，即对病房里一个现象的初步归因。这些观点在收集时一律去标识化，凡涉及具体患者、具体床位、具体医护人员的内容均须删除或替换，材料的使用须经科室与伦理委员会许可。医学部分还有一条特殊的约束：无论哪一组系统的回应，都不得被学生用来修改医嘱或作临床决策，实验只观察系统对一条观点的回应方式，以及学生在问对中形成的判断是否以本病房假说的身份交科室验证。如果某一组系统在回应中给出了具体的临床处置建议，这一点本身应当被记录，作为该组行为特征的一部分。

化学的观点，取自课程型研究中学生对失败批次的初判，类型与沈澄的“产率低是因为试剂纯度不够”相近。化学部分的特殊约束是安全：材料中不得包含危险操作的细节，任何一组系统的回应若建议改动危险步骤，学生都不得照做，实验只记录这一回应。多轮层的化学部分，还要把可复做性纳入完成度的考察：一条判断如果在子本里写得很新，却无法交给下一位同学按记录复做，它的I项分数再高，也不能单独作为工序层成功的证据。化学题在预试里有一个值得留意的迹象：顺前提展开的几条回复集中在化学题上，所以化学部分的辅助读数“顺前提展开”要单独列出，看它在其他底座上是否重现。

法学的观点，取自《司法实践研讨》课学生围绕本地基层法院公开裁判文书与旁听记录写下的初判。这些材料以已公开的文书为主，但旁听记录与学生的笔记可能涉及当事人信息，同样须去标识化处理，并遵守司法公开的有关规则。法学观点的一个特点是，它们常常已经带着某种规范性立场，比如“调解难是因为当事人不讲理”，系统对这类观点的回应，既可能质疑事实前提，也可能质疑规范前提；判别说明里需要写清两种质疑是否都算“碰到前提”，也需要写清替学生写出一份规范论证算不算代答，以免评审者各按各的理解计分。

三个学科共同的还有一条对学生的伦理要求。参与多轮层的学生应当事先知情并同意，知道自己所在的课堂正在做一项比较，知道问对记录将被换手评审者阅读；被分到哪一组，不应当影响他在这门课上的成绩，换言之，课程成绩按第五编的评估方案照常评定，实验所用的读数只服务于检验工序层，不进入学生的成绩单。这样安排，一是为了公平，二是为了防止学生为了成绩去迎合某一组系统，把实验本身变成另一种回音。

三个学科合起来是不少于30条观点，这个规模只够作为一次初步检验，不足以支撑大范围的推论。本书之所以仍然把它写进设计，是因为一次规模不大、但事先登记、换手评分、写明失败条件的检验，比一百个精心挑选的成功案例更有说服力。如果初步检验显示出值得追究的差别，再扩大学科与样本；如果没有，那就按下一节写好的失败条件办。

◆ 预期与失败条件：什么结果出现，主张即撤回

先把预期写出来，再把失败条件写出来，两者都要在实验之前定下，不能看了结果再改。预期有两类来源。一类是预试已经给出方向的：在单回合层，本书预期裸用组与提示批判组的代答率都很高，工序层组很低；预期提示批判组与裸用组在首回合质疑率上略有差距，在代答率上差距不大。另一类没有先验，是真正待检验的：在多轮层，预期工序层组的核心判断首发率更多落在学生一侧，撤机之后的判断分和子本I项分数在换手评审下不低于另三组；预期学习模式组介于裸用组与工序层组之间。这些预期写在这里，是为了让读者可以拿未来的结果来对照，而不是为了说明它们已经成立。

“看不出可区分的差别”这句话本身也需要事先说清，否则它就成了一个可以事后随意解释的弹性词。在实验登记时，研究者应当写明采用什么统计方法比较各组，事先确定多大的差距才算有意义，以及样本量在这一标准下能否给出可靠的判断；如果规模不足以检出事先认定有意义的差距，就应当如实写明“本次检验力不足”，而不是把“没看到差别”说成“证明了没有差别”，也不是把“看到一点差别”说成“证明了有差别”。下面的阈值是本书给出的最低门槛，是设计值，不是已有数据，实施者可以收紧，不可以放宽。

第一条，关于代答率，这是本章的总条件。在不少于三个厂商的底座上，每个底座取不少于30条真实初判，两名换手评审独立编码，代答一项一致率不低于90%。若在某个底座上，工序层组与裸用组的代答率相差不足20个百分点，则该底座上“工序层改变了代答”不成立；若这种情形出现在多数底座上，本书第8章关于“两种默认”的主张、第10章关于“差别住在工序层”的主张即撤回。预试里两者的差距是从全部代答到一份也没有（每组各30份，30对0），所以20个百分点是一个宽松的门槛，没有越过它，说明预试的差别在更真实的输入上基本消失。若裸用组在多数底座上的代答率低于50%，则第8章“默认替学生说破”的主论断撤回，改为只就附录二所测的家族陈述。

第二条，关于提示。若工序层组与提示批判组的代答率相差不足20个百分点，那么工序层的分量被高估了，本书需要改写为“一句批判提示即可达到同等的不说破”，并相应修改第10章与第26章关于工序与护城河的论述。反过来，若提示批判组与裸用组的代答率相差超过20个百分点，则预试里“一句批判提示改变不了代答”这一条也要撤回。

第三条，关于多轮层。连续两届、每届不少于30名学生，若工序层组与学习模式组的核心判断首发率，其差距小于两名评审之间的分歧幅度，或者小于15个百分点，取较大者，则题闸、第零轮、选刀和回写没有带来可检验的区分，本书关于ChatSDE差别的主张撤回，只保留“包装层决定行为”这一条。若工序层组的核心判断首发率落在学生一侧的比例不高于裸用组，则宪法没有被执行，需要回到工序层本身去修。

第四条，关于代答是否有害，这条直接承接Bastani等（2025）的先验。若撤机之后的判断分上，裸用组与工序层组相差在评审误差之内，则“代答损害判断发生”没有得到支持，第8章只保留“默认配置的行为有此特征”这一描述性判断，不再把它当作教育上的问题。反过来，若裸用组的撤机判断分显著低于工序层组，同时代答率也显著更高，这一组读数合起来，才是本书所说的“默认值就是教育”的第一份直接证据。

还有两种情形要预先说明。一种是批判变成了模板：若工序层组的代答率很低，但在套路化检测上同样呈现高度套路化，或者首发率明显偏向系统一侧，那就说明工序层制造的是另一种回音，批判变成了模板，判断被换了一种方式说破，第10章的宪法没有被执行，需要回到工序层本身去修。另一种是结果混杂，有的指标支持，有的不支持，有的学科或底座支持，有的不支持。这时不能只挑支持的那几项来报告，而要把全部结果按事先登记的方案如实列出，并对不支持的部分作出说明。比如只在化学部分看到差别、在法学部分看不到，那么本书的主张就要收窄为“在以客体条件为主的课堂里成立”，而不能再以三个学科的名义说话。主张随证据收窄，本身就是一次回写，它比一个从不改变的主张更接近知识发生。

最后要说明这些失败条件对课堂的意义。如果主张被撤回，九问模式的其余部分并不因此全部作废：母本、九问、改题、现场、子本、评估这些环节，有许多并不依赖ChatSDE这一个产品，它们可以用别的工序层，甚至不用模型来实现。被撤回的，是“ChatSDE这一工序层在知识发生上带来了可检验的差别”这一条具体主张。一本书能够写清楚自己的哪一部分可以被推翻、被推翻之后哪些还站着，它的论证才算有了边界。

◆ 本章怎样算错

本章的主张不是关于ChatSDE的，而是关于检验方法的：“ChatSDE与大模型不同”须由同底座对照、代答率与首发率两项主指标、撤机之后的判断、换手评分与事先写定的失败条件来检验，否则它只是一个未被批判的前提。这一方法主张自己也有三处会出错，下面都写成可以核对的条款。

其一，指标的可靠性。预试里不同的指标一致性差别很大，有的完全一致，有的低到几乎无法使用。若在正式实验的训练之后，两名评审在代答率或核心判断首发率的编码上，一致率仍低于90%，经过两轮判别说明修订后依旧如此，那么这两项主指标就不能承担检验的任务，本章的指标设计需要重做，而不能带着不可靠的指标去下结论。其二，同底座对照的可行性。若工序层的实现必然改变了生成参数、上下文长度或检索内容，且这些差别无法分离，那么本章“只差工序”的设计前提就不成立，需要改为更细的拆解实验，逐一检验工序层的各个部件；在拆解完成之前，不得把整组差别归给工序层。其三，独立性。若实验的实施、评分与解读始终只由学派内部完成，没有任何外部复核，那么无论结果如何，本章所要求的独立性都没有达到，结果只能作为内部参考，不能作为本书主张的证据。此外，若只在一个模型家族上跑完了正式实验，结论只能写成对该家族成立，而不能写成对通用大模型成立，这一点与附录二的限定一脉相承。

第三编到这里告一段落。四章所讲的，是同一个问题的四个侧面：两种默认值怎样不同，差别落在配置的哪一环，那一环怎样由工序层组织起来，以及这一切怎样被检验、怎样认错。其中第一个侧面，经过附录二的预试，已经从“批判与不批判”改写成了“说破与不说破”。剩下的问题回到学生身上：假定问对确实把学生带到了一个母本所无的判断跟前，这个判断接下来怎样被写成一本属于学生自己的书？把九份问对成果写成子本，是把它们逐章翻译一遍，还是要重写？下一编从子本的打造讲起。

> **带走的话**
>
> **一个主张只有事先写明了怎样算错，又交给不相干的人去评，才配叫可检验；附录二的预试只给了方向，没有给出结论。**

◆



第　四　编

E：打造子本与落壤（铺土壤）

编　序

苏文把子本第一稿交上来，老师读到一半停下，在页边写了一句：这一稿里，哪一个判断是母本里没有的？稿子很顺，章章对得上母本，他只是把母本的话换成了家乡的人名、地名和年份。九问走完，课堂的产出若只是一份换了例子的作业，土壤就没有铺成；子本要怎样才算重写，又怎样才带得出这间教室，是本编的问题。

第12章讲子本是母本加九问研究成果的重写，不是翻译，里面须有母本所无的判断。第13章讲子本进入底盘，成为个人的检索库与技能，毕业后还能继续生长，并说明它是思考的底盘，不是临床决策工具。第14章讲同学碰撞场：只有一对一的线，是高级个别辅导；课堂之为课堂，在于同学互撞。

三章从一个人的稿子，到这份稿子带得走，再到稿子与稿子相撞，土壤由一个人厚到一个班。子本落了壤，后面的评估与回写才有东西可看。

12

第12章

子本打造：重写不是翻译

◆ 一份像极了母本的子本

苏文把子本第一稿交上来的那天，《地方史料研讨》的老师读到一半就停下了。稿子写得很顺，章节整齐，引注规范，每一章都对得上谱系母本里的一章：母本讲族谱的编修体例，他就讲家乡族谱的体例；母本讲方志与谱牒互证的方法，他就拿家乡县志与族谱互证一遍；母本讲口述史料的可靠性，他就引了几段村里老人的话。老师在页边写了一句批语：“这一稿里，哪一个判断是母本里没有的？”苏文回去翻了两天，找不出来。每一个结论都能在母本里找到原型，他只是把母本的话换成了家乡的人名、地名和年份。

这份稿子并不差，换在讲授课堂里，它几乎是一份好作业：读懂了母本，会用母本的方法处理新材料，文字也干净。可它恰好不是子本。苏文在九问与问对里其实已经碰到过一个母本没有的东西——他追问“你家族谱断了的那一代，县志同年记了什么”，在县志里发现，断代那一年正是户籍登记口径改变之年。这个判断在问对记录里清清楚楚，到了子本第一稿里却只剩下一个脚注，夹在“方志与谱牒互证”那一章的末尾，像是母本方法的一次成功应用。新判断被旧章节吞掉了。

这个场景提出了本章要回答的困惑。第二编讲改题与问对，那里发生的判断散落在九份问对记录里，有的成形，有的半成形；第四编要把它们落进土壤，第一步就是打造子本。可是“打造子本”很容易被理解成“把母本按自己的情况改写一遍”，而这样做的结果，往往就是苏文的第一稿：形式上是自己的书，判断上仍是母本的书。本章的核心判断是：子本是母本加九问研究成果的重写，不是逐章翻译；子本里必须有母本所无的判断。下面依次讲打造流程、翻译与重写的判别、一次逐章改写的失败教训、两份子本的骨架，以及子本的学术规范。

之所以把“母本所无的判断”定为子本的硬要求，与全书的母题有关。导论说，九问课堂的产出从“掌握了同一本书的人”变成“建成了自己子本的人”。如果子本可以没有新判断，那么一个把母本读得很熟、又把它译成了自家现场版本的学生，就同时满足了两种产出：他既掌握了同一本书，又“建成了”一本子本。这样一来，课堂典范的转变就只剩下了作业形式的变化。新判断的硬要求，正是把两种产出区分开的那条线；没有它，“子本”只是“读书报告”的一个新名字。母本掌握仍然重要，但它在评估里另有位置，即第①项母本底盘，以复述加反例口试来检验，而不由子本来证明。

> **这一章的问题**
>
> 一本子本和母本的逐章翻译怎样分得开，它凭什么算“重写”？

◆ 与《子本打造法》的关系：沿用什么，改写什么

在讲流程之前，要先交代本章与作者旧作的关系，免得读者把旧话当新话。子本的概念和做法并非本书首创：王德生在《子本打造法》（德麦国际专著第527号）里已经提出，读者以自己的画像和现场材料，从母本得到属于自己的整本子本，其中“子本是重写而不是翻译”这一条也已立定。本书沿用的，是这条流程的骨架：母本、加上九问、加上问对、再重写。

改写之处在于子本的身份。在《子本打造法》的情境里，子本主要是读者给自己读的书，读得对不对，损失由读者自己承担，一份贴近母本、只是换了人名地名的子本，对那位读者完全可能是有用的个人阅读。在课堂里，子本成了学业产物，要被评分，要在同学之间比较，要决定一个学生是否通过。身份一变，旧作无须回答的三个问题就逼了出来：其一，这本子本里哪一句话出自谁，是母本、问对还是学生，需要可核对的归属（见下文第二条规范）；其二，怎样的子本算“重写”，需要别人也能复核的判别（见下文三条检验）；其三，评分时看什么，需要一把分得开翻译与重写的尺子。三条检验和四条规范，就是对这三个问题的回答，它们是本书新增的部分。因此，本章不是对旧作流程的复述，而是在旧作流程之上加了一层“可被他人检验”的约束；旧作里读者自用的子本，仍然可以走旧作的路，不必受这里的限制。

◆ 从九份问对成果到子本：打造流程

本书所说的子本，专指九问子本：它以谱系母本为出发点，以学生九问的研究成果为材料，经过重写而成（概念见第3章、第4章）。打造流程分三步：先把九份问对成果整理成判断清单，再从判断清单长出子本大纲，最后按大纲重写成文。三步的次序不能颠倒，尤其不能跳过第一步直接从母本目录开始。从母本目录开始，得到的就是一份带着家乡口音的母本；从判断清单开始，得到的才可能是一本自己的书。

第一步是整理判断清单。九问走完，学生手里有九份问对记录、一叠现场材料和一份题的版本史。整理时只问一件事：每一问最后落下了什么判断，这个判断是谁先说出来的，靠哪一份现场证据撑住，在问对里经受过哪一次批判。判断要写成完整的句子，而不是关键词；“交接断点”是一个词，“本病房夜间跌倒的风险更多落在交接断点上，而不只是年龄”才是一个判断。九问不一定产出九个判断，有的问走进了死胡同，有的两问合成一个判断，这都正常。清单里要如实标出哪些判断成形、哪些还是猜测、哪些被现场推翻了。

第二步是从判断清单长出大纲。这一步最容易出错，因为大纲有一个现成的模子摆在旁边——母本的目录。子本大纲的骨架应当是判断之间的关系：哪个判断是承重的，哪些判断为它提供证据，哪些判断是它的推论，哪些判断与它相冲突而需要专门处理。母本的核心判断在大纲里仍有位置，但它的位置由新判断决定：新判断从母本的哪一句话出发，又在哪一处与母本分道。这就好比自家装修，图纸要从这一家人怎么过日子画起，而不是拿样板间的图纸把墙挪一挪。样板间可以参考，但起点不能是它。

第三步才是重写。重写的意思是：用自己的判断作主语，把母本的知识与九问的材料一起组织进去，写成一个有完整论证的文本。母本在这里是底盘，不是蓝本——它提供概念、方法与已有结论，让新判断有处可站；但文本推进的动力来自新判断要求被证明、被限定、被检验。打个大白话的比方：翻译是把一首歌从普通话唱成方言，调子、歌词次序都不变；重写是在这首歌的旋律上发现原作没写出的一段，把它写成新的一节，旧旋律在新歌里仍听得见，但整首歌是为那一节而重新编排的。

三步之中，ChatSDE与学生的分工需要说清。判断清单与大纲由学生自己写，ChatSDE的角色是对照者与追问者：它可以把清单里的每一条判断拿去与母本精要库比对，提示“这一条在精要库里已有等价表述”，也可以追问“这一条靠哪一份现场材料撑住”，但它不替学生决定哪一条承重。重写阶段同样如此，学生执笔，ChatSDE按章读稿，指出哪一段又滑回了母本的叙述次序，哪一个新判断被埋进了脚注。这样的分工并非出于对模型能力的怀疑——让大模型直接生成一本文从字顺的子本并不困难——而是因为子本要记录的是学生自己的发生，代笔得来的通顺稿件，恰好抹掉了评估所要看的那条轨迹。就像学游泳，教练可以在岸边喊，可以下水托一把，但不能替你游完这一趟。

◆ 翻译与重写的判别

翻译与重写在外观上可以非常接近：两者都引用母本，都使用学生自己的材料，篇幅、体例、文风都可能相似。要分辨它们，需要几条可以操作的检验。这里给出三条，它们与第4章的题闸同出一个道理：题闸问“拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案”，子本的判别则问“拿掉学生的现场材料，子本的结论是否在母本里原样存在”。前者检验一道题是不是谱系性的，后者检验一本书是不是谱系性的。

第一条是来源检验。把子本里的每一个结论性判断摘出来，逐条到母本和精要库里去找。找得到原句或者等价表述的，标为“母本判断的应用”；找不到的，再看它是不是从九问的某一问、问对的某一轮里长出来的。一份子本可以有大量母本判断的应用，这是底盘；但如果全部结论都能找到原句，它就是翻译。苏文的第一稿正是如此：互证方法是母本的，可靠性讨论是母本的，连“族谱断代往往与战乱、迁徙有关”这样的推测也是母本里的现成话，只有那个户籍口径的发现找不到出处——可它被放在脚注里，没有当成结论。

第二条是结构检验。看子本的章节次序跟着谁走。如果子本的第一章对应母本第一章，第二章对应母本第二章，新判断只能挤进某一章的某一节里，那么结构本身就在宣告：这是母本的一个地方版本。重写的结构跟着新判断的论证次序走，母本的章节被拆开，按新判断的需要重新调用。这有点像搬家：翻译是把旧房子的家

具按原来的摆法原样搬进新房，重写是先看新房的门窗朝向，再决定哪件家具放哪里，有的家具留下，有的不再需要，还要添几件旧房没有的。

第三条是追溯检验。从子本里任取一个新判断，反向追它的来路：它最早出现在哪一问的哪一轮问对里，是学生先说的还是ChatSDE逼近之后学生说破的，靠哪一份现场材料撑住，经受过哪一次批判而修改过表述。追得通的，是重写的产物；追不通的，要么是从别处搬来的，要么是写作时临时生出来的修辞，两者都不算子本里的新判断。追溯检验同时也回应了AI代写的疑虑：一个能被逐轮追溯到问对记录与现场材料的判断，很难由旁人或模型在最后一夜代写出来（轨迹作为证据的计分办法见第17章）。

这条检验的思路，与写作过程研究里常用的“版本史”视角相通。研究者研究学生怎样写作，不只看交上来的终稿，也看文稿怎样一版一版长出来：哪一段先写、哪一段被推倒重来、哪些句子是整块贴进来的、修改发生在哪些时刻。终稿相同的两篇文章，修订过程可以截然不同，所以过程比终稿更能说明文字的来路。子本的判断清单、各版大纲、重写稿的几次修订，合起来就是一份子本的版本史；追溯检验读的就是这份版本史。本书只取这一思路，不引用任何具体数字。但也要看到它的限度：版本史可以被刻意制造，一个学生完全可以把代写稿分几次贴进去，造出一条看似渐进的轨迹。所以追溯检验不单独使用，要与来源检验、结构检验合用，并且要与问对记录里的轮次、现场材料的采集记录互相对得上，换手评分人读的是几条互相印证的线索，而不是任何一条孤立的痕迹。

三条检验要合起来用。只用来源检验，学生可以把母本的话换个说法逃过去；只用结构检验，学生可以把章节打乱而判断不变；只用追溯检验，又可能把问对里一句漂亮话误当成判断。三条都过，才能说子本里有一个母本所无、来路清楚、结构上承重的判断。需要强调的是，这三条检验检的是“是不是重写”，不是“写得好不好”；一份通过检验的子本，其新判断仍可能是错的，它的质量要交给创新智商与完成度去评。

三条检验做全，评分人的工作量不小，所以要有一个合理的次序，省力而不失准。入口是判断清单：先看清单里承重的那一条，只对它做三条检验，通过了再往外扩到其余几条；承重判断都过不了，其余几条再漂亮，子本也谈不上重写。这样的次序还有一个好处，评分人读的是几页清单和一份追溯记录，而不是几十页成

文，成文越流畅，越容易让人把“读得顺”当成“写得新”。好比验收一栋楼，先查承重墙，再看装修，装修再精致，承重墙有裂缝也不能交付。

三条检验在实际使用中会遇到一个灰色地带：母本判断的限定。子本里常常出现这样的句子——“母本说的那一条在一般情形下成立，但在我的现场里只在某种条件下成立”。这算不算母本所无的判断？本书的处理是：如果限定只是给母本判断加上一个泛泛的“视情况而定”，它仍属于应用；如果限定指出了一个母本没有命名的条件，并且这个条件由现场证据撑住、能被下一届学生检验，它就算新判断。苏文的发现正是如此：母本说族谱断代多与战乱迁徙有关，他的判断没有否定这一条，而是指出了另一个母本未曾命名的条件——户籍登记口径的改变。日常里也有类似的事：老中医说“这个方子入冬喝”，徒弟发现“家里有地暖的人入冬喝反而上火”，后一句没有推翻师父，却说出了师父没说的那个条件。

苏文的第一稿是个人的失误，还有一种失误是设计上的，三条检验正是从这样一次失败里磨出来的，值得单独讲一讲。设想一门课在试行九问模式的第一学期，为了让学生早点交出成形的子本，老师给了一个看起来很高效的办法：把谱系母本按章拆开，让学生对每一章都问一句“这一章放在我的现场里会怎样”，再请大模型按母本章节逐章改写，把学生的现场材料填进去。结果是一个学期下来，全班交上来的子本整齐划一，篇幅都够，文字都通，引注都在，几乎没有一份需要退回重做。到了换手评分时，评分人却发现，这些子本的结论与母本的结论高度重合，彼此之间也高度重合，不同学生的子本只是换了现场的名字。

这次失败的教训有三层。第一层是流程层：逐章改写把起点放在了母本的目录上，九问研究成果只能作为“材料”被填进母本的格子里，判断清单那一步被省掉了。一旦起点是母本目录，后面再勤奋也只能产出母本的变体，就像拿着一张别人的地图走自家后山，走得再仔细，也只会走到地图上已经标出来的地方。第二层是工具层：通用大模型接到“按母本第三章改写成适合我现场的版本”这样的指令，在默认配置（未开启学习模式）下，往往会把“母本第三章是对的、只需要换个现场”这个前提轻轻质疑一句，随即替学生把改写后的版本一并写出来，也就是第8章所说的替学生说破，而不是让学生自己去现场里找这一章哪里会坏。即便开启了学习模式，它追问的也多是“你的现场有什么不同”，而不是“这一章放在你的现场

里可能坏在哪里”。指令里没有“这一章放在我的现场里可能坏在哪里”，模型也就不会主动去问。

第三层是评估层，也是最隐蔽的一层。这批子本之所以几乎没有被退回，是因为评估实际上在问“像不像答案”：结构是否完整，概念是否用对，引注是否齐全。按这个标准，翻译稿天然占优，因为它从母本那里继承了完整的结构与正确的概念。评估的方向如果不从“像不像答案”翻转为“离已有有多远”（越像已有的东西，越说明没有发生），逐章改写就会被评估反复奖励，下一届学生也会学着这样做。这次失败之后，课程把流程改成先交判断清单、经老师与同学过目后再交大纲，大纲通过了才开始重写；子本评分前先过三条判别检验，翻译稿不进入创新智商评分。

从这次失败里还能看出一件事：重写之所以比翻译难，不在文字功夫，而在于它要求学生先认下“我有一个母本没有的判断”。这一认，意味着要为这个判断承担证据责任，要准备好被同学和老师追问，也要准备好它被现场推翻。翻译是安全的，因为它的每一句话都有母本作保；重写是有风险的，因为新判断只有自己的现场作保。九问模式把这份风险当作课堂的正当部分，而不是需要被技术手段消除的麻烦——第7章讲的“接住批判”，到了子本这里，就变成了“为自己的判断立字据”。

还需要为翻译说一句公道话。逐章翻译并不是没有用处的劳动：把母本的方法逐章搬到自己的材料上走一遍，正是熟悉母本、检验自己是否真读懂了的好办法。在第1章所说的练手阶段，终点老师事先写得出，学生把统一性知识在自己的小实验里再发生一遍，形式上就很接近翻译，而这正是练手的本分。问题不在于翻译本身，而在于把练手的产物当成战场的产物交上来。大学课堂的子本要打向连老师也写不出终点的判断，翻译稿在这里只能算作母本底盘那一项的证据，不能算作子本。这就像学书法，临帖是必经之路，临得像是本事；可若是参加创作展，交上去一幅临得再像的帖，也只能退回。

另一个反对意见来自相反的方向：硬性要求子本必须含母本所无的判断，会不会逼着本科生去标新立异，在三年级的课程里造出一批站不住的“新发现”？这个担忧有道理，本书的回应有三点。第一，“母本所无”不等于“学科所无”：苏文的户籍口径发现，也许地方史学者早已注意过，只要它在他的现场里由他的材料撑

住、母本精要库里查不到，就够格作为子本的新判断；文献检索只用来保证它没有被当成学科首创来宣称。第二，要求的是一个承重判断，而不是一部创新专著，并且允许它写成“待检验”，门槛并不高。第三，防止造假的机制不在要求的宽严，而在追溯检验：追不到现场材料和问对来路的新说法，一律不算。这就像要求烹饪课的学生交出一道“在自家灶上改过的菜”，并不等于要求他发明一个新菜系，要的只是那一勺糖、那一把火里有他自己的试错。

◆ 两份子本的骨架：苏文与沈澄

回到苏文。在老师的批语之后，他重做了判断清单，从九问里整理出几条判断，其中真正承重的只有一条：家族族谱断了的那一代，断代处恰是户籍登记口径改变之年。其余几条要么是它的证据，要么是它的推论，要么是它需要排除的竞争解释。按照这份清单，他的重写稿换了一副骨架。全书不再从“族谱的编修体例”讲起，而是从那个断代处讲起：先摆出族谱在哪一代断开，再摆出县志同年记了什么，然后才调用母本里方志与谱牒互证的方法来论证两者的关联。母本那一章在重写稿里被拆开，一部分用来建立证据，一部分用来处理反例。

重写稿的中段专门处理竞争解释。母本里“族谱断代往往与战乱、迁徙有关”的现成说法，在第一稿里被当成结论，在重写稿里成了必须正面回应的对手：那一年家乡有没有战乱和迁徙的记载？如果有，户籍口径改变的解释是否多余？如果没有，为什么断代偏偏落在这一年？苏文在这一段里引了老人口述，但不再把口述当成补充材料，而是把它与县志、族谱三方对勘，标出哪里一致、哪里互相打架。重写稿的末段写判断的边界：它目前只在一个家族、一部县志上成立，要成为谱系性的判断，需要更多家族族谱与同期户籍档案的对照，这一条他写进了“待下一届检验”的清单。

沈澄的子本还在打造中，她交上来的是大纲。她的判断清单里承重的一条是：本组某一步的成败依赖未入规程的环境湿度与操作时长——规程里没写的那一句，在师门口头里传了十年。这个判断目前仍是“待复做验证”的状态，所以大纲的骨架是按“假说—证据—检验”排的，而不是按“结论—论证”排的。第一部分从她自己复做的那一批讲起：每一步的时间、温度、湿度都记了下来，与十年档案比，

差在哪一步。第二部分回到档案：产率最低的20批落在哪些月份，“室温”一词十年里是否写过具体温湿度，失败批次里哪一步的时长记录最不一致。

沈澄大纲的第三部分处理口头传统与成文规程的缝：师兄师姐说的“那一步要快”在规程里写了没有，别的实验室做同样的反应，规程里多了哪一句，那个反复出现、没人解释的小峰最早出现在哪一年。第四部分是一个她自己提出的检验设计——如果把湿度写进本组规程，第一个要改的是哪一行，改了之后应当在下一批里看到什么。这份大纲里，母本的合成方法学与课程型本科研究的方法论被分配到各部分作底盘，没有一部分是“母本第N章的本组版本”。由于化学子本的完成度以可复做性为核心，大纲的每一部分都标明了复做所需的记录项；涉及实验操作的地方，只写需要记录的变量与对照方式，不写也不改动任何危险步骤。

两份骨架还有一处不同，值得留意。苏文的判断在他的现场里已经得到了初步支持，他的重写稿可以按“结论—论证—边界”的次序写；沈澄的判断还停在假说阶段，她的大纲只能按“假说—证据—检验”的次序排。这提示子本的形态应当诚实地跟随判断的成熟度，而不是一律写成“我发现了什么”的样子。一个尚待复做的假说，硬写成已证实的结论，既违背化学的可复做性要求，也违背本书的证据责任；反过来，一个已有现场支持的判断，若因怕担责任而写得吞吞吐吐，也是没有把发生交代清楚。子本里允许“待检验”三个字出现，而且应当在它该出现的地方出现，它们本身就是给下一届学生、给课题组留下的接力棒（见第24章）。

把两份骨架放在一起看，可以看到重写的一个共同特征：子本的起点都是一处“断”。苏文的是族谱的断代处，沈澄的是档案与规程之间的空白。这并非巧合。母本是连续的叙述，它把已知编织成一个整体；新判断往往从整体里的一处断口长出来，那里母本说不清，学生的现场却恰好有材料。翻译会把断口抹平，好让子本像母本一样连贯；重写则把断口当作全书的入口，让连贯性在新判断周围重新长出来。

◆ 子本的学术规范

子本是一本书，不是一份心得，它要在学术共同体里站得住，就得守学术规范。这里的规范有一部分与一般学术写作相同，例如概念先定义后使用、论证与结论相称、引注准确可查；另一部分是子本特有的，因为子本同时连着三样东西——

母本、问对记录与学生的现场。连着母本，就有“哪些是母本的、哪些是我的”的边界问题；连着问对记录，就有“哪些是我说的、哪些是ChatSDE说的”的归属问题；连着现场，就有证据真实与伦理许可的问题。下面分四条来讲。

第一条是母本标注。子本里凡是调用母本判断的地方，要标明出自母本或精要库的哪一条，用的是原判断还是经过限定的版本。这样做不是为了显示谦虚，而是为了让读者能做上文所说的来源检验：只有母本的部分标清楚了，新判断才能被看清。这就像一份家常菜谱，哪些是外婆的做法、哪些是自己改的，写明白了，下一个做菜的人才知道该在哪里试。母本标注也服务于后面的回写：若子本里的某一处限定被证明成立，它将来可能进入精要库，标注清楚了，回写时才知道改的是哪一条。

第二条是问对归属。子本里的判断，凡是在问对中形成的，要能指出它在哪一份记录的哪一轮出现，由谁先说出。ChatSDE在问对里会逼近判断、亮出选刀的依据，但按工序宪法不替学生说破；即便如此，问对记录里仍会有ChatSDE提出的反例、追问与竞争解释。子本可以使用这些内容，但要如实写明它们来自问对，而不是把它们写成学生自己的发现。许宁的子本里，“同样年纪的患者，为什么白天跌得少”这一问是ChatSDE第一回合提出的，子本应如实记作问对中的批判；而“风险藏在交班本里被省略的那一句”是她在问对与夜班记录之间往返后自己说出的，才可以记作她的判断。归属清楚，是问对深度与子本创新智商能够分开评分的前提。

第三条是现场证据的呈现。子本里的现场证据要给出来源、采集方式与局限：数据从哪里来，怎么记录的，谁看过，样本有多小，哪些可能有偏。苏文的口述材料要写明访谈对象的身份类别、访谈的时间与方式，并说明老人的记忆可能与县志互相影响；沈澄的复做记录要写明仪器、记录频率与环境条件，使他人可以复做。证据的局限写得越清楚，判断的边界就越可信。一份把小样本写成普遍规律的子本，即便判断新颖，在完成度上也要被扣分；所有人物与数据在本书中均为构造案例，但规范本身是真实课堂必须执行的。

第四条是伦理与安全的门槛。医学子本涉及患者的，一律去标识化，并须有科室与伦理委员会的许可，否则整份不计分；化学子本不写危险操作细节，不改动危险步骤；历史与文学子本涉及家族与口述材料的，要征得当事人或家族同意，必要

时隐去姓名与可识别细节。苏文在重写稿里把族谱中涉及近世族人的私事一概略去，只保留与断代判断相关的世系信息，这一处理既是伦理要求，也让论证更干净。门槛不是加分项，而是前提：过不了门槛的子本，判断再新也不进入评分。

最后谈一谈篇幅与形式。子本不以页数计分，但一本子本至少要能容下三样东西：一个承重的新判断，撑住它的现场证据与母本底盘，以及它的边界与待检验之处。缺了任何一样，它都不成其为书，而只是一份报告或一篇感想。在形式上，子本可以是论文体，也可以按学科习惯采用其他体例：医学子本可以写成病房研究报告加讨论，化学子本须附完整的复做记录，历史子本可以附史料对勘表，文学子本可以是理论阐释与自己的写作样本合编。形式随学科走，但上面四条规范不随学科改变，它们是不同学科的子本能放在同一把尺子下评分的前提。

四条规范合起来，回答的是一个问题：一本子本里，每一句话的来路是否都交代清楚了。来自母本的，标母本；来自问对的，标问对；来自现场的，标现场；剩下的，才是学生自己在三者之间做出的判断。翻译稿在这套规范下会显得很尴尬——它几乎每一句都要标母本，标完之后，属于自己的部分所剩无几。规范在这里同时起了判别作用：它不是额外加在重写之上的形式要求，而是让“重写不是翻译”这一判断能被检查的工具。

◆ 本章怎样算错

本章的主张是：子本须是重写而非翻译，须含母本所无的判断；先整理判断清单、再长大纲、后重写的流程，能比逐章改写更可靠地产出这样的判断。下面三种读数可以推翻它。

第一，连续两届、每届不少于30份子本，随机分成两组，一组按“判断清单—大纲—重写”的流程打造，另一组按逐章改写打造。由两名互相独立、不参与问对的评分人，按来源、结构、追溯三条检验逐份计数能通过全部三条的新判断。若流程组平均每份通过的新判断，比逐章改写组多不出1条，那么本章关于流程次序的主张站不住，应当撤回，转而查找真正决定新判断出现的因素。

第二，同一批子本由两名评分人各自独立判定“翻译”还是“重写”。若两人判定一致的比例低于80%，这套判别就只是修辞，需要重新设计可操作的检验；若一致率达标，却只在同一底座上的多个智能体之间达标，则不能算独立证据。

第三，从通过检验的子本里随机抽20条被标为“母本所无”的判断，反向追溯来路。若其中有5条以上追不到现场证据或问对记录，就说明学生是在用表述上的新说法代替判断上的创新，本章对新判断的界定要收紧。

这些条款要靠记录来检验，不能靠宣布。一门课自称“我们的学生都在重写”，并不构成证据；能作为证据的，是可供换手评分人逐条核对的判断清单、大纲版本与追溯记录。若这些材料拿不出来，本章的主张就只停留在说明层面，尚未经受过检验。

◆ 子本写成之后

苏文的重写稿交上去之后，老师的批语换了一句：“这个判断有了，下一步，它要去哪里？”这句话把本章引向一个新问题。子本写成，在传统课堂里意味着作业完成，论文入档，学期结束；可是一个从断口里长出来的判断，本身还在生长——苏文需要更多家族的族谱，沈澄需要下一批复做，许宁的“被省略的那一句”需要交科室验证。如果子本只是放上书架的论文，这些生长就随学期结束而中止。

子本写成之后，它能不能继续长？它以什么形式进入学生此后的学习、工作与研究？一门课的子本与下一门课的子本之间是什么关系？毕业之后，学生与大学之间还剩下什么联系？第13章将讨论子本怎样进入学生的个人知识底盘，通过个人检索增强生成（RAG）与Skill成为可调用、可回写的数据基因，并以许宁从实习到住院医的经历为例，说明子本怎样长成一个人的思考底盘。

> **带 走 的 话**
>
> **子本不是把母本换成自己的口吻再写一遍，而是从自己的判断清单里长出来的书：里面至少有一个母本所无、来路查得清的判断。**

◆

13

第13章

子本进入底盘：RAG、Skill与子本系列

◆ 书架上的子本

《临床推理与病房研究》研讨课的最后一次课，老年病科的带教老师没有讲新内容，只问了全组一个问题：“一年以后，你们的子本在哪里？”大家互相看了看，答案差不多：在学院的课程档案里，在自己电脑的某个文件夹里，运气好的话改成一篇小论文投出去。许宁想了想，说她不知道。她的子本里那个判断——本病房夜间跌倒的风险更多落在交接断点上，风险藏在交班本里被省略的那一句——还挂着“本病房假说，待检验”的标签，已交科室讨论；下个月她就要离开老年病科，转去别的科室实习。她问老师：一个还没检验完的判断，人走了，它怎么办？

这个问题在讲授课堂里几乎不会被问出来。讲授课堂的产物是“掌握了”，掌握的东西随人走，期末考完，课程就结束了；论文写完，交上去，就入档了。九问课堂的产物却是一本含有新判断的子本，而新判断的特点是它还在长：它需要下一次检验，需要别的现场来试，需要被推翻或被加厚。一个还在长的东西若被放上书架，生长就停了，就像把一株正抽芽的苗连盆搬进仓库，盆还在，苗却慢慢枯了。

本章的核心判断是：子本不是放上书架的论文，而是个人知识发生平台的数据基因，毕业后继续生长。所谓“数据基因”，是说子本里存着的不只是结论，还有判断怎样发生的来路——母本出发点、九问、改题史、问对记录、现场证据与边界——这些来路可以在新的现场里被再次调用，长出新的判断。下面先讲子本怎样转成个人检索增强生成（RAG）的资料库与可调用的Skill，再以许宁为例讲它怎样成为思考底盘而不是临床决策工具，以及什么叫知行一体；然后与默会知识和SECI对话，再讲子本系列，并以周川毕业五年后回来更新子本为例，讲大学与人的关系怎样变为终身；最后说底盘的边界。

这一章的问题

一个还没检验完的判断，人离开课堂之后它怎么办？

◆ 从子本到底盘：RAG与Skill

先说两个技术名词在本书里的用法。检索增强生成，是让大模型在回答之前先到一个指定的资料库里检索相关内容，再结合检索结果生成回答的做法，最早由Lewis等（2020）作为一种系统设计提出，如今已是大模型应用里的常见配置。个人RAG，在本书中指以学生自己的子本及其来路材料为资料库的检索配置：学生向它提问时，它先查的是“我自己曾经怎样想过这件事”，而不是互联网上的通用答案。Skill在本书中指一套被打包、可反复调用的工序说明：它不存结论，存的是“遇到这类情况，先问什么，再查什么，怎样判断自己的判断可能坏在哪里”。

两者的分工可以用一个日常比喻说清。RAG像是一个人的老笔记本，翻开能找到自己当年记下的事实、证据与判断；Skill像是这个人练成的手艺，比如一位老裁缝量体时总先看人的站姿再拿尺子，这个次序不写在任何一本笔记里，却每次都在用。子本同时是两者的原料：子本的正文、判断清单、问对记录与现场材料进入RAG，成为可检索的资料；子本里那条判断是怎样发生的——从哪一个初判出发，被哪一个反问翻过来，在现场里查了什么——被提炼成Skill，成为可迁移的工序。

转化的过程本身也要守第12章讲的学术规范。进入RAG的材料，要保留来源标注，母本的归母本，问对的归问对，现场的归现场，否则日后检索时，学生会把ChatSDE当年提出的一个反例误记成自己的结论，或把母本的一般说法误当成自己现场的证据。提炼Skill时，要写明它是从哪一本子本、哪一个判断里来的，它在原现场里是否检验过，检验到什么程度。一条只在一个病房、一个学期里得到初步支持的工序，打包成Skill时就要带着这个限定，就像药品说明书要写清适用人群，不写清，就容易被用错地方。

个人底盘与通用大模型自带的“记忆”功能也要分开说。一些通用助手会记住用户说过的偏好与事实，下一次对话时调出来，这对日常使用很方便。个人底盘与之不同的地方，不在技术上能否记住，而在记住什么、按什么结构记：通用记忆记的多是“这个用户是谁、喜欢什么”，个人底盘记的是“这个人的某个判断是怎样

发生的、靠什么证据、在哪里还没检验”。前者让助手更贴合用户，后者让用户在新现场里能调出自己的来路。两者可以并存，但不能互相代替；把底盘降格为偏好记忆，子本里最值钱的那条来路就丢了。这也不是对通用大模型的贬低，差别落在配置与工序上，而不在底座能力。

这里需要划清一条界线：个人RAG与Skill构成的是学生的思考底盘，而不是一个替学生做决定的系统。底盘的意思是，学生在新的现场里思考时，它提供可调用的来路——我曾经怎样把一个判断翻过来看，我曾经在哪里找到过被省略的东西；它不提供“这里应当怎么做”的答案。这一界线在各学科都成立，在医学里则是硬性的伦理要求。下一节就以许宁为例展开。

◆ 许宁的交接班底盘：不是决策工具，而是有回写的知

许宁离开老年病科之前，在带教老师的指导下把子本转成了个人底盘。转化时做的第一件事，是把一切可能识别患者的信息剔除干净：子本本身已经去标识化，并经科室与伦理委员会许可（这是医学子本的计分门槛，见第18章），但问对记录与夜班笔记里仍可能有床号、日期与细节的组合，足以让熟悉病房的人认出是谁。这些内容在进入个人RAG之前一律删去或改写为类别描述，只留下与判断相关的结构信息，例如“某次交接前一班的记录里未提及当日药物调整”。原始记录按科室规定留在科室，不随个人带走。

她提炼出的Skill只有一句核心工序，外加几条展开：交接时，不只看交班本写了什么，还要问没写什么；被省略的那一句，是谁、在什么情况下省略的。这句工序来自她九问里的第2问及其提升版，但它已不再是关于跌倒的判断，而是一种看交接的眼光。Skill的说明里写明了它的来路与限定：形成于老年病科一个病房、一个学期的夜班观察，判断的内容（风险落在交接断点上）仍是本病房假说，待科室检验；工序本身（问没写什么）可以迁移，迁移后的结论要在新现场重新取证。

转到急诊科实习时，这条工序第一次在新现场里被调用。急诊与病房之间也有交接，患者从急诊收入病房，一份记录交到另一个科室手里。许宁在实习日志里注意到，自己读急诊转入记录时，习惯性地在问“这里有没有被省略的那一句”。她把这个观察写进个人底盘，向自己的RAG提问：老年病科那次，被省略的句子多半

出现在什么情况下？检索出的是她当年的问对记录与夜班笔记，提醒她当时的发现是“忙乱的交接时段前后”。她没有据此对任何一位急诊患者下判断，只是在实习日志里记下一个新的待问之题：急诊与病房之间的交接，被省略的句子是否也集中在某个时段？这道题她交给带教老师，由老师决定是否值得在科室里作为研究问题提出。

到了住院医阶段，许宁有了自己的夜班，也有了自己写交班的责任。这时底盘的用法又变了一层：她每次写完交班，会用自己那条Skill回头看一遍——我这一份里，省略了什么？是因为太忙，还是因为我默认下一班“都知道”？这个动作不改变任何医嘱，也不替代科室的交接规范，它只是让她在既有规范之内，对自己的记录多一道反省。临床决策照常按诊疗规范、上级医师指示与科室制度执行；底盘出现的位置，是在决策之外、在她反思自己怎样看问题的地方。用一个大白话的比方：底盘像是司机自己养成的、并线前多看一眼后视镜的习惯，它不是导航，不告诉你在哪里转弯，更不替你踩刹车。

把底盘与临床决策工具分开，不只是为了合规，也是为了保住底盘本身的用处。一个被当作决策工具的系统，会被要求给出确定的答案，而许宁底盘里存着的恰恰是不确定的东西：一个待检验的假说，一种看交接的眼光，几道还没回答的题。如果把这些东西包装成“风险提示”去指挥临床，假说就被偷换成了结论；如果因为它们不确定就丢掉不用，那条从病房里长出来的眼光也就丢了。思考底盘的位置正好在两者之间：它让不确定的东西继续以“问题”的形式活着，并在新的现场里被重新检验，而不是在任何人的病床前替人做主。

许宁的经历里还有一个容易被忽略的转折：她的子本在离开课堂之后，没有停在“知道”这一层，而是变成了一种做事的方式，并且做事的结果又回过头来改写了底盘。急诊实习时的那道新题，住院医时对自己交班的反省，都被她记回个人底盘；底盘因此不再只是一本子本的镜像，而是一份在行动中被改写的记录。本书把这种状态叫作知行一体，并给它一个限定的定义：有回写的知。知若没有回写，就是存放着的知识；知被行动调用、行动的结果又改写了知，两者才连成一体。

这个说法与杜威在《我们怎样思维》《民主主义与教育》中对经验与反思的论述有相通之处：思维起于疑难情境，在行动中检验假设，经验的意义在于它改变了

后来的经验。舍恩在《反思性实践者》里讲专业人员“在行动中反思”，也描述了相近的现象——有经验的从业者在做事时会察觉意外，并在当下调整自己的理解。本书不把子本底盘说成是这些理论的实现，只是指出：九问模式为“在行动中反思”提供了一种可以被留下痕迹的载体。舍恩所说的反思常常发生在从业者头脑里，事后难以追溯；个人底盘让这种反思有了来路标注、有了版本，可以被自己和他人看见。

回写在这里同样要被观察，不能被宣布（回写的三层结构见第24章）。一个学生说“我的子本已经融入了我的工作方式”，这只是一句宣布；能作为观察的，是底盘的版本记录：哪一天加入了哪一条新的待问之题，哪一条Skill的说明被修改过，修改的理由是什么现场材料。许宁的底盘里，那条“问没写什么”的工序在住院医第一年被加了一条限定：“在本人书写的交班中，省略多发生于自认为下一班已知情的事项。”这条限定有日期、有来路、有她自己的夜班记录作证，它是可观察的回写。反过来，如果一个人的底盘两年没有任何改动，那么无论他怎样描述子本对自己的影响，都只能说底盘处于停滞，回写没有发生。

还需要指出一种伪回写：底盘内容被大量增加，但增加的全是从外部搬来的资料，而不是自己在现场里检验过的判断。这就像一个人的笔记本越来越厚，却全是剪报。个人RAG在技术上很容易被做大——把指南、综述、教材一股脑儿放进去，检索起来似乎什么都有——可这样做恰恰稀释了它作为思考底盘的特性，检索到的多半是别人的结论，而不是自己的来路。本书主张，个人底盘的增长应以经过自己现场检验的判断与工序为主，外部资料可以作为母本层另行存放，并与个人判断层分开标注。

◆ 外化了多少：与默会知识和SECI的对话

子本进入底盘，在知识管理的语言里有一个现成的位置。Polanyi（1966）在《默会维度》里提出，人的知识总是多于他所能言说的；Nonaka与Takeuchi（1995）在《创造知识的企业》里，把默会知识与显性知识的往复转化排成SECI四步：共同化、外化、组合、内化。照这个框架读许宁的经历，轮转之前她看交接的那点眼光是默会的，说不清为什么每回看交班本总觉得哪里少了点什么；把它写成九问、问对记录和子本里的判断，是外化；子本的材料与其他资料在个人RAG里被

调出、并排比对，近于组合；她在急诊实习里不自觉地问“有没有被省略的那一句”，是内化；至于科室讨论与带教老师的点拨，则是新一轮的共同化。本书承认这一对应，也不主张子本底盘是一种全新的知识转化，它是SECI里“外化”这一步在课堂上的一种具体做法。

承认之后，先要守住Polanyi的警告：能写进Skill的，只是知识里能说出来的那一部分。一份食谱写得再细，也写不出老厨师凭火候翻锅的那一下；许宁的Skill写下“问没写什么”，写不进去的是何时该问、在什么神色的交班者面前该多问一句。倘若把Skill当作能力本身，就会生出一种错觉，以为谁拿到这条说明谁就拿到了这双眼睛。所以底盘里除了工序说明，还要保留它的痕迹——问对记录、现场材料、当时的判断被翻过来的那个回合——因为痕迹是下一个人学着“重走一遍”的路标，而说明只是路标上的字。

那么本书在SECI之上多出的一格是什么？SECI描述的是转化的流向，对外化之后的东西落在哪里、怎样继续使用，说得很少；而组织知识管理里常见的做法，是把外化的成果放进一个共享库，由库里的条目代表知识。子本底盘的一格多出，有两处。一处是外化之后立刻接入两个会被反复调用的设施：个人RAG使它可以按“我曾经怎样想过这件事”被检索，Skill使它可以按“遇到这类情况先问什么”被调用，二者缺一，外化成果就是静态存档。另一处是外化后的东西要接受版本记录：每一次增补、限定与撤回都带日期、带理由、带现场材料，使一个外人不必相信本人的叙述，也能核对这条工序是怎样改变的。前文所说的“问没写什么”被限定在“本人书写的交班中，省略多发生于自认为下一班已知情的事项”，就是这样一条能被核对的改动。这两处合起来，才是本章所说的“数据基因”，它比一份外化后的文档多出来的，是可调用与可追溯。

这里有一个设想的反例，用来说明外化的限度。设想一位学生把“问没写什么”当作万能问句，到了一个交接记录本来就格式固定、逐项勾选的科室，她仍然对每一个空格追问“这里省略了什么”，结果淹没在不构成省略的空格里，反而漏掉了真正出问题的那一项。这不是Skill写错了，而是Skill里能外化的是问句，不能外化的是分辨空白与省略的判断。所以本章要求Skill带着限定出门，并且主张用迁移检验来看外化究竟传出去多少：让一位没有做过这门课、但基础相当的同学，拿

着同一条Skill到一个新现场去找省略，与没有这条Skill的同学比较，看有无差别。如果有Skill的人并不比没有的人找得更准，那么外化出来的只是一句好听的话，没有传出什么。

◆ 子本系列与终身关系：一门课一卷，一生一套

如果一门课产出一本子本，那么一个学生在大学里会积累若干本子本，毕业之后，在工作与研究中还会继续产生新的子本。本书把这一串子本称为子本系列：一门课一卷，一生一套。系列不是简单的合订本，它的各卷之间有谱系关系——后一卷的出发点里，常常含有前一卷的判断或工序，就像一棵树的新枝从老枝上长出来，而不是另插一根。

以许宁为例。她的第一卷是老年病科轮转的跌倒子本；假如此后某一门研究方法课让她以急诊与病房之间的交接为现场打造第二卷，那么第二卷的母本除了该课的谱系母本之外，还会包括她自己第一卷里的那条工序与那个待检验的判断。第二卷的九问可能问：急诊转入记录里，被省略的句子集中在什么情形下？与老年病科病房的交接相比，省略的内容有什么不同？这样，第一卷里的判断就在第二卷里得到一次新的检验：它可能被加厚（交接断点在不同科室都是风险所在），也可能被限定（只在某类交接里成立），也可能被推翻。无论哪一种，都是谱系在生长。

子本系列对大学的课程设计提出了一个新要求：课与课之间要允许学生把前一卷带进后一卷。在传统课程体系里，每门课是一个独立单元，前一门的作业很少被后一门调用；在九问模式里，老师出九问时，学生画像里应当包含学生已有的子本卷目。这有点像一位好的家庭医生，看病时会翻一翻病人过去的病历，而不是每次都从头问起。一门课的老师若能看到学生前几卷子本里那几条待检验的判断，就可能把其中一条转成本课九问的一问，让它在新的谱系母本下接受检验。

子本系列还改变了“学历”这个词的含义。传统的学历记录的是一个人通过了哪些课程、拿到了哪些学位，它是一串已完成事项的清单；子本系列记录的则是一个人在哪些现场里发生过哪些判断，这些判断此后又被怎样加厚、限定或推翻。前者像一张盖满章的通关文牒，后者更像一本仍在续写的行医日志或航海日志。本书不主张以子本系列取代学位制度，那是制度层面的问题；但在课堂层面，老师认识

一个学生，可以从读他的子本卷目开始，而不只是看成绩单上的课程名（成绩单的新格式见第15章）。

子本系列也带来一个需要警惕的风险：学生可能一辈子在同一条判断上打转，后一卷只是前一卷的重复。判别的办法与第12章的翻译检验相同，只是对象从母本换成了自己的前一卷：后一卷里，有没有前一卷所无的判断？如果每一卷都只是把第一卷的判断搬到新现场里再说一遍，那么系列就成了自我翻译，它与逐章翻译母本一样，看起来勤奋，实际上没有发生。

周川的经历可以说明子本系列怎样把大学与人的关系变为终身。他在《小微经营者的动力与节律》研讨课上打造的子本，承重判断是：父亲的动力被账本提前收走；在同学碰撞场里，他和林晓又撞出了班级判断——小微经营者的动力被现金节律收走。毕业时，他按自选权重拉高了现场证据一项，因为他准备回家接手饭馆。毕业后的头几年，他确实接手了饭馆，子本转成的个人底盘也跟着他回了家：每到月底，他会翻一翻自己当年关于“父亲为什么一到月底就泄气”的问对记录，看看自己是不是也在泄气，泄气是不是又落在交房租、付货款的那几天。

毕业五年后，周川回到学院，提出要更新自己的子本。他带回来的不是一篇论文，而是五年的饭馆流水账、一本自己写的经营日志，以及一个他自己也说不清的新困惑：他按子本里的判断调整了付款节奏，把几笔大额支出从月底挪开，月底的泄气确实轻了，可另一种疲惫出现在了别处——不在月底，而在每个季度报税之前。他问当年的老师：这是同一个判断的延伸，还是另一个判断？老师没有直接回答，而是让他先写第零轮初判，再进入问对（第零轮见第6章）。周川的初判是“报税太麻烦”；ChatSDE的第一个反问是：“五年前你父亲也要报税，那时为什么疲惫没有落在季度上？”

这个反问把周川带回了他自己的现场。他翻出父亲当年的账本，发现父亲那时用的是另一种纳税办法，申报的节律与今天不同。于是新的题出现了：动力被现金节律收走，那么“现金节律”是由什么排定的？房租、货款、报税，它们的周期是谁定的，经营者能挪动哪一个，挪不动哪一个？这个问题已经超出了五年前那本子本的范围，周川的更新稿因此不是对原子本的修补，而是系列的第二卷。原子本的判断在第二卷里被限定了一层：在他自己的饭馆里，能被经营者挪动的那部分现金

节律，挪动后泄气减轻；挪不动的那部分，泄气转移到了新的时点上。这仍是一家饭馆的构造案例，作为待检验的判断，而非对小微经营的一般结论。

周川这次回来，还让老师看到了一件在学期内看不到的事：子本里的判断，有一部分只有在学生真正承担后果之后才能被检验。在课堂上，“动力被账本提前收走”是从父亲的流水账与访谈里推出来的，周川是观察者；五年后，他自己成了那个在月底面对账本的人，判断第一次落在了他自己的身上。他发现挪动付款节奏确有效果，这是判断得到的一次支持；他发现疲惫转移到了报税前，这是判断遇到的一次限定。两者都不是课堂上能得到的。这就像学开车的人在驾校里学会了看后视镜，真正上路几年之后，才知道哪种路况下最容易忘记看。大学若在学生毕业时就把子本封存，这些只有在承担后果之后才出现的检验，就永远进不了课程的谱系。

周川这一次的经历，可以用本书对123原理的限定用法来描述：D与E之间出现了矛盾——他按子本的判断调整了付款节奏（D），而现场（E）却给出了一个不合预期的结果，疲惫没有消失，只是挪了位置；这个矛盾逼出了S的改变，他对“现金节律”的理解从“经营者面对的一组付款日期”变为“由不同主体排定、有的能挪有的挪不动的一组周期”；改变了的S又回写到D与E上，他开始重新安排可以挪动的那部分，也开始在自己的流水账里专门标出那些挪不动的日子。这个循环之所以可以被说出来，是因为它在他的经营日志与底盘版本里留下了可以对照的痕迹，而不是因为他事后这样总结。倘若日志与底盘里找不到这些痕迹，那么同样的故事讲得再圆满，也只能算作一次宣布。

周川的更新同时回写进了课程。他的第二卷经他同意、去除经营隐私后，进入《小微经营者的动力与节律》的部落档案；第二年，老师给一位家里开杂货店的学生出九问时，把“哪一笔支出的周期是你家能挪动的，哪一笔不能”转成了其中一问。一位毕业五年的学生，就这样成了下一届学生九问的一个来源。这正是本章所说的“大学与人的关系变为终身”：不是校友会式的联络，也不是终身教育式的再培训，而是一个人的子本系列与一门课的谱系母本互相回写，人在课程的谱系里留下痕迹，课程也在人的一生里继续出现。打个比方，它不像毕业后偶尔回母校参观，而更像一棵树与它扎根的那块地：树长大了，落叶回到土里，地也因此变了。

这种终身关系对大学提出了制度上的要求，本章只点到为止。学院需要为毕业生保留子本更新的入口，需要有人在毕业生带着新材料回来时，替他组织问对、安排换手评分、判断哪些内容可以进入部落档案；部落档案需要有处理隐私与授权的规则，毕业生的经营数据、病房经历、家族史料，进入档案前都要经过本人同意与去标识化。这些制度如何安排，属于兄弟书《大学的典范革命：从统一性到谱系发生》所讨论的宏观层面；本书只指出，课堂层面的九问模式一旦把子本视为会继续生长的数据基因，就必然向制度提出这一要求。

◆ 底盘的边界：归属、隐私与换底座

把子本做成会继续生长的底盘，会带出几处张力，需要在课堂层面先说清。第一处是归属。子本是学生写的，但它的出发点是谱系母本，它的问对发生在学院提供的ChatSDE上，它的现场材料有时属于科室、课题组或家庭。底盘应当归谁？本书的主张是分层归属：学生自己的判断、工序与写作归学生，学生毕业后可以带走、更新、继续使用；母本与精要库归课程与课题组，学生可以调用但不能据为己有；现场原始材料按其来源方的规定处理，医学的原始记录留在科室，化学的原始实验记录留在课题组，家族史料与经营数据归当事人。这有点像一个学徒出师，手艺是自己的，可以带走；师父的作坊、工具和账本，不能一并搬走。

第二处是隐私。个人底盘一旦可以被检索、被调用，它所存的东西就比一份锁在抽屉里的论文更容易外泄，也更容易在多年后被用于当初没有同意过的目的。因此进入底盘的材料，必须比进入子本的材料多过一道筛：子本是一份要给评分人看的文本，底盘是一个要陪人很多年的系统，后者在去标识化上应当更严，而不是更松。许宁的做法——床号、日期与细节组合一律删去，只留结构信息——在医学里是底线；周川的经营数据进入课程档案前须经他本人同意，并去除可识别的商户信息；苏文的族谱材料，涉及近世族人的私事一概不入底盘。凡是拿不准能否放进去的，就不放。

第三处是换底座。个人RAG与Skill要靠某个大模型来运行，而模型会更新换代，学生毕业后所用的模型也未必与学院的相同。如果底盘的价值全系于某一个模型，那么模型一换，底盘就作废了。本书的设计是让底盘的核心以与模型无关的形式保存：子本正文、判断清单、来源标注、工序说明，都是人能读懂的文本，任何

一个能做检索与生成的模型都可以把它们装进去。这与第10章的判断一致——模型是发动机，工序是驾驶术，底座可换，工序跟走。个人底盘是这条判断在个人层面的版本：发动机可以换，一个人自己的路书与驾驶习惯留在手里。

还有一处张力较为微妙：底盘会不会反过来把人框住？一个人长年调用自己的Skill，也可能使那条工序变成新的惯性，就像一个只会用锤子的人看什么都像钉子。许宁若在任何场合都只问“被省略的那一句”，她也可能错过那些并非出在省略上的问题。对这一风险，本书不给技术解法，只给一个课堂上的提醒：底盘里的每一条Skill都要带着它的来路与限定，并且像第7章所说的那样，时时问一句“这一条工序可能坏在哪一处”。底盘是用来让人看得更多的，一旦它让人看得更少，就该被改写。

◆ 本章怎样算错

本章的核心主张是：子本可以通过个人RAG与Skill转成个人思考底盘，在毕业后继续生长，形成子本系列，并使大学与人的关系变为终身。这一主张跨越了课堂以外的时间，许宁与周川的经历都是构造案例，只能说明这条路径如何可能，不能证明它会普遍发生。要检验它，需要跨越多年的追踪，以及对底盘版本记录的独立审读。下面四条各自带着样本和时间窗，任何一条成立，对应的论断就要退让。

第一条看生长是否发生。追踪不少于30名完成九问课程的毕业生，毕业后每满一年，由不参与教学的审读者核对其底盘版本记录；若满两年时，超过一半的人没有任何一条带日期、带现场材料的新增判断或被修改的工序说明，那么“毕业后继续生长”只是一个愿望，应当撤回，转而研究是什么阻断了生长。

第二条看底盘是否滑成决策工具。对高风险学科的在读学生，每学期抽查不少于20条其调用底盘的记录，若其中有任何一条被直接用作临床处置或医嘱的依据，且课堂层面的界定与伦理门槛无法使下一学期的抽查重新归零，那么“思考底盘不是决策工具”的设计就失败了，需要重新划定底盘可以进入的场景，乃至暂停在高风险学科里推行个人底盘。

第三条看子本系列是否只是自我翻译。取不少于15位已有两卷以上子本的学生，对后一卷逐一做来源检验；若超过一半的后一卷找不出前一卷所无的判断，那

么“一门课一卷，一生一套”所描述的谱系生长并未发生，系列只是自我翻译的累积，本章需要修改对子本系列的论述。

第四条看外化出来的东西是否传得出去。取同一门课的毕业生，把其中一条带限定的Skill交给不少于10名没有学过这门课、基础相当的同学，到同一个新现场做一次找省略的作业，与不少于10名没有这条Skill的同学对照；若两组找到的有效项数在盲评下没有可辨的差别，那么本章所说的外化只是一份存档，不构成对默会知识的任何转化，“数据基因”这个说法就要缩回到“个人备忘”。

◆ 从一个人的底盘到一群人的碰撞

本章讲的都是一个人：许宁带着自己的底盘走过急诊与住院医，周川带着自己的底盘回到饭馆又回到学院。可是细心的读者会注意到，周川子本里最有分量的那个判断——小微经营者的动力被现金节律收走——并不是他一个人说出来的。他自己的判断只到“父亲的动力被账本提前收走”为止；“现金节律”这个说法，是他和林晓在同学碰撞场里撞出来的，在他们各自的子本里都找不到原型。

这就提出了一个新问题：如果九问模式只是让每个学生与ChatSDE一对一地走完九问，打造自己的子本，转成自己的底盘，那么它与一位水平很高的私人导师有什么区别？课堂之为课堂，难道只是许多一对一辅导恰好安排在同一个教室里？第14章将讨论同学碰撞场：在什么时点、以什么方式组织同学互撞，怎样打造个人说不出的班级子本，以及沈澄与许宁这两个毫不相干的学科之间，怎样撞出“没写下来的那一句”。

> **带 走 的 话**
>
> **子本不是放上书架的论文，而是毕业后继续生长的思考底盘：它存着判断怎样发生的来路，用来让人看得更多，从不替人做决定。**

◆

14

第14章

同学碰撞场：班级子本的涌现

◆ 两份各自成立的子本

《小微经营者的动力与节律》研讨课进行到第11周，老师把周川和林晓的子本草稿并排投在屏幕上。两份稿子都过了第12章所说的三条判别检验，都有母本所无的判断。周川的判断是：父亲的动力被账本提前收走——每到月底，账本上那几笔必须付出的钱还没付，父亲的劲头就先泄了。林晓的现场是母亲的理发店，她的问题是“母亲为什么交房租前不爱说话”，她在九问里一路追下去，落下的判断是：母亲的沉默不是因为累，而是在交租的日子临近时，心思全被那一笔钱占住了。老师没有点评任何一份，只问全班：“这两份子本，说的是一件事还是两件事？”

教室里先是一阵各说各的。有人说是两件事，一个是饭馆，一个是理发店，一个是泄气，一个是沉默；有人说是一件事，都是小本生意的压力。周川自己起初也觉得是两件事，他的判断落在“账本”上，林晓的判断落在“房租”上，账本和房租不是一回事。林晓却注意到一个细节：周川父亲的泄气与林晓母亲的沉默，在各自的日历上都不是随机的，而是落在钱要出去的那一段。她问周川：“你父亲月中进货付款那几天，会不会也有一点泄气？”周川翻了翻自己的流水账与访谈记录，发现确有几次，只是不如月底明显，他当时没有当回事。

讨论一直持续到下课。最后落在白板上的那句话，是两个人谁都没有在自己子本里写过的：小微经营者的动力被现金节律收走。“账本”是周川的词，“房租”是林晓的词，“现金节律”这个词出现在两人一来一回的追问之后，起初是林晓说出半句“是不是跟钱出去的节奏有关”，周川接上“节奏不止房租，进货、工资、税都是”，老师把它写上白板时，两人都愣了一下。这个判断在周川那里找不到，在林晓那里也找不到，在谱系母本里同样找不到——它是在两份子本相撞的那一刻才出现的。

这个场景提出了本章的困惑。第12章、第13章讲的都是一个学生与一本子本，九问、改题、问对、打造子本，整条流程看起来都可以由一个学生与一个ChatSDE完成。若是如此，九问模式与一位水平很高的私人导师有什么区别？课堂为什么还要把几十个学生聚在一起？本章的核心判断是：只有一对一的线是高级个别辅导；课堂之为课堂，在于同学互撞，涌现个人说不出的判断。下面依次讲一对一的天花板、碰撞场的时点与组织、二阶碰撞方法怎样下沉课堂、班级子本怎样打造并接入谱系链、与知识建构的对话，以及跨学科碰撞的一个样例，最后把“没写下来的那一句”与默会知识划清界线。

> **这一章的问题**
>
> 如果每个学生都能和ChatSDE一对一走完九问，课堂为什么还要把几十个学生聚在一起?

◆ 一对一的天花板

先给一对一辅导一个公允的评价。布卢姆（1984）提出的“两个西格玛问题”指出，在他所报告的研究里，接受一对一辅导的学生，成绩明显高于常规课堂教学的学生，差距约为两个标准差；他由此提出的问题是，怎样找到一种在常规班级规模下也能接近这一效果的教学办法。此后关于辅导效果的研究有不少讨论与修正，一对一辅导的价值本身却很少被否认。大模型出现之后，这个问题有了新的答案：像可汗学院的Khanmigo这样的AI辅导，让一对一在成本上第一次有了大规模铺开的可能。九问模式里的问对，也是一种一对一。

这里须补上一笔修正。VanLehn（2011）在《教育心理学家》上，回顾了人类一对一辅导与各类计算机辅导的比较研究，指出人类辅导的效应远小于布卢姆当年报告的两个标准差，约为0.79个标准差。数字掉下来，一对一并没有因此失色，0.79在教育干预里仍是很大的效应；但它改变了本章论证可以怎样使用布卢姆。倘若以两个标准差为天花板，人们很容易说：一对一已经把能得到的都得到了，班级不过是出于成本的无奈折中。真实的上限既然低得多，剩下的差距就更可能出在“一对一在结构上给不出来的东西”，而不是“一对一还不够好”。因此本书不再借布卢姆的数字渲染一对一多么神奇，只借他提出的那个问题——怎样在班级规

模下得到辅导的效果——并指出它的答案不止一个：大模型把辅导的成本降下来是一个，同学互撞提供辅导给不出的第二块现场是另一个。还须留意这类研究的共同局限：它们汇集的多是以事先编定答案的测验为成绩的研究，衡量的是已知终点的学习，对“连老师也写不出终点的判断能否发生”几乎没有说话的余地。

2025年的两项研究把这一点推得更清楚。Kestin等（2025）在哈佛一门物理课里做了随机对照，发现精心设计的AI导师带来的学习收益高于课堂上的主动学习；Bastani等（2025）的高中数学现场实验（见第8章）则表明，护栏的有无决定了AI辅导会帮学习还是损害学习。把两者合起来读：一对一的辅导——无论由人还是由设计良好的AI来做——在课程既定内容的学习上确有实效，这正是本书愿意把讲授与辅导的发现一侧交给AI的理由；但两项研究的因变量都是既定内容的掌握，它们没有也不可能回答，两位带着不同现场的同学相撞，能否撞出两人都说不出的判断。碰撞场要守的，正是这一块一对一的实验设计没有覆盖的地方。

然而本书要指出的是，一对一有它的天花板，而且这个天花板不在辅导水平上，而在结构上。一对一的线，无论导师多么高明，参与者都只有两方：学生的现场与导师的知识。新判断只能从这两方之间长出来。ChatSDE的批判性激发可以把学生的前提翻过来看，可以逼近判断，可以亮出工序的依据，但它所能调动的另一方，始终是母本、精要库与它所接触过的材料；它没有第二个学生的现场。周川与ChatSDE再磨十轮，也磨不出林晓母亲交房租前的那几天沉默，因为那几天只在林晓的现场里。

用一个日常比喻来说：一位好厨师可以把一种食材做到极致，一个人在自家厨房里也可以把拿手菜反复改良；可是有一类味道，只有在两家的厨房偶然搭伙、你家的酱碰上我家的料时才会出现，这味道既不在你的菜谱里，也不在我的菜谱里。一对一的线，是一个人的厨房配了一位顶级的厨艺顾问；同学碰撞场，是许多人的厨房同时打开，彼此端着各自的半成品走来走去。顾问再好，也替代不了另一个厨房里的那一罐酱。

这一判断还有一个教育学上的旁证。维果茨基的“最近发展区”概念指出，学习者在更有能力的他人协助下能完成独立时完不成的任务，这个“他人”既可以是成人，也可以是能力更强的同伴。九问模式在这里取的是一个更窄的用法：碰撞场

里的同学之所以重要，主要不在于谁比谁更有能力，而在于每个人都带着一块别人没有的现场。周川与林晓在能力上大致相当，他们的碰撞之所以有产出，是因为饭馆的流水账与理发店的房租日历是两块互不相同、却在某处同构的现场。这是一对一在结构上给不出来的东西。

Chi与Wylie（2014）提出的ICAP框架，可以把这层意思说得更精确。该框架把学习者的投入方式分为被动、主动、建构、互动四种，学习效果依次递增；其中“互动”一级的要求，是两方都在生成对方没有的内容，并且这些内容相互推动，使双方得到单独一人得不到的东西。有人会据此反驳：学生与ChatSDE的问对，ChatSDE也会提出反问、翻转前提，难道不已经是“互动”？本书的回答是，就投入方式而言，问对确实可以达到“建构”乃至互动的上沿，ICAP并不能把一对一与碰撞场区分开；区分它们的是互动另一方带来的材料来自哪里。ChatSDE的反问来自工序与母本，同学的追问来自另一块只有他才摸得到的现场。ICAP回答的是学生怎样投入，而本章回答的是对方手里有什么，两者不在同一条轴上，碰撞场多出的一格就落在后一条轴上。

◆ 碰撞场：时点、组织与二阶碰撞

碰撞的时点要选对。太早，学生的子本还没有成形，手里只有九问与一些零散的初判，碰撞会退化成“互相介绍自己的题目”；太晚，子本已经定稿，学生会本能地守护自己的判断，碰撞会退化成“互相评审”。本书建议把碰撞场放在两处：一处在问对进行到中段、学生已有初步判断而尚未成本之时，一处在子本大纲完成、重写尚未开始之时（一学期的具体日程见第25章）。前一处的碰撞主要帮助学生发现自己判断的边界，后一处的碰撞主要用来撞出班级判断。周川与林晓的那一次，正是在第二处。

碰撞的组织要防两种退化。第一种是“汇报会”：每人上台讲十分钟自己的子本，台下提两个问题，下一个。这种组织看起来人人发言，实际上每一份子本仍是一条孤立的线，同学之间只有观看，没有相撞。第二种是“辩论赛”：把两份子本当作正反方，比谁的判断更有理。这种组织有冲撞，但冲撞的目的是分出胜负，最好的结果也只是一方被另一方说服，而不会出现两方都说不出的第三个判断。碰撞

场要的既不是展示，也不是胜负，而是让两份各自成立的判断在某个共同的结构上咬合，咬出一个新的东西。

具体的组织办法可以很朴素。老师在碰撞前读过全班的判断清单，按“现场不同、判断看似无关、却可能有共同结构”的标准把学生配对或配成三人组，这是老师在碰撞场里不可替代的工作，正如出题台上老师认得学生的现场一样。配对时尽量避开现场相同的组合：两个都以自家饭馆为现场的学生，撞出的多半是彼此的修补，而不是新判断；饭馆配理发店、配杂货店、配菜摊，现场不同而节律可能相通，才是好的配对。每一组碰撞时，先由各自用一句话说出承重判断，再互相追问对方现场里有没有自己判断所指的那种东西，最后合写一句“我们两人都没写过的话”。写不出也无妨，写不出本身也要记录。

ChatSDE在碰撞场里的角色要收窄。它可以在碰撞前帮老师比对判断清单，提示哪几份子本的判断在结构上可能相通；它可以在碰撞中充当记录者，把双方的追问与回答按轮次记下；它也可以在碰撞后对写出的那一句做批判性激发，问“这句话在你们俩以外的第三个现场里还成立吗”。但它不应当在碰撞中抢先说出那一句。理由与问对中“不替学生说破”相同，而在碰撞场里更为要紧：班级判断的价值，在于它是同学之间相撞的产物，如果是ChatSDE看了两份子本之后替大家总结出来的，它就成了一次一对多的辅导，碰撞场也就名存实亡。检验的办法依旧是问“核心判断谁先说”，碰撞记录里要能看到那句话最早出自哪位同学之口。

组织之外还有方法。周川与林晓的那一次碰撞，事后回看，有一个可以说清楚的结构。两人的判断各自成立，但都默认了一个没有说出来的前提：泄气与沉默的原因，要到某一样具体的东西上去找——周川找到账本，林晓找到房租。正是这个前提让两人起初都觉得说的是两件事，因为账本和房租确实是两样东西。碰撞时，林晓用周川自己的材料（月中进货付款那几天的轻微泄气）把这个前提撬开了：如果泄气不止落在月底，那么原因就不是账本或房租这一样东西，而是钱出去的那种节奏。新判断由此出现，它不是两人判断的平均，也不是其中一方胜出，而是两人的证据合在一起、却谁也没有单独说出的那一句。

这一结构可以写成课堂上可用的方法。在SDE学派的写作工作中，有一种被称为“二阶碰撞”的做法：把两三个领域或两三篇研究放在一起，不是取长补短，而

是找出它们共有而从未说出的那个前提，用其中一方自己的材料把它推翻，撞出一个各方单独都不会得出、但各方证据合起来只能得出的新判断。这原本是研究者撰写论文与专著时使用的方法；九问模式把它下沉到课堂，简化为三步：各说一句承重判断，找出两句话共同默认的那个前提，用对方的现场材料去试这个前提。三步的名称不必向学生强调，重要的是让学生习惯于问：“我们俩都以为理所当然的，是什么？”

二阶碰撞与一般的小组讨论有一个可以检验的区别。一般讨论的产物，通常能在参与者各自的观点里找到出处，是观点的汇总或折中；二阶碰撞的产物，在参与者各自的子本里找不到出处，用第12章的来源检验去查，它是双方都“所无”的。这就像两个人合伙做生意，汇总是把两家的货摆在一个摊上卖，折中是两家货各卖一半；而碰撞出来的东西，是两家都没有、却必须用两家的货才能做出来的一样新货。来源检验在这里同样是判别工具：如果一份“班级判断”经检查原来就是某一位同学子本里的原话，那么它只是被全班认同的个人判断，而不是碰撞的产物。

还要说明的是，碰撞不保证产出。现场不同、判断看似无关的两个学生配在一起，多数时候撞不出新东西，或者撞出的是一句漂亮而空洞的话，例如“小本生意都很难”。后者比撞不出更需要警惕，因为它看上去像一个班级判断，实际上任何人不下现场也说得出来。检验的办法借用题闸的思路（见第4章）：拿掉这两位同学的现场，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的这句话？能给出，就不算碰撞的产物，退回去再撞，或者如实记作“未撞出”。“小微经营者的动力被现金节律收走”之所以过得了这道闸，是因为它的支撑来自周川父亲月中与月底的泄气对比、林晓母亲交租前的沉默，这些都只在两人的现场里。

碰撞的单位也值得斟酌。两人一组最容易推进，因为追问可以一来一回地进行，谁先说了什么也最容易记录；三人一组的好处是多一块现场，常常能把两人之间的共同结构再试一次，坏处是容易出现两人对话、一人旁观的局面。本书的建议是以两人组为主，在两人组撞出一句之后，再请第三位现场不同的同学加入，专门做“试错者”：用自己的现场去试那一句，看它在第三块现场里是成立、需要限定还是根本不成立。周川与林晓撞出“现金节律”之后，老师请一位家里开菜摊的同

学加入，她说自家菜摊按天结算，没有月底，泄气也就不落在月底，而落在批发市场涨价的那几天。这一试没有推翻“现金节律”，却给它加了一条边界：节律的周期由什么决定，因经营方式而异。这条边界后来写进了班级子本的第四部分。

◆ 班级子本与谱系链

碰撞撞出的判断，要落成文本，才能进入评估与回写。本书把碰撞场中涌现的、个人说不出的判断所成之本，称为班级子本。班级子本的打造方式与个人子本相似：先整理碰撞记录，列出班级判断清单，再长大纲，后重写；判别标准也相同，须含母本与各成员个人子本都没有的判断。不同之处有两点。其一，班级子本的撰写者是参与碰撞的几位同学，大纲要由他们共同商定，每一部分标明由谁撰写、依据谁的现场；其二，班级子本里的每个判断都要标明它在碰撞记录里的出处——哪一轮、谁先说出、谁的材料撑住——这是评估第⑦项碰撞贡献的依据（见第15章）。

周川与林晓那一组的班级子本，骨架就从“现金节律”这个判断展开：第一部分摆出周川父亲泄气的时点与流水账上现金流出的时点，第二部分摆出林晓母亲沉默的时点与交租、进货的时点，第三部分讨论两者共同的结构——经营者的劲头在现金流出的日子之前被预先收走——以及这一结构与周川原判断“动力被账本提前收走”的关系。第四部分写边界：这是两家小店、两段构造的观察，属于待检验的班级判断；它是否适用于现金流出节律不同的经营者，例如按日结算的菜摊，留作下一届的问题。班级子本里还如实记下一个分歧：周川认为“账本”是现金节律被经营者看见的那个形式，不应被“节律”完全取代；林晓则认为账本只是记录，真正收走动力的是节律本身。这个分歧没有被强行统一，两种读法并列写入。

班级子本在谱系里处在一个中间位置，于是形成一条链：谱系母本→班级子本→个人子本。母本是一门课的出发点，个人子本是每个学生从母本出发、经过九问在自己现场里长出的书，班级子本则是若干个人子本相撞后长出的、比任何一本个人子本都更接近谱系层的判断。这条链有两个方向。向下，班级判断会回到个人：周川在定稿自己的子本时，用“现金节律”重新读了一遍自己的材料，补上了月中进货那几天的分析，他的个人子本因此比碰撞之前更厚；但他在子本里如实标注，这一层读法来自碰撞，不是他独自发生的。向上，班级判断比个人判断更有可能进

入母本精要库，因为它已经在两个现场里得到了初步支持，是谱系化走出的第一步（三层回写见第24章）。

班级子本还面临一个个人子本没有的问题：班级是会解散的。一学期结束，周川、林晓与那位开菜摊的同学各奔东西，班级子本若只存在课程档案里，就会像第13章开头许宁担心的那样，成为一本放上书架的书。九问模式对此的处理，是让班级子本以两种方式延续。其一，它的待检验问题进入下一届九问的出题素材：“按日结算的菜摊，泄气落在哪一天”这样的问题，可以由老师转成下一届某位学生的一问。其二，参与者各自把班级判断连同出处标注一起带进个人底盘，日后在新现场里调用时，记得这一层读法来自谁的现场、由谁先说出。这样，班级虽散，它撞出的判断仍在谱系里有下落。

把这条链说成“链”而不是“层级”，是为了避免一种误解：班级子本并不比个人子本更高明，个人子本也不是班级子本的原料。两者是不同性质的产物。个人子本的价值在于它把一个人的现场挖到了底，班级子本的价值在于它在几个现场之间找到了共同结构。就像一张地图集，每个县的分县图画得越细，全省图才有可能画对；全省图画出来了，回头再看某个县，又会发现县与县交界处原先没注意的一条河。没有一张分县图是为全省图服务的附件，全省图也不能代替任何一张分县图。

◆ 与知识建构的对话：多出的一格在哪里

班级子本最近的邻居，是Scardamalia与Bereiter（2006）在《剑桥学习科学手册》里系统阐述的知识建构。这一传统把课堂看成一个共同体：学生对观点的持续改进共同负责，观点被当作可以被检视、被改造的真实对象；共同体的贡献沉积在一个社区数据库（早期的Knowledge Forum是其典型形态）里，后来者在先前的贡献上继续写；而当共同体把若干分散的观点合成一个更高层次的说法时，称为“超越”（rise-above）。班级子本与它有明显的同构：班级判断超越了周川与林晓各自的判断，班级子本又像一个小型的数据库，把撞出的判断与出处留给下一届。承认这一点并不丢人，本书的部件大多有来历，主张落在组合与证据责任上。

在此基础上，本书说的“多出一格”有三处，且须分开说。其一是上游：知识建构强调学生是认识论上的主体，观点的走向应由共同体自己决定，教师的介入宜轻。本章的碰撞场则由老师在碰撞前读全班的判断清单，按“现场不同、判断看似无关、却可能有共同结构”配对。这是一个有成本的取舍：它牺牲了一点自发性，换来的是碰撞的材料确有差异，不至于撞成彼此的修补。其二是闸口：知识建构里“超越”的好坏，主要靠共同体对观点的批评与改进来把关，一个看起来高于各方、实则谁都说得出来的说法，未必会被拦下；本章要求碰撞产物过题闸，即拿掉这几位同学的现场，一个未接入这些材料的通用大模型能否给出说得过去的这句话，同时过来源检验，看它是否原本就是某位成员的原话。“现金节律”之所以立得住，是因为它同时通过了这两道，而“小本生意都很难”通不过。其三是账：社区数据库记录贡献的积累，班级子本则要求每个判断带出处——哪一轮、谁先说出、谁的材料撑住——并带版本史与撤回记录，被后来的现场推翻的判断留下被撤回的痕迹，而不是悄悄消失。这第三处是证据责任上的差别：一个数据库可以只增不减而显得越来越丰富，一份有撤回记录的班级子本却能让人看到，哪些判断是在什么证据下被放弃的。

这个对话也带来一条自我限制。如果有一天对比表明，不设上游配对、不设闸口、不记撤回的普通知识建构课堂，其“超越”性的说法在来源检验与题闸检验下并不更差，那么上述三格就只是仪式。这一条列入本章末的失败条件。

◆ 跨学科碰撞：没写下来的那一句

同一门课里的碰撞，现场不同而学科相同，共同结构往往比较容易浮现；跨学科碰撞的难度更大，产出也更难预料。下面讲一个构造的样例。某校在学期中段组织了一次跨院系的碰撞场，各门九问课程推荐一两位学生，带着子本大纲参加。化学系的沈澄与临床医学院的许宁被分在了一组。组织者这样配对，是因为在两人的判断清单里读到了同一种气味：沈澄的判断是，本组某一步的成败依赖未入规程的环境湿度与操作时长——规程里没写的那一句，在师门口头里传了十年；许宁的判断是，本病房夜间跌倒的风险更多落在交接断点上——风险藏在交班本里被省略的那一句。

跨学科碰撞与同学科碰撞在组织上有一处不同：它需要一个“翻译层”，但翻译的不是判断，而是现场。沈澄说“失败批次”“谱图里的小峰”，许宁说“交班本”“护理记录”，这些词对对方来说都是陌生的。碰撞前，组织者请两人各用五分钟、不用本学科术语，讲清自己现场里的一件具体事：沈澄讲的是三月与六月同一瓶试剂用出不同产率，许宁讲的是夜班里一次跌倒前后交班本上写了什么、没写什么。这五分钟的作用，是把两块现场摊成对方能摸到的样子，就像两个说不同方言的人做生意，先把货摆上桌，比先对词表要快得多。没有这一步，跨学科碰撞很容易滑向两种退化：要么各讲各的术语，谁也听不懂；要么急于找共同点，把两边的判断都抽象成“细节很重要”这样谁都能说的空话。

碰撞开始时，两人都觉得对方的世界离自己很远。沈澄的现场是实验记录本、失败批次与谱图，许宁的现场是交班本、护理记录与夜班；一个面对的是反应瓶，一个面对的是病床。按碰撞场的步骤，两人各自说出承重判断之后，去找两句话共同默认的前提。他们发现，两人起初都把“没写下来的那一句”当作一个记录上的疏漏：沈澄以为规程漏写了湿度，补上就好；许宁以为交班本漏写了一句，提醒大家写全就好。两人的判断都默认了一个前提——书面记录本应是知识的完整载体，缺了的部分是偶然的遗漏。

撬开这个前提的，是彼此的材料。许宁问沈澄：那一句为什么十年没写进规程？沈澄答：因为师兄师姐都“知道”，口头说一句“那一步要快”就够了，没人觉得需要写。沈澄反问许宁：交班时省略的那一句，是谁、在什么情况下省略的？这正是许宁提升过的第2问。许宁说，她跟夜班时注意到，被省略的常常是交班者以为接班者“都知道”的事。两人在这里停住了。没写下来的那一句，在实验室和病房里都不是随机的遗漏，而是在书写者以为读者已经知道的地方，系统地出现。差别在于，实验室里那一句被口头传统接住了，传了十年，只是在人员轮换、季节变化时会断；病房里那一句在交接的忙乱中，有时没有被任何东西接住。

从全书的三领域对应来看，这次跨学科碰撞也有它的位置。化学的现场侧重知识发生的客体属性，湿度、时长、谱图都是物的读数；医学横跨三者，交班本上的省略属于互动属性，是一班人与另一班人之间的事。两人撞出的结构恰好落在客体与互动的交界上：实验室里那一句没写下来的话，表面上关乎物（湿度），实际上

是靠人与人之间的口头传承维系的；病房里那一句被省略的话，表面上关乎人（交接），实际上关系到患者身体在夜间的状况。这提示跨学科碰撞的价值，未必在于两个学科“找到相似之处”，而在于各自学科里原本被归到另一种属性去的东西，在对方的映照下被重新看见。沈澄此前没有把师门口头传统当作自己课题的一部分，许宁此前也没有把交班本当作一种需要被读懂的“文本”。

两人合写的那句话是：“没写下来的那一句”是跨实验室与病房的共同结构——书面记录在书写者以为“对方已知”之处留下空白，风险与成败就落在这个空白能否被别的东西接住上。这句话在沈澄的子本里没有，在许宁的子本里也没有；它不是化学的判断，也不是医学的判断，而是两个学科的现场合在一起才看得见的一个结构。它同时给两人各自的工作带回了新的问题。沈澄回去后给自己的九问追加了一层追问：口头传统在哪些情况下会断——新生入组、老生毕业、换了实验室？许宁则在自己的问对里写下：被省略的那一句，是不是集中在交班者自认为接班者已知情的事项上？这一问后来成了她住院医阶段底盘里的那条限定。

这次跨学科碰撞的成果，必须带着严格的边界。它是一个化学课题组与一个病房、两段构造的观察所撞出的结构性假说，在两个现场里各自仍待检验：沈澄的湿度假说待复做验证，许宁的交接断点判断是本病房假说，已交科室讨论。许宁在碰撞中只使用了去标识化、经科室与伦理委员会许可的材料，没有带入任何可识别患者的信息；碰撞的结论也不构成对病房交接或临床处置的任何建议，更不改动医嘱，是否以及如何在科室里检验，由科室决定。沈澄方面，碰撞只讨论记录与规程的关系，不涉及任何危险操作的细节，也不改动任何实验步骤。跨学科碰撞最容易出的错，就是把一个在两个现场里初现的结构，说成了适用于一切实验室与一切病房的规律；本书在这里只能说，这是一个值得下一届在更多现场里检验的问题。

◆ 与默会知识的划界：这一句算不算“默会”

读到这里，熟悉知识管理的读者大概已经按捺不住：“没写下来的那一句”不就是默会知识吗？Polanyi（1966）说“我们所知多于所能言”，Nonaka与Takeuchi（1995）把默会知识与显性知识的转化排成SECI四步，上一章讲子本进入底盘时已经用过。照这个框架读，沈澄师门十年的口头传统是典型的共同化，把湿度与操作时长补进规程则是外化。这一联系应当承认，而且承认得彻底些：

Nonaka与Takeuchi本人就指出，外化常借助隐喻、类比与对话，沈澄与许宁隔着学科一来一回的追问，正是一次借对话完成的外化。本书并不发明一个新概念去替代这些旧概念。

需要划出的界线，在于“这一句”到底在说什么。默会知识说的是一种认知状态：知道的人说不全，不是不愿说，而是语言够不着，比如一位老师傅凭手感判断“那一步要多快”，你让他写下来，他写出“快”字便写不下去了。本章的结构判断说的却是空白出现的位置：书面记录的空白，系统地出现在书写者以为读者已经知道的地方。这是关于书写者怎样想象读者的判断，而不是关于知识能否被说出的判断。沈澄的师兄师姐完全说得出“那一步要快”，只是“都知道”，没人觉得需要写；这一类是“能说而未写”，严格讲并不是默会的。至于“快到什么程度”，说到最后说不清的那一部分，才是Polanyi意义上的默会。所以同一个没写下来的句子里，往往同时含有两层：上层是以为对方已知而省略的，可以靠追问补成文字；下层是确实说不尽的，补写的结果只能是一个近似。

这一划界给出一个可操作的分流办法。碰撞时请书写者当场补一句，限定在一分钟之内：补得出的归“以为已知”，补出来后仍说不到位、需要示范或在场才传得了的归“说不尽”，写不出又找不到理由的归“遗忘”。三类的处置各不相同：第一类只要把读者的假定改成显性，补进规程或交班本就够了；第二类补写无济于事，要靠示范、同班并肩或设置一个接住它的环节；第三类属于一般的疏漏，与本章的结构判断无关。我们并不预设沈澄与许宁的那一句属于哪一类，这正是需要数出来的东西。如果数出来的结果是，空白绝大多数落在“说不尽”类，那么本章在这里不过是把Polanyi的结论换了一套病房与实验室的材料说一遍，所谓“多出的一格”就应当撤回。

还有一层区别在于用途。外化的目标是让知识可传递，SECI关心的是组织怎样把个人的默会转成组织的显性。本章的结构判断把方向反过来，用在读者一侧：既然空白会集中出现在书写者以为已知之处，那么读一份记录的人，就可以带着“哪一处是写的人最可能觉得不必写的”去找空白，而不必等事故发生之后才知道哪里缺了。许宁后来在自己的交班里回头问“我这一份省略了什么”，用的就是这个方

向；它是一种读法，不是一种转化。这一格与默会知识理论并不冲突，只是落在它们没有专门讨论的位置上。

◆ 本章怎样算错

本章的核心主张是：课堂之为课堂，在于同学互撞，能涌现个人说不出的判断；只有一对一的线，无论多高明，都是高级个别辅导。这一主张是比较性的，它的反面不是“碰撞没有用”，而是“碰撞的产出可以由一对一替代”。本章的案例同样是构造的，碰撞的实际增益需要在真实课堂里以换手评分来检验。下面四条，只要有一条在规定的样本与时间窗里成立，对应的论断就要收回。

第一条检验产出是否真是涌现。连续两届、每届不少于30名学生、不少于12个按本章办法组织的碰撞组，对全部班级判断做来源检验与题闸检验；若两项检验中任何一项的未过比例达到一半以上——即一半以上的班级判断在某位成员的个人子本里找得到原话，或拿掉现场后一个未接入材料的通用大模型（记录所用版本）就能给出——则“涌现个人说不出的判断”没有发生，碰撞场只是个人判断的展示与汇总，本章主张撤回，课堂的价值需要另找依据。

第二条检验碰撞相对一对一的增益。同一门课、同一届里，随机抽出不少于10名学生不参加碰撞、只与ChatSDE一对一深度问对，其余学生参加；若换手评分下，两组子本中的新判断不相上下，并且班级子本在整学期内没有带来任何可观察的向上回写（进入母本精要库）或向下回写（个人子本因碰撞而加厚），则碰撞场的增益不成立，“课堂之为课堂”的论断要降级为“碰撞可选”。

第三条检验跨学科碰撞。至少6组跨学科碰撞、由两个学科各自独立的检验者在随后一学期的各自现场里复查：若撞出的“共同结构”在两个现场里都能找到对应证据的不足三分之一，其余只在一个学科里成立，则跨学科碰撞产出的更多是比喻而不是结构，本章对跨学科碰撞的期待收紧为“启发问题”，不再主张“共同结构”。

第四条检验与默会知识的划界。对跨学科碰撞中涉及书面记录空白的判断，按前述分流办法取不少于10处省略项逐项归类；若“说不尽”类超过半数，则“没写下来的那一句”的结构判断不再构成默会知识之外的多出一格，本章在这里只作

Polanyi论断的又一份材料。与此并行，前述与知识建构的三格比较若在来源检验与题闸检验下不显示差别，也同样撤回。

◆ 判断撞出来之后，怎样被计分

到这里，第四编铺土壤的工作告一段落：子本被重写出来，进入个人底盘，在同学碰撞场里又长出了班级子本。可是本章留下了一个评估上的难题。周川在班级子本里如实标注，“现金节律”这个判断最早由林晓说出半句、由他接上；他的个人子本也因碰撞而加厚。那么在评分时，这份加厚算谁的？林晓的那半句，算她的碰撞贡献，还是也要计入周川的创新智商？沈澄与许宁的跨学科碰撞，产出的是一个双方都还在检验的结构性假说，它在两人各自的成绩里应当怎样体现？

这些问题不能靠一个总分回答。一个学生可能现场证据极厚而碰撞贡献很弱，另一个学生可能自己的子本平平却在碰撞场里撞出了关键的一句。九问流程的每一段——母本、改题、问对、现场、子本、碰撞、回写——都留下了不同的证据，也就需要不同的评估项目来承接。第五编将转向学业评估，第15章先讨论学业成绩的结构：八个评估项目怎样与九问流程逐段对应，怎样组合成一张发生剖面，以及沈澄与周川的剖面会长成什么样子。

> **带走的话**
>
> **一对一的辅导再好，对面也只有母本和工序；两块不同的现场相撞，才撞得出谁也说不出的那一句。**

◆



第　五　编

学业评估：多项目与权重组合

编　序

期末评分会上，两位老师各拿着一份成绩单：沈澄和周川总分都在75分上下，旧制度里同属“良”。可一个把反应从头复做，每一步记得清清楚楚，却几乎没在碰撞场里开过口；另一个把题改得更近，问对一轮比一轮逼人。一个数字把两个人抹平了。本编要问的是，这门课的成绩该怎样如实地说话。

第15章提出学业成绩是八个项目加一套权重组成的发生剖面。第16章讲核心项，子本创新智商以五维、160分制来量，完成度作门槛，评分须换手。第17章讲过程项，改题、问对、现场证据按质不按量计分，轨迹是不可还原性的证据，拿不出轨迹的高分子本要被追问。第18章讲权重怎样随学段、课型、学科变化，并在区间内交给学生自选。

先有剖面，再有量尺，再有过程里的证据，最后才是权重。次序不能颠倒：权重先定，剖面就成了凑分。

15

第15章

学业成绩的结构：八项目与发生剖面

◆ 两张总分相近的成绩单

期末评分会上，化学系的任课老师和经济管理学院的任课老师恰好坐在一起，各自拿着一份学生成绩单。化学这边是沈澄，《合成条件的发生》课程，总分75分上下；经管那边是周川，《小微经营者的动力与节律》研讨课，总分也是75分上下。按旧的成绩制度，两人都落在“良”这一档，教务系统里两个数字并排放着，看不出任何差别。可是两位老师都知道，这两个“良”几乎没有一处相同：沈澄把一批反应从头复做，每一步的时间、温度、湿度都记得清清楚楚，子本写得规整，却几乎没有在同学碰撞场里开过口；周川把“哪个伙计干得最起劲”改成“父亲为什么一到月底就泄气”，问对一轮比一轮逼近，家里三年的流水账翻得透透的，可子本交上来时论证还没收拢，母本口试也答得磕磕绊绊。两个数字一样，两个人在发生上走过的路完全不同。

这个场景提出的困惑并不新鲜，但在九问课堂里变得尖锐。讲授课堂的总分之所以还能凑合用，是因为那里的学习只有一条考核线：掌握同一本书，考卷量你离标准答案有多远，一个数字大体就说清了“离多远”。九问课堂把一个学期拆成了母本、九问、改题、问对、现场、子本、碰撞、回写这样一条发生链，每个学生在链上的强弱处各不相同。把整条链压成一个数字，就像体检报告只写一句“健康指数75”，血压高还是血糖高、该看心内科还是内分泌科，全都看不出来。病人拿着这个数字既不知道哪里要改，医生也没法据此开下一张处方。

本章的核心判断是：学业成绩是八个项目加一套权重组合成的发生剖面，不是一个总分。总分仍然要算，因为学校需要排序、需要授予学位、需要一条及格线；但总分只是剖面的一个投影，成绩单的主体应当是剖面本身。下面先把八个项目逐一对上九问流程，再说剖面怎样读，然后讨论评估方向为什么必须翻转，再给出成

绩单的新格式，并说明它与档案袋评价等既有做法的关系，最后回到沈澄与周川的两张剖面，讲清剖面自己的限度。

> **这一章的问题**
>
> 两个学生的总分都是75分上下，却几乎没有一处相同，成绩单为什么不该只交出那个总分？

◆ 八个项目怎样对上九问流程

九问模式是S→D→E这一条发生路径在课堂中的实现：谱系母本精要→九问→改题→问对→打造子本→评估→回写（导论与第1章已有交代）。八个评估项目不是另外拼出来的一张清单，而是沿着这条路径一段一段取证。S段是立骨，对应第①项母本底盘，量学生对谱系母本精要与核心判断的掌握，证据是复述加反例口试；在大学阶段它是门槛项，不是得分主力。S到D的转折处是改题，对应第②项改题质量，量学生改动九问时是在提升还是在逃离，证据是题的版本史（见第5章）。D段是问对，对应第③项问对深度，量轮次之间有无推进、接不接得住批判、关键判断是否由学生先说，证据是问对记录（见第6、7章）。

E段是下到自己的现场，对应第④项现场证据，量观察、数据、访谈的真实与厚度，证据是田野记录、实验记录与原始材料。之后进入产物段：第⑤项子本创新智商是全部八项的核心，按S、D、E、I、F五维评定，满分160，再按档位分段映射到百分制；第⑥项子本完成度量论证是否完整、引注是否无误、结构是否成形，化学类课程另加可复做性。协作段对应第⑦项碰撞贡献，量学生在同学碰撞场中有没有撞出班级判断，证据是碰撞记录与班级子本（见第14章）。回写段对应第⑧项回写证据，量学完之后学生自己被改写了什么、有没有内容贡献进母本精要库，证据是第零轮初判与期末自述的对照，以及精要库的采纳记录。

把这八项排成一列，可以看到一个容易被忽略的结构：过程项（②③④）与产物项（⑤⑥）并不重复计分。产物项看的是子本这件成品本身的质地，过程项看的是这件成品是怎样一步一步长出来的。打个日常的比方，评一道菜，可以只尝味道，也可以连厨房一起看：食材是不是当天买的，火候是不是自己掌握的，中途试味改了几次。只尝味道的评委分不清现做的菜和买来的预制菜，连厨房一起看的评

委才分得清。九问课堂两样都要，原因在第17章还会细说：过程留下的轨迹，正是产物不可还原的证据。

还有两项要单独交代。第①项在研究生课里权重为零，但它仍然存在，作为门槛：母本口试不过，后面七项不予评定，因为一个连谱系母本的核心判断都复述不出来的人，他的“新判断”是否新，根本无从核对。第⑧项在各种组合里权重都只有5%，看起来分量轻，却是整条链能否闭合的检验。按本书的口径，123原理所说的循环是D—E矛盾逼出S改变、S再回写D与E；回写如果只被宣布而不被观察，这个循环就只走了一半。第⑧项的权重小，是因为它难以在一学期之内充分显露；它必须存在，是因为没有它，评估就退回到只看一次性产出。

八项之间的分量并不平均，这一点由几条配套规则共同决定。第一条，大学阶段第⑤项的权重不低于25%：在第18章给出的各组合里，本科研讨课取30，研究生课取40，应用实践课取25，只有大一补练手阶段取20，因为那时的学生还在补中学本该练熟的发生技术，这一列的由来在第25章交代。第二条，第③项与第⑤项必须换手评分并写利益冲突声明，因为这两项最容易被同一套智能体既参与生成又参与评定。第三条，全部八项按质不按量，问对轮次、子本页数、现场照片张数都不直接计分。这三条规则合起来，给八项定了一个骨架：创新智商居中承重，过程项围绕它提供可核对的来历，门槛项与回写项守住两端。骨架定下之后，具体的数字才交给学段、课型、学科与学生自选去调整，第18章专讲这一步。

◆ 发生剖面：长在哪里，缺在哪里

八个项目各自给出一个百分制分数，再按权重加权，得到总分；而把八个分数按流程次序排开，得到的就是发生剖面。剖面可以画成折线，也可以画成雷达图，本书建议优先用折线，因为八个项目本身有先后次序，折线能保留“从S走到回写”的方向感，雷达图会把这个次序抹平成一圈。读剖面不是读最高点和最低点那么简单，要读的是形状：哪一段隆起，哪一段塌陷，塌陷是发生在链条的前端、中段还是末端。形状比单个分数更能告诉老师下一学期该从哪里下手。

一个日常的对照能说明形状为什么重要。体检时，心率75次每分钟，这个数字本身很正常；可心电图上同样平均75次，有的人波形整齐，有的人时快时慢、中间

夹着早搏。只看平均心率的医生会放行两个人，看波形的医生会把第二个人留下来再查。总分就是那个平均心率，剖面就是那张心电图。两个总分都是75的学生，一个可能是八项都在75上下的“平台型”，另一个可能是现场证据90多、碰撞贡献40多的“陡峭型”；前者下一学期要找的是哪一项能先拔起来，后者下一学期要补的是那一处塌陷。同一个总分，开出的是两张完全不同的处方。

在九问课堂里，剖面有几种常见形状，值得老师先熟悉。第一种是“前高后低”：母本底盘很高，改题和问对平平，子本创新智商偏低，这往往是讲授课堂里的好学生刚进九问课堂时的样子，复述能力很强，却还不习惯把题改离母本、把判断带到自己的现场。第二种是“中间高、两头弱”：改题、问对、现场都走得很深，可母本底盘薄、子本完成度低，回写证据也拿不出多少，这样的学生往往发生得很真，却没把发生收拢成一件立得住的成品。第三种是“产物高、过程空”：子本创新智商与完成度都很高，改题版本史、问对记录、现场材料却稀薄或对不上，这是本书要求老师特别警惕的形状，第17章将专门讨论。第四种是“过程厚、产物平”：轨迹扎实，现场材料丰厚，子本却没有撞出母本所无的判断，常见于勤奋而尚未学会取舍的学生。

需要说明的是，这几种形状是读剖面的辅助说法，不是给学生贴的类型标签。给学生贴类型恰恰是本书第2章所说已经被大模型接管的发现一侧的工作，做得快、做得全，却把人固定在一个格子里。剖面描述的是这一个学期、这一门课、这一份九问上走出的路，下一学期换一份九问，形状完全可能翻过来。所以成绩单上只记剖面本身，不记“你是某某型”的结论；形状的名字只在老师与学生面谈时作为对话的抓手使用，用完即放下。读剖面的目的是找到下一步，而不是给出一个关于学生是谁的定论。

剖面与权重的关系也要先说清，因为两者常被混为一谈。剖面是八个未加权的原始分数，记录的是学生实际走出的形状；权重是学校、课型、学科以及学生自己对这条路的哪几段更看重所作的选择。同一张剖面，放进本科研讨课的权重里算出一个总分，放进应用实践课的权重里会算出另一个总分：现场证据在前者占15，在后者占25，一个现场特别厚的学生在应用实践课里总分会更高。这好比同一份体检数据，交给心内科医生和交给运动医学医生，两人关心的指标不同，给出的综合判

断也不同，但体检数据本身不会因为看的人不同而改变。所以成绩单上剖面写在前、权重写在后，剖面是事实，权重是对事实的一种读法；把这个次序弄反，就会出现为了总分好看而反过来去“做形状”的倾向。

剖面还承担一项旧总分承担不了的职能：让评分本身可以被讨论。总分是一个结论，学生对它的异议只能是“我觉得应该更高”；剖面是八个带出处的分数，学生的异议可以具体到“第③项的换手评分人没有读到我第5轮回到病房之后补写的那一段”。异议一旦具体，就可以核对，核对之后要么改分、要么说服学生，两种结果都比含混的申诉更有教育意义。从这个角度说，剖面不只是给学生的一张报告，它也把评分这件事本身变成了一次可以被批判、可以被修订的判断，这与九问课堂要求学生对待自己初判的态度是同一种态度。

◆ 评估方向的翻转：从“像不像答案”到“离已有有多远”

剖面之所以必须取代总分，还有一个比信息量更深的理由：九问课堂的评估方向翻转了。讲授课堂的考试，量的是学生的回答“像不像答案”，越像越高分；标准答案事先写好，阅卷就是比对，比对得越严，评估越公正。这个方向在统一性知识的传递里是合理的，因为课堂要交付的正是那本人人相同的书。第1章已经讲过，老师能事先写出终点的再发生是练手，应在中学完成；大学是战场，连老师也写不出终点。终点写不出来，“像不像答案”这把尺子就失去了刻度，因为根本没有那份答案可以去比。

战场上的评估只能换一个方向，去量学生的产物“离已有有多远”：像母本，像精要库里已收的判断，像全球文库里已经被说过的话，像一个未接入这些材料的通用大模型在拿掉学生现场之后张口就能答出的内容。越像已有，创新智商里的不可还原性越低。这个方向与第4章的题闸是同一把刀：出题时问“拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案”，评估时问“拿掉学生的现场与轨迹，这份子本里的判断能否被已有材料重新生成”。出题台把住入口，评估把住出口，两处用的是同一个检验，课堂的前后才连成一条线。许宁关于“交接断点”的判断之所以值得给高分，不是因为它“正确”，而是因为任何老年跌倒防治指南里都写不出本病房交班本里被省略的那一句。

但翻转不是全面倒转，这一点必须讲清，否则评估会滑向“为不同而不同”。八个项目里，第①项母本底盘量的仍然是“像不像”：复述得准不准，反例举得对不对，这是发现一侧的底盘，发现没有作废，只是降为底盘，这是第2章的判断。第⑥项子本完成度量的也是“够不够规范”：论证是否完整，引注是否无误，化学子本能否被别人复做。真正翻转方向的是第⑤项，以及在较小程度上的第②③⑦项。于是一张剖面里同时存在两种方向的尺子：一种量学生有没有站稳在已有之上，一种量学生离开已有走出了多远。这就是剖面不能被压成单个总分的结构性理由——两种方向的分数相加，得到的数字既不表示“站得多稳”，也不表示“走得多远”。

用学书法作比。初学临帖，评的是像不像帖，笔笔有来历；入帖之后要出帖，评的是有没有写出帖里没有的东西。一个只会出帖、从未入帖的人，写出来的往往只是乱涂；一个只会临帖、从不出帖的人，写得再像也只是复印。好的书法老师会分开看这两件事，再决定下一阶段让学生多临还是多写。九问课堂里的母本底盘与子本完成度对应“入帖”，子本创新智商对应“出帖”，改题与问对是从入帖走向出帖的那段路。评估方向的翻转，是在“出帖”这一段上翻转，在“入帖”那一段仍保留旧尺子。防止学生为了显得新而丢掉根，还有两道约束：改题时的两锚之一“根还在”，以及创新智商里的可证伪性F，它要求新判断能说清在什么情况下会被推翻。

◆ 成绩单的新格式

成绩单要把剖面交给学生、交给下一位老师，也要交给学校的学籍系统。本书建议的新格式由五部分组成。第一部分是课程信息：课程名称、学段与课型、学科、所用权重组合，以及权重是否由学生在区间内自选（若自选，附理由书编号）。第二部分是门槛状态：母本口试是否通过，医学课程的伦理门槛（去标识化、许可齐全）是否满足，化学课程的安全门槛是否满足；任何一道门槛未过，整份不计分或后续项目不予评定，而不是扣几分了事。第三部分是八项剖面，每一项给出百分制分数、权重和证据索引。第四部分是总分与创新智商原始分。第五部分是评分程序声明与一句面向下一步的评语。

证据索引是这张成绩单与旧成绩单最大的不同。旧成绩单上的“85分”是一个不需要出处的数字；新成绩单上每一项分数都要指向一份可调阅的材料：改题质量指向题的版本史，问对深度指向问对记录，现场证据指向实验记录或田野笔记，碰撞贡献指向碰撞记录与班级子本，回写证据指向第零轮初判与期末自述。这就像审计报告的每一个数字都要附底稿编号，审计师说“存货账实相符”，必须能翻出盘点表来。一个分数一旦有了出处，学生可以复核，下一位老师可以接着读，学校也可以抽查评分是否一致。第⑤项与第③项还须注明换手评分的情况与利益冲突声明，其做法见第16、17章。

以沈澄为例，下面是一张按本书建议格式写成的成绩单示意。表中数据为构造，只作说明。

项目	权重（化学·本科）	得分	证据索引
① 母本底盘	5	80	母本口试记录（复述＋反例）
② 改题质量	10	75	九问版本史（老师版→学生改版→定版）
③ 问对深度	10	72	问对记录（含第零轮），换手评分
④ 现场证据	15	92	复做批次实验记录（时间、温度、湿度逐步记录）
⑤ 子本创新智商	30	70	子本；换手评分；原始分约120（本科分档线110）
⑥ 子本完成度	15	90	子本（含可复做性核查）
⑦ 碰撞贡献	10	45	碰撞记录、班级子本
⑧ 回写证据	5	70	第零轮初判与期末自述对照
总分	100	75.0	安全门槛：已满足

这张表的权重取自第18章给出的化学本科示例，加权总分为75分整。在表下，成绩单还应附一句评语，例如“现场与完成度已成熟；下一学期建议在碰撞场中先把复做数据拿出来给同学撞”，这句话就是剖面开出的处方。还要说明一点：成绩单上的总分保留，是为了服务学籍管理与学位授予，它的地位从“成绩本身”降为“剖面的摘要”。学生、导师、用人单位若只看总分，等于只读了摘要不读正文，这是他们的选择，但成绩单本身必须把正文交出来。

新格式上还有几样东西是刻意不写的。它不写问对的轮次总数，不写子本的页数，也不写学生在ChatSDE上花了多少小时，因为按质不按量是本书评估方案的配套规则，把量写上成绩单，就等于暗示量有分量，学生很快就会去刷量。它也不写班级排名百分位，因为剖面要回答的是“这个学生往哪里走”，排名回答的是“这个学生在别人中间站在哪里”，后者在九问课堂里意义有限：每个人的九问是按自己的画像出的，现场各不相同，把子本排成一列，就像把各家各户的家常菜排一个名次。学校若出于管理需要仍要排名，可以由总分另行生成，但不进入交给学生的那张剖面。

多个学期的成绩单排在一起，还会出现单张成绩单看不到的东西：剖面序列。一个学生第一学期“前高后低”，第三学期变成“中间高、两头弱”，第五学期变得较为平整，这条序列本身就是第⑧项回写证据最有力的旁证——它显示学生在发生上被改写了。学籍系统若只存总分，这条序列就丢了；存剖面，序列就自然保留下来。这也是第13章所说“大学与人的关系变为终身”在评估上的落点：剖面序列跟着学生走，成为他个人知识发生平台的一部分履历。

◆ 剖面不是新发明：与档案袋评价和ICAP的对话

在评价研究里，用一个数字压住一条学习路径的不安由来已久，档案袋评价（portfolio assessment）是其中最成形的回应：把学生一学期的作品、草稿和反思收进同一个夹子，让评价者读到的是一段成长，而不是一场考试。八项剖面与它同向，都承认过程材料本身就是评价的正文。本书不宣称自己发明了过程性评价；如果有读者把本章读作“给档案袋加了八栏表格”，这个读法有一半是对的，另一半要靠下面三处收紧来区分。

档案袋评价长期被两件事困扰。一是放什么由学生挑，挑选本身就是一次展示，评价者读到的是学生愿意让人看见的部分，就像相亲时递上去的照片，总是挑拍得最好的那几张。二是各人夹子里的东西不同，评分者之间缺少共同的锚，评分的一致性容易成为争议。本书的做法在三处收紧。第一，证据索引把每一项分数钉在一份指定的材料上，夹子里必须有什么，由八项各自的取证要求决定（版本史、问对记录、现场记录），而不由学生的展示意愿决定。第二，第零轮初判在问对开始前定格，档案因此有了不能事后补称的起点，学生无法在期末回头把成长的故事

编得更漂亮。第三，第③、⑤项换手评分并写利益冲突声明，评分者与材料的牵连被摆到明面。三处合起来，才是相对档案袋多出的一格：多出来的不是更多材料，而是材料与分数之间可以被他人核对的关系，以及起点不可补造。承认这一点也就承认了代价，核对要花评分者的时间，第16章和第17章会处理这笔工作量。

另一个值得对话的框架是Chi与Wylie（2014）的ICAP：学习者的投入方式分为被动、主动、建构、互动四种，学习效果依次递增。把八项放进这个框架，大致可以看到：母本口试里的复述偏在主动一端，改题、问对、子本的撰写偏在建构一端，碰撞场则是互动。这给了八项一个来自学习科学的旁证，说明过程项排在产物项旁边不是随意的。但ICAP给出的是投入方式的分类，没有回答“怎样证明这个学生此刻真的在建构”；证据索引恰好补上这一问：问对记录里每一轮比上一轮多出的东西，是建构留下的痕迹；碰撞记录里被同学接过去又改写的判断，是互动留下的痕迹。还要承认限度，ICAP排的是投入方式的次序，不是产物的质量，一个学生可以投入得很“建构”而写出平庸的判断。所以第⑤项仍须独立量产物，过程项不能替它打分。

◆ 两张剖面：沈澄与周川

回到开头那两张总分相近的成绩单。沈澄的剖面已在上一节的表中给出，现在读它的形状。最高的两点是第④项现场证据（92）与第⑥项子本完成度（90），这与她的九问高度相关：第3问追问实验记录里的“室温”十年里有没有写过具体温湿度，第6问要她复做一批并把每一步的时间、温度、湿度都记下来与档案比对，这些问题本身就把她推向厚实的现场；化学子本又以可复做性为完成度的核心，她的复做记录恰好让子本经得起别人照着做。剖面最低的一点是第⑦项碰撞贡献（45），碰撞记录显示她在两次碰撞场（第9周与第11周，见第25章与附录一的日程）里多是报告自己的数据，很少接过同学的判断去撞。第⑤项创新智商原始分约120，越过本科分档线110，映射到百分制约70分：“规程里没写的那一句”这一判断有母本所无的内容，但在待复做验证之前，可证伪性一维还只是写出了检验方案，没有走完。

这张剖面开出的处方是具体的。沈澄不缺发生的真实度，她缺的是把自己的发生拿到别人的发生旁边去撞的习惯。第14章讲过一个跨学科碰撞的构造场景：沈澄的“规程里没写的那一句”与许宁的“交接断点”撞出了“没写下来的那一句”作

为跨实验室与病房的共同结构；若沈澄在本门课的碰撞场里也这样做，她的第⑦项会上去，第⑤项的跨域厚度也可能随之上去，因为E维量的正是纠缠深度与跨域厚度。总分75只会告诉她“良”，剖面却告诉她下一学期把力气花在哪里，以及为什么花在那里会同时带动另一项。这就是剖面比总分多出来的那部分信息：它不仅显示长短，还显示项目之间的牵连。

这里顺带说明剖面里第⑤项的两个数字为什么要同时出现。原始分120是在160分制上量出的，它和资深学者、和全球文库用的是同一把尺子；70分是按本科分档线分段映射之后的百分制，它让第⑤项能与其他七项放在同一张剖面上、按同一套权重相加。只给百分制，下一位老师就看不出这份子本在全尺度上站在哪里；只给原始分，又无法与剖面上其他项目对齐。两个数字同时写，就像体温既写摄氏度、又注明是否超过发热线，一个告诉你绝对位置，一个告诉你相对于这一学段的要求意味着什么。映射怎样分段、为什么按档位而不按比例折算，是第16章的任务。

周川的剖面形状完全不同。按经济管理本科的权重示例（①5、②10、③15、④15、⑤30、⑥10、⑦10、⑧5），下面是一张构造的示意。

项目	权重（经济·本科）	得分	读法
① 母本底盘	5	55	口试复述零散，反例举得不准
② 改题质量	10	88	“伙计起劲”→“父亲月底泄气”：根还在，离现场更近
③ 问对深度	15	85	轮次推进清楚，接住了“月初也忙为什么不累”
④ 现场证据	15	90	三年流水账逐月整理，附与父亲的访谈记录
⑤ 子本创新智商	30	65	原始分约115，“动力被账本提前收走”有新意，论证未收拢
⑥ 子本完成度	10	55	结构未成形，引注有缺
⑦ 碰撞贡献	10	85	与林晓撞出班级判断“动力被现金节律收走”
⑧ 回写证据	5	60	期末自述有改写，未见精要库采纳
总分	100	74.3	—

这正是“中间高、两头弱”。中间的改题、问对、现场与碰撞都很高，说明周川的发生是真的：他的题是从自己家的小饭馆里长出来的，问对里那个“月初也忙为什么不累”的追问他接得住，流水账翻得透，碰撞场里他和林晓一起撞出了两人单独都说不出的判断（见第14章）。两头弱则说明两件事：前端母本底盘薄，他对谱系母本的核心判断掌握不牢，这使他在写子本时不知道自己的判断究竟比母本多出了什么，也就难以把“多出来的那一点”说清；后端子本完成度低，发生了的判断没有被收拢成论证完整、引注无误的成品。对周川来说，下一学期的处方与沈澄恰好相反：不是去碰撞场里多开口，而是回到母本多读，并在子本写作上下功夫。

这张表用的是经济本科的基准权重，周川在第18章里实际选的却不是它。他在理由书中把④现场证据从15拉到25，把⑤从30降到硬底25，把③问对深度从15降到10，其余五项（①5、②10、⑥10、⑦10、⑧5）不动，八项合计仍是100。八个分数是剖面，不因权重而变；权重换了，只是对同一张剖面换一种读法。把同一张剖面放进他自选的权重重新核算，结果如下。

项目	基准权重	自选权重	得分	自选下的加权分
① 母本底盘	5	5	55	2.75
② 改题质量	10	10	88	8.80
③ 问对深度	15	10	85	8.50
④ 现场证据	15	25	90	22.50
⑤ 子本创新智商	30	25	65	16.25
⑥ 子本完成度	10	10	55	5.50
⑦ 碰撞贡献	10	10	85	8.50
⑧ 回写证据	5	5	60	3.00
总分	100	100	—	按基准权重74.3；按自选权重75.8

两个总分的差来自三处移动：④多了10个权重点，乘上他的90分，加9.0分；⑤少了5个点，乘上65分，减3.25分；③少了5个点，乘上85分，减4.25分；合计加1.5分，74.3变成75.8。这个核算里有一点要读出来：周川在第④项上最强、在第⑤项上最弱，所以把权重从⑤挪向④，总分必然上升，他的选择与他的剖面是对得上

的；但这并不说明他“选对了”，只说明他在选权重时看清了自己的形状，而这正是第18章所说“选择权重本身是一次判断”的一个侧面。

把两张剖面并排放，总分只差不到一分（沈澄75.0，周川按基准权重74.3、按自选权重75.8，名次还会随权重翻转），形状却几乎互为镜像：沈澄两端稳、中间的协作一处塌陷，周川中间满、两头塌陷。旧成绩单把他们放在同一档“良”里，等于对两个症状相反的病人开同一张药方。剖面的价值也在这里显露出来：同一套八项、同一种百分制，只是不再把八个数压成一个，学生、老师和下一门课就都拿到了可以行动的信息。还应当注意，剖面没有给出“谁更好”的判决，它只给出“各自往哪里走”的方向，这与九问课堂不交付同一本书、而交付各自子本的母题是一致的。

◆ 剖面的三个限度

剖面不是万能的，有三个反例需要先摆出来。第一个是小班。研讨课若只有十来名学生，碰撞场里凑不齐能互相撞的人，第⑦项的分数更多反映的是同桌是谁，而不是学生有没有撞出判断。这像足球里的助攻数，一支小球队里五个人上场，助攻多寡大半取决于队友站位，不取决于传球的人。这样的班里，第⑦项应当并入第③项，或者改为跨班碰撞，否则剖面里就混进了一个与学生无关的项目。第二个是评分误差。八个分数各带各的误差，尤其是第⑤项，两位换手评分者差出几分是常事，读形状时不能把相邻两项差几分读成“塌陷”。本书暂定的读法是：一项分数要低于相邻各项约10分以上，才在面谈中作为塌陷提出，否则只当作波动。这像测血压，医生不会因为两次读数差一点就改药，要看趋势。

第三个是时间窗。一份剖面只记这一学期、这一门课的路，一个学生在某门课上呈现“产物高、过程空”，也可能只是那个学期他同时赶着另一门课的现场工作，过程材料没来得及留。所以对这种形状，第17章的补证程序在先，剖面上的警惕只是请老师去问一句“这个学期你的时间花在哪里了”，而不是给学生下一个结论。三个反例合起来说明一件事：剖面比总分多出信息，同时也多出需要被诚实对待的不确定性；把它当作面谈的起点，它就有用，当作判决的依据，它就和总分一样危险。

◆ 本章怎样算错

本章押在三件可以查的事上，每一件都给出读数。第一件是八项有没有区分力。取连续两届、每届不少于30名学生的八项分数，把八项两两求相关，共28对；若其中超过半数（15对以上）的相关系数高于0.8，说明八项在系统地同升同降，剖面并不比总分多出信息，八项结构就是冗余的，应当合并为四项左右。第二件是剖面有没有改变行动。期末面谈之后追踪下一学期的学习安排与教学安排：若连续两届、每届不少于30名学生中，有一半以上的人下一学期的着力点与上一份剖面里最低的两项毫无对应，“剖面开出处方”的主张就不成立，剖面只是一张更花哨的总分。

第三件是证据索引能不能维持。每学期随机抽查不少于10份成绩单，逐项调阅其索引的材料；若任一项的可调阅率低于90%，分数就重新变成了没有出处的数字，该项应先退出权重，保住可索引的几项，而不是硬撑八项。此外，本章关于评估方向翻转的论述依赖第4章题闸的有效性：如果某个学科连续两届，两位独立评分者对“拿掉现场之后，核心判断能否被已有材料重新生成”的判定一致率低于80%，那么该学科第⑤项的方向翻转就缺少操作基础，需要另寻检验。

剖面八项里，分量最重、也最难打准的是第⑤项。它要把一份子本按五个维度打到160分制上，再折算成百分制；它要求换手评分，要求利益冲突声明，还要防止分数越打越高。一个本科生的子本凭什么能和资深学者用同一把尺子量？叶青写外婆的方言与高远量校园老桥的振动，两种“新判断”又怎样放在同一个五维里比？这些问题留给下一章。

> 带走的话
>
> **总分只是剖面的摘要，成绩单的正文是八项分数和各自的出处；两个“良”之所以不同，是因为它们走过的路不同，而路只有剖面才画得出来。**

◆

16

第16章

核心项：子本创新智商与完成度

◆ 同一把尺子量两份子本

学期末的换手评分会上，评分老师手里同时摊着两份子本。一份来自文学院《写作与阅读的发生》，作者叶青，写的是外婆，正文里夹着她草稿史的修改痕迹，核心判断是“不敢写的词才是叙述的承重词”；另一份来自土木学院《结构健康监测研讨》，作者高远，满纸是校园老桥的振动曲线，核心判断是“回落迟缓与一处被忽略的伸缩缝维修记录有关（待检）”。评分表是同一张：S、D、E、I、F五维，满分160。一位旁听的同事忍不住问：一份讲方言里的一个词，一份讲桥面上的一条缝，凭什么用一把尺子量？量出来的120分和120分，难道是同一种120分吗？

这个疑问还有另一层。同一场评分会上还摊着另外两份同一门课的子本，暂且称作甲与乙。甲把谱系母本复述得滴水不漏，章节齐整，引注规范，读起来像一本压缩得很好的教材；乙的子本论证还有毛边，却在自己的现场里撞出了母本所无的判断。按讲授课堂的习惯，甲是优等生，乙是“有想法但不够扎实”的学生。可是如果第⑤项子本创新智商在本科研讨课里占30，甲的总分会掉得很厉害。这是不是对认真学习的学生不公平？

本章回答这两层疑问。核心判断是：子本创新智商是学业成绩的核心指标，它的五个维度与九问课堂的各段一一对应，题闸就是I维在出题端的预演；子本完成度是门槛，它保证创新智商量的是一个立得住的东西，而不是一团还没成形的灵感。下面先讲五维与课堂各段的对应，再讲160分制怎样按档位分段映射到百分制，然后讲换手评分与利益冲突声明，以及反膨胀规则在课堂里的用法，最后回到叶青与高远、甲与乙。

> **这一章的问题**
>
> 一份讲方言里的一个词，一份讲桥面上的一条缝，凭什么用同一把尺子量，量出的分数又凭什么算数？

◆ 五维与课堂各段：题闸即I

本书采用的创新智商是五维160分制，五个维度及其权重是：S显露精确度0.20，D路径组织与差异锐度0.25，E纠缠深度与跨域厚度0.20，I不可还原性0.20，F可证伪性0.15。前三维量的是发生的构造，后两维像两道闸，查的是发生的真伪。标尺上有几个锚点需要记住：130上下10分是资深学者档，150是提智达标线；I或F任一低于120，则综合分封顶145。这把尺子原本用于评审论文与专著，本书把它搬进课堂，不另造一把“学生专用”的尺子，理由在下一节讲；这里先讲五个维度怎样落在九问课堂的各段上，因为只有对上了流程，评分人才知道每一维去哪里找证据。

S维显露精确度，对应课堂的立骨段，也就是谱系母本精要与九问。它问的是：子本有没有把自己的判断相对于母本的位置精确地显露出来，说清自己的判断落在母本哪一条核心判断上、改动了它的哪一部分。一份S维高的子本，读者能一眼看出“母本说的是什么，我说的是什么，差在哪一格”；S维低的子本，新旧判断混成一片，读者分不清哪句是母本的、哪句是作者的。这像修房子的施工图，好的图纸会把原有承重墙和新砌的墙用不同线型画出来，差的图纸全用一种线，施工队拿到手就不敢动锤子。

D维路径组织与差异锐度，对应课堂的改题与问对段，权重最高，是因为九问模式本身就是一条路径。它问两件事：从母本到新判断，这一路是怎样组织起来的，每一步凭什么走到下一步；新判断与母本、与通行看法之间的差异有多锐利，是含混地“补充了一点”，还是清楚地“在这一处不同”。许宁从“因为年纪大”走到“同样年纪为什么白天跌得少”，再走到交接断点（见第7章），这条路本身就是D维的材料。差异锐度可以用一个日常的比方理解：两张照片“有点不一样”，说不出不一样在哪里，是钝的差异；指着某一处说“这里多了一个人”，是锐的差异。

E维纠缠深度与跨域厚度，对应课堂的现场与碰撞段。纠缠深度问的是判断与学生现场的材料咬合得有多紧：拿掉一份交班本、一批复做记录，判断还站不站得住；跨域厚度问的是判断有没有触到不止一个领域的材料，例如第14章那次跨学科碰撞里，“没写下来的那一句”同时咬住了实验室规程与病房交班本。E维高的判断像扎了根的树，拔不出来也挪不走；E维低的判断像插在花瓶里的花，好看，但换个瓶子照样插。

I维不可还原性，对应的正是第4章的题闸。题闸在出题端问：拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案？能答即不合格，打回重出。I维在评估端问的是同一件事：拿掉学生的现场与轨迹，这份子本的核心判断能否从母本、精要库与全球已有文献里被重新生成？能被重新生成，I维就低。所以说题闸即I，前者把住入口，后者把住出口。出题时过了题闸的问题，并不保证子本的I维就高，因为学生完全可能在一个好问题上给出一个从指南里抄来的答案；但出题时没过题闸的问题，子本的I维几乎注定上不去，因为那道题本身就是可以被还原的。还要补一句限度：题闸只检验“必须依赖现场”，不检验“值得研究”（见第4章与附录二），一道题完全可以因为写了“本病房”而过闸，却仍不值得做；值不值得，要靠I维的评审和老师对学生现场的了解来回答，不能指望题闸。

F维可证伪性，对应课堂的回写段，也对应全书每章末那一节“本章怎样算错”。它问的是：这个判断说清了在什么情况下会被推翻吗？检验方案写出来了吗，能执行吗，执行了吗？沈澄的判断标着“待复做验证”，许宁的判断标着“待检验”的本病房假说，这些标注不是谦辞，而是F维的入口：一个判断如果怎样都推不翻，它在课堂里就没有学习价值，因为学生无从知道自己错在哪里。F维权重最低，但它与I维并列为两道闸：任何一道低于120，综合分封顶145，这一条规则在下一节会看到它对研究生和博士的特殊意义。

◆ 160分制与分档线：同一把尺，不同的及格线

回答旁听同事的第一层疑问之前，要先回答一个更基础的问题：学生的子本，凭什么用评审学者论文的尺子量？一种直觉的做法是为学生另造一把宽松的尺子，让本科生也能拿到“140分”，看上去皆大欢喜。本书不这样做。理由很朴素：尺子

一旦分学段另造，同一个分数在不同学段就意味着不同的东西，学生从本科走到研究生、再走到博士，手里的分数无法前后比较，第15章所说的剖面序列也就断了。更要紧的是，九问课堂打向的是谱系知识，学生的子本与学者的论文面对的是同一个“已有”，不可还原性只能对着同一个全球文库来量，没有一个只属于本科生的“已有”。

所以本书的口径是：同一把尺子，分学段设合格线。在160分制上，本科的分档线是110，研究生是125，博士是140。这像跳高比赛用的是同一根标杆、同一把卷尺，少年组、青年组、成年组的及格高度不同；一个少年跳过一米六，在少年组是好成绩，在成年组还够不上及格，但那一米六在两个组里是同一个一米六。对照标尺锚点看，这三条线的意思也就清楚了：本科分档线比资深学者档的下沿还低10分，要求的是“在自己的现场里有了真东西”；研究生分档线落在资深学者档之内的下半段；博士分档线落在资深学者档的上沿，离150的提智达标线只差10分。

然后是映射。剖面上的八项都用百分制，第⑤项必须从160分制折算过来。最省事的办法是按比例折算，原始分除以160再乘以100，可这样做没有意义：120分折成75分，这个75与分档线、与资深学者档都没有关系，老师和学生都不知道它代表什么。本书的建议是按档位分段映射：先把各学段的分档线钉在百分制的60分上，作为及格；再把150的提智达标线、160的典范级钉在高端；中间各段按不同的斜率连接。分档线附近斜率最大，因为大多数学生落在这一带，评分人要能把115与120的差别显示出来；高端斜率较小，因为越往上，评分人的把握越弱，过陡的映射会把评分误差放大成成绩差距。

下面是一张示意对照表，标明为本书建议口径，不是已经验证的标准。各学校可以调整斜率，但不宜改变“分档线对60”这一锚定方式。

创新智商原始分（160分制）	本科（分档线110）	研究生（分档线125）	博士（分档线140）
90	40	40以下	40以下
110	60（分档线）	45	40以下
120	70	55	40
125	75	60（分档线）	45

创新智商原始分（160分制）	本科（分档线110）	研究生（分档线125）	博士（分档线140）
130	80	约67	50
140	85	80	60（分档线）
145	87.5	85	70
150（提智达标线）	90	90	80
160（典范级）	100	100	100

这张表的分段规则可以这样读。本科：90到110每一分对百分制一分，从40升到60；110到130仍是每一分对一分，从60升到80；130到150每两分对一分，升到90；150到160每一分对一分，升到100。研究生：105到125对40到60，125到140对60到80，140到150对80到90，再到160对100。博士：120到140对40到60，140到150每一分对两分，升到80，150到160每一分对两分，升到100。分档线以下更低的区段按比例压缩到40以下，示意从略。第15章沈澄的原始分约120，映射为70；周川约115，映射为65，都是按本科一列读出的。

封顶规则在这张表上显出了它对高学段的分量。I或F任一低于120，综合分封顶145。对本科生，145已在资深学者档之上，这条封顶几乎不起作用；对研究生，它把百分制上限压在85；对博士，它把上限压在70，只比及格线高10分。这意味着一个博士生的子本无论写得多么宏阔，只要不可还原性或可证伪性有一道闸没过，就拿不到好成绩。这一安排是有意的：学段越高，学生越有能力写出结构漂亮、路径完整、跨域厚重的文本，也就越需要两道闸来查这份漂亮是不是真的发生了，而不是把已有的东西重新排了一遍，或者说成了一个怎样都推不翻的判断。

还要说一句150这条提智达标线在课堂里的位置。它不是任何学段的及格线，连博士的分档线都比它低10分；它是一个方向标，告诉学生这把尺子的上端在哪里。本科生的子本偶尔越过150，并不奇怪，因为一个人在自己独有的现场里撞出的判断，完全可能在不可还原性上超过很多学者的论文；但这种情况必须经过最严格的换手复核，因为越过达标线的分数也最容易是膨胀的结果。把150当方向标而不当及格线，课堂既保留了对真正出色发生的承认，又不会让学生误以为每一份子本都该冲着它去写，以至于为了显得新而丢掉根。

◆ 换手评分与利益冲突声明

九问课堂里，学生的子本是在与ChatSDE的多轮问对中长出来的，老师又用装有SDE工序层的大模型出了九问。如果评分时再请同一套智能体来打第⑤项，就出现了一个结构性的问题：参与生成的系统在评定自己参与生成的东西。这像让一家装修公司自己给自己做验收，不必假设它有意作弊，它的眼光、它的偏好、它在施工中做过的每一个判断，都会让它在验收时更容易看到“做得好”的地方。本书第11章在讨论对照实验时已经立下一条口径：同底座的多个智能体不能互作独立证据。这条口径搬到评分上，就是换手评分：第③项问对深度与第⑤项子本创新智商，必须由没有参与该学生问对与子本生成的评分者来打。

换手可以有几种做法，按可靠程度从高到低排列。第一种，换人：由本课程以外、未参与该生九问与问对的教师评定，必要时两人独立评分后取均值，分差过大则请第三人。第二种，换底座：若用智能体辅助评分，须用与问对所用ChatSDE不同底座的模型，并且由人复核；同一底座换一个提示语、换一个名字，不算换手。第三种，换时：在人力不足时，至少由原任课老师在隔开一段时间、先读完学生轨迹之后再打分，并在声明中如实写明这是同手评分。第三种只能作为过渡，不能成为常态，因为任课老师正是出九问、读问对、给批判的人，他与学生的子本之间有最深的牵连。

“换底座”这一条还要补一层警惕。迎合倾向不是某一家模型独有的毛病：Sharma等（2024）在多个主流助手上都观察到，模型倾向于给出与用户已表达的观点相一致的回答。移到评分场景里，意思是评分智能体若在提示里读到“这位学生很用功”“老师觉得这份不错”，分数就会向上靠。所以智能体辅助评分时，评分提示不得带入学生姓名、成绩史和老师的预判；评分结果只作为人的初稿，由人对着材料复核；使用之前先让它给几份已知定分的锚样打分，看它偏高还是偏低，再决定它的初评能被采信到哪一步。换一个底座只能减少同源偏差，消除不了迎合，这一点与其说是缺陷，不如说是提醒评分卡必须留出人的签字位置。

换手解决的是“谁来评”，利益冲突声明解决的是“评的人与这份子本有什么牵连”，两者缺一不可。声明要写清：评分者是否参与过该生的出题、问对、碰撞或子本修改；评分时是否使用了智能体，用的是哪个底座，与问对所用系统是否同

源；评分者与学生之间是否有导师、课题组、亲属等关系。凡有牵连而未能换手的，评分须标为偏严一侧的自评、非认证分，也就是说评分者应主动往严里打，并在成绩单上注明这一项不作为认证依据。这种做法有点像体育裁判：裁判与运动员同乡并不必然判罚不公，但必须回避或者公开，公开之后观众才知道该怎样看待那一记判罚。

还要强调的是，换手评分与利益冲突声明的对象，不只是第⑤项的总分，也包括五维各自的分数。因为封顶规则落在I与F上，这两维的分数一旦被同手抬高，封顶就形同虚设。评分卡上除了五维分数与综合分，还应写明两项必填信息：参照文库与评分截至时间。参照文库默认取全球最新，跨语种、跨邻近学科，时间窗推到当期；只用本课程的精要库或只用中文文献作参照，只能作为对照组并列写出，不能替代全球文库。评分截至时间之所以必写，是因为“已有”在不断增长：今天判为不可还原的判断，一年之后可能已被别人发表；写清截至时间，这份评分就能被事后核对，而不必因为后来的发表被推翻。

换手在操作上最常被问到的是工作量。一门40人的本科研讨课，若每份子本都请两位课外教师独立评五维，确实负担不轻。一个可行的安排是院系内课程之间对换：医学的老师评化学的子本并不合适，但同一学院不同研讨课的老师互评，既保证了换手，又保证了评分者懂得这一学科的母本。另一个安排是分层抽样：全部子本由一位换手教师初评，其中靠近分档线的、触发封顶的、五维分项差异异常大的，再请第二位教师独立复评。这样做的道理与工厂质检相同，不是每一件都拆开检查，而是把最容易出问题的那几类挑出来重点查。无论采取哪一种安排，课程大纲都应事先写明，学生在交子本之前就知道谁会来评、按什么程序评。

◆ 反膨胀规则与完成度门槛

任何评分制度运行久了，都会出现分数往上漂的倾向。学生希望分高，老师希望自己班的分数好看，智能体在问对中被反复要求“再好一点”，也会倾向于给出更高的评价；前文与第8章援引的Sharma等关于大模型迎合倾向的研究，讨论的正是这一类现象。九问课堂的评估若不设防，几个学期之后，130分就会变成本科子本的常见分数，资深学者档这个锚点也就失去了意义。所以本书沿用评分标尺本身的

一条铁律：反膨胀规则先于一切酌情。评分人在任何一维上想要“酌情加分”之前，先要过一遍反膨胀规则，看这一维有没有触发封顶或扣减。

在课堂里，反膨胀规则至少有三条常用的用法。第一条是两闸封顶，即前面讲过的I或F任一低于120则综合分封顶145；在课堂上还可以由任课教师在课程大纲中事先写明本课程采用的更细的封顶条款。第二条是最强竞争解释未被排除时F维封顶：许宁的判断如果没有说清“交接断点”与“夜间人手少”这一竞争解释怎样区分，F维就不能给高分，因为一个没排除最强对手的判断，还谈不上可证伪。第三条是坏引注扣S：子本中引用了不存在或与原文不符的文献，S维要扣分，因为显露的前提是“母本说的是什么”要说准，引错了母本，后面的差异就失去了参照。这些规则的共同点是：它们都是可以事先写出、事后核对的硬条款，而不是评分人临场的感觉。

除了条款，反膨胀还需要一个看得见的参照物，这就是锚样。每门课可以从往届子本中挑出若干份，经多人换手评定、分数稳定的，作为本届评分前的校准样本：评分者先独立给锚样打分，与锚样的定分比对，偏高或偏低的先自我校正，再开始评本届。这像称重之前先放一个标准砝码，秤若显示1020克，就知道它偏重了20克。锚样还要逐年更新，因为“已有”在增长，去年的锚样今年再评，I维理应略降；若评分者给出的分数反而更高，那恰是分数在漂移的信号。锚样的选取与更新，也是第24章所说课程成为谱系的一部分。

完成度在第⑤项旁边扮演门槛的角色。第⑥项子本完成度量论证是否完整、引注是否无误、结构是否成形，化学类课程再加可复做性。它本身有权重，在本科研讨课里占10，在化学本科占15；但它的更深作用是保证第⑤项量的是一件成形的东西。一份论证还没收拢的子本，创新智商量出来的往往是评分人自己对它的想象：评分人读到一个闪光的句子，脑子里替学生把论证补全，然后给这个被补全的版本打了分。这像评一道还没出锅的菜，评委闻到香味就给了高分，可那份香味里有多少是火候到了，有多少是评委自己的想象，谁也说不清。完成度门槛的意思是：子本若在论证完整、引注无误、结构成形三项中有一项明显不达标，第⑤项的评分应当暂缓或打折，而不是越过完成度直接去量创新。

化学的可复做性是完成度门槛最清楚的一个例子。沈澄那句“规程里没写的那一句”，如果不能被别人照着子本里的记录复做出来，它就既不完整，也无法证伪：别人复做失败，是判断错了，还是记录缺了，分不清。所以化学子本的完成度把可复做性放在核心，它与F维是咬合在一起的。在医学，与之对应的是伦理门槛，去标识化与许可不齐，整份不计分，连创新智商都不进入评定；在人文学科，对应的是史料与口述来源的可核对性，苏文的族谱与县志对照、叶青的草稿史，都必须能被评分人调阅。

完成度门槛还要面对“AI在场”这个现实压力。Bastani等（2025，PNAS）在高中数学课上做的现场实验发现，学生在不加护栏的通用大模型辅助下练习时，表现更好，可是撤掉AI之后的考试反而更差；加了教学护栏的辅导版本，基本消除了这种损害。把这个结论搬到子本上：一份子本在AI全程在场时写得完整漂亮，并不保证学生手里真有这个判断，就像练习时有人扶着自行车，骑得稳不等于会骑。所以完成度只证明“成形”，不证明“归属”，归属要由第17章的轨迹与答辩去证。九问课堂的几项安排，ChatSDE的追问不替学生说破，第零轮初判在先，既是教学设计，也是评分设计：它们让“AI撤出之后学生还剩什么”成为可以被检验的事。当然，Bastani等的研究对象是高中生的数学练习，不是大学生的研究型写作，迁移到子本只能算一个先验，不是证明。

◆ 两个反对意见：参照文库查不全，160分制太精细

对这把尺子，有两个反对意见值得认真对待。第一个是：I维要对着“全球最新、跨语种、跨邻近学科”的文库去查不可还原，评分人哪里查得全？这个反对成立。评分人实际能做到的，是在有限的检索里没有找到可以重新生成该判断的文献，这只能支持“在所查范围内不可还原”，不能支持“绝对不可还原”。举个例子：高远的伸缩缝判断，在另一个国家的桥梁监测报告里也许早有类似记录，只是评分人没有搜到。所以评分卡要求写明参照文库与截至时间，I维的结论是带范围的；日后发现已有文献，修订的是这份评分的读数，不是对学生的指控，这一点要在课程大纲里先讲清，免得学生把I维失分读成抄袭嫌疑。

第二个反对意见是：评分人的分辨率恐怕只有10分上下，给本科生子本一个“120”这样精确的数字，是不是虚假的精细？这个反对同样成立，所以本章主

张：原始分用于记录和对齐，决策用档位。具体做法是评分人在给出五维分和综合分的同时给一个区间，例如“综合分118至124”，映射到百分制时取区间中点，并在成绩单上同时标出区间；分档线上下各留一个缓冲带（暂定各5分），落入缓冲带的子本不单凭一位评分者的读数断及格与否，而交第二位换手评分者复评。这样牺牲一点果断，换来的是评分误差不会冒充学生之间的差异，好比体重秤只能称到半斤，就不该报到两位小数。

◆ 叶青与高远，甲与乙

现在回到开头的两层疑问。先看叶青与高远。两份子本的“新判断”确实长得很不一样：叶青的判断是“不敢写的词才是叙述的承重词”，它的证据是她自己的草稿史，那个写外婆时一直不敢写进去的方言词，在哪一稿被删、在哪一稿换成了普通话、在哪一稿终于留下，留下之后整段叙述的重心怎样移动；高远的判断是“回落迟缓与一处被忽略的伸缩缝维修记录有关（待检）”，它的证据是校园老桥的振动监测数据，那一段桥面在早高峰后迟迟不回落的曲线，以及与之对应的维修档案。一个落在主体属性一侧，一个落在客体属性一侧（三领域对应见第19章），旁听同事的疑问正来自这里。

五维把这两种判断放到了可以比较的位置上，但比较的不是内容，而是发生的构造与真伪。S维上，叶青要说清她的判断落在写作学母本哪一条核心判断上，例如关于叙述视角或细节选择的那一条，并显露她在哪一格上改动了它；高远要说清他的判断落在结构健康监测母本的哪一条上，例如关于振动响应与损伤识别的那一条。D维上，叶青要显示她怎样从“写外婆”走到“哪个方言词不敢写”，再走到“承重词”；高远要显示他怎样从“哪一段振动不回落”走到“维修记录”。I维上，两人面对同一个检验：拿掉叶青的草稿史，拿掉高远的监测数据与维修档案，能否从已有文献里重新生成这个判断？两人都不能，因为那份草稿与那座桥都只在他们自己的现场里。

差别主要落在F维与E维。高远的判断天然带着检验方案：去查那一处伸缩缝的维修前后数据，比较维修节点前后的回落时间，判断就会被支持或被推翻；它标着“待检”，F维的入口是清楚的。叶青的判断要拿到同样的F维分数，就得多做一步：说清什么样的证据会推翻“承重词”这一判断，例如把那个方言词替换掉之

后，请不同读者读两个版本，叙述的重心如果没有移动，判断就要修改。人文学科的“新判断”并不天然不可证伪，但它的可证伪性需要学生自己把检验构造出来，这正是第23章将讨论的人文新判断与创新智商的适配问题。E维上，叶青的纠缠深度可能更高，因为一个人的草稿史与她的语言、家族、记忆紧紧咬在一起；高远的跨域厚度可能更高，如果他把维修档案这种管理记录与振动数据这种物理信号撞在了一起。所以两个120分不必是同一种120分，五维分项会把它们的不同长处写出来，而综合分保证它们在同一把尺上可以对齐。

再看甲与乙。甲的子本把母本复述得滴水不漏，这一能力在剖面上已有它的位置：第①项母本底盘会很高，第⑥项完成度也会很高。但在第⑤项上，甲的S维也许不低，因为他把母本显露得很准；D维的差异锐度却很低，因为他的子本与母本之间几乎没有差异；I维更低，因为拿掉他的现场，他的子本几乎可以被完整地重新生成，事实上他的子本里本来就没有现场。I维很可能跌到120以下，综合分便被封顶；更直接地说，甲的子本按第12章的判别是一份翻译，不是重写，它没有满足子本“须含母本所无的判断”的定义。乙的子本论证有毛边，第⑥项会扣分；但只要完成度没有跌破门槛，她的I维与D维会把第⑤项拉上去。

这对甲是不是不公平？换一个角度看就清楚了。甲所擅长的复述与组织，正是第2章所说已被大模型接管的发现一侧的工作，大模型做得更快、更全、更耐心。如果大学的评估仍以复述为核心，等于在奖励学生做一件机器已经做得更好的事，并且让最擅长这件事的学生在毕业时发现，自己最强的能力恰恰最不稀缺。剖面并没有抹掉甲的长处，第①项与第⑥项如实记下了它；创新智商作为核心项，只是把评估的重心放到了机器替不了的那一侧。对甲最有帮助的，不是在第⑤项上给他同情分，而是在剖面上清楚地告诉他：你的底盘已经很稳，下一学期请把你的九问带到一个只有你能去的现场里。

◆ 三种难例：闪光未成形、现场真而判断旧、检验跨学期

五维加完成度门槛的组合，在三种难例面前最容易露出缝隙，不妨先摆出来。第一种是闪光但未成形：一份子本里有一两句真正新的话，论证却只写了一半，引注也有缺。评分人最容易犯的错，是替学生把论证补完，再给被补完的那个版本打高分。本书的处理是暂缓，不是低分：第⑤项在剖面上标“暂缓”，给限期补交，

补交之后再评；暂缓期间不记零分，因为零分会让学生以为那句新话不值钱，也不记高分，因为高分评的是评分人自己的想象。这像一道没熟的菜，厨房不该直接端上桌，也不该倒掉，而是退回灶上。

第二种是现场很真而判断是旧的。设想一位学生在病房里连着三个月记夜班数据，记录扎实，真实性毫无疑问，结论却是“夜班人手少所以容易跌倒”，这条在任何防跌倒指南里都找得到。这时第④项会很高，第⑤项的I维会很低，剖面呈现的是第15章所说“过程厚、产物平”。处方不是安慰他“现场很扎实”，而是让他回到九问里去找母本所无的变量，例如同样人手少的夜晚，为什么有的床位出事、有的不出；现场的厚度要换一个问法，才会长出新判断。第三种是判断新、可证伪，但检验需要超出一学期的时间，高远的伸缩缝判断就是这样：维修前后的回落时间要对比，也许要等下一个汛期之后的数据。对这种情形，F维评的是检验方案是否可执行、是否已经启动（数据来源、所需样本、判定标准、预计完成时间是否写明），不要求学期内走到结果；方案写不出来，或者写出来根本无法执行，F维才低。否则人文与工程里凡是需要长周期数据的判断，都会在课堂里被系统性地压低，课堂就会鼓励学生去选那些一学期内能验完的小问题。

◆ 本章怎样算错

这一章押在尺子的信度、区分力和预测力上，每一处都能给出读数。信度先查：取连续两届、每届不少于30份子本，由两名换手评分者独立评五维；若两人综合分之差的中位数超过8分（160分制），或者有四分之一以上的子本两人分差超过15分，创新智商在课堂里就还担不起核心指标的分量，应暂不以它排定学位等级，先改进评分卡与评分人培训，直到连续两届的复测回到这两个阈值之内。区分力再查：同样取连续两届、每届不少于30份子本，若某一学段第⑤项百分制分数的中间一半（第25到第75百分位）宽度小于10分，或者超过六成的子本落在同一个10分区间里，那么分档线或斜率就标错了，示意对照表需要按实际分布重标；这张表本来就是建议口径，被修订不算意外。

预测力最难，也最要紧。把两届学生各自按第⑤项排序，毕业后满两年追踪，样本合计不少于60人：若前四分之一与后四分之一在后续的新判断产出上（继续写的子本、参与课题或项目里提出的有出处的新判断，见第13章）看不出可辨认的差

别，那么五维量到的多半是一次性写作的技巧，而不是发生的能力，创新智商作为核心项的地位需要重新论证。反膨胀规则也有读数：每学期用不少于5份锚样对评分者做校准，若连续三个学期，同一批锚样的复评均值比首次定分高出5分以上，说明分数在漂移，须追查是换手失效、参照文库收窄，还是评分截至时间没有写。

五维量的是产物，换手保证评的人与产物无牵连，完成度保证产物已经成形。但还有一个问题本章没有回答：评分人怎样知道这份子本是学生自己发生出来的，而不是在别处写好、拿来交差的？一份五维俱佳、轨迹却拿不出来的子本，该给多少分？这要从过程项去回答，也就是改题、问对与现场证据怎样计分，这是下一章的任务。

带走的话

同一把尺子，分学段设合格线：量的不是内容像不像，而是判断能不能从学生的现场之外被重新生成；换手评分与完成度门槛，保证这个分数评的是一件成形的、与评分人无牵连的东西。

◆

17

第17章

过程项：改题、问对与现场证据的计分

◆ 一份拿不出轨迹的高分子本

《司法实践研讨》的期末评分里，有一份子本让换手评分的老师停了很久。它写的是本地基层法院物业纠纷的调解进程，结构严整，引注规范，核心判断也不俗：调解进程受到物业收费节点的牵引，当事人往往在账单周期的某个节点前后松口。按第16章的五维去量，S维与D维都不低，I维看上去也站得住，因为它指向的是本地法院的案件，而不是物业纠纷的一般规律。如果只看产物，这份子本可以进入本科130分以上的区段。可是评分老师调出它的过程材料时，发现题的版本史只有一个版本，就是老师发下的原题；问对记录只有两轮，第1轮学生问“物业纠纷调解有什么规律”，第2轮问“请帮我把以上内容整理成论文结构”；现场材料一栏写着“已阅公开裁判文书若干”，没有文书清单，没有旁听记录，没有任何一份原始材料。

同一门课里，韩正的子本形成了鲜明的对照。他的问题是“本院近三年物业纠纷里，调解成功的案子大多在立案后第几周结束？那一周发生了什么？”，判断是“调解进程被物业的账单周期推着走”。题的版本史有三版，问对记录里他至少有两次在ChatSDE追问之后回到法院去补材料，现场证据里有一份逐案整理的公开文书清单，标着立案日期与结案日期，还有几次公开庭审的旁听记录。两份子本的结论相近，一份背后是一整条可以翻查的路，一份背后是一片空白。如果评分只看产物，两者分数相差无几；如果评分把过程计进来，两者的差距就清楚了。（这两份子本和其间的评分场景都是构造的，用来说明过程项怎样计分，不是真实课堂的记录。）

这个场景把本章要回答的问题摆了出来：过程项怎样计分，才能既奖励真实的发生，又不被刷量所骗？核心判断是：过程项按质不按量；改题、问对与现场证据

留下的轨迹，是子本不可还原性的证据，它也正面回答了“AI代写怎么办”的问题。下面依次讲改题质量、问对深度、现场证据三项的计分要点，再讲反刷分规则，然后回到这份拿不出轨迹的法学子本，讨论它应当怎样评，最后与学习分析、写作过程研究对话，说明本书的轨迹相对既有做法多出了哪一格。

> **这一章的问题**
>
> 一份结构严整、引注规范却拿不出任何轨迹的子本，该不该得高分？过程项怎样计分，才既奖励真实的发生，又不被刷量所骗？

◆ 改题质量：看版本史，不看改动幅度

第②项改题质量量的是学生改动九问时是在提升还是在逃离，证据是题的版本史，也就是老师版、学生改版、问对后定版三个版本（见第5章）。计分的第一要点是看两锚：根还在吗，也就是改过的题是否仍连着母本的核心判断；离现场更近了吗，也就是改过的题是否比原题更深地落在学生自己能去的那块现场上。两锚都满足，是提升；丢了根而只剩现场，往往是把研究题改成了一件私事；丢了现场而只剩根，往往是把题改回了教科书。这像给一棵果树嫁接，接穗要接在原来的砧木上，根才还在；新枝要朝着有阳光的一侧长，才是离现场更近；把枝子剪下来插到别处，或者干脆把新枝剪掉只留老桩，都不叫嫁接。

周川的改题是一个清楚的高分样例。原题问“哪个伙计干得最起劲”，他改成“父亲为什么一到月底就泄气”。根还在，因为两题都连着课程母本关于小微经营者动力的核心判断；离现场更近了，因为“伙计”是他看得见却未必问得透的人，“父亲”与“月底”则是他每天都在场、并且有三年流水账可以对照的现场。同一门课里还有一个被打回的方向：“AI会不会取代饭馆”。这个改法看上去更宏大、更时髦，却两锚俱失，它离开了母本关于动力与节律的核心判断，也离开了周川自己的现场；拿掉周川的现场，一个未接入这些材料的通用大模型张口就能谈上一大篇。改题质量要给前者高分，给后者低分，即便后者的题面听起来更“有意思”。

计分的第二要点是看提升的方向是否让题更锐、更可检。许宁把第2问“交班本上，跌倒前一班对该患者写了什么、没写什么？”提升为“交班时被省略的那一

句，是谁、在什么情况下省略的？”。原题已经很好，它指向本病房的交班本，过了题闸；改过的题更进一步，从“写了什么”推进到“谁省略、在什么情况下省略”，把一个描述性的问题变成了一个可以追到人与情境的问题，也就为后来“交接断点”这条判断铺好了路。这一类改题，改动的字数不多，提升却很大。所以改题质量不看改动幅度，不看学生改了几问，九问里只改一问而改得准，可以高于九问全改而改得散。

计分的第三要点是版本史本身的可信度。版本史要能显示改题发生在什么时候、因什么而改：是读了母本精要之后改的，是第零轮初判之后改的，还是问对中被ChatSDE的某一句追问逼出来的。苏文的改题之所以有说服力，就在于版本史能把每一次改动对到一份具体的材料上，族谱的断代处或县志的某一年。一个版本史如果只有首尾两个版本、中间没有任何时间与理由，评分人就无从判断这是一次提升，还是学生在交作业前一天对着原题重写了题面。版本史不是越长越好，而是每一个版本都要有来由。

◆ 问对深度：推进、接住与首发

第③项问对深度量的是问对记录里的三件事：轮次之间有没有推进，学生接不接得住批判，关键判断是不是学生先说的。这三件事对应着问对之所以为问对的三个理由（见第6、7章）：问对是来回磨，所以要看推进；ChatSDE的默认是批判性激发，所以要看接住；问对的目的是让学生自己发生判断，所以要看首发。三者缺一，问对就会退化成别的东西：没有推进，是闲聊；接不住批判，是被动听讲；判断总由ChatSDE先说，是请人代写。

推进的计分看的是每一轮比上一轮多出了什么。多出来的可以是一条新材料，一个被修正的前提，一个更窄更准的问题，一次回到现场之后带回来的观察。沈澄围绕失败批次的七轮问对（第零轮加第1—6轮）是一个推进清楚的样例：第零轮她写下初判“试剂纯度不够”，ChatSDE问“同一瓶试剂，为什么三月好、六月差”，她回实验室去翻档案、去复做，带回每一步的时间、温度、湿度，再进入下一轮，最后走到湿度假说。每一轮都能指出“这一轮比上一轮多了什么”，评分人只需顺着记录读，推进就自己显露出来。这像下棋复盘，好的一盘棋每一步都改变

了局面，复盘时能说出每一步的意图；一盘乱棋走了一百步，复盘时却说不清哪一步改变了什么。

接住批判的计分看的是学生被翻过前提之后怎么办。许宁的第零轮是“夜间跌倒是因为患者年纪大”，ChatSDE第一回合问“同样年纪的患者，为什么白天跌得少？”。接得住的表现，是她没有为“年纪大”辩护，也没有立刻放弃自己的观察，而是把这个追问当成一个可以去病房里查的问题，接下来去看交接时段、看交班本。接不住有两种样子：一种是硬顶，重复自己的初判并要求ChatSDE“按我的思路展开”；另一种是全盘投降，ChatSDE说什么就是什么，自己不再提出任何判断。前一种把问对拉回只要答案、不碰前提的老路，后一种把问对变成另一种讲授。两种都不能给高分，接住意味着既不硬顶，也不投降，而是把批判变成下一步的行动。

在这三件事之前，还有一份不容易被注意的证据：第零轮。第零轮是学生在与ChatSDE开口之前先写下的初判，许宁的“因为年纪大”、沈澄的“试剂纯度不够”、方舟的“服务器不够”都是第零轮。第零轮本身不计分，初判错得再远也不扣分，因为它的用处是做基线：没有它，评分人就无从判断后来的推进有多大，也无从判断最后的判断是不是学生自己从初判出发走出来的。这像减重之前先称一次体重，那个数字高低无所谓，要紧的是它必须在开始之前就记下来；事后补称的“起点体重”，谁也不会相信。所以问对记录若缺了第零轮，或者第零轮的时间晚于第1轮，问对深度一项不能进入高档，第⑧项回写证据也失去了对照的起点。

首发的计分看的是关键判断第一次出现在谁的话里。这一项在全书中有一个专门的检验：核心判断谁先说（口径见第6章）。首发是相对ChatSDE而言的：ChatSDE的追问不得出现核心判断的关键变量词，它可以问“同样年纪的患者，哪几个时段出事更多”，却不能替学生说出那个时段叫什么；而现场里他人说的话、档案里已有的记录，是学生取得的现场材料，不算替学生说破。比如沈澄在实验室里听师兄随口说过一句“免得受潮”，她把这句话带回问对，它是她取得的现场材料，不影响她对判断的首发；若是ChatSDE自己先说出了那个变量，判断就不是学生首发。即便追问干净，学生仍可能顺势把判断“认领”过去，比如ChatSDE问“同样的患者，为什么这个时段出事更多”，学生答“对，应该是交接前后”。

评分时要分辨：学生是在这一轮之前已经从交班本里看到了线索、只是借这一问说了出来，还是完全由这一问提供了判断、学生只是点头。前者算学生首发，后者不算。为了让这种分辨可行，问对记录必须完整保留，包括第零轮，并且记录中学生回现场的时间与ChatSDE追问的时间要能对得上。

问对深度是必须换手评分的两项之一，这与第16章讲的道理相同：问对是学生与ChatSDE共同进行的，参与问对的系统去评问对深度，等于对弈的一方去写这盘棋的评语。换手评分者读问对记录，最好与读现场证据交替进行，因为推进与首发的判断常常需要对照现场材料才能做出：学生说“我回去查了交班本”，现场证据里就应当有那一次查阅的记录。利益冲突声明的写法也与第⑤项相同，任课老师若因人力不足未能换手，须标明为同手评分、偏严一侧。

还有一件事不进入问对深度的计分，那就是问对的风格。第6章讲过，问对的风格是自由的：有的学生习惯追问型，一轮一轮往下钻；有的学生习惯故事型，先把现场里发生的事讲一遍，再从故事里挑出疑点；有的学生习惯争辩型，偏要和ChatSDE顶几个回合才肯松口。评分人不能因为某种风格看起来更“学术”就给它加分，也不能因为学生用了大量口语、问对记录读起来像聊天就扣分。风格是学生自己的发生节奏，计分只看推进、接住、首发这三样落在记录里的东西。这像评一位登山者，有人走之字形，有人走直线，有人中途绕到另一面，评的是他登上了哪一处、每一段是否在向上，而不是他的步态好不好看。

◆ 现场证据：真实与厚度

第④项现场证据量的是学生下到自己现场的深度，证据是田野记录、实验记录与原始材料。计分要看两样东西：真实与厚度。真实问的是这份材料是不是学生在现场亲手得来的，能不能被核对；厚度问的是材料是否足以撑起子本的判断，是否不止一种来源、不止一个时点。真实是底线，厚度是高下。一份材料再厚，若不真实，这一项不能给分，并且要追查子本其他各项；一份材料真实而单薄，可以给及格，但给不了高分。

许宁的夜班记录是一个真实而有厚度的样例。她的第8问是“跟一次夜班，记下护士从听到动静到到达床旁的时间，哪一段最长？”，现场证据里就应当有这一次

夜班的逐段计时记录：什么时候听到呼叫铃或动静，护士在做什么，从护士站到床旁走了多久，中间被什么事情打断。这份记录的真实性可以核对，因为它附带了夜班的日期与科室许可，并且与交班本、护理记录的时间相互对得上；它的厚度来自它与其他几问的材料互相咬合：第1问的交接时段统计，第2问的交班本比对，第6问陪护与无陪护夜晚的呼叫铃使用，第8问的计时，四份材料指向同一处交接断点。医学案例的伦理门槛在这里同样适用：所有记录须去标识化、经科室与伦理委员会许可，学生在夜班中只观察记录，不改医嘱、不作临床决策；伦理门槛不过，这一项连同整份成绩都不计分。

韩正的旁听记录是社会科学一侧的样例。他的问题要求统计调解成功的案子大多在立案后第几周结束、那一周发生了什么，这需要两类材料：一类是本院近三年公开裁判文书的逐案整理，标出立案与结案日期；另一类是公开庭审或调解相关公开活动的旁听记录，看那一周里当事人之间实际发生了什么。前一类给出节律，后一类给出机制，两类合起来，才撑得起“调解进程被物业的账单周期推着走”这个判断。只有前一类，判断停在相关；只有后一类，判断停在印象。这里也有一条隐私与司法公开的边界：只用依法公开的文书与允许旁听的场合，记录中不出现可识别当事人的信息。

厚度不等于数量，这一点要特别讲清。一份现场证据附了300张照片，却没有一张能对上子本中的某一句判断，它的厚度是零；一份现场证据只有一页，却是沈澄复做那一批时逐步记下的时间、温度、湿度，与档案里同一步骤的记录逐项对照，它的厚度足以撑起湿度假说的一半。评分人看厚度，要看的是材料与判断之间的咬合：子本里的每一个关键判断，能否在现场证据里找到它的来处；现场证据里的每一份主要材料，能否在子本里找到它所支撑的那一句。这像审计里的双向追查，从账到凭证查一遍，再从凭证到账查一遍，两头都对得上，账才算实。

不同学科的现场长得不一样，真实与厚度的具体形态也随之变化，评分人要按学科去找。化学的现场是实验记录本、失败批次、仪器谱图与实验室温湿度，真实性落在原始记录能否与仪器导出的数据、与课题组档案对上，厚度落在复做与档案比对的逐步程度；化学现场证据同样受安全门槛约束，记录里不出现危险步骤的操作细节，涉及危险操作的部分按实验室规范由带教老师在场完成并签字。历史的现

场是县志、族谱与老人口述，苏文那一问“你家族谱断了的那一代，县志同年记了什么？”要求的是两份文献的同年对照，真实性落在文献出处可查、口述有录音或签字的访谈记录，并尊重家族隐私。计算机的现场是系统日志，方舟那一问“崩溃前十分钟最先变慢的是哪一类请求？”要求的是从选课系统故障日志里取出的原始时间序列，真实性落在日志来源与访问权限的说明。不管学科怎样变，判断规则不变：材料能被核对，并且能与子本中的判断双向对上。

在大模型可以生成看似真实的记录、表格乃至访谈稿的时代，现场证据的真实性还需要一层额外的保障，就是留痕的时点。一份夜班计时记录如果在夜班当天就上传到课程平台，带着平台的时间戳，它的可信度远高于期末前一周一次性补交的一叠表格。一份复做记录如果每一步都写在实验室的纸质记录本上，带教老师当天签字，它比一份打印整齐的电子表格更难伪造。所以本书建议，九问课堂的现场证据应当随做随交，而不是期末汇总；课程平台记录每一次上传的时间，评分人在评厚度时把时间线一并读进来。随做随交还有一个附带的好处：老师每周读学生轨迹时，能及早发现哪个学生迟迟没有下到现场，在问题变成期末低分之前就去帮他。

◆ 反刷分规则：十轮打转不如三轮推进

只要过程被计分，就会有人去刷过程。问对计分，就有人把问对拉到30轮；现场计分，就有人把照片拍到300张；改题计分，就有人把九问每一问都改一遍。本书评估方案的配套规则第一条就是按质不按量：轮次、页数不计分。落到过程项上，就是几条具体的反刷分规则，它们应当写进课程大纲，在学期一开始就让学生知道。

第一条规则：十轮打转不如三轮推进。问对深度只计有推进的轮次，没有推进的轮次不加分，反复出现同一类打转的，还要在“接住批判”一项上扣分，因为打转往往正是没有接住批判的表现：ChatSDE翻过一次前提，学生换个说法把原来的前提再提一次，ChatSDE再翻，如此往复。这像第6章用过的中医复诊的比喻，复诊的价值在于每一次都根据上一次服药后的变化调整方子；如果病人每次都不按方子吃药，回来说同样的症状，复诊十次也只是同一次门诊重复了十遍，大夫不会因此多算他十次疗效。

第二条规则：改题只计有来由的版本。版本史中没有时间、没有理由、对不上任何材料的版本，一律不计；为了显示“我改过题”而做的字面改动，例如把“集中在哪个交接时段”改成“主要分布于哪些交接时段”，不算一个新版本。第三条规则：现场证据只计与判断咬合的材料。与子本判断无关的材料，无论数量多少，不进入厚度评定；明显来自网络或他人的材料冒充现场材料的，现场证据一项计零，并且触发对整份子本的复核。第四条规则：过程项之间互相印证，不能互相孤立。问对记录说回过现场，现场证据里就要有那一次；版本史说某次改题是因为问对中的某一问，问对记录里就要有那一问。三项之间对不上的地方，是评分人最该停下来细看的地方。

为了让评分人有一个共同的起点，三项过程项可以各分几档描述。下面是一张示意的档位描述表，标明为本书建议口径，各课程可以按学科细化，但档位之间的分界应当始终落在“质”上，而不是落在“量”上。

档位（百分制）	② 改题质量	③ 问对深度	④ 现场证据
85—100	两锚俱在，提升使题更锐、更可检；每个版本有来由	轮轮有推进；接住批判并转为行动；关键判断由学生首发	材料真实可核；多源、多时点；与子本判断双向咬合
70—84	两锚俱在，提升幅度有限；版本来由基本清楚	多数轮次有推进；接住批判但转化不充分；判断首发存疑	材料真实可核；来源较单一；主要判断有来处
60—69	一锚松动，或基本未改但原题已合格	推进少、打转多；对批判或硬顶或投降	材料真实而单薄；部分判断找不到来处
低于60	两锚俱失（逃离），或版本史无来由	无推进；判断由ChatSDE说出、学生认领	材料不可核或与判断无关；冒充现场者计零

表中“基本未改但原题已合格”一格需要单独说明。九问是种子不是考卷，改题权是权利，不是义务；如果学生认真检查之后认为老师的原题已经两锚俱在、无需改动，并在版本史里写明理由，这本身也是一次判断，可以落在及格档。但它拿不到高档，因为第②项量的就是学生把谱系化接到自己手里的那一步，没有走这一步，就没有可供评定的提升。这像学徒拿到师傅画好的图纸，看过之后说“这张图不用改”，可能是对的，但他还没有在这张图上留下自己的一笔。

这几条规则的共同道理是：过程项要量的不是学生付出了多少时间，而是发生在多大程度上真的走过了一条路。时间与轮次只是路的长度，推进、接住、首发、咬合才是路的方向。一个学生用三轮问对、一次复做、一次改题就走到了判断，与另一个学生用30轮问对兜圈子最后被ChatSDE推着走到同一个判断，前者的过程项应当更高。这对勤奋的学生并无不公，因为勤奋若是有方向的，它自然会在推进与咬合上留下痕迹；勤奋若是无方向的，评分就该如实告诉他：力气花了，路还没有走出来。

◆ 轨迹作为不可还原性的证据：那份法学子本怎样评

现在回到开头那份拿不出轨迹的法学子本。它的产物质量不低，可以设想它在五维上量出130分以上；它的过程项几乎是空的：版本史只有原题一版，问对记录两轮且没有推进、没有接住、没有首发，现场证据没有任何可核对的原始材料。评分人面对的问题是：这份子本的第⑤项还能不能按产物给高分？

本书的回答是：不能直接给。理由要从第16章的I维说起。I维问的是拿掉学生的现场与轨迹，这份子本的核心判断能否被已有材料重新生成。对一份有轨迹的子本，评分人可以沿着版本史、问对记录与现场证据，一步步看到判断是怎样从学生的现场里长出来的，于是可以确认，拿掉这条轨迹，判断就无处生根。对一份没有轨迹的子本，评分人无法做这一步确认：它的判断也许真的来自学生的现场，只是学生没有留下记录；也许来自某次未留档的对话，或者来自别人。评分人不能证明它是代写的，但也无法证明它不可还原。在这种情况下，I维不能按文本表面给分，因为不可还原性本来就是要靠证据撑住的主张，而不是一个文本自带的性质。

所以，拿不出轨迹的高分子本，在本书的口径下按以下程序处理。第一步，过程项三项按实际证据计分，版本史、问对、现场三项都会很低。第二步，第⑤项的I维按“未能确认不可还原”处理，评在120分以下，由此触发第16章所说的两闸封顶，综合分封顶145；在本科一列里这条封顶本身不算严，但I维占五分之一的权重，它的低分会把综合分明显拉下来。第三步，给学生一次补证的机会：在限定时间内补交能够核对的原始材料与过程记录，或者参加一次围绕子本核心判断的口头答辩，由换手评分者提问，看学生能否当场说出这个判断从哪份材料里来、在问对

中第一次在哪里出现、哪一种证据会推翻它。补证或答辩通过，I维重新评定；不通过，维持原评。

这个程序同时回答了“AI代写”的问题。在大模型能写出结构严整、引注规范的文本之后，只看产物的评估已经无法分辨谁写的；查重软件对生成文本的作用有限，靠文风猜测又极易误伤。九问课堂的回答不是去猜文本，而是要求文本带着它的来历一起交上来。一份子本的来历包括它的题怎样从老师版改到定版，它的判断在问对的哪一轮由学生先说出，它的证据在现场的哪一次记录里。这条来历本身就是最难代写的东西：不是因为技术上写不出一份看似合理的版本史，而是因为版本史、问对记录、现场材料三者要互相咬合，并且要经得起换手评分者的追问与答辩。这像古董鉴定，看器物本身固然重要，但真正让人放心的是一条完整、可查的流传记录；一件器物精美绝伦而来历全无，行家不会因为它精美就为它出高价。

回看韩正，他的子本结论与那份法学子本相近，但他的过程项会显著更高，他的I维也能被确认，于是在第⑤项上同样可能高过对方。两份子本在产物上的差距不大，在剖面上的差距却很大，这正是第15章所说剖面比总分多出来的信息。还需要说明一点：这个程序并不预设拿不出轨迹的学生一定作弊。有的学生确实做了现场工作，只是不习惯留记录，补证程序就是为他们设的；九问课堂也应当从第1周起就教学生留轨迹，把版本史、问对记录、现场笔记当成研究本身的一部分，而不是交作业时的附件。

补证与答辩也会遇到公平性的质疑：口头答辩是不是偏向能说会道的学生，让不善表达却做了真工作的人吃亏？这个质疑有道理，所以答辩的问题被严格限定在来历性事实上：这个判断从哪份材料里来，在问对的哪一轮第一次出现，哪一种证据会推翻它。这三问考的是记忆与材料的咬合，不考口才；一个真做过现场的学生，即使讲得磕绊，也答得出“第三周在哪一页笔记里看到了什么”，一个没做过的学生，再流利也会在这样的细节上露出空档。同时答辩不是唯一出口，学生可以选书面补证，在限期内补交带时间痕迹的原始材料，由换手评分者核对，两条路径的结论效力相同。这像驾校的路考，考官不看你说得多漂亮，只看你在路口有没有打灯。

◆ 与学习分析、写作过程研究的对话：轨迹多出的一格

过程性评价在教育研究里并不陌生。学习分析（learning analytics）长期把学习者在平台上的行为轨迹，登录、停留、提交的先后，拿来推断学习的状态，与终结性成绩互为补充；写作过程研究也早就用修订历史和版本演进观察一段文字是怎样长成的，近年更有人尝试用修改轨迹辨认文本的来历。本章所说“轨迹是不可还原性的证据”与这两条线同向：它们都认为，只看终点产物，会丢掉最能说明这件产物来历的那部分信息。本书不是第一个把轨迹当证据的，也不该假装是。

多出的一格有三处。第一是取什么。学习分析多取后台行为日志，轨迹是平台替人记下的；本书取的是三类指定材料，版本史、问对记录、现场记录，由学生自己留下并注明来由。日志告诉你谁在什么时间敲了键，指定材料告诉你一个判断是怎样从现场长到纸面上的。一个现成的反例：学生可以把别处生成的整段文字逐字敲进去，键击和停留的日志与真人写作几乎无异，日志在这里证明的只是“有人在敲”，证明不了“判断从哪里来”。第二是用轨迹回答什么问题。过程数据常被用来预测成绩或预警辍学，本书只用轨迹回答第16章I维那一问：拿掉学生的现场与轨迹，这个判断能否被已有材料重新生成。轨迹在这里不是行为画像，而是一份来历证明，并与题闸前后呼应：题闸在入口问拿掉现场通用大模型能不能答，轨迹在出口问拿掉轨迹判断还能不能被重新生成。

第三是起点与首发。第零轮在问对开始前定格，ChatSDE的追问不出现核心判断的关键变量词，所以轨迹里“谁先说”可以被判读，而不只是“有没有记录”。写作过程研究有版本史，但未必有一个在动笔之前就被封存的初判基线，这是本书的轨迹相对既有做法多出的又一格。三处合起来，说的不是本书的轨迹比别人的更多，而是它被设计成可以回答一个具体的问题，并且这个问题与题闸、首发口径是联动的。

当然会有反对意见。既然轨迹这么重要，干脆装一个全程监控，把学生所有的键击和对话都记下来，岂不一劳永逸？本书不取这条路，理由有二。其一，全程监控把研究变成被盯着的作业，学生会为了“看起来有推进”去表演，这正是反刷分规则要防的；其二，如前所说，它记录得了动作，记录不了判断的来源，监控越细，评分人越容易把动作的密度当作发生的深度，也就是把量重新请进了按质不按

量的评估里。本书的选择是随做随交的指定材料加答辩，成本落在学生每次写下几行来由说明上；监控则把信任成本压在整个班级身上，这个账不划算。

◆ 本章怎样算错

这一章押在轨迹难以伪造、推进可以判读、过程项不是产物项的影子这三件事上，各给一个读数。先查伪造：组织一次红队检验，请不少于5名教师或研究生在一周内借助通用大模型，为同一份九问伪造版本史、问对记录和现场材料，再把伪造件混入真样本交换手评分者在不知情时评分；若伪造件被评进过程项85分以上的比例超过30%，或者其中有三成以上通过了围绕核心判断的口头答辩，那么“轨迹是不可还原性的证据”这一主张站不住，过程项需要换成更难伪造的证据形式，例如在问对与现场工作中引入实时留痕。

再查可判读：取连续两届、每届不少于30份问对记录，由两名换手评分者各自逐轮标注“推进”或“打转”；若两人标注的一致率低于75%，说明按质不按量的计分要点还不能稳定执行，第③项的要点需改写，或降低权重。最后查独立性：对同样的学生，若连续两届、每届不少于30人，过程项（②③④）均分与产物项（⑤⑥）均分的相关系数高于0.9，过程项只是产物项的影子，八项结构应相应简化；反之，若过程项均分比产物项均分高出20分以上的学生超过四分之一，则需回头检查是过程项计分过宽，还是子本打造环节（见第12章）出了问题。

过程项按质计分，产物项按五维计分，八项各有其证据与方向。但还有一个问题没有回答：八项的权重由谁来定，定到多少？医学的学生与化学的学生、准备接手饭馆的周川与准备写作的叶青，是否应当用同一套权重？如果允许学生在区间内自选权重，选择本身又该怎样看待？这是下一章的任务。

> **带 走 的 话**
>
> **过程项量的是一条路有没有走过，不是走了多久：十轮打转不如三轮推进；一份拿不出来历的好文本，来历补齐之前，不能算作不可还原。**

◆

18

第18章

权重组合：学段、课型、学科与自选

◆ 一张表格交到学生手里的那一刻

开学第2周，《小微经营者的动力与节律》研讨课的老师把一张表发到每个学生手里。表上列着八个评估项目，每一项后面不是一个固定的百分数，而是一个区间，最下面留了半页空白，标题是“权重理由书”。周川拿着表看了很久，举手问了一个问题：成绩不是老师来定的吗，为什么要我自己选权重？老师没有直接回答，只是请他先想一想，这门课结束以后他打算拿学到的东西去做什么。周川想了一会儿说，家里的小饭馆父亲已经撑不太动了，他毕业后多半要回去接手。

同一个星期，另一栋楼里的《写作与阅读的发生》研讨课上，叶青面对的是一张结构相同的表。她的回答和周川完全不同：她准备以写作为业，最在意的是能不能写出一个别人没写过的判断，至于下现场的厚度，她觉得自己的现场就是自己的草稿史和方言，用不着跑多远。两个学生拿到的是同一套八项目，填出来的却是两种剖面的权重。这一幕让不少第一次接触九问模式的老师感到不安：如果每个人的权重都不一样，成绩还有可比性吗？如果学生可以自己挑，会不会人人都往自己最强的那一项上堆分？

本章要回答的就是这种不安。本编前三章已经讲清了八个项目各量什么、怎样计分（见第15—17章），这一章只讲最后一步：八个项目怎样加权合成一份剖面和一个总分。本章的核心判断是：权重随学段、课型、学科而变，并在老师给定的区间内交给学生自选——选择权重本身就是一次判断。权重不是阅卷时才拿出来的系数，而是课程开始时就摊在桌面上的一份关于“这门课为何而开”的声明；当它的一部分交到学生手里，它也就成了学生关于“我为何而学”的第一份声明。

在讲授课堂里，权重通常是一行很少有人细看的说明文字：平时成绩占30%，期末考试占70%。这种权重之所以不引人注意，是因为它背后的假设很统一——所

有学生学的是同一本书，考的是同一套题，权重只是在“平时努力”和“最后答案”之间做一点调和。学生不需要理解它，也没有理由去选它，因为不管怎么配，衡量的都是同一件事：像不像标准答案。

九问模式下情况变了。八个项目量的是发生的不同环节：①看母本底盘是否扎实，②看改题是提升还是逃离，③看问对有没有推进，④看现场证据的真实与厚度，⑤看子本里有没有新判断，⑥看子本是否成形，⑦看碰撞中有无撞出班级判断，⑧看学完以后人有没有被改写。这八项彼此不能互相替代，一个现场记录极厚的学生未必能写出新判断，一个判断锋利的学生也可能从没认真下过现场。于是权重就不再是调和，而是取舍：把三成给⑤还是给④，说的是这门课最想让学生长出什么。

打一个日常的比方。同一家饭馆，早市档和晚宴档考核厨师的重点并不一样：早市要的是出餐快、口味稳，一笼包子和昨天一模一样就是好包子；晚宴要的是一道菜能不能让客人记住，稳当只是底线。如果拿早市的标准去考晚宴厨师，最后留下来的都是不出错的人；拿晚宴的标准去考早市，后厨会乱成一团。课堂的权重也是这样：它把“这门课的好”写成了数字，写错了，整门课的方向就会被成绩悄悄拉走。所以权重组合要分三层来定：先按学段与课型定大框，再按学科调细部，最后在区间内交给学生。

> **这一章的问题**
>
> 权重若随学段、课型、学科乃至学生自己的选择而变，成绩还有可比性吗？让学生自选，会不会人人都往自己最强的那一项上堆分？

◆ 学段与课型：四列权重怎样读

本书采用的第一张权重示例表按学段与课型列出了四种组合。它不是唯一正确的配方，而是四个有代表性的示例，供院系据此调整。读这张表时，最好一行一行地横着读，看同一个项目在四种课里怎样移动，而不是竖着把一列加起来；每一列合计都是100，真正说话的是行与行之间的落差。

项目	大一补练手	本科研讨课	研究生课	应用实践课
① 母本底盘	20	5	0（门槛）	5

项目	大一补练手	本科研讨课	研究生课	应用实践课
② 改题质量	10	10	10	10
③ 问对深度	20	15	15	10
④ 现场证据	10	15	10	25
⑤ 子本创新智商	20	30	40	25
⑥ 子本完成度	10	10	15	10
⑦ 碰撞贡献	5	10	5	10
⑧ 回写证据	5	5	5	5

先看两行不动的。②改题质量四列都是10，⑧回写证据四列都是5。这两行不动并不是因为它们不重要，而是因为它们是九问模式的两个关节：改题是谱系化由学生自己完成的那一步，回写是这门课是否真的改写了人的那一个证据，无论在哪个学段、哪种课型，这两处都必须被看见。它们的权重不高，是因为它们更适合做“有没有”的判断而不是“多少分”的竞赛；一旦把改题的权重抬高，学生就会为改而改，把题改得面目全非来换分，那恰恰是逃离而不是提升。

再看变化最大的两行：①母本底盘与⑤子本创新智商。在大一补练手里，①占20，⑤也只占20；到本科研讨课，①降到5，⑤升到30；到研究生课，①变成0并注明“门槛”，⑤升到40。这条斜线就是本书“练手与战场”的判断在成绩上的投影。大一补练手是过渡期的产物：今天的大学新生大多没在中学练过发生技术，只好在大一用一两门基础课补一遍，医学院的吴静用“血压调节”走一遍九问，就是这种课的典型。补练手的终点老师事先写得出，母本底盘本身就是要掌握的东西，所以①给到20，与⑤持平；问对深度③也给到20，因为这时候要练的正是怎样和ChatSDE来回磨。到了研究生课，母本底盘不再是计分项，而是进门的条件：底盘不过关，后面的分都不计，过了关也不再加分。这好比驾照考试，场地考试在驾校里要一项项计分，上了路以后没人再给你的倒车入库打分，但没有驾照你根本上不了路。

应用实践课的特点在④：现场证据给到25，与⑤持平，是四列中最高的。应用实践课的学生往往是要回到具体场所去解决具体问题的，他们的子本价值很大一部分取决于有没有把那个场所真正读透；一个判断再锋利，如果没有扎在现场的材料

里，拿回去也用不上。相比之下，研究生课的④只占10，并不是说研究生不需要现场，而是研究生课的现场证据会在⑤与⑥里被更严格地追问——没有现场的新判断过不了题闸，也过不了换手评分对I维的审查。权重的分配从来不是孤立的，同一种能力可能在不同的项目里被重复检验，权重表要避免的是在同一处重复计分，而不是在不同处重复要求。

这里有一个必须正面处理的张力。配套规则写明：大学阶段⑤的权重不低于25。可是第一列大一补练手的⑤只有20。两者并不冲突，但需要说清楚。补练手课做的是中学本该完成而尚未完成的练手，它的性质是再发生，终点老师能事先写出；⑤不低于25的硬底管的是战场上的课，也就是九问打向连老师也写不出终点之判断的那些课。补练手是借大学的教室补中学的课，所以它适用练手的权重；等中学的九问练手普及以后，这一列就该从大学的权重表上取消（见第25章）。反过来说，任何一门大学的研讨课、研究生课或应用实践课，如果给⑤的权重低于25，就等于把战场又退回了练手场，那就不是在调整权重，而是在改变课程的性质。

◆ 学科组合：医学重现场，化学重可复做，人文重新判断

学段与课型定下了大框，学科再在大框之内调细部。本书采用的第二张示例表以本科研讨课为底，列出四类学科的权重。这张表同样要横着读：它比的不是哪个学科的总分更难拿，而是同一个项目在不同学科里分量为什么不同。

项目	医学	化学	经济／法学	历史／文学
① 母本底盘	5	5	5	5
② 改题质量	10	10	10	10
③ 问对深度	15	10	15	15
④ 现场证据	20	15	15	10
⑤ 子本创新智商	25	30	30	35
⑥ 子本完成度	10	15	10	10
⑦ 碰撞贡献	10	10	10	10
⑧ 回写证据	5	5	5	5

这张表里，①②⑦⑧四行在四个学科之间完全相同，差别集中在③④⑤⑥四行。这本身就说明了一件事：学科之间的差异不在发生的骨架上，九问模式的立骨、走来路、铺土壤、回写四段在每个学科都一样；差异在于每个学科的“新判断”从哪里来、靠什么站住。本书的三领域对应（详见第19章）可以帮助读这张表：自然科学侧重知识发生的客体属性，社会科学侧重互动属性，人文侧重主体属性，医学横跨三者。权重的学科差别，大体就是这三种属性在计分上的投影。

医学一列的④现场证据是20，四个学科里最高，⑤创新智商是25，恰好落在硬底上。这样配的理由在病房里。许宁在老年病科轮转时发生的“交接断点”判断，之所以算得上一个判断而不是一句漂亮话，全靠她跟夜班记下的时间、翻过的交班本和患者自己的讲述。医学的判断一旦脱离现场就会立刻退化成指南里的常识，“老年人跌倒有哪些危险因素”这种问题，拿掉现场材料，一个未接入这些材料的通用大模型张口就能答，题闸早已把它打回。所以医学把现场抬高，并不是轻视新判断，而是承认在病房里，新判断的分量首先取决于它扎得有多深。医学的⑤停在25而不更高，还有一层考虑：医学子本里的判断只是本病房假说，要交科室验证，课堂不宜鼓励学生为了追求“新”而把一个尚未检验的假说说得过满。

化学一列的特点在⑥：完成度给到15，是四个学科里最高的，而③问对深度降到10。化学的完成度含可复做性，沈澄“规程里没写的那一句”，如果别人照着她的记录复做不出同样的差异，就只是一次偶然。打个比方，一份菜谱写得再动人，别人照着做不出那道菜，它就不算一份菜谱。化学把完成度抬高，是把“能不能被别人复做”当成判断成立的一部分；问对深度略降，则是因为化学的推进更多发生在复做批次与谱图之间，而不完全在对话里，问对记录只是轨迹的一部分，实验记录本是另一部分，后者已经在④和⑥中被看见。

经济／法学一列是四个学科里最接近本科研讨课基准的一列，③④都是15，⑤是30。社会科学的现场是交易与裁判，周川的饭馆流水账、韩正的旁听记录，都需要下现场；但社会科学的判断又常常落在“看不见的节律”上，例如“动力被账本提前收走”“调解进程被物业的账单周期推着走”，这类判断的价值主要在于它把散落的互动连成了一个此前没人说出的结构，所以⑤与④之间保持了二比一的关系。历史／文学一列把⑤推到35，是四列中最高，④降到10。这并不意味着人文学

生可以不下现场，苏文的县志与族谱、叶青的草稿史与方言，都是不折不扣的现场；而是因为人文的主体属性决定了它的现场往往就在学生自己身上，取证的工作量不如病房或实验室那样外显，而判断的锋利程度才是它与已有解释拉开距离的地方。“不敢写的词才是叙述的承重词”这样的判断，现场材料可能只是十几页草稿，但它能不能站住，几乎全看它有没有说出别人没说过的东西。

两张表怎样叠用，是院系最常问的问题。学科表以本科研讨课为底，如果要给一门医学研究生课定权重，不能把两张表的数字简单相加再平均，而应当先从学段表定下这门课的性质：研究生课，母本底盘改为门槛，⑤至少要在研讨课之上；再看学科表里医学相对本科基准移动了哪几行——④上调、⑤压在硬底附近——把这种移动的方向带进研究生课的框里。换句话说，学段表给的是“打多深”，学科表给的是“往哪边偏”，叠用的是方向，不是数字。这好比裁缝做衣服，尺码定了身长，体型定了哪里放宽、哪里收紧，两样不能互相顶替，也不能简单折中；尺码是一件事，版型是另一件事，最后的衣服要两样都对。

需要强调的是，这张学科表和上一张学段表一样，是示例而不是规定。一个法学院的应用实践课，完全可以从“应用实践课”一列出发，再参照“经济／法学”一列调整，最终的配方由院系与任课老师据课程目的定下。读这两张表的正确方式不是查表取数，而是追问每一个数背后的理由：为什么这里高，为什么那里低，如果我的课程目的不同，哪一行应该随之移动。

◆ 门槛不是权重：伦理与安全的一票否决

两张表的格子里都没有出现的，是两条门槛。学科表下另有注明：医学另设伦理门槛，去标识化、许可齐全，否则整份不计分；化学另设安全门槛。门槛和权重是两种性质不同的东西，混在一起就会出大问题。权重回答的是“这一项占多少”，门槛回答的是“这份作业算不算数”。如果把伦理做成一个占10%的计分项，就等于告诉学生：你可以在伦理上丢掉10分，换取别处多拿20分。这种交换在病房和实验室里是不可接受的。

门槛与权重必须分开，还有一层认知上的根据。塞勒关于心理账户的研究（Thaler　　1985）指出，人们并不把钱当作可以任意互换的整体，而是分进不同

的“账户”，每个账户有自己的规则。本书借这一思路作类比，并不是说学生的算分习惯已被实验证明如此：分数很容易被学生当作一个可互换的整体，在这里少拿一点、在那里补回来，似乎天经地义。伦理与安全之所以要从分数的账户里整个取出来，正是为了不让它进入这种互换的算术，因为一旦它标了价，学生就会自动把它和别的项目放进同一个账户，开始问“丢这10分划不划算”。这好比食堂里把“不许浪费”写成扣分条款，很快就会有人算出多吃一份的收益超过扣分；写成“浪费者不得再领餐”，这笔账根本打不起来。

日常生活里有现成的例子。驾驶考试里，闯红灯不是扣几分的事，而是当场不合格；一个人倒车入库再漂亮，也补不回一次闯红灯。伦理门槛就是课堂里的红灯。许宁的九问几乎每一问都碰到真实的人：跌倒的患者、写交班本的护士、陪护的家属。她的子本里如果出现一个可以被认出的患者，或者她跟夜班、访谈患者的安排没有经过科室与伦理委员会的许可，这份子本就整份不计分，不论它的创新智商评得多高。这样设计，是为了让伦理在学生的权衡中根本不进入交换，学生在选择权重时，伦理从来不是选项。

化学的安全门槛同理。沈澄的第6问要她复做一批，把每一步的时间、温度、湿度都记下来，这里的复做必须在实验室安全规范之内进行，危险步骤不改动，复做方案须经指导老师审核。一个为了验证湿度假说而擅自改动危险步骤的学生，即便拿到了漂亮的数据，这份工作也不能计分。门槛还有一层不太被注意的作用：它保护老师。九问模式把题交给了学生的现场，现场越真，风险越真；门槛把风险的底线写在课程开始之前，老师就不必在期末面对一份“成果很好但来路不正”的子本时左右为难。

门槛在程序上要放在评分之前，而不是评分之后。具体做法是：子本提交时，先由课程助教核对两样东西——许可文件是否齐全、文本中是否残留可识别的个人信息或危险操作细节——核对通过，才进入八项评分；核对不通过，退回补正，补正不了就不计分。把门槛放在前面，评分人在读子本时就不必一边欣赏判断、一边担心来路，两件事不会互相污染。这和医院手术前的核对单是同一个道理：核对单不评价手术做得好不好，只决定手术能不能开始。法学与历史虽然没有在表中另设

门槛，但司法公开规则、家族隐私与口述伦理同样必须遵守，各学科样板章会分别说明。

◆ 自选区间与理由书：周川与叶青的两张表

现在回到开头那一幕。配套规则写明：学生可以在老师给定的区间内自选权重，并写理由书。这条规则有三个要件：区间由老师给，选择由学生做，理由要写出来。三者缺一，自选权就会变味——没有区间，自选就成了刷分；没有学生的选择，区间就只是摆设；没有理由书，选择就只是一次勾选，而不是一次判断。

先说区间。在经济管理研讨课上，老师以“经济／法学”一列为基准，固定①②⑧三项和两条门槛，其余五项各给一个上下浮动的区间，⑤的下限钉在25。这里的设计逻辑是：固定的三项是九问模式的骨架，不能随个人偏好挪动；浮动的五项，是学生可以按自己的去向调整着力点的地方。区间的宽窄由老师据课程目的定，本章不给统一数字，只给一条原则：区间要宽到让不同去向的学生能选出明显不同的剖面，又窄到任何一种选择都不会让某一项实际上被放弃。

周川的选择是把④现场证据从15拉到25，⑤从30降到硬底25，③问对深度从15降到10，八项合计仍是100。他在理由书里写：毕业后要接手家里的饭馆，这门课对他最重要的，是把三年流水账和父亲月底泄气的那几天真正读透；他知道这样选会让自己的⑤只占25，但他愿意让成绩更多地取决于自己有没有把现场摸清，而不是判断是否足够新。这份理由书最有意思的地方，是它和周川的改题史互相照应：他把题从“哪个伙计干得最起劲”改成“父亲为什么一到月底就泄气”，本身就是一次向自家现场的靠近，权重的选择不过是把这次靠近又说了一遍。

叶青在写作课上做了相反的选择。以“历史／文学”一列为基准，她把⑤从35拉到45，③从15降到10，⑦碰撞贡献从10降到5，同样合计100。她在理由书里写：她准备以写作为业，最想检验的是自己能不能写出一个别人没写过的判断；“你写外婆时，哪个方言词一直不敢写进去”这一问已经把她带到了承重词上，她愿意把成绩押在这个判断能不能站住。老师在批阅理由书时提出了一处保留意见：⑦降到5，意味着她不太在意班级碰撞，而写作课的同学里恰有几位也在写方言与家族，碰

撞或许正是她的判断被撞得更硬的地方。叶青保留了自己的选择，但在理由书上补了一句，说明自己会参加碰撞场，只是不想让成绩为此加码。

老师审读理由书，不是给它打分，而是问三个问题。第1问，这个选择与你的去向对得上吗？周川要接手饭馆，偏重现场就对得上；如果一个准备读博的学生把⑤压在硬底，却说不出为什么，就需要谈一谈。第2问，这个选择与你的改题史对得上吗？权重偏向哪里，改题往往也已经朝那里走过一步，两者对不上，常常说明学生对自己的题还没有想清楚。第3问，你知道这样选会失去什么吗？一份只写“我擅长这一项”的理由书，说的是偏好而不是判断；一份写得出“我这样选，就会在哪一项上少得，我愿意承担”的理由书，才把选择当成了取舍。三问之后，老师可以提出保留意见，但不改学生的选择，除非它越出了区间或碰到了硬底。

这三问的用意，可以借阿吉里斯与舍恩的双环学习（Argyris & Schön 1978）来说明。单环学习只调整行动去达到既定目标，就像恒温器把室温拉回设定的度数；双环学习则追问设定的度数本身是否合适，改的是支配行动的变量。一个学生若只是揣摩“怎样选才拿分多”，他走的就是单环：目标（总分）不动，只调权重这个手段。理由书要他做的是第二环：先说出支配他选择的那个变量，比如周川的“毕业后要接手饭馆”，叶青的“想检验自己能不能写出别人没写过的判断”，再让这个变量接受老师三问的质询。所以理由书的价值不在字数，而在它能否让支配变量显形：一份只写“我擅长现场”的理由书停留在单环，而一份写出“我把成绩押在现场上，是因为我不确定自己将来是否真有判断可说”的理由书，已经碰到了第二环。当然，双环学习在实践中并不容易发生，这一传统自己就强调人们往往在捍卫支配变量而不是检验它；这也是本书要让学期末的自述回头对照第3周写定的理由书的缘故，对照的对象是一份自己写下、无法抵赖的旧稿。

这两份理由书说明了“选择权重本身就是一次判断”是什么意思。周川和叶青选的不是更容易拿分的配方，而是对“我为何而学”的一次明确表态，这个表态可以被老师质疑，也会在学期末被学生自己检验。选权重的时点要放在学生提交改题版本之后、问对展开之前，并且选定后锁定，不得在看到阶段分数后再改，这样它才是一次事前的判断，而不是事后的找补。理由书本身不单独计分，但它会归入问对档案；学期末学生写期末自述时，回看自己第3周写下的理由，常常是第⑧项回写

证据里最诚实的一份材料：周川也许会发现，真正改写他的不是流水账本身，而是他在流水账里看见了父亲；叶青也许会发现，她押注的那个判断，是在一次她原本不想加码的碰撞中才站住的。

◆ 自选的限度：三种反对意见

自选权并非没有对手，至少有三种反对值得正面回答。第一种来自最小指导批评。Kirschner、Sweller与Clark（2006）认为，对新手而言，几乎不给指导的探究或发现式教学效率低下，因为新手缺少可调用的图式，认知资源被耗在盲目搜索上。自选权重看起来正是一种“少给指导”的做法：把一个有后果的决定交给对课程还不熟悉的人。本书的回答分两层。其一，自选不是在空白处自选：区间由老师划定，硬底和门槛钉死，选前有一份写明各项含义的说明，选后有三问，这更接近Lazonder与Harmsen（2016）所说的适度指导，他们对探究式学习的元分析显示，效果取决于指导量，有适度指导的探究优于无指导的探究。其二，指导量应当随学生的新手程度而变：大一补练手课里的学生多数还没见过八个项目怎样咬合，区间宜收窄到只开放一两项，甚至先由老师定权重，学期中再做一次“如果让你重选，你会怎样选”的回顾练习，让他们先看过权重怎样起作用，再在下一门课里真正自选。如同教人游泳，先在浅水区扶着走几趟，才谈得上让他自己选泳道。

第二种反对是前文提到的“堆分”：学生会不会把权重都堆到自己最强的项目上？用周川的数字可以做一次检验。他的八项得分在第15章已经给出：①55、②88、③85、④90、⑤65、⑥55、⑦85、⑧60。按基准权重，总分是74.3；按他自选的权重（①5、②10、③10、④25、⑤25、⑥10、⑦10、⑧5），总分是75.8，只多了1.5分（逐项核算见第15章）。拆开来看，他把10个点的权重挪给了得分90的④，贡献增加9.0；又从⑤拿走5个点，少算3.25，从③拿走5个点，少算4.25，两头相抵，净增1.5。这个小数字说明两件事。一是堆分的上限被区间和硬底锁住了，最强的一项抬到顶，也抬不出几分。二是他的选择不像一条按得分高低排出来的曲线：③得85，是他的强项之一，他却削了它；⑥只有55，是他最弱的两项之一，权重原封不动。单纯为了刷分的人会往强项上堆、从弱项上撤，周川抬高了④，这一步像刷分，其余几步却是由去向而不是由分数驱动的。更根本的是，权重在学生提交改题版本之后、问对展开之前就已锁定，那时他不知道自己最终会在哪一项得多

少分；所谓“押最强项”，押的其实是他对自己的预判，预判准不准，本身就是第⑧项里一份有价值的材料。

第三种反对是公平：两个学生剖面相同、权重不同，总分就不同，排名怎么办？这个反对有真实的重量，因为奖学金、保研、评优常常要一个可排序的数字。本书的回答不是否认它，而是承认总分在这门课里是剖面的投影，投影的角度是学生自己选的。必须排序时，院系有两条路：一是只在同一权重配置的学生之间排序；二是统一用基准权重重算一个仅供排序的总分，自选权重的总分只写进成绩单。周川的两个数字，74.3与75.8，恰好说明这两条路的差别很小：自选并没有把他挪进另一个名次段。反过来，如果两个数字的差别大到足以改变名次，那是需要老师介入的信号，它往往意味着区间开得太宽，而不是学生耍了花招。

◆ 硬底与分档线：为什么⑤不低于25

自选权给出去以后，必须有一条谁也不能越过的线，这就是⑤子本创新智商的权重在大学阶段不低于25。这条硬底的理由，在于九问模式的母题：未来大学的课堂交付的是发生，产出是“建成了自己子本的人”。子本里须有母本所无的判断，⑤量的正是这一点。如果允许学生把⑤压到10，然后把权重全部堆到①母本底盘和⑥完成度上，一个把母本复述得滴水不漏、子本写得四平八稳却没有一个新判断的学生，就能拿到一个漂亮的总分。那样的课堂在成绩上就退回了讲授课堂，只不过换了一套更复杂的表格。

硬底还防住另一种风险，那就是“现场崇拜”。周川把④拉到25时，⑤恰好停在硬底上，这不是巧合，而是区间设计的结果：老师可以允许学生偏重现场，但不允许他把现场当作终点。现场证据再厚，如果最后没有长出一个判断，就只是一份详细的调查记录；饭馆的流水账抄得再全，不说出“动力被账本提前收走”这样的判断，它对周川将来接手饭馆的帮助也有限。⑤的硬底保证了无论学生怎样调整着力点，判断始终是这门课的重心。

与硬底并列的是分档线：按创新智商口径，本科110、研究生125、博士140。分档线管的不是权重，而是合格。它回答的问题是：在这个学段上，一份子本的新判断要达到什么程度才算过关。五维160分制如何按档位分段映射到百分制，第16章已

经说明，这里只强调一点：分档线与权重是两把尺子，前者定“够不够”，后者定“占多少”。一个本科生把⑤的权重拉到45，并不会让他的分档线随之提高；他押得越多，只意味着这一项的得失在总分里影响越大。这也是叶青的选择有风险的地方：如果她的判断最终没有跨过110这条线，她押上的那45%会让总分下滑得比别人更明显。自选权从来不是免费的，它带着后果，而承担后果正是判断的一部分。

权重越往⑤和③倾斜，评分的来路就越要干净。配套规则规定，③问对深度与⑤子本创新智商须换手评分并写利益冲突声明，这条规则在自选权重之下分量更重。叶青把⑤拉到45，比基准的35多出近三成，这一项的评分误差传到总分上也就放大了近三成；如果评分人恰好是指导她改题的老师，或者评分只靠一个与问对同底座的智能体，她押上的那45%就押在了一个未经换手的读数上。同底座的多个智能体不能互作独立证据，这一点在第11章已经讲过，放到权重组合里，它的含义很具体：权重高的项目，必须由与学生的问对过程无利益关联的人或配置来评。与此同时，按质不按量的原则也不因权重变化而松动，学生把③拉高，并不意味着问对轮次越多越好，十轮打转仍然不如三轮推进。

最后，成绩单给八项剖面加总分。这回答了开头那个关于可比性的不安：总分在不同权重之间确实不能直接比较，但剖面可以。看周川的成绩单，用人单位或下一阶段的导师看到的不是一个孤零零的总分，而是一张八项剖面——哪一项厚、哪一项薄一目了然（周川剖面的具体形状见第15章）——外加一份说明他为什么这样选的理由书。这比一个可以横向排名的总分说出了更多真话。可比性没有消失，它从总分转移到了剖面与理由上。

◆ 本章怎样算错

要推翻本章，最直接的办法是做一次对照。取连续两届、每届不少于30名学生、实行自选权重的课程，把每份理由书与同一学生的改题史、去向声明逐份对照。若能找到明确对应关系的理由书不足一半，同时选定的权重较基准权重使总分抬升超过3分的学生超过全班的三分之一，就说明学生是在事后为堆分找说法，“选择权重本身就是一次判断”不成立，自选权应收回老师手里，或把区间收窄到几乎不可选。周川抬升1.5分、理由与改题史对得上，是本章设想的正常形态；抬升超出这个量级而理由对不上，才是堆分。

硬底是否守得住，可以用另一个读数检验。同样的两届里，把总分排名前五分之一的子本取出，若其中⑤原始分低于本科分档线110的超过三分之一，就说明总分仍然可以绕开新判断，硬底不足以守住九问模式的重心，应当上调硬底，或把⑤从权重项改为与分档线挂钩的门槛项。

最后看门槛。以一个学年为窗口，把全部被判伦理或安全不合格的子本调出来，若有超过10%最终是以扣分而不是整份不计分处理的，门槛就事实上退化成了权重，本章关于两者必须分开的设计没有落地；此时该改的是执行办法，而不是门槛本身。以上三条都要靠院系真实的成绩档案来检验，而不是靠本章的论证；书中的人物与课堂均为构造案例，只作说明。

本章把第五编的评估方案收了口：八个项目是什么，各怎样计分，又怎样按学段、课型、学科与学生自己的选择合成一份剖面。但所有这些表格里的数字，都还悬在半空——医学一列为什么把现场证据抬到20，伦理门槛为什么一票否决，只有回到病房里才能真正看清。从下一章起，本书进入第六编“学科九问样板”，逐个学科把九问模式从母本走到子本、再走到评估。第一站是医学：许宁所在的老年病科，那里的夜班、交班本和一个被省略的句子，会把本章的每一个权重都放到真实的重量下面去称一称。

带走的话

权重不是阅卷时才拿出来的系数，而是开课时摊在桌面上的声明：学段、课型与学科定框架，学生在区间内自选并写下理由，硬底与门槛守住底线，可比性从总分转到剖面与理由上。

◆



第　六　编

学科九问样板

编　序

老年病科的晨间交班，夜班护士念到一位高龄患者凌晨在床旁跌倒。许宁答，年纪大、肌力差，每个词都有出处；走廊里，两位护士却在小声说，这个月第三次了，又是快交班的时候。教材里那个“没有错”的回答，解释不了这句话。同样的空白，在实验台前、在装置旁、在判决书后、在县志里，也都有。

本编给出五个学科的样板，每章同一结构：从该学科发现版课堂的样子写起，经谱系母本、完整九问、改题与问对、子本形态、评估权重与门槛，写到学科特有的风险。第19章医学，战场在病房与交接；第20章化学，在失败批次与规程空白；第21章物理、工程与计算，在装置与系统里；第22章经济与法学，在交易与裁判里；第23章历史与文学，在史料与语言里。

前几编的框架在这里落进具体的学科。读者可以先读本学科的一章，再回头读其余。各学科的风险不同，伦理门槛、实验安全、隐私与口述伦理，各守各的底线。

19

第19章

医学：病房里的九问

◆ 指南背得出，病房看不见：发现版医学课堂

老年病科的晨间交班会上，夜班护士念到一位高龄患者凌晨四点多在床旁跌倒，所幸只是皮肤擦伤。主任转头问轮转的五年级学生：你们怎么看？许宁几乎不假思索地回答，患者八十多岁，肌力下降，又有骨质疏松，年纪大是最主要的危险因素。这个回答没有错，每一个词都能在老年医学教材和跌倒防治指南里找到出处，带教老师也点了点头。可是交班会散了以后，许宁在走廊里听见两位护士在小声说：这已经是这个月第三次了，又是快交班的时候。她愣了一下——自己刚才那个“没有错”的回答，对这句话一个字也解释不了。

这一幕是医学课堂上典型的困惑。学生背得出指南，答得出危险因素，考试能拿高分，可站到具体的病房里，面对“为什么偏偏是这几次、偏偏在这个时候”，却发现手里的知识只能给出一个放之四海皆准的解释，而放之四海皆准的解释，恰恰说不出这间病房的任何事。本章是第六编“学科九问样板”的首章，要把九问模式在医学里从母本走到子本、走到评估，完整走一遍。本章的核心判断是：医学课堂的战场在病房与交接里；九问把临床推理从“套指南”变成“在本病房里发生判断”。

先看医学的发现版课堂长什么样。传统的医学教育沿着讲授体系展开：先学解剖、生理、病理、药理，再学内外妇儿，最后到病房里实习，把学过的知识对上真实的病人。这条路的功绩毋庸置疑，它让一代代医生掌握了统一的疾病知识，使得一位在北方受训的医生和一位在南方受训的医生，面对同一种疾病能说同一套语言。它的方向是S→D→E的发现版：先立起体系，再按体系推理，最后到病人身上去找体系预言的那个终点。许宁在交班会上的回答，就是这条路的标准产物——她

从“老年人跌倒”的知识体系出发，推到眼前这位患者，找到了体系早已写好的答案。

二十世纪后半叶，医学教育内部对这条路提出了最有力的改造，这就是问题导向学习（PBL）。Barrows与Tamblyn（1980）在《问题导向学习：医学教育的一种方法》中系统阐述了这一方法：不再先讲完基础再碰病人，而是从一个临床问题出发，学生在小组里分析病例、找出自己不懂的地方、自主学习、再回到病例上验证，导师从讲授者变为引导者。这个方法的功绩至少有三处。它把“为什么要学这个”放在了学习之前，学生是带着问题去查书的；它把推理过程本身变成了学习对象，学生要说出自己是怎样从症状走到诊断的；它让小组讨论成为课堂的主体，打破了一个人讲、一群人听的格局。九问模式在不少地方受惠于这一传统，问题在先、学生主动、导师退后，这些都是PBL早已证明可行的东西。

但PBL的边界也在这里看得清楚。PBL的病例通常是教师事先写好的“纸上病人”，病例背后有一个明确的学习目标，学生走完讨论，最后要到达的是教师预先设定的那组知识点。用本书的判别第1问来检验：终点能否事先写出？在PBL里，答案是能。这就是说，PBL改造了发现的路径——从讲授驱动变为问题驱动——却没有改变发现的方向，学生仍然是在去找一个先在的终点，只不过走的路更主动、更像临床。打个比方，PBL像是把驾校的场地考试从“背交规”改成了“模拟路况练车”，比死背强得多，但模拟的路况再逼真，也是考官事先布置好的场地。九问模式要把学生带到的，是没有人布置过的那条路：这间病房、这几个夜班、这本交班本，连带教老师也说不出这里的跌倒究竟落在哪里。

指出这条边界，并不是要用九问模式取代PBL。在医学教育的前几年，学生还在建立疾病知识的底盘，纸上病人恰恰是最合适的练手场：终点事先写得出，学生可以在一个安全的情境里反复练习从症状走到诊断的推理，练熟的正是改题、追问、查证这些发生技术的雏形。按本书“练手与战场”的划分，PBL更接近练手的一种成熟形态，而五年级的病房研讨课才是战场。两者的关系不是新旧替换，而是前后相接：一个在PBL小组里练过“先找出自己不懂的地方”的学生，到了病房里更容易接受“同样年纪为什么白天跌得少”这样的追问；反过来，一个只在纸上病人里练过的学生，第一次面对一本真实的交班本时，常常会不知道从哪里下手，因

为那里没有人替他划好学习目标。九问模式要补上的，正是从纸上病人走到真实病房的这一步。

> **这一章的问题**
>
> 学生背得出指南，却解释不了“为什么偏偏是这几次、偏偏在这个时候”，九问怎样让临床推理在本病房里发生？

◆ PBL的证据：它支持什么，不支持什么

PBL有比较厚的实证积累，本书不能绕过它。Strobel与van Barneveld（2009）把此前的多项元分析放在一起做了一次元综合，结论大致是：在长期保持和技能应用方面，问题导向学习优于传统教学；而在短期的知识保持与标准化考核上，传统教学并不吃亏。这个结论对本书有两层含义。第一层是支持：它说明把学习放进问题里、让学生带着问题去查书，并不是一种只图热闹的做法，长在应用里的知识确实更耐放，这正是本书把PBL看作练手的成熟形态的依据，也是医学院前几年应当保留它的理由。第二层是划界：这些证据衡量的，是PBL有没有把学生带到事先设定的那组知识和技能上，终点都是教师预先写好的；它们回答的是“走向先在终点的路，哪一种走得更牢”，没有回答“终点没有人写过时，学生能不能走出一个新判断，这个判断能不能被检验”。用学车来比，证据告诉我们模拟路况练车比背交规学得牢，却没有告诉我们这样练出来的人，能不能在一条没有人画过线的新路上辨出方向。

一个自然的追问是：PBL本来就有“学习议题自生成”的机制，小组要自己说出“我们还需要知道什么”，这难道不是学生在出题？确实有这一步，这也是PBL留给九问的重要启示。但这个议题空间是被病例设计限定的：病例的写作者在文本里埋好了需要被发现的线索，小组生成的议题大体会落在埋线的位置上，导师的职责之一，正是在议题跑偏时把小组拉回来。九问的议题来自学生的现场，没有人事先知道线埋在哪里，甚至不保证埋了线；同时九问的产物是一个本病房假说，要附上“被什么推翻”的条件，交科室验证。另一个追问来自临床轮转：床旁教学本来就在真实病人身边进行，它算不算战场？本书的回答仍然落在那个判别问题上：终点能否事先写出。床旁教学查房讲“跌倒的评估与预防”，带教老师心里是有答案

的，要检验的是学生有没有走到；九问问“本病房的跌倒究竟落在哪里”，没有人心里有答案，要检验的是这个答案经不经得起科室验证。真实的病人是战场的必要条件，不是充分条件。

◆ 谱系母本：老年医学与轮转档案

许宁所在的是临床医学五年级的《临床推理与病房研究》研讨课，附属医院老年病科轮转。按九问模式，这门课的出发点不是一本教材，而是谱系母本：老年医学中关于跌倒与衰弱的前沿专著，加上本科室历届轮转学生的问对档案（见第3章）。前一源保证学生站在学科的前沿上，知道关于跌倒，学界已经确认了什么、还在争论什么；后一源保证学生站在本病房的历史上，知道前几届学生在这里问过什么、发现过什么、哪些判断后来被科室证实或推翻。按第3章的口径，这两源的比例会逐年向部落档案倾斜：第一年，轮转档案还很薄，母本主要靠专著撑起；五年以后，档案里积累了几十份问对记录和子本，它对本科室的解释力会远远超过任何一本通用专著。

从母本提炼精要库时，带教老师整理出三类核心判断。方程判断说的是关系的结构：跌倒风险是多种因素相互作用的结果，而不是某一种因素单独决定的，年龄、药物、环境、照护安排之间存在着此消彼长的关系。路径判断说的是事情走到跌倒这一步的次序：许多跌倒并不发生在风险最高的那一刻，而发生在风险从一个人手上移交到另一个人手上的那一刻，例如体位改变、离床如厕、人手交替。动力判断说的是什么在推动局面变化：病房里的风险评估往往是一次性的，评估结果被写下以后，就不再随患者状态与照护安排的变化而更新。这三条判断都还只是精要库里的条目，它们是出题的根，而不是现成的答案；许宁要做的，不是证明它们，而是在本病房里看它们怎样落地，或者怎样落不下去。

医学在三领域对应中的位置，决定了这份母本的特殊性。本书把自然科学对应知识发生的客体属性，社会科学对应互动属性，人文对应主体属性，而医学横跨三者：病理是客体，骨质疏松、肌力下降、药物的镇静作用，都是可以被测量的身体事实；医患与交接是互动，护士与护士之间、护士与医生之间、患者与陪护家属之间，风险在人与人的交接中流动；患者叙事是主体，患者自己怎么讲跌倒前那一刻，是任何量表都替代不了的材料。一份只覆盖客体属性的母本，培养出来的是会

背危险因素的学生；只有三属性都在场，学生才有可能在病房里看到病理、交接与叙事是怎样缠在一起的。许宁后来的判断之所以能长出来，正是因为她的九问把三属性都打到了。

这里还要说明一件常被忽略的事：母本底盘在医学课堂里虽然权重不高，却是一道真门槛。九问打向的是新判断，但新判断要站在扎实的常识上，一个连跌倒的常见危险因素都说不全的学生，到病房里去找“指南之外的东西”，多半只会找到指南之内、却被他自己漏掉的东西。所以在九问发下之前，带教老师会用复述加反例的口试检验每个学生的底盘：先请学生复述精要库里的核心判断，再给一个反例，看学生能不能说出这个反例为什么不推翻、或者在哪里推翻了那条判断。许宁在口试中被问到：有一位年轻的术后患者也在夜间跌倒，这是否说明年龄不重要？她回答，这恰好说明跌倒风险不是单一因素决定的，术后的体位性低血压和镇痛药物同样会把一个年轻人放到风险里。这个回答让她过了门槛，也在不知不觉中为后来的前提批判埋下了伏笔。

◆ 许宁的九问：从交接断点出发的出题逻辑

带教老师为许宁出题时，用装有SDE工序层的大模型走了两步（见第4章）：先从精要库里选出要打的核心判断，再按许宁的学生画像把它具体化到她能下到的现场上。许宁的现场是本病房、交班本、护理记录和夜班，这四样东西，未接入它们的通用大模型一样也看不到，这就是九问能过题闸的根本原因。九问的每一问都被要求落到这四样东西中的至少一样上，落不下去的问，再漂亮也不发。以下是许宁的九问定本，原文照录（其他章引用时一字不改）：

1. 本病房近三个月的夜间跌倒，集中在哪个交接时段前后？
2. 交班本上，跌倒前一班对该患者写了什么、没写什么？
3. 跌倒的患者中，有几位当天刚调整过药物？调整是在哪一班记下的？
4. 同样高龄、同样用药，未跌倒的患者床位离护士站多远？
5. 患者自己怎么讲跌倒前那一刻？与护理记录的说法差在哪里？
6. 有家属陪护的夜晚与无人陪护的夜晚，呼叫铃的使用有什么不同？
7. 本科室已有的防跌倒措施里，哪一条在夜班最常被跳过？为什么？

8. 跟一次夜班，记下护士从听到动静到到达床旁的时间，哪一段最长？
9. 如果把跌倒风险评估从“入院一次”改成“每次交接一次”，本病房第一个会被改变的是什么？

第4章已经逐问讲过出题逻辑，这里换一个角度，看这九问怎样把医学的三属性分配到不同的问上，又怎样围着同一个根转。第1问和第8问是时间之问：一个问跌倒落在哪个交接时段前后，一个问护士从听到动静到到达床旁哪一段最长。它们都打向精要库里的路径判断——风险在移交的那一刻最脆弱——但一个看病房的整体分布，一个看一次夜班的细部，一粗一细，互相校准。第2问和第3问是记录之问：交班本写了什么、没写什么，药物调整在哪一班记下。这两问打向互动属性，风险在人与人之间的交接中流动，而交接留下的痕迹，就在交班本和医嘱记录里。

第4问和第6问是对照之问。第4问在“同样高龄、同样用药”的条件下看床位离护士站的距离，等于把年龄和药物这两个指南里的首要因素按住不动，看剩下的差异从哪里来；第6问把有无陪护的夜晚放在一起，看呼叫铃的使用有什么不同。对照之问的力量在于，它让学生自己的现场替她做了一次粗糙的控制实验：如果年龄真是全部原因，同样高龄的患者就不该因为床位远近而有差别。第5问是主体之问，它把患者自己的讲述和护理记录放在一起比较，差在哪里，往往就是记录体系没有接住的那一部分。第7问是制度之问，问的是已有措施里哪一条在夜班最常被跳过、为什么，它不指责任何人，而是看规则在真实的人力条件下怎样变形。第9问是回写之问，它请许宁设想一个具体的改变——把风险评估从入院一次改成每次交接一次——并追问本病房第一个会被改变的是什么。这一问没有标准答案，它要求许宁在前八问的基础上，把她发现的东西推到一个可以交给科室去检验的位置上。

九问的排列也有讲究。从第1问到第8问，学生的现场一步步变近：先看三个月的统计，再翻交班本，再查药物记录，再看床位，再听患者讲，再看呼叫铃，再问措施，最后亲自跟一次夜班。这是一个从档案走进现场、从别人的记录走到自己的眼睛的过程。第9问则从现场回到制度，把发现转化为一个可检验的设想。日常生活里修一台老是半夜跳闸的电闸，有经验的电工也是这样做的：先翻跳闸的时间记录，再查最近接了什么新电器，再看线路走向，最后半夜守在电箱前等它跳一次，

弄清楚了才谈要不要换线。九问不是九道互不相干的题，而是一条从记录走向现场、再从现场走回制度的路。

题闸在医学里有一个特别清楚的对照。合格的问是“交班本上写了什么、没写什么”，拿掉许宁的现场材料，一个未接入这些材料的通用大模型（记录所用版本）看不到本病房的交班本，这一问就无从回答；被打回的问是“老年人跌倒有哪些危险因素”，同一个大模型张口就能答出一份比任何学生都全的清单。后一问并非没有价值，它属于母本底盘，许宁在口试时已经答过；它的问题在于它不是战场上的问。更要警惕的是，这份清单完整、准确、引用规范，学生拿到手，很容易误以为自己已经“懂了”这间病房。医学的题闸还有一个附带的作用：凡是能过题闸的问，几乎都要求学生进入一个有伦理要求的现场，所以过题闸的那一刻，也就是伦理程序必须启动的那一刻。同时要记住题闸的限度：它只检验“必须依赖现场”，不检验“值得研究”，一道问可以因为写了“本病房”而过闸，却仍不值得问；值不值得，要靠带教老师对许宁现场的了解，以及第16章对不可还原性一维的评审来回答（见第4章与附录二）。

出这九问，大模型能做的和带教老师必须亲手做的，分得很清楚。它可以从精要库里把三类核心判断一一调出，可以对每一问做全球敌拓，确认诸如“交接断点与夜间跌倒”这样的问法在本病房的条件下还没有人答过，也可以检查九问之间有没有重复、有没有漏掉某一属性。但它不认得这间病房：它不知道本科室的交接在几点，不知道交班本是纸本还是电子记录，不知道哪几张床离护士站最远，更不知道许宁是一个在交班会上听到护士私语会愣住的学生。第4问里“床位离护士站多远”、第8问里“跟一次夜班”，这些具体到可以下脚的地方，都是带教老师凭着对病房和对学生的熟悉补进去的。老师不可替代的，正是认得学生的现场；在医学里，这份认得还包括知道哪些现场学生可以进、哪些现场必须先走完伦理程序才能进。

九问也并非处处可用，一个反例可以说明它的边界。设想另一家医院的交班已经全部电子化，交班内容由系统按模板带出，护士只勾选、不手写，那么第2问问的“写了什么、没写什么”就要换一种读法：被省略的不再是一句没写下的话，而是模板里没有对应勾选项的那一类信息。这时九问要改的是取证的位置而不是问的

方向，学生要去看模板本身缺了哪些字段，以及护士在勾选之外用口头或便签补了什么。如果连口头补充也被系统取消，第2问就失去了现场，带教老师应当在发题时就把这一问撤下，而不是勉强留着，让学生去做一份没有材料的作业。

◆ 改题与问对：“因为年纪大”被翻过来

九问发下以后，许宁先做的是改题。她最在意的是第2问。按老师版，这一问问的是“跌倒前一班对该患者写了什么、没写什么”，她翻了几份交班本以后发现，“没写什么”远比“写了什么”难找，因为没写的东西不会留下痕迹，只能从下一班的补记、护士的口头交接和事后的护理记录里倒推出来。翻得越多，她越觉得问题不在于漏了什么内容，而在于漏是怎样发生的。于是她把第2问提升为：“交班时被省略的那一句，是谁、在什么情况下省略的？”这次改题过了两锚：根还在，它仍然连着精要库里“风险在移交时最脆弱”的路径判断；离现场更近，它从一份文件追到了一个具体的交接动作和做出这个动作的具体情境。它不是为了找人追责——许宁在改题说明里特意写明，这一问关心的是“什么情况”，而不是“哪一个人”——而是为了看清省略在什么条件下最容易发生。

问对从第零轮开始。按九问模式的规程，学生在与ChatSDE对话之前，先写下自己的初判（见第6章）。许宁写的是：夜间跌倒是因为患者年纪大。这句话和她在晨间交班会上的回答一模一样，说明经过几天的翻档案，她的底层判断还没有动。ChatSDE的第一回合没有替她列出“高龄患者跌倒的危险因素清单”，而是先接住它，承认年龄确实是被大量研究确认的危险因素，然后把前提翻过来问了一句：同样年纪的患者，为什么白天跌得少？

这一问的工序选择值得说明（见第7章）。按本书的判别细则，判断一条初判坏在哪里，要按次序查：先看学生在自己的现场里是否已经接触到、并与初判直接相撞的事实，有，就是断拍，刀是点火，让矛盾被看见；没有，再看初判的关系里是否少了一项被当作常数的因素，少了，就是缺式，刀是补式，追问被漏掉的那一项；关系齐全、也无矛盾，而起手次序与任务不配，才是偏路，刀是排路。许宁的“因为年纪大”落在第二档：她的关系里漏了一项被当作常数的东西，也就是事情发生的时间次序，年龄被默认为在任何时刻都以同样的方式起作用。所以ChatSDE要做的是补式，把缺掉的那一项逼出来，而不是直接告诉她缺的是什

么。“同样年纪，白天跌得少”这个对照，把年龄按住了，剩下的差异只能来自白天与夜间之间别的东西：人手、光线、睡眠、用药时间、陪护安排。ChatSDE没有替她说破是哪一个，只是把问题留在了她必须回现场去答的位置上。

有一处容易读错，需要正面说明：开头那一幕里，许宁在走廊里已经听到“又是快交班的时候”，这算不算她“已经接触到的与初判相撞的事实”，从而让这条初判成了断拍？按细则逐项检验，不算。理由有二。其一，相撞要求事实与初判不能同时成立，而护士那句话与“年纪大是危险因素”并不冲突：高龄患者完全可以既因为年纪大而脆弱，又恰好在某个时段出事，那句话只是提示关系里还缺一项，并没有推翻已有的一项。其二，那是一句耳闻的感叹，不是她自己记下的数据；她当时愣了一下，手里却没有可以摆在初判旁边的统计；后来翻档案的头几天，她翻的也只是一份份单独的交班本，还没有把时段连成分布，这也解释了为什么她写第零轮的初判时仍是原来的那句话。对照断拍的典型就更清楚：如果许宁已经翻完了三个月的记录，知道跌倒大多集中在同一个时段，却仍然坚持“主要是年纪大”，那才是断拍，刀就要换成点火，把她手里本来就握着的两件事摆到一起。

再看与通用大模型的对照。默认配置（未开启学习模式）下的通用大模型，面对“因为年纪大”的前提，往往也会先质疑一句，比如说年龄只是背景、不是机制；可是在同一回合里，它随即把夜尿、用药、视力、肌力、照明这一串替代的原因列好，再建议用跌倒风险量表逐项评估。前提被翻过来了，新的答案也一并交到了许宁手里，这就是附录二所说的“默认替学生说破”（见第8章）。她得到的是一份比自己周全的清单，失去的是自己从病房里把那一项找出来的机会。需要老实承认的是，2025年起，OpenAI的ChatGPT Study Mode（2025年7月）、Google的Gemini Guided Learning（2025年8月）和Anthropic的Claude学习模式（2025年4月）都已把“先追问、不直接给答案”做成了可以开启的选项，开启之后，通用模型同样可能问出“同样年纪，白天为什么跌得少”。这并不削弱本章的叙述，反而说明追问这一层壳并不难抄。它们没有的是更靠近病房的几样东西：没有题闸，不知道这一问是否只有许宁的现场才答得出；没有第零轮的基线，无从判断她的话被问过几轮之后到底变了没有；不会按初判的坏法在补式、点火与排路之间选刀；也不会把结果回写给下一届学生。

接下来的几轮问对，许宁是带着现场材料回来的。第2轮，她带回了第1问的统计：近三个月的夜间跌倒，大多落在凌晨交接前后的一段时间里。ChatSDE追问的是：那一段时间前后，病房里什么东西在变？第3轮，她带回了跟夜班的记录，第8问的答案是，从听到动静到到达床旁，最长的一段不是走路，而是护士正在交接、手上的事放不下的那几分钟。第4轮，她带回了改题后的第2问的初步发现：交班本上被省略的，常常是一句关于“这位患者今晚可能要起夜”的提醒，省略多发生在交接任务集中、时间紧的那一班。到这一步，判断已经呼之欲出，ChatSDE仍然没有替她说，只问了一句：如果把你这几轮带回来的东西连起来，你现在会怎样改写第零轮的那句话？许宁写下的是：本病房夜间跌倒的风险更多落在交接断点上，而不只是年龄——风险藏在交班本里被省略的那一句。这句话是她先说出来的，这一点在第⑤项和第③项的评分里都会被记录下来。

关于这句话算不算许宁“首发”，要把口径说清：首发是相对ChatSDE而言的。第1问的题面里本来就写着“交接时段”，护士在走廊里说过“快交班的时候”，这些都是许宁从老师手里和现场里取得的材料，不算被人替她说破；ChatSDE的四轮追问里，没有一句先说出“交接”两个字，“交接断点”四个字是她把几轮带回来的东西连起来之后自己写下的。这个口径也有限度：九问的题面确实预载了方向，所以首发只能证明她没有被工序层喂出判断，不能证明她没有受到出题的影响，后一层要靠第零轮的基线与她的改题说明去分辨。

需要强调的是，这个判断从诞生的那一刻起就被标明为本病房假说，待检验。它是从一间病房、一段时间、有限的几个夜班里发生出来的，样本小、混杂多，许宁不能、也不应把它说成一个关于老年跌倒的普遍结论。问对的最后一轮，ChatSDE请她列出这个假说最可能被什么推翻：如果交接时段的跌倒集中只是因为那段时间恰好是患者起夜的高峰，与交接本身无关，假说就要修改。这一列，把她的判断从“我发现了”推到了“这里可以被检验”，这正是医学判断应有的姿态。

◆ 子本形态与评估：一份交给科室的本病房研究

九份问对成果做完以后，许宁开始打造子本。按第12章的口径，子本是母本加九问研究成果的重写，不是翻译，子本里必须有母本所无的判断。医学子本在这一点上的检验特别直接：如果许宁的子本只是把老年医学专著里关于跌倒危险因素的

章节换成自己的话复述一遍，再附上几个病房里的例子，那就是一份翻译，哪怕例子再真实也是如此。她的子本之所以算重写，是因为它的骨架换了：专著的骨架是“危险因素—评估工具—干预措施”，她的骨架是“交接断点—被省略的那一句—省略发生的条件—可检验的改动”。母本的知识没有被丢掉，年龄、药物、环境都还在，但它们被放进了一个以交接为中心的新结构里，成了这个结构的背景条件而不是主角。

医学子本的形态通常由四部分组成。第一部分是母本精要的重述，交代学界已经确认了什么，这一部分要准确、要有引注，它也是第⑥项完成度检查的重点。第二部分是现场证据的整理，包括跌倒时段的分布、交班本的比对、跟夜班的时间记录、患者讲述与护理记录的差异，所有材料都必须去标识化，不出现可以识别患者或医护人员的信息。第三部分是判断及其论证，许宁要说清楚她的判断怎样从九问里一步步发生，哪些证据支持它，哪些证据与它不一致。第四部分是检验建议，这是医学子本区别于其他学科子本的地方：子本的结尾不是结论，而是一份交给科室的验证方案，例如建议科室在一段时间内对交班本中“起夜提醒”一项做专门记录，再比较跌倒是否随之变化。这份方案交给科室以后，做不做、怎样做，由科室依据医疗规范决定，学生不参与执行决策。写法上也有一条纪律：许宁的结论只能写成本病房的假说，凡把它写成关于老年患者的普遍结论的子本，在⑤和⑥上都要被追问，这不是压制新判断，而是守住新判断的可信度。

医学子本还有一项常被低估的分值，这就是⑦碰撞贡献。许宁所在的研讨课上，同学们分在内科、外科、老年病科等不同病区轮转，各自带着自己病房的九问回到碰撞场。有同学在外科病房发现，术后患者的意外多发生在从手术室转回病房的那一次交接上；许宁的判断与它一碰，班级里便浮出一个比两人各自的判断都更宽的说法：在病房里，风险最脆弱的时刻是它换手的时刻。这类班级判断在跨学科的碰撞场里还能走得更远，第14章讲过许宁与沈澄的那次碰撞：病房交班本里被省略的那一句，与实验室规程里没写进去的湿度，被看作“没写下来的那一句”的共同结构。医学的⑦给到10，就是要让学生知道，一间病房的判断只有被别的病房、别的学科撞过，才知道它到底是本病房独有的，还是更普遍的结构在本病房的一次显露。

许宁的子本完成以后，会进入两条路。一条进入她自己的知识发生平台，成为她从实习到住院医阶段交接班思考的底盘。需要再强调一次，这里的底盘是思考的底盘，不是临床决策工具：它提醒她在交接时多问一句“有什么没写下来”，但不替她做任何诊疗决定。另一条进入本科室的轮转档案，成为下一届学生母本的一部分。如果科室的验证支持了她的假说，下一届学生的九问就不必再问“风险是不是落在交接上”，而可以直接问“交接中的哪一类省略最危险”；如果验证推翻了它，这个被推翻的假说同样是档案里有价值的一页，它告诉后来者这条路已经走过、走不通。回写要被观察，不能被宣布，第24章会专门讲这一点。

评估按本科医学的示例权重进行：①母本底盘5，②改题质量10，③问对深度15，④现场证据20，⑤子本创新智商25，⑥子本完成度10，⑦碰撞贡献10，⑧回写证据5。第18章已经讲过这一列为什么把④抬到20、把⑤压在硬底25上，这里只补充它在许宁身上的具体含义。④的20分看的是她的夜班记录、交班本比对与患者讲述的真实与厚度，不看她跟了几次夜班、翻了多少页，而看这些材料能不能撑起判断；⑤的25分看的是她的判断有没有说出母本所无的东西，本科分档线是创新智商110。③和⑤都要换手评分并写利益冲突声明，评许宁的问对深度的人，不能是全程陪她改题的带教老师。

在这一切之上，是伦理门槛。医学子本凡涉及患者，一律去标识化，并须经科室与伦理委员会许可；去标识化不彻底或许可不齐全，整份子本不计分，不论其他各项多么出色。许宁的九问几乎每一问都碰到真实的人，第5问要访谈患者，第8问要跟夜班观察护士，第6问要涉及陪护家属，这意味着她在开始取证之前，就必须把访谈的知情同意、观察的告知、资料的保存与销毁办法一一落实。门槛不是事后的检查，而是事前的通行证：没有许可，九问里的大部分问根本就不能开始做。这一点要在课程开始时向学生讲清楚，免得学生把伦理程序看成拖慢进度的手续，而不是医学研究的一部分。

◆ 医学独有的两道线：医嘱与追责

九问模式进入病房，带来的风险比进入任何其他学科的课堂都更直接，因为医学现场里的每一个判断，都可能被误当成对真实病人的处置意见。这使病房里有两道线是别的学科所没有的。第一道是医嘱：学生不改医嘱，不作临床决策，判断作

为本病房假说交科室验证。这条规定要放在医学九问的第一页，而不是最后一页。许宁在跟夜班时，如果看见一位她认为风险很高的患者，她能做的是按照实习规范向带教老师和值班护士报告，而不是凭自己的研究判断去调整任何照护安排。她的子本可以建议科室改变风险评估的频次，但这个建议能不能实行，要经过科室的讨论和医疗规范的审查。这好比学开车的人在副驾驶上看出了前面的路况有问题，他可以提醒司机，但不能伸手去抢方向盘。

第二道线是追责。许宁改题后的第2问问的是“交班时被省略的那一句，是谁、在什么情况下省略的”，在病房这样一个人命关天的场所，这一问如果处理不当，很容易被读成对某几位护士的指责。医院里处理不良事件，通常遵循对个人不追责、对系统追根的原则，一旦护士觉得有人在翻她们的交班本找人，之后的交接就会更谨慎地少写、不写，研究的对象反过来被研究本身改变了。所以医学九问在处理涉及医护人员行为的问题时，要始终把焦点放在“什么情况下”而不是“哪一个人”，子本里涉及医护的材料同样要去标识化。许宁在改题说明里写明了这一点，这不只是伦理上的谨慎，也是方法上的必要：她要找的是省略发生的条件，而条件只有在不追责的氛围里才可能被如实讲出来。

◆ 本章怎样算错

本章的核心判断可以用三个读数来检验。第一个读数针对“九问把临床推理从套指南变成在本病房里发生判断”。取连续两届、每届不少于20名轮转学生的同一科室，把他们交科室验证的本病房判断逐条与指南已列的危险因素条目对照；若超过三分之二的判断能够还原为指南已知因素的换一种说法，例如把“高龄”换成“夜间高龄”，就说明九问没有越出套指南，核心判断在医学本科阶段要降为“九问帮助学生更好地应用指南”。同一次对照里还要看第零轮：若多数学生的初判在问对结束后与第零轮没有实质差别，而改写只是措辞上的，那么问对的推进作用同样要打折扣。

第二个读数针对病房作为战场的设计。以一个轮转周期为窗口，统计九问里经科室与伦理委员会许可、实际可以开始取证的问数；若平均不足5问，学生最终只能回到纸上病例，那么“病房是医学战场”就要修改，或者先在制度上为本科生的病房研究建立可行的通道，再谈九问。第三个读数是边界：只要在九问课堂的任何一

个学期里出现一起学生因研究判断而自行调整照护安排的事件，课程设计就必须先暂停取证、重新核对“学生不改医嘱”的执行，再谈新判断；这一条不设比例，因为病房里一起就已经太多。以上三条都要靠真实的科室档案去检验，而不是靠本章的论证；所有人物、病房与判断均为构造案例，只作说明，不是调查结果，许宁的判断本身也只是一个待检验的本病房假说。

许宁在病房里看到的，是风险藏在一句没写下来的话里。下一章要去的地方看起来与病房毫不相干：沈澄所在的化学课题组，一本写了十年的实验记录本和一批批失败的反应。可是那里同样有一句没写下来的话——规程里那一步为什么“要快”，师兄师姐口头说了十年，却从没写进规程。从病房的交接断点到实验室的规程空白，第20章会把九问模式带进知识发生的客体属性最集中的地方，看它在化学里怎样让规程里没写的那一句显露出来。

> **带走的话**
>
> **指南回答的是“老年人为什么跌倒”，九问问的是“这间病房为什么偏偏在这个时候跌倒”；前者背得出，后者要学生自己下到交班本里去发生，并且只当作待检验的本病房假说交给科室。**

◆

20

第20章

化学：实验室里的九问

◆ 照着规程做的那一个下午

周四下午的有机合成实验课，二十几个学生站在通风橱前，讲义摊在台面一角，每一步都写好了：称量、加料、升温、搅拌、淬灭、萃取、过柱。沈澄那一组做的是课题组正在用的某个偶联反应，产率比讲义上的参考值低了一截。实验报告最后一栏是“误差分析”，她几乎没有犹豫就写下了“试剂纯度不够”。这句话写得很顺，顺到她自己都没有想过去核对：这瓶试剂在别的月份、别的人手里用出过什么产率，那些数字躺在别人的记录本里，不在她手边，她也没有理由去翻。助教在报告上打了个勾，批语是“分析合理，注意试剂保存”。

这一幕在化学系的实验课上每学期都会重演，它让人困惑的地方不在学生偷懒，而在整个课堂的设计本来就指向这个答案。讲义规定了路线，参考产率规定了终点，“误差分析”一栏预设了误差来自几个熟悉的去处：纯度、称量、损失、操作不熟练。学生要做的，是把自己的偏差归进其中一个格子。格子是老师和教材事先备好的，学生的批次、学生那天的实验室、学生手上那一步究竟多快多慢，都不在格子里。于是一个本来可能显露新东西的失败批次，被一句顺口的归因收进了已有的知识。

这就是化学课堂的发现版：S→D→E一条路的讲授形态落在实验室里，先给出体系（反应机理与标准规程），再给出路线（按步骤操作），最后用实验结果去“验证”体系。它有不可抹杀的功绩。化学是一门动手的学科，规程是无数前人用失败换来的安全与可重复性，一个本科生先学会照规程做、做得干净、记录得规范，是进入这门学科的第一道门。就像学厨先跟着菜谱做，火候和刀工没练稳之前，谈不上改菜谱。问题在于，大学实验课常常停在这一道门口，把“照规程做出参考产率”当成全部，而失败批次只是被扣分的对象，不是被追问的现场。

把这个困惑说得更尖锐一些：化学知识的发生，恰恰常在规程没写到的地方。一个反应在甲实验室好用、在乙实验室不灵，一个课题组的老成员“手感好”、新成员总做不出来，这些现象每个做过合成的人都见过，却很少进入课堂。本章的核心判断是：化学课堂的战场在失败批次与规程空白里；九问把“照规程做实验”变成“让规程里没写的那一句显露出来”。下面先看已有的最好改革走到了哪里，再看九问在它的边界上多做了什么。

在全书的三领域对应里，自然科学侧重知识发生的客体属性，社会科学侧重互动属性，人文侧重主体属性，化学属于第一侧。所谓客体属性，是说这里的判断最终要向物质本身交账：反应发生了没有，产率是多少，谱图上有没有那个峰，不由任何人的叙述决定。这一点让化学九问与人文九问、社会科学九问有一个根本的区别：学生可以改题、可以在问对里争辩，但最后裁决判断的是下一个批次。然而客体属性并不意味着化学的现场里没有人。规程是人写的，记录本是人记的，“室温”这个词是人选的，十年档案是一届又一届学生的手留下的痕迹。客体在说话，但它是通过人的记录说话的；记录不写的东西，客体就算说了，也没人听见。

这正是化学课堂最容易被忽略的一道缝。教科书把反应写成物质与物质之间的关系，规程把操作写成人对物质的指令，两者之间那些属于“这一个实验室”的条件——空气湿度、某一步拖了多久、谁的手更快——既不属于教科书的理论，也没进入规程的文字。用日常话说，这就像一道家常菜，书上写的是食材和步骤，奶奶做得好吃靠的是灶台和她那双手，而灶台的脾气、手上的节奏，没有一本菜谱写得出来。九问要做的，是让学生站在这道缝上，把它当作研究对象，而不是当作误差的垃圾桶。

> **这一章的问题**
>
> 化学知识常常发生在规程没写到的地方，九问怎样让失败批次里“没写的那一句”显露出来，又不越过安全规程？

◆ 课程型本科研究：功绩与边界

在讲授式实验课与九问课堂之间，已经有一个成熟而有力的中间形态，即课程型本科研究（course-based undergraduate research experience，简称

CURE）。Auchincloss等（2014）把这类课程的要素概括为几项：学生使用科学实践、做出发现、工作有更广泛的相关性、彼此协作、经历迭代。它的核心转变，是把一门课整体嵌进一个真实的研究课题，让学生做的不再是答案已知的验证实验，而是结果连老师也不知道的真实实验。更早，博耶委员会（1998）在《重塑本科教育》中就呼吁研究型大学让本科生在研究的基础上学习，CURE可以看作这条呼吁在课程层面的落实。

CURE的功绩至少有三处，九问模式全部继承。第一，它把“答案已知”换成了“答案未知”，这正是本书所说练手与战场的分界：老师写不出终点，课堂就离开了练手。第二，它让一个班的学生都能进入研究，而不只是少数被导师挑进课题组的人，研究机会从稀缺变成课程配置。第三，它承认迭代，失败与重做被写进课程设计，而不是被当作事故。沈澄所在的《合成条件的发生》本身就是一门课程内嵌真实课题的课，没有CURE这一代改革开出的路，九问在化学系连落脚的课程形态都没有。

但CURE也有它的边界，这条边界不是它的缺陷，而是它的设计重心。CURE的“未知”通常由课题给定：课题组要筛一批底物、测一组条件、扩一个反应范围，学生在这个框架里贡献数据。框架本身——哪些变量值得测、规程里记什么不记什么、失败算在谁头上——仍是课题组既有的。打个比方，CURE像是让学徒进了真厨房，做的是真菜、端给真客人，但菜单是主厨定的，学徒的发现是“这道菜换这个火候更好”；它很少让学徒去问“这本菜谱为什么从来不写那天的湿度”。学生做出了真实数据，却未必发生了自己的判断。

九问模式在CURE的边界上多做的，是三件事。其一，题不是课题组分派的任务，而是老师从谱系母本精要里按学生画像打造的谱系性研究课题，并且每一问都要过题闸：拿掉学生的现场材料，一个未接入这些材料的通用大模型（须记录所用版本）能否给出说得过去的答案。其二，学生有改题权，可以把题提升到离自己的现场更近的地方。其三，学生的初判要在第零轮写下，交给ChatSDE做批判性激发，而不是交给助教打勾。这三件事合起来，把“贡献数据”推进到“发生判断”。下面这张对照只为看清差别，不为分高下。

维度	讲授式实验课	课程型本科研究（CURE）	九问课堂
终点	参考产率已知	课题结果未知	判断连老师也写不出
题从哪里来	讲义	课题组任务	母本精要按学生画像出题，过题闸
失败批次	扣分对象	迭代的一环	首要现场
学生的产出	实验报告	数据与报告	九问子本，须含母本所无的判断

◆ SEA-PHAGES：与“回写”最近的对手

在CURE的家族里，有一个样板和本书所说的“回写”靠得特别近，值得单独对话，它就是SEA-PHAGES。Jordan等（2014）发表于mBio的论文报告了这个项目的做法：多所院校的本科生在同一门课程里，分别从自己所在地的土壤中分离噬菌体，纯化、鉴定、测序并注释其基因组；学生的工作成果进入一个共享数据库，其中一部分还以论文的形式发表。换句话说，学生的课堂产物不是交给助教打分就沉底，而是成了一个跨校共同体的公共资产。若问一门课怎样才算“回写”，这是目前最成熟的答案之一。谈回写而绕开它，等于关起门来称第一，本书不这样做。

同样要对话的，是Corwin、Graham与Dolan（2015）为CURE提出的研究与评估议程。大意是：不能停在对这类课程的一般性赞许上，要弄清楚课程里究竟是哪些要素在起作用，并用经得起比较的评估设计去检验。这套议程对本书是一种逼视。九问课堂的成效论断，若没有可比较的评估设计，只能算构造案例的说明；书末附录二的小样本试跑，距离Corwin等所要求的比较设计还差得很远，只是朝这个方向迈出的第一步。

两者的回写形态却并不相同，多出来的那一格就落在这里。SEA-PHAGES回写的是数据：一条一条的基因组注释，进入共享数据库之后，任何一所参与院校的学生都可以再用。统一是它的力量，也划定了它的对象。课程的方案由项目统一设计，学生的问题落在噬菌体身上，而不落在方案本身；某所学校的实验室在这套方案上有什么走样、有什么口传的窍门，不是它的回写对象。打个比方，它像一家全国连锁的标准化后厨，菜谱统一，各家出品可以汇总比较；九问关心的却是一家老店的灶台，那个只有这家店才有、没有写进总部菜谱的火候。九问课堂回写的，不

是更多的数据，而是对本组记录语言和规程的一处修订，连同这处修订是怎样来的（问对档案）、改过什么（版本史）、后来被撤回过什么（撤回记录）。

这一格要说得谨慎。首先，它不是SEA-PHAGES的缺陷：标准化为的是汇总，汇总为的是统计力量，这是单一课题组的九问做不到的，在回写的规模和可汇总性上，九问远不如它。其次，一个实验室里长出来的“规程里没写的那一句”，可以被这个课题组采纳，却不能自动推广到另一个实验室，这是九问的代价。最后，两者并不互斥。最自然的组合是：在SEA-PHAGES式的课程里，让每所院校的学生在统一方案之外，再各写一份“本校本实验室方案没写到的一句”，交给带教老师建档，统一的好处与本地的好处就叠在了一起。如果这个组合在实践中只增加了学生的文书负担，而没有产出被采纳的修订，它就不值得做；本章末尾的第三条失败条款，正是照这个标准写的。

◆ 谱系母本与沈澄的九问

沈澄这门课的谱系母本有两源：一是课程型本科研究与合成方法学的前沿专著，二是本课题组十年的反应条件档案。前者给出学科在这类反应上已经说清的东西，后者给出这个课题组在这类反应上真实做过、做坏过的东西。按本书的口径，大学母本的比例要逐年向部落档案倾斜，化学课尤其如此：教科书讲的是反应在理想条件下应当怎样，档案记的是反应在这一栋楼、这几台仪器、这几届人手里实际怎样。专著是地图，档案是这条路上十年的车辙。

从母本提炼精要时，老师要的不是“偶联反应的机理要点”这一类条目，那是统一性知识，大模型讲得更全。精要库里要的是核心判断，分三类。方程判断问的是条件与结果之间的关系式里有没有缺项，例如“规程把环境写成常数‘室温’，而档案显示产率随季节波动”。路径判断问的是操作次序与记录次序是否错位，例如“操作时长靠口头传授，而记录本只记投料量”。动力判断问的是什么在推动一个课题组的规程更新，例如“规程只在出过事故或换了导师时才修订”。这几条判断本身就是待检验的，它们把老师领向出题的方向，而不是给学生的答案。

有了精要，老师用装有SDE工序层的大模型按沈澄的画像出题：三年级、做过两学期合成、细心而不太敢问师兄师姐、初判习惯归到试剂。下面是沈澄的九问（定本原文）：

1. 本组十年档案里产率最低的二十批，各在哪个月份？
2. 同一瓶试剂在三月和六月用出的产率差多少？
3. 实验记录里“室温”一词，十年里有没有一次写过具体温度和湿度？
4. 失败批次里，哪一步操作的时长记录最不一致？
5. 师兄师姐口头说的“那一步要快”，规程里写了吗？
6. 你复做一批，把每一步的时间、温度、湿度都记下来，和档案比，差在哪一步？
7. 谱图里那个反复出现、没人解释的小峰，最早出现在哪一年？
8. 别的实验室做同样的反应，他们的规程里多了哪一句？
9. 如果把湿度写进本组规程，第一个要改的是哪一行？

这九问的出题逻辑可以分三段看。第1至3问是“翻档案”：先把失败按时间排开，再把试剂这个学生最熟悉的解释固定住，最后去查记录语言本身的空白。第2问是有意设计的对照，它把“试剂纯度”这个变量钉死在同一瓶上，只让时间变动；如果差别仍在，原来的归因就站不住。第3问最朴素也最锋利，它不问化学，只问记录本上的一个词，而“室温”这个词在十年里是否从未被展开，恰恰是规程空白最直接的痕迹。好比一家老店的账本，十年都写“杂费若干”，你不翻开它，永远不知道钱漏在哪里。

第4、5、7问转向“人与仪器”。第4问去找操作时长的不一致，第5问把口头传承与书面规程并排放，第7问追一个没人解释的谱图小峰的起点。这三问的共同点，是它们都指向课题组内部那些“大家都知道、没人写下来”的东西。第5问尤其需要老师认得学生：沈澄不太敢问师兄师姐，这一问就是逼她去问，而问的方式也规定好了，不是质疑前辈，而是核对“那一步要快”在规程里有没有对应的一行字。第7问则提醒学生，自己的判断不必一次说尽，一个延续多年的异常本身就是一份时间档案。

第6问是全套九问的支点：亲手复做一批，记全每一步的时间、温度、湿度，再与档案对照。这一问把前五问在纸上找到的空白，带回学生自己的手和自己那一天的实验室，复做的意义不是重复，而是第一次以完整记录的方式做一遍别人做了十年的反应。第8问向外看，第9问向前看。第8问是一次小规模的敌拓：别的实验室做同样的反应，规程里多了哪一句。第9问则把整个探究压到一个可落地的动作上：如果要把湿度写进规程，第一个要改的是哪一行。一个判断如果说不出它会改动哪一行规程，它就还停在议论里。

再用题闸复核一遍。合格的样子是第3问：“实验记录里‘室温’一词，十年里有没有一次写过具体温度和湿度？”通用大模型看不到这个课题组的记录本，拿掉沈澄的现场材料，这一问无从回答。打回的样子则是“偶联反应产率低的常见原因有哪些？”，通用大模型张口就能列出一页。第8问处在边缘：别的实验室公开发表的规程，大模型可能检索得到；但这一问的落点是“他们多了哪一句”，比较的另一端是本组规程，所以仍然过闸。题闸不是要求每一问都与外界隔绝，而是要求每一问的答案都必须经过学生的现场。

◆ 改题、问对与复做

九问发下去之后，第一件事不是做实验，而是改题。九问是种子不是考卷（见第5章），学生可以改、可以提升，但改题有两锚：根还在，即仍连着母本的核心判断；离现场更近，即更深地扎进自己能触到的记录、批次和操作。化学课里分辨提升与逃离，有一个特别直观的办法：看改过的题会不会把学生更早地带到通风橱前和记录本前。举个假设的例子，如果有学生把第4问改成“失败批次里时长最不一致的那一步，是不是总落在同一个人或同一个时段？”，这是提升，题更窄、更贴现场，根仍连着“操作时长未入规程”这条判断。如果改成“怎样提高偶联反应的产率”，那就是逃离，题回到了通用大模型能直接回答的地方，过不了题闸。

改题之后进入问对。沈澄的第零轮写的就是她在实验报告里写过的那句话：“产率低是因为试剂纯度不够。”第零轮的意义，是让学生的初判先落在纸上，成为后面一切推进的起点和对照。ChatSDE的第一回合没有反驳“纯度”，而是先接住：纯度确实是常见原因，而且你注意到了产率偏低，这本身就是一个值得追的现象。然后它把前提翻过来看：“同一瓶试剂，为什么三月好、六月差？”这

一问没有说纯度不对，它只是把纯度这个变量固定住，让时间显露出来。就像一个人说“我睡不好是因为枕头不舒服”，朋友问他“同一个枕头，为什么出差时反倒睡得好”，问题本身就在提示另一个方向。

这一回合背后是工序选择。沈澄的初判坏在哪里？不是推理错了，而是关系式里少了一项：她把环境当成了常数。按本书的判别次序，先查有没有一件她自己已经接触、与初判直接相撞、判断却没有改的事实：没有。三月与六月的对照躺在旧年的档案和别人的记录本里，不在她手边，所以这不是断拍。再查关系里是否少了一项被当作常数的因素：少了，所以是缺式，ChatSDE选的刀是补式，即把被默认为常数的那一项拿出来变成变量；同时它指出立的方向，不是“你错了”，而是“去档案里看月份”。后面几轮她先翻档案、再问师兄、再回通风橱复做，查证的次序几度变换，容易让人误读为偏路；其实那次序是补式追问一路带出来的结果，并不是起手的次序与任务不配。这正是第1问和第2问在问对中被激活的样子。随后几轮的推进，第6章已经按七轮完整展开，这里只摘与本学科特有的部分。问对中最关键的转折，不发生在对话框里，而发生在沈澄回到实验室复做一批之后。

复做是化学九问里不可替代的一环。沈澄按第6问的要求，把每一步的时间、温度、湿度都记下来，然后和档案比。她发现档案里大多数失败批次缺的不是投料量，而恰恰是这些她刚刚记下的东西，而在可以比对的少数批次里，差别集中在某一步操作的时长上。她把复做记录带回问对，ChatSDE没有替她把结论说出来，而是追问：“你把这一步的时长、失败集中的月份，还有师兄那句‘要快’放在一起看：这一步越拖越坏，拖的时候，除了时间在走，这一步里还有什么也跟着变了？”这一问里没有出现任何指向具体变量的词，它只是指出，“除了时间”之外还有一项没被点名。这时她想起，复做之前为第5问去问师兄时，师兄转述过更早的师姐当年丢下的半句：“快一点，免得受潮。”这句话是沈澄在现场取得的材料，出自一位做了好几年的师兄之口，它进入问对的方式，与档案里的月份分布、她自己的复做记录没有区别：都是她从现场带回对话框的东西。当时她追问受的是什么潮、潮到哪一步，师兄答不上来，只说夏天那一步慢了容易坏，再往下他也说不清。十年口传的那一句，第一次和一份完整的复做记录对上了。

沈澄把这几条线连了起来，说出了本章的那条判断，它目前的地位是待复做验证的假说：那一步的成败依赖未入规程的环境湿度与操作时长，也就是“规程里没写的那一句”，在师门口头里传了十年。请注意这条判断里有两个“未入”：湿度没有进入记录语言，时长要求没有进入书面规程；而它们彼此相连，口头说“要快”其实是在替规程补一个没写出的湿度条件。这条判断是沈澄先说出来的，ChatSDE只是在她接近时把两条线递到她手边。这一点对评估很要紧，因为问对深度要看核心判断是谁先说。如果是ChatSDE先说破，沈澄得到的只是一条信息；她自己说出来，才是一次发生。

这里有一处必须老实交代：“受潮”两个字是师兄说的，“湿度”两个字还出现在老师发下的九问里（第3、6、9问），那么“核心判断由谁先说”到底怎么算？本书的口径分两层。第一层，首发是相对ChatSDE而言的：现场里他人说的话、档案里的记录、老师出的九问，都是学生取得的材料，不算替学生说破；而ChatSDE的追问不得出现核心判断的关键变量词，上面那一问里没有“空气”，没有“湿度”，没有“受潮”之类的词，就是照这条规则写的。第二层，词的出现不等于判断被说出。师兄的“受潮”是一句手艺人的习惯语，没有说清受的是哪一步的潮、季节怎样改变它、本组规程为什么没有它；九问里的“湿度”是作为“要记下的量”出现的，不是作为“原因”被断言的。判断是一个关系：把季节分布、档案里从未记录的温湿度、某一步的时长、口头的“要快”连成一条关于“本组规程缺了什么”的主张，并附上什么结果会让它作废。这个关系是沈澄连起来的。评审据此要查两件事：ChatSDE的每一句追问里有没有核心判断的关键变量词；判断的每一次改向，触发它的材料来自哪里、动手连线的是谁。两项都是问对记录里查得到的，不靠印象。

问对中也有沈澄接不住的时候。第7问追谱图小峰，她翻到最早出现的年份，想把小峰也归进湿度假说，ChatSDE提醒她：小峰最早出现的那一年，档案里还有一次仪器更换，两件事在时间上重叠，不能只凭先后断定因果。沈澄一度很沮丧，觉得自己刚建起来的判断又被拆了一角。这正是“一破一立”在化学里的样子：破的是过度延伸，立的是判断的边界。她最后在子本里把小峰列为“与湿度假说相容但尚未区分的现象”，留给下一届去做区分实验。一个敢于写出“这里我还说不清”的判断，比一个什么都解释得通的判断更像化学。

不妨把这段问对与同一个学生直接问通用大模型的情形放在一起看（第8章已作一般讨论，这里只看化学的特殊处）。如果沈澄在默认配置、未开启学习模式的通用大模型里提交“产率低是因为试剂纯度不够，怎样提高”，她大概率会得到一份周到的清单：试剂重结晶或换批次、惰性气氛保护、无水溶剂、控制加料速度，等等。这份清单并不错，甚至可能包括“注意除水”，但它是替沈澄说破的一份清单：纯度只被轻轻质疑了一句，各种可能的原因就被一并摆好，哪一项在本组这个反应里真的起作用，不必她自己去找，沈澄的十年档案、那瓶三月和六月用出不同产率的试剂、那句口头的“要快”，都不在它的视野里。

2025年起，情形有了变化。OpenAI的ChatGPT Study Mode（2025年7月）、Google的Gemini Guided Learning（2025年8月）、Anthropic的Claude学习模式（2025年4月），都把“先追问、不直接给答案”做成了可以开启的学习模式。开启之后，沈澄大概会被问到“你怎么确定是纯度的问题？有没有做过对照？”，这是个好问题，本书不否认它的价值，而且它恰好支持本书的判断：差别住在包装层，不住在底座。同时它也说明，“苏格拉底式追问”这一层壳并不难抄。壳以外的东西，学习模式并没有给：它不知道同一瓶试剂三月和六月用出过不同产率，因为那是本组十年档案里的事；没有题闸和改题权，把问题卡在只有现场才答得出的地方；没有第零轮写下的初判作基线，也不会按她这一次的坏法（缺式）去选刀。它追问得再得体，问的也是一个通用的“你怎么知道”，而不是在接入本组档案之后才问得出的“同一瓶试剂，为什么三月好、六月差”。通用大模型缺的不是化学知识，也不全是追问的姿态，而是把学生的现场接进来、把前提翻过来看的那套工序。

还有一层要说清：沈澄的问对并不只在人与ChatSDE之间。第5问逼她去问师兄师姐，第8问逼她去读别的实验室的规程，第6问逼她回到通风橱前。ChatSDE在这里的角色像一位好的带教，它不替学生做实验，也不替学生和师兄谈话，它做的是在每次学生回来时，把他带回的东西放到判断上掂一掂：这一条撑住了什么，又松动了什么。问对的轮次之所以要多、要回现场往返，原因就在化学判断的材料不在对话里，在实验室里。十轮不离开对话框的问对，在化学课上往往不如三轮各回了一次实验室的问对。

◆ 子本：可复做性是完成度的核心

九问做完，沈澄要打造的是子本：母本加九问研究成果的重写，而不是把专著和档案逐章翻译一遍（见第12章）。化学子本的大纲大致是这样展开的：先交代谱系母本里这一类反应已有的条件判断，再交代本组十年档案的结构与它的记录空白，然后用九问的成果重写“这个反应在本组为什么时好时坏”，最后给出一份修订规程的建议稿和一份留给下一届的待区分清单。母本里没有的判断，就是那条关于未入规程的湿度与时长的假说；没有这一条，子本就只是一份整理得很好的档案摘要。

化学子本与其他学科子本最大的不同，在于它必须可复做。人文子本的判断要经得起史料与读者的检验，医学子本的判断要交科室验证，而化学子本的判断要交给另一双手、另一天、另一个批次。所以本书把可复做性放在化学完成度的核心位置：子本里的复做记录，要详细到一个没见过沈澄的同学，拿着这份记录、在同样的安全条件下，能够做出可以比对的那一批。这就像一张真正好用的家传菜谱，不是写“盐少许、火候到了就出锅”，而是写清楚哪一步需要多久、厨房那天潮不潮；写不清楚的地方，也要老实标出“这里靠手感，尚未量化”。

可复做性在课堂上要落到几个可检查的地方。其一，记录是否完整：每一步的时间、温度、湿度是否都有，缺项是否注明原因。其二，比对是否诚实：复做批次与档案批次的比较，是否只挑了支持假说的那几批，被排除的批次是否写明了排除理由。其三，判断与证据的距离是否标清：哪些是复做直接看到的，哪些是推测，哪些像谱图小峰那样暂时说不清。其四，是否给出了失败条件：如果下一届控制了湿度、时长仍然不稳，这条假说要怎样修改。一份子本如果只有一个漂亮的结论而没有这四样，它在化学里就不算完成。

还要强调一点：可复做性不是要求学生一学期内把假说证实。沈澄的判断在这门课结束时仍是“待复做验证”，这不是子本的缺陷，而是子本的诚实。子本交出去，进入课题组的部落档案，下一届学生可以拿它出题、复做、修正；第24章会讲到这条湿度假说怎样被下一届复做后写进课题组规程。课程不是每年重做一遍同一个实验，而是一条会长的谱系，每一届的子本是下一届的母本的一部分。可复做性，就是让这条谱系能够接得上的那个接口。

子本的学术规范在化学里也有特别处。引证要分清三种来源：公开发表的文献与专著，课题组内部档案，以及学生自己的复做记录与访谈记录。三种来源的证据地位不同，子本里不能混写；尤其是“师兄说”这一类口头材料，要写明是访谈所得、何时何人，并与书面规程并列呈现，而不是悄悄当作事实写进论证。课题组档案属于部落财产，子本引用它时要遵守课题组的数据使用约定，涉及未发表的结果，须经指导老师同意。学生不是在写一份给自己看的实验笔记，而是在为下一届的母本写一段可以被引用、被质疑、被复做的文本。

◆ 权重、安全门槛与化学特有风险

化学课在本科阶段的权重示例如下（%）：母本底盘5、改题质量10、问对深度10、现场证据15、子本创新智商30、子本完成度15、碰撞贡献10、回写证据5。与本科研讨课的通用示例相比，化学把问对深度的5个百分点挪给了子本完成度；与医学相比，它的现场证据权重低一些，完成度高一些。这个配法背后的理由是：化学的判断最终要回到可复做的实验，问对再精彩，复做不出来就等于没有发生；而现场证据的一部分，已经通过可复做性被吸收进完成度里了。子本创新智商仍是核心项，30的权重高于大学阶段不低于25的硬底。

沈澄的发生剖面可以作为一个示例：现场证据与完成度高，她的复做记录在全班最完整；碰撞贡献弱，她在同学碰撞场里话少，班级子本里属于她的那一笔不多。剖面不是要她把短板补成一样高，而是让她和老师都看见自己长在哪里、缺在哪里。值得一提的是，第14章那场跨院系的碰撞里，沈澄与许宁相撞，撞出“没写下来的那一句”作为实验室与病房共同结构的判断；那是她少有的一次接过别人的判断去撞，不在本门课两次碰撞场的记录之内，也说明她在本门课里碰撞弱未必是能力弱，可能只是同学科的场没有给她合适的对手。

按第18章的规则，学生可以在老师给定的区间内自选权重并写理由书，化学课也不例外，但区间的设定要考虑这门学科的特性。完成度在化学里不宜被压得太低，因为它承载着可复做性，一份不可复做的化学子本，创新智商再高也难以进入部落档案；创新智商也不得低于25的硬底。在这两道约束之内，一个准备读研、打算继续做方法学的学生，可以把现场证据略微拉高，用更厚的复做记录为自己的判断加重；一个打算转向化学教育的学生，则可以把碰撞贡献拉高，因为他未来的工

作正是组织别人碰撞。理由书本身就是一次判断：学生要说清楚，在自己这一份九问里，哪一项最能说明发生了什么。

化学课的分档线与全书一致，本科以创新智商110为线。这条线对化学学生意味着什么，可以用沈澄的子本来想：如果她只是把档案整理清楚、复做一批、写出“湿度可能有影响”，那多半落在线下，因为“湿度影响某些反应”是教科书里已有的判断，母本里就有；她的子本之所以可能过线，在于她把湿度与操作时长、口头传承与书面规程连成了一条关于“这个课题组的规程为什么缺这一句”的判断。换句话说，化学子本的新，不一定是新反应、新机理，更常见的是对一个具体实验室的知识怎样被记下、怎样被漏掉的新判断。这一点也提醒评分者，不要只拿合成方法学期刊的创新标准去衡量本科化学子本。

化学另设安全门槛，其地位与医学的伦理门槛相同：不过门槛，整份不计分。门槛的内容不在于学生懂多少危险化学，而在于九问设计与执行全程不越过既有的安全规程。具体说，复做只在课题组现行规程与安全评估覆盖的范围内进行；九问可以要求学生多记录，不可以要求学生改动操作中的危险步骤；凡涉及条件变动的设想，比如第9问所问的“把湿度写进规程”，在子本里只能以建议稿的形式出现，须经指导老师与实验室安全负责人审定后才可能付诸实验。用开车打比方，九问教学生把仪表盘上每一个读数都记下来，但不教学生在路上拆刹车做实验。

十年档案是这门课最诱人的矿，也是最容易让学生滑脚的地方。设想沈澄的同桌拿到同一份档案，直接对产率和所有能提取的字段做相关性筛查，挑出两三个显著的变量，当天就宣布发现。这样做的毛病不在统计，而在材料：档案里能被提取的字段，是当年的人愿意记的字段，而湿度恰恰一次都没有记过，筛查筛得出来的，只能是记下来的东西。九问把第3问放在前面，就是要先问记录语言缺了什么，再谈从记录里挖规律。这个风险有一个读得出来的征兆：如果子本结论所依赖的字段全部来自档案的现有列，没有一项是学生新增记录的量，评审就应当追问，这份子本是不是只在已有记录里找规律。

复做也有它的滑口。学生一旦知道老师和问对都期待湿度假说，记录就可能不自觉地向假说靠拢：湿度计的读数顺手取整，漏记的时长凭印象补齐。好比称体重的人知道自己在减肥，总会在读数时多信一分少信一分。防它的办法有三层：原始

记录随子本提交；换手评分时由另一位评审抽取复做批次的原始记录独立复核（见第16章）；让记录先于假说，也就是第6问要求只记时间、温度、湿度，而不预告要检验什么。

还有一处在于对口传的态度。“那一步要快”可能藏着真实的湿度条件，也可能只是某位前辈的习惯。一律当迷信丢掉，会把十年的手艺一起丢掉；一律当真理收下，会把一个习惯悄悄写进规程。九问的做法是先让它被复做核实，才有资格进规程，就像老中医传下来的口诀，可以采信，但要先在自己的病人身上验一验。

◆ 本章怎样算错

先说最容易检的一条，也是首发口径的底线：抽查问对记录，ChatSDE的每一句追问里出现核心判断关键变量词的回合数应当为0；任何一名学生的记录里出现一次，该生“核心判断首发”一项即作废，并复核该轮之后的全部推进。这一条不需要等多届数据，一个学期就能查完。

第二条看九问有没有多做那一步。取连续两届采用九问的化学课，每届不少于20名学生，由换手评分者对每份子本里“母本所无的判断”做核查：拿掉学生的现场材料，交给一个未接入这些材料的通用大模型（记录所用版本，默认配置与学习模式各试一次），若超过三分之一的判断能被它给出说得过去的复述，或评分者认定超过半数的判断只是档案里已有规律的重新表述，那么“九问在CURE的边界上多做了发生判断这一步”就要收窄：差别可能只在形式，不在发生。

第三条回到与SEA-PHAGES的对话。同一课题组连续三届采用九问之后，三届合计若没有至少一条修订建议经指导老师与实验室安全负责人审定后写进规程，或者已写进规程的修订在下一届复做后被撤回的超过半数，那么本章所说“多出的一格”（修订、问对档案、版本史与撤回记录）在化学里就只是文书，不是回写。

第四条看可复做性是否误伤了大胆的判断。同一批子本里，如果创新智商进入分档线（本科110）以上的子本有一半以上因复做记录不全而完成度低于及格，同时完成度最高的子本又几乎都只是改动一处记录栏的小判断，说明完成度的核心位置压低了判断的胆量，权重配法与完成度门槛就要重设。

第五条看安全门槛。如果指导老师与安全负责人在审定建议稿时，有一半以上的设想因触及危险步骤被退回，即只要守住安全规程，学生就触不到任何规程空白，那么“化学的战场在失败批次与规程空白里”就要收窄为“在可安全记录的规程空白里”，并承认有一部分化学现场不适合本科九问。这几条之中任何一条出现，都应回到沈澄这个构造案例本身：它只是说明，不是证据；真正的证据要从多届学生的版本史、问对记录、复做记录和下一届的复做结果里来。

沈澄的判断最后落在一句没写进规程的话上。化学的规程空白藏在通风橱和记录本里，而到了物理、工程与计算课堂，空白换了地方：它藏在仿真模型与实测数据之间，藏在一座桥的伸缩缝和一套选课系统的排队环节里。装置与系统不像化学反应那样可以在一个下午里复做一批，它们常常只能被观测、被记录、被回放；那么，当学生无法“复做”一座桥时，九问靠什么让装置自己说话？下一章转向高远与方舟。

带走的话

规程写下的是前人已经知道的，失败批次留下的是规程还没写到的；化学九问把失败当作现场，让“没写的那一句”先被复做核实，再有资格进规程，而且全程不越过安全规程。

◆

21

第21章

物理、工程与计算：装置与系统里的九问

◆ 验证模型的课堂

土木学院的《结构健康监测研讨》第一次课，老师在投影上放出校园那座老桥的有限元模型，彩色的振型图一层层展开：一阶竖弯、一阶扭转、二阶竖弯。作业要求写得很清楚：用课程提供的监测数据识别前几阶频率，与模型计算值比较，误差在规定范围内即为“模型验证通过”。高远很快交了作业，频率对得不错，误差都在范围内。同一周，软件工程的《真实项目研讨》上，方舟的小组拿到学校选课系统去年的故障日志，作业是“分析崩溃原因并提出改进方案”；方舟写的第一句话是“崩溃是因为服务器不够”，改进方案是扩容。两份作业都写得规范，都能拿到不错的分数，也都没有让任何新东西发生。

困惑在于，这两门课都已经用上了真实数据——一座真实的桥，一套真实的系统——按说已经离开了“答案已知”的验证实验，可学生做的事仍然是验证。高远拿实测去对模型，对上了就结束；方舟拿日志去对一个通行的解释，对上了就结束。数据是真实的，问题却是现成的；学生的眼睛被作业要求引向“模型和数据合不合”，而不是“数据里有没有模型没想到的东西”。就像拿着地图走一条熟路，每到一个路口就核对一下地图，核对上了心里踏实，但路边新开的岔道、塌了一半的台阶，因为地图上没有，也就没有被看见。

这就是物理、工程与计算课堂的发现版。大学物理实验测重力加速度、测杨氏模量，给定装置、给定方法、给定公认值，学生的任务是把误差压小；工程课按规范校核构件，代入公式、查表、判断满足与否；计算机课按需求文档实现功能，通过测试用例就算完成。它们都是S→D→E讲授路径在装置与系统上的形态：先给模型，再给方法，最后用数据去验证模型。这套训练的功绩毋庸置疑，一个工程师如果不会校核、不会建模、不会写过测试的代码，就没有资格碰真实的桥和真实的系

统。问题仍然在于停在门口：真实装置与系统最有价值的信息，恰恰在它不服从模型的那一刻。

本章的核心判断是：工程与计算课堂的战场在真实装置与系统的异常里；九问让学生从“验证模型”走到“让装置自己说话”。所谓让装置自己说话，不是说装置有什么神秘的意志，而是说学生不再拿模型去审问数据，而是让数据里那些模型解释不了的段落成为问题的出发点。在全书的三领域对应里，物理、工程、计算与化学同属侧重客体属性的一侧（见第20章），判断最终要向装置与系统本身交账。但与化学不同，一座桥不能在一个下午里复做一遍，一次选课崩溃也不能为了做实验在真实系统上重演一遍。装置与系统通常只能被观测、被记录、被回放，这决定了本章两套九问的出题方式，也决定了它们的风险与门槛。

工程教育里并不缺少走出讲授的改革。毕业设计、项目式课程、校企合作的真实课题，都把学生带到了真实的装置与系统面前；第20章讨论过的课程型本科研究（CURE），在工程与计算领域也有大量对应形态。它们的功绩与在化学中一样：答案未知，人人可进，失败可迭代。它们的边界也一样：题目多半由项目方或课题组给定，学生在既定框架里交付成果，“桥该测什么”“故障该归哪里”的框架本身很少被学生碰到。九问在这条边界上多做的，仍是那三件事：题闸、改题权、第零轮的初判交给批判性激发。区别只在于，工程与计算的现场是装置与系统，九问要把学生的目光从“交付一个满足需求的东西”挪到“看见这个东西不服从模型的那一刻”。

还要说一句物理。本书的贯穿案例中没有为物理设专门人物，这并不意味着物理课堂不适用九问，而是物理的九问形态与工程高度相近：大学物理实验室里那台用了多年的装置，它的系统误差为什么总偏向一边，实验室的老教师为什么叮嘱某个旋钮“别拧到底”，这些问题与高远的老桥同构。本章以高远和方舟两套九问为样板，物理课的老师可以把“老桥”换成“老装置”，把“养护档案”换成“仪器维修与校准档案”，出题逻辑不变。

> **这一章的问题**
>
> 学生用的已是真实的桥和真实的系统，做的却仍是拿数据验证模型，九问怎样让装置与系统自己说话？

◆ 谱系母本与高远的九问

高远这门课的谱系母本，一源是结构健康监测与桥梁工程的前沿专著，另一源是本课题组历年对这座校园老桥的监测档案，以及后勤部门保存的养护记录。前者告诉学生，一座桥在理想模型里应当怎样振动，异常通常意味着什么；后者告诉学生，这一座桥在这些年里实际怎样振动、被谁修过、修了什么。从母本提炼精要时，方程判断可以写成“模型把边界条件写成定值，而实测显示某些段落的响应随时段变化”；路径判断可以写成“养护记录与监测模型由两个部门分别维护，维修事件未必进入模型参数”；动力判断可以写成“模型只在监测数据明显偏离时才更新”。这些判断都是给老师出题用的方向，不是给学生的结论。

按高远的画像出题：三年级，有限元做得熟，对数据处理有信心，习惯把与模型不合的数据当作噪声剔除。他的初判是“振动大是因为车流量大”。这一初判不算错，早高峰车多，振动当然大；它的问题在于把车流当成了唯一的输入，而把桥自身当成了不变的被动对象。ChatSDE后来在第一回合问他的，正是“早高峰的车流过去了，为什么那一段还不回落？”。下面是老师为高远打造的九问：

1. 监测系统上线以来，各测点在早高峰之后回到平日基线，各需要多长时间？
2. 哪一段桥面的振动在早高峰后迟迟不回落？
3. 迟迟不回落的那一段，实测的振动特征与仿真模型的预测在哪里对不上？
4. 那一段附近的传感器，安装、校准与更换的记录写了什么、缺了什么？
5. 后勤养护档案里，这一段桥近十年做过哪些维修？哪一次维修没有进入监测模型的参数？
6. 跟一次早高峰，在人行道上记下那一段的车流、行人与伸缩缝处的声响，和同时段数据比，对得上吗？
7. 同一座桥上结构位置相似的另一段，同样的车流过后，为什么回落得快？
8. 后勤老师傅口头怎么说这一段桥？他的说法在档案里有没有？
9. 如果把那次维修写进仿真模型，第一个要改的参数是哪一个？

出题逻辑分三段。第1、2问先把“回落”这个概念立起来：学生熟悉的是振幅、频率，而“回到基线要多久”是一个模型通常不关心、数据里却天天都有的量。第2问是整套九问的轴心，它把高远从“哪里振得大”引到“哪里回不来”，问

法一变，车流量这个解释就不够用了，因为车流过去之后，车流不再是原因。第3、4问是两道排除题：先问与仿真不合在哪里，把异常定位到模型与实测之间的缝上；再查传感器自身的记录，排除“只是传感器坏了”这一最省事的解释。就像家里的电子秤忽然显示体重涨了两斤，先换块电池再秤一次，才谈得上怀疑自己吃多了。

第5、8问转向桥的“病历”。第5问让高远去翻另一个部门的档案，看维修事件是否进入了监测模型；第8问让他去问那位每天巡桥的老师傅，把口头经验与书面档案并排放。这两问共同指向一个工程里常见而课堂里罕见的事实：一座桥的知识分散在几个互不相通的地方，设计院的图纸、监测组的模型、后勤的养护本和老师傅的记忆，各自都是真的，却从未被放到一处。第6问是现场题，学生要亲自站到桥上去，让耳朵和眼睛与传感器同时工作；第7问是对照题，找一个结构相似、表现不同的段落，让差别自己显露。第9问与沈澄的第9问同构，把判断压到模型里的一个参数上：说不出改哪个参数的判断，还没有落地。

用题闸复核。第2问合格：通用大模型没有这座桥的监测数据，拿掉高远的现场材料，它答不出哪一段回落迟缓。打回的样子是“桥梁振动异常的常见原因有哪些？”，任何大模型都能列出疲劳、支座退化、伸缩缝损坏、基础沉降等一长串。有意思的是，“伸缩缝”正在这串常见原因里。这并不让高远的判断失去价值：知道“伸缩缝可能导致异常”是统一性知识，确认“这一座桥的这一处伸缩缝、因为这一次没进入模型的维修、导致早高峰后回落迟缓”是谱系知识。前者是教科书的一行，后者要经过学生的数据、档案、老师傅和那一个早高峰。

◆ 方舟的九问

方舟的课堂在软件工程的《真实项目研讨》里，谱系母本一源是软件可靠性与分布式系统的前沿专著，另一源是本课程历届学生在校内真实项目上留下的排障档案，加上学校信息中心授权提供的选课系统故障日志与复盘记录。精要库里的方程判断可以写成“容量规划把负载当作总量，而故障发生在某一类请求的局部堆积上”；路径判断可以写成“系统的每个组件有开发负责人，组件之间的交接环节却未必有人负责”；动力判断可以写成“每次故障之后的改进，多半落在最容易申请预算的那一项上”。最后这一条尤其值得学生体会：扩容之所以总是第一个方案，不只是因为技术上合理，也因为它最容易写进报告、最容易获批。

方舟的画像：三年级，编程能力强，熟悉云服务和压力测试工具，遇到性能问题的第一反应是加资源。他的初判是“崩溃是因为服务器不够”。ChatSDE第一回合问他：“去年服务器加了一倍，为什么还崩？”这一问没有否认资源的作用，它只是把一个已经发生过的事实摆出来：如果资源是唯一的瓶颈，加倍之后至少应该明显好转。老师为方舟打造的九问如下：

1. 近三个学期选课开放日的崩溃，各发生在开放后的第几分钟？
2. 崩溃前十分钟最先变慢的是哪一类请求？
3. 去年扩容前后，这一类请求的排队长度和等待时间各有什么变化？
4. 日志里哪一段处理环节的耗时，在上线前的压力测试里从未出现过这么长？
5. 那个最先变慢的环节，在设计文档、运维值班表和外包合同里，各写的归谁负责？
6. 选课开放那天，教务老师和同学各怎么描述“崩了”？与日志里的时刻对得上吗？
7. 开发、运维、教务三方历次的故障复盘记录里，哪一句话每次都出现，却从未变成一张工单？
8. 别的学校公开的选课系统故障说明里，有没有提到类似的排队环节？他们怎么分的责？
9. 如果给这个排队环节指定一个负责人，第一条要改的配置或值班规则是哪一条？

出题逻辑与高远的九问在结构上相似，在对象上不同。第1、2问把“崩溃”这个笼统的词拆成时间序列：先问崩在第几分钟，再问崩之前最先变慢的是哪一类请求。第2问是整套九问的轴心，它把方舟的目光从“系统总负载”移到“某一类请求的先兆”，就像医生不只问病人哪里疼，还问“最先是哪里不舒服”。第3问把ChatSDE的那一句追问变成可以在日志里查的对照：扩容前后，总吞吐变了，那一类请求的排队变没变。第4问指向本章的关键主题，仿真与实测之间的缝：压力测试是对真实负载的一种仿真，它没有测出来的那段耗时，正是仿真漏掉的现实。

第5、7问转向组织。第5问把技术问题翻成责任问题，问那个变慢的环节在三份文件里各归谁；第7问去找复盘记录里“每次都说、从未落实”的那一句。这两问在

软件工程课上往往最令学生意外，因为他们习惯把故障当作代码或硬件的问题，很少想到一个环节可能在技术上完全正常、在组织上却无人照看。好比一栋楼的楼道灯，每家都觉得应该有人换，物业以为是业主的事，业主以为是物业的事，灯坏了半年，谁也没有错，可楼道一直是黑的。第6问让方舟离开日志，去问真正使用系统的人；第8问是一次小规模敌拓；第9问照例把判断落到一条可改的配置或规则上。

题闸复核。第2问合格：通用大模型没有这套系统的日志，拿掉方舟的现场材料，答不出哪一类请求先变慢。打回的样子是“高并发系统崩溃的常见原因有哪些”，答案张口就来。第5问也值得一说：它看似是管理学的问题，但它的答案只存在于这所学校的设计文档、值班表和合同里，拿掉方舟的现场就无从回答，所以过闸。方舟这套九问最终发生的判断是：瓶颈在一个没人负责的排队环节。这条判断同时含着技术的一半和组织的一半，单看日志得不到后一半，单看合同得不到前一半。

◆ 改题、问对与仿真实测之间的缝

两套九问发下去，改题先行。高远改的是第2问。原题问“哪一段桥面的振动在早高峰后迟迟不回落”，他在翻完第1问的数据后发现，迟迟不回落的段落不止一处，但只有一处在晚高峰之后同样迟缓，于是把题提升为“哪一段桥面在早晚两次高峰后都迟迟不回落，而它在平日夜间的基线是否也在逐年抬高？”。这是提升而不是逃离：根还在，仍连着“边界条件被模型写成定值”这条母本判断；离现场更近，题从一个时段扩到两个时段，又从一天扩到逐年，逼他去读更长的数据序列。方舟则没有大改题，他只在第2问后面加了半句：“……那一类请求在进入处理之前，排在谁的队里？”这半句把技术追问直接接上了第5问的责任追问。

高远的问对，第零轮写的是“振动大是因为车流量大”。ChatSDE先接住：车流确实是最主要的激励，你对早高峰振幅的识别也很准确。然后把前提翻过来：“早高峰的车流过去了，为什么那一段还不回落？”工序选择上，高远的初判坏在把桥当成不变的被动对象，按三大方程查缺式：他的作业只要求识别频率，“回落要多久”不在他碰过的量里，手边没有一件与初判相撞而被他视而不见的事实，所以不是断拍；缺的是“桥自身状态”这一项，补式的方向是把边界条件从常数变成变量。接下来几轮，高远先怀疑传感器，查第4问的记录排除了；又怀疑

温度，ChatSDE问他：“同一天、同样的温度，另一段为什么回落得快？”这把他推向第7问的对照段。真正的转折出现在他去后勤翻档案、又在桥上跟了一次早高峰之后。

他在第5问的养护档案里找到一条多年前的伸缩缝维修记录，写得很简略，只记了日期与“已处理”，没有记更换的部件型号；而监测组的仿真模型里，这一处伸缩缝的参数仍是建桥时的设计值。第6问的现场记录里，他在那一段的伸缩缝处听到车辆驶过时有一声比别处更沉的闷响。第8问里，后勤老师傅说：“那一段修过，修完以后过车声音就不一样了，不过也没出过事。”三条线放到一起，高远在第5轮写出了自己的判断：回落迟缓与一处被忽略的伸缩缝维修记录有关。ChatSDE没有替他确认，而是追问他：这个判断如果成立，仿真模型里改了那一处参数之后，预测的回落时间应当朝哪个方向变？如果改了之后不变，你的判断要怎样修改？这条判断因此带着它的检验方式进入子本，地位是待检。

方舟的问对节奏不同。第零轮“崩溃是因为服务器不够”，ChatSDE接住之后问“去年服务器加了一倍，为什么还崩？”。方舟起初的回应是“可能加得还不够”，这是一个很典型的接不住：把批判吸收进原来的前提，再要更多资源。这里的诊断要说清楚，它与高远不同，也与许宁不同。扩容一倍而崩溃照旧，是方舟自己经历过、复盘记录里写着的事实，它与“资源不够就是瓶颈”的初判直接相撞，判断却没有因此改动。矛盾就摆在眼前，循环在逼判断改变的那一拍上断了。按本书的判别次序，这是第一步就成立的断拍，不必再去查少了哪一项，刀是点火，而不是补式。

点火不是争论，而是让矛盾被他自己看见。ChatSDE没有和他辩，也没有告诉他“还缺一个什么因素”，而是把他那句“可能加得还不够”变成一个可以拿日志检验的预测：如果只是不够，那么崩溃前最先卡住的那一类请求，在扩容之后等待时间应当随资源成比例地缩短；请你先找出崩溃前十分钟最先变慢的是哪一类，再把扩容前后它的等待时间并排摆出来。这正是第2问和第3问。数据一摆出来，反驳就不是ChatSDE的，而是日志的：那一类请求的等待时间几乎没有变。到这时，方舟的问题才真正变了，他不再问总量够不够，而是从崩溃前十分钟的一条条请求往上看，找到了第2问的答案：最先变慢的是某一类提交请求，它们在进入数据库之前

排在一个中间队列里。第3问的对照还显示出扩容为什么没有用：扩的是应用服务器，而这个队列由一个早年配置、此后无人调过的组件承担。

判别次序的用处，可以用一个对照看出来。设想另一位同学接手这个课题，没有经历过去年的扩容，也没读过当年的复盘，他写下同一句初判“崩溃是因为服务器不够”。他手边没有与初判相撞的事实，断拍这一步不成立；他的关系里漏掉的是“哪一类请求、卡在哪一环”，那就是缺式，ChatSDE的首问也不再是“去年加了一倍为什么还崩”，而是“崩溃前十分钟最先变慢的是哪一类请求”，刀是补式。同一句话，诊断不同，因为判别看的是学生的现场里接触过什么，不看句子本身。这也回应了一个合理的质疑：任何单因归因都“少了一项”，缺式岂不是处处成立？判别次序的回答是先查断拍，再查缺式；有相撞的事实在手而未改，就先点火，不急于补式。

方舟的判断在第5、7问之后成形：瓶颈在一个没人负责的排队环节。设计文档里，这个组件被列为“公共基础设施”；运维值班表上，它不属于任何一个班次的巡检清单；外包合同的维护范围里，也没有点到它的名字。复盘记录里，“中间队列有积压”这句话出现过不止一次，每次都被写在“其他观察”一栏，从未变成工单。这个判断同样是方舟先说出来的。ChatSDE在他说出“没人负责”之前，只问到“这个组件在值班表上归谁巡检”，把说破的那一步留给了他。九问第5问里本有“归谁负责”的字样，那是老师发下的种子，属于学生取得的材料；ChatSDE自己的追问里，从头到尾没有“没人负责”“排队”“环节”这类核心判断的关键词；“排队”二字最早出现在老师发下的第3问里，后来才由方舟自己从日志里连成判断。

两段问对里有一个共同的东西，正是本章要专门讲的仿真与实测之间的缝。高远的仿真模型是一座理想化的桥，方舟的压力测试是一场理想化的选课；两者都在设计阶段被认真做过，也都在某个具体时刻与实测分道扬镳。工程教育长期把这道缝当作误差，要求学生用更精细的模型去缩小它。九问的态度不同：这道缝首先是信息，它告诉学生，现实里有一件事没有进入模型。缝里藏着的，往往不是物理学或计算机科学的新原理，而是一次没写进参数的维修、一个没写进值班表的组件。

打个日常比方，导航软件规划的通勤时间总比实际少十分钟，与其不断改进算法，不如先问一句：那十分钟是不是每天都耗在小区门口那个没画进地图的路障上。

这道缝还提醒我们，仿真本身是一种母本。模型里写进去的，是设计者、建模者在某一时刻认为重要的东西；模型里没写的，是他们当时没想到、或者想到了却无从获得数据的东西。学生用实测去对仿真，如果只看合不合，就是在做验证；如果去问“这个模型是在哪一年、由谁、基于哪些资料建的，此后哪些事没有回写进去”，就是在做谱系。九问把学生从前一种位置挪到后一种位置。这也解释了为什么两套九问的第9问都落在“改哪一个参数”“改哪一条规则”上：一个判断的最终去处，是回写进模型或规程，让下一次仿真不再漏掉这件事（见第24章）。

这道缝可以和舍恩（1983）《反思性实践者》里的一个说法对读。舍恩描写的建筑师、工程师在做设计时，把一个设想投向情境，情境会“回话”：线条画下去，场地的坡度就以意想不到的方式回应；好的实践者听得懂这种回话，并据此重新框定问题。高远在桥上听到的那声闷响，方舟在日志里看到的排队积压，正是装置与系统的回话。九问与舍恩相同的地方，是都把回话当作研究的起点，而不当作干扰；多出来的一格有两处。其一，回话在舍恩笔下常随实践者的当场判断而消散，九问要求把它落成可回放的材料：附数据片段、处理脚本、档案摘录，让别人能重走一遍。其二，实践者的反思往往停在自己心里或同行的闲谈中，九问要把判断交给责任方验证，并留下采纳与否的记录。回话若不被记录、不被验证，就仍只是一位高手的手感，进不了课程的谱系。有人会反驳：这样要学生写太多文书。回答是，只有当一个判断需要离开写它的人、去被另一些人检验时，这些文书才是必要的；一个只给自己用的发现，不必套上九问的全套手续。

◆ 子本形态与评估权重

工程与计算的子本，形态上与化学子本相近，都要求可检验，但可检验的方式不同。化学子本的判断交给下一个批次去复做；桥梁与系统无法复做，它们的子本要做到的是可回放：任何一个评审者，拿着子本附带的数据片段、处理脚本、档案摘录和现场记录，都能重新走一遍学生的推理，并在关键处得到同样的中间结果。高远的子本大纲大致是：母本中关于桥梁健康监测与边界条件退化的已有判断；这座桥的监测档案与养护档案的结构及其断裂处；九问的发现，以伸缩缝维修为中心

重写“这一段桥为什么回落迟缓”；对仿真模型的参数修订建议与检验方案；留给下一届的待检清单。母本所无的判断，就在“维修记录未回写进监测模型”这一环。

方舟的子本则多了一层组织分析。他的子本要同时呈现日志证据链（哪一类请求最先变慢、在哪个队列积压、扩容前后队列能力如何）与责任证据链（三份文件各怎么写、复盘记录里那句话出现过几次、为什么从未成为工单）。子本的结尾不是一个扩容方案，而是一条可执行的建议：为该组件指定负责人，把它纳入巡检清单，并给出一条具体的配置修改建议，交信息中心评估。可回放性在这里落实为：处理日志的脚本要随子本提交，任何人在同样授权的数据上运行，都能得到同样的请求时序图。这既是完成度，也是防止“看图说话”的保险。

权重方面，本书第五编没有为工程与计算单列权重示例，课堂可以从已有的两列出发。像高远这样以观测、建模、实测比对为主体的研讨课，可参照化学列（本科，%）：母本底盘5、改题质量10、问对深度10、现场证据15、子本创新智商30、子本完成度15、碰撞贡献10、回写证据5。理由与化学相同：判断最终要回到可回放、可检验的数据与模型，完成度承载着这一要求。像方舟这样嵌在真实项目里、要与信息中心协作落地的课，则可参照应用实践课列：母本底盘5、改题质量10、问对深度10、现场证据25、子本创新智商25、子本完成度10、碰撞贡献10、回写证据5。应用实践课把现场证据拉到25，因为在真实系统里，能不能拿到并读懂一手日志与文件，本身就是这门课最难的一段。无论哪一列，创新智商都不得低于25的硬底，本科分档线为110。

评分时要注意工程类子本的“新”长什么样（第16章已以高远与叶青作对照）。高远的判断里没有新的力学原理，方舟的判断里没有新的算法，评分者若拿顶刊的创新标准去量，两份子本都显得平淡。但它们都满足本书对子本的要求：含有母本所无的判断，而且这个判断只有经过学生的现场才能发生。五维评分中，I项要看的正是题闸意义上的不可还原：拿掉学生的现场，这个判断能否由大模型直接生成。另一个要提防的是工程学生特有的膨胀：一份子本附了漂亮的模态图、时序图、架构图，形式完整，容易让评分者高估。反膨胀规则在这里的用法很简单，图越多，越要追问哪一张图里有学生先说出的判断。

母本底盘在工程与计算课里虽然只占5%，却是不可跳过的门槛。复述加反例口试在这里有一个很具体的样子：请高远讲清楚结构模态识别的基本做法，再请他举一个“频率对上了、结构却已经有问题”的反例；请方舟讲清楚排队论里到达率与服务率的关系，再请他举一个“总资源充足、局部仍然堵死”的反例。答不出反例的学生，往往会把自己九问里的发现误认为是学科的新原理，或者反过来，把真正的新判断当作教科书常识而轻轻放过。底盘不是九问的对立面，它是学生判断“我发现的东西是不是早就有人说过”的那把尺子；没有这把尺子，敢拓与题闸都会落空。

回写证据在这两门课上也有可观察的痕迹。高远的第零轮写的是“振动大是因为车流量大”，期末自述里他写道，自己以后再看监测数据，会先问“这一段桥最近被谁动过”。方舟的期末自述则说，他第一次意识到，一个系统的架构图上画着的方框都有主人，方框之间的箭头却常常没有。这两句话不能当作回写的宣布，只能当作线索；真正的回写要看下一次课题里他们出题、问对的方式有没有变，看他们的子本建议有没有被监测组或信息中心采纳进模型与值班表。若被采纳，这一届的判断就成为下一届九问的母本精要，课程开始像一条谱系那样生长。

◆ 数据权限、安全与特有风险

工程与计算九问的门槛，不在伦理委员会，而在数据权限与安全。高远用的监测数据、养护档案，方舟用的故障日志、设计文档、值班表与合同，都不属于学生，也不全属于课题组。它们分别归属监测组、后勤部门、信息中心乃至外包方。九问课堂的规矩是：每一份进入九问的数据，都要在开课前由老师与数据所属部门办好使用授权，写明可用范围、可用期限、能否进入子本、能否公开。学生不得自行抓取、扫描或探测任何未授权的系统；方舟的九问里没有一问要求他去测试线上系统，第6问问的是用户的描述，第3、4问用的是已归档的日志，这是有意为之。日常里也是一样，修车师傅可以看你的行车记录，但不能自己把车开去高速上试刹车。

数据进入子本时还要过一道脱敏与保密的关。选课日志里有学号、姓名、选课记录，任何一条进入子本的日志都要先去标识化；系统的内部地址、组件版本与配置细节，涉及安全漏洞的部分只能写进交给信息中心的内部报告，不能写进准备公

开或进入部落档案的子本正文。桥梁监测数据一般不涉及个人信息，但涉及校园设施的安全状态，若学生的判断提示某处结构存在隐患，应第一时间经老师转交后勤与专业检测单位，由有资质的人员处置，学生不得自行在桥上做任何加载或拆检。学生的判断在这里与医学课一样，是交给责任方验证的假说，不是处置决定（见第19章）。

这两门课最容易出错的地方，各有各的样子。先说剔除离群值。工程训练教会学生把离群点平滑掉，这在建模阶段常常必要，但九问恰恰要从离群处出发。设想高远在预处理时，把“早高峰后迟迟不回落”的几个测点当作漂移，用滑动平均抹平了，第2问便无从谈起，而且抹平之后谁也看不出抹过。对此的规矩是子本必须保留原始数据与预处理前后的对照图；凡被剔除的数据点超过该测点记录总量的5%，须在子本里逐类说明剔除理由。

再说两种相反的偏执。仿真崇拜认为模型对、数据错，实测崇拜认为数据对、模型无用，两头都是在躲开那道缝。九问把缝本身当研究对象：第3问追问实测与仿真在哪里对不上，是在给仿真崇拜降温；第4问查传感器的安装与校准记录，是在给实测崇拜降温。

方舟那类判断还有一个特别的陷阱，是技术与组织各只做了一半。若只有一句“没人负责”的感叹，而没有队列数据，那是牢骚；若只有队列数据，而没有设计文档、值班表与合同的对照，那是又一次技术归因，只是换了更细的部件。检验的读数很具体：子本里的每一条责任论断，能否指到某份文件的某一句；每一条技术论断，能否指到某段日志的某个时间窗。指不到的，一律降为待查，不进结论。

◆ 本章怎样算错

先看题闸。连续两届、每届不少于15名学生的工程与计算九问课堂里，由换手评审对每份子本中“由仿真与实测之间的缝提出的判断”做核查：拿掉学生的现场材料，交给一个未接入这些材料的通用大模型（记录所用版本，默认配置与学习模式各试一次），若超过三分之一的判断能被它给出说得过去的复述，就说明工程与计算的现场比本章设想的更容易被通用知识覆盖，“九问让装置自己说话”的主张要收回，至少收窄为只在少数现场成立。

再看诊断能否识别，这是方舟的断拍判定要过的关。取两名受过训练的评审，独立对每届不少于20条初判，按判别次序标注断拍、缺式、偏路；若两人的标注一致率低于70%，或者对方舟这类初判的标注总是随评审而变，那么“工序选择”这一构念在工程课堂里就不可识别，本章把方舟判为断拍，也就只是作者的口径，不是方法。

再看数据权限。如果一个学期里，开课前能办妥授权、拿到一手数据的班级，不足开课班级的一半，学生大多只能用公开数据集做九问，那么“战场在真实装置与系统的异常里”就只在少数有条件的课堂成立，需要改写为有前提的主张，并单独讨论公开数据集上九问的降级形态。

最后看回写。像高远、方舟这样的子本交给监测组、信息中心之后，若连续两个学期都得不到书面回应（采纳、否决或要求补充，都算回应），判断永远停在“待检”，回写链条就是断的。此时问题不在学生，而在课程与职能部门之间缺少承接的渠道，“判断回写进模型与规则”要降格为愿景，并到第24章所说的回写痕迹中寻找能被观察的替代证据。高远和方舟都是构造案例，他们的判断只作说明；真正能证伪或支撑本章的，是多届学生的版本史、问对记录和责任方的验证结果。

高远的伸缩缝、方舟的排队环节，都是人造装置里的“无人之处”：维修了却没人回写，存在着却没人负责。它们已经透出一点社会科学的味道。到了经济与法学课堂，装置换成了交易与裁判，缝隙换成了节律：父亲为什么一到月底就泄气，物业纠纷的调解为什么总在某一周了结。互动被看不见的周期推着走，那么九问要怎样让学生看见这些周期？下一章转向周川、林晓与韩正。

带 走 的 话

模型与实测之间的缝首先是信息，不是误差：一次没写进参数的维修、一个没写进值班表的组件，都是装置自己说出的话；判断要落成可回放的材料，交给责任方去验证。

◆

22

第22章

经济与法学：交易与裁判里的九问

◆ 案例讨论课上的那一阵沉默

《司法实践研讨》的旧版课堂是这样上的。老师在课前发下一份打印好的物业纠纷判决书，事实部分已经删繁就简：某小区业主以服务不到位为由拒交物业费，物业公司起诉，法院部分支持原告诉请、酌减了滞纳金。讨论题写在最后一页：“本案裁判是否妥当？请结合民法典关于物业服务合同的规定作答。”同学们轮流发言，有人引合同条款，有人谈服务瑕疵的举证责任，有人比较另一地法院的类似判决，讨论热烈而有条理。轮到韩正时，他问了一个让教室安静下来的问题：“这个案子判决之前，调解谈了多久？谈崩在哪一周？”老师翻了翻材料，说案例里没有写。那一阵沉默不是因为问题太难，而是因为问题落在了案例的边框之外——案例从一开始就只给了“值得讨论的那部分事实”，而时间、节奏、谁在什么时候松口，早在编写案例时就被当作噪声剪掉了。

这一幕在经济管理的课堂上也有一个对应版本。《小微经营者的动力与节律》开设之前，同类课程的标准做法是讨论一份连锁餐饮企业的激励制度案例：门店店长的绩效怎样考核，员工提成怎样设计，学生分组给出改进方案，再由老师点评哪一组更接近“成熟企业的通行做法”。周川在这样的讨论里总是走神，他想到的是自家小饭馆——没有店长，没有绩效表，只有父亲、两个伙计和一本记了三年的流水账。案例里的激励问题有一个已知的终点，即教科书里的激励理论；而他家饭馆的问题，连老师也说不出终点在哪里。案例讨论课的热闹与那一阵沉默放在一起，正好照出本章要处理的张力：社会科学课堂的发现版并不只有讲授，它最成熟的形态是案例教学法，而九问模式要回答的，恰恰是案例教学法做得最好的地方之外还缺什么。

先要承认案例教学法的功绩。法学院的案例教学一般追溯到19世纪70年代兰德尔（1871）在哈佛法学院的改革：不再由教授从体系讲到条文，而是让学生直接读上诉判决，在教授的苏格拉底式追问下自己归纳规则。商学院后来发展出另一种案例法，以一份企业决策情境为中心组织讨论。两种案例法都把真实的裁判与真实的经营带进了教室，让学生练习从事实中抽出要点、在对立论证中站稳立场，这比单纯背诵条文或模型前进了一大步。打个大白话的比方，讲授课堂是看菜谱，案例课堂是看厨师做菜的录像——录像里有真实的火候和手法，学生看得到判断是在什么情境下做出的。就训练“在给定事实上推理”而言，案例教学法至今仍是社会科学教育里最有效的方法之一，九问模式不是来取代它的，而是把它放回底盘的位置。

案例教学法的边界也正在这里。录像里的菜是别人做的，镜头对准哪里、剪掉哪一段，都由编写者决定；学生在案例课上推理得再好，推理的对象仍然是一个被预先裁剪过的现场。按本书的路径口径，案例教学是S→D→E发现版的一个精致变体：先给出规则或理论（S），再让学生把它用到案例上（D），案例本身是一个预制的现场（E），而且这个现场通常配有一个“标准答案”——法院的终审判决，或企业后来的实际选择。终点能否事先写出？在案例课上，答案多半是能（见导论的判别两问）。更要紧的是，社会科学研究的对象是互动，而互动最重要的维度之一就是时间：谁先开口、谁在哪一周松口、钱在哪一天进出。被剪成案例的事实，恰恰最容易丢掉这个维度，因为编写者关心的是“争点”，而争点在案例里是静止的。

法学内部对这条边界并不陌生。庞德（1910）发表《书本中的法与行动中的法》，批评的正是只读上诉判决的法学教育：判决书里的法，是法官写成条文形态之后的法；案子在基层怎样被拖、被谈、被撤，是另一个世界。此后一个多世纪，法学界一直用诊所式教学和实务实习去补这块短板，可见短板是公认的。但这些补法的共同点，是把学生送到现场去“见习”，见习的产出是技能与见闻，而不是一个母本里没有的判断。加兰特（1974）的论文又给这条边界添了一层：重复博弈者倾向于把对自己不利的案子在判决之前谈掉，把有利于自己立规则的案子推到判决，于是能进入上诉判决书、进而进入教材案例的那一批案子，本身就是被诉讼生态筛过的。读上诉判决的学生，读到的是重复博弈者愿意让人读到的那部分法。这就像只在橱窗里看一家店的货——橱窗摆得再好，也看不到后仓里没上架的那些。

案例的“预裁剪”因此不只来自编者，也来自诉讼本身；九问把学生送去的，恰是被筛掉的那一半：没有走到判决的案子，以及走到判决之前的那些周。

最强的辩护来自案例教学法自己的改良者：把立案、调解、结案的完整时间线附进案例，把企业的现金流日历附进餐饮案例，问题不就解决了吗？这个反驳应当认真对待，因为它恰好是一个可以被检验的替代方案（本章末的失败条件第二条专门检验它）。本书的回应有两层。第一，附了时间线的案例仍由编者挑选：时间线上画哪些节点、不画哪些，是编者对“什么重要”的预判，学生拿到的是一份已替他回答了“看哪里”的材料。周川的“预扣”之所以能被看见，是因为它在账本上写在当天收入的旁边，一个编案例的人不会知道这个记法值得抄录。第二，案例课的学生不为材料的缺口负责：缺的那一页，在案例里是编者的疏漏，在自己的现场里却是自己必须去找的下一步。这像是考驾照与上路的差别——科目二的场地线画得再准，也教不会人在雨天的路口判断前车会不会突然刹车。

还有一个2025年以来的新事实不能绕开：ChatGPT的Study Mode、Gemini的Guided Learning和Claude的学习模式（日期与出处见第8章）都已能对一份给定的案例材料做苏格拉底式追问，案例课上“老师逼问一句、学生答一句”的那一层，AI现在可以陪着任何学生做。这对案例教学法是帮助而不是威胁，它让案例课的发现一侧更便宜；同时也把分界线划得更清楚：这些追问的对象仍是别人预先准备好的现场，谁来决定现场里放什么、缺什么，仍在出题台上。

于是本章的核心判断可以写下来：社会科学课堂的战场在交易与裁判的节律里；九问让学生看到互动怎样被看不见的周期推着走。这里的“周期”不是宏观经济学意义上的经济周期，而是嵌在具体交易与具体程序里的小节律——月底结账、季度交租、物业公司年底对账、法院的审限与排期。这些节律在案例里看不见，在统计年鉴里也看不见，只有当学生下到自己的现场、翻开一本三年的流水账或坐进一个基层法庭的旁听席，它们才显露出来。AI可以在几秒钟内把激励理论、物业服务合同的裁判规则讲得清清楚楚，这一侧已经移交出去（见第2章）；它讲不出的是周川父亲每个月20号以后的那张脸，也讲不出本院物业纠纷在立案后某一周集中结束时那一周发生了什么。

这一章的问题

案例课已经把真实的裁判与经营带进了教室，社会科学的九问还能让学生在自己的账本和旁听席上，找回被案例剪掉的什么？

◆ 谱系母本与精要：互动属性与看不见的周期

按全书的三领域对应，自然科学侧重知识发生的客体属性，人文侧重主体属性，社会科学侧重互动属性（医学横跨三者，已在第19章讨论）。所谓互动属性，是说社会科学要解释的东西大多不在单个人身上，也不在物上，而在人与人、人与制度的往来之中：一笔交易是买卖双方的往来，一次调解是当事人、调解员与法官的往来，一个小饭馆的经营是老板、伙计、供货商、房东与食客的往来。互动有一个容易被忽视的特征，它总是在时间里展开，而且常常被某一方的时间表所支配。就像一桌麻将，牌面看起来是四个人在出牌，真正定节奏的却可能是那个要赶末班车的人——他什么时候着急，整桌的打法就跟着变。社会科学课堂的九问，要训练的正是看出“谁的时间表在定节奏”的那双眼睛。

《小微经营者的动力与节律》研讨课的谱系母本由两部分组成。前沿专著一侧，老师选入了行为经济学中关于心理账户的研究（塞勒在20世纪80年代提出的这一概念，讨论人们怎样把钱分进不同的“账户”并区别对待），关于稀缺心态的研究（穆来纳森与沙菲尔的《稀缺》，讨论资源紧张怎样占据人的心智带宽），以及小微企业现金流管理与非正规经济的若干研究。部落档案一侧，是本课往届学生的问对档案与他们家里、亲戚、邻里小店的经营材料摘录——这门课开得不算久，档案还薄，比例上仍以前沿专著为主，但每一届都在加厚。母本里的这些研究为学生提供了解释“钱怎样影响人”的已有框架，它们本身就是统一性知识与谱系前沿的交界地带。

把周川的判断放回这两条母本研究里，才看得出它多出的一格是什么。塞勒的心理账户说的是钱被分进不同账户之后会被区别对待，却默认账户的边界是画在人心里的；穆来纳森与沙菲尔说的是稀缺占据带宽，却多从实际短缺来理解稀缺。周川流水账里的“预扣”同时动了这两个默认：账户的边界写在纸上，而且写在短缺发生之前。这就给出一对可以分开的预测。按稀缺研究的读法，泄气的强度应当随

抽屉里实际余额的下降而加深；按“账本提前收走”的读法，泄气应当跟着预扣条目出现的日子走，哪怕那几天抽屉里的现金还很厚。两种读法在大多数月份里重合——月底一般既缺钱又有预扣——只有在少数偏离的月份才分得开，所以第7问去找“月底没有泄气的月份”才有判决意义。这也是九问和“用心理账户分析我家饭馆”的区别：后者拿理论去套现场，前者在理论默认值松动的地方下刀。

从母本提炼的精要库，按第3章的三类判断组织。方程判断一类，例如“同一笔钱进入不同心理账户，会被区别对待”“资源稀缺会改变人的注意分配”；路径判断一类，例如“小微经营者的决策常常先看现金、后看利润”；动力判断一类，例如“经营者的投入意愿随预期而不随当期收入波动”。这里要注意的是精要库的措辞层级：它记录的是母本里已有的、可被复述和反驳的核心判断，而不是教条。学生在母本底盘口试中要能复述这些判断，并给出一个反例——例如说出一种“稀缺并没有占据带宽”的经营情境。精要库在这门课里的作用，是给九问提供“根”：每一问都必须连着其中某一条判断，否则就不是谱系性课题，而只是一个家庭生活的好奇。

《司法实践研讨》的谱系母本，前沿专著一侧有两条主要来源。一条是法社会学关于“书本上的法”与“行动中的法”之分的传统，这一区分追溯到前文提到的庞德（1910）的论文；另一条是关于诉讼参与者结构差异的研究，最常被引的是前文提到的加兰特（1974）的论文《为什么“有者”占先》，他区分了反复参与诉讼的“重复博弈者”与偶尔涉诉的“一次性参与者”，指出前者在规则形成与程序运用上的系统优势。国内关于法院调解、诉讼与非诉讼纠纷解决机制的研究也在母本之列。部落档案一侧，是本研讨课多年积累的本地判例库与历届学生的旁听记录。精要库里的路径判断有“法院调解在立案后的推进受审限与排期约束”，动力判断有“重复博弈者更能承受程序时间成本”。韩正后来的九问，接在后一条判断上，但打向了它没有说到的地方：重复博弈者的时间成本，究竟是被什么样的具体节律定出来的？

两门课的精要库有一个共同的空白，这个空白恰好是互动属性最难被书本捕捉的部分。母本里的研究大多是在“结构”层面说话：稀缺占据带宽，重复博弈者占先，心理账户区别对待。它们说清了“谁更有利”“什么会被影响”，却很少说

到“在哪一天、哪一周、被谁的哪一张账单推了一把”。这不是前沿研究的缺陷，而是研究尺度的自然结果——要在可推广的样本上得出结论，就必须把具体的日历抹平。九问模式在这里的位置就清楚了：母本给出结构判断，学生在自己的现场把被抹平的日历找回来。日历找回来之后，结构判断会被细化、修正，有时被局部推翻，这就是谱系往前长出的那一截。

◆ 周川与林晓：饭馆账本与理发店房租的九问

周川的九问由老师借助装有SDE工序层的大模型，从精要库出发、按周川的画像打造。画像的要点有三：家里经营一家小饭馆，周川从小在店里帮忙；他手里有父亲三年的流水账，记账方式是手写，按日记收入、按笔记支出；他自己对“激励”问题的直觉偏向“人干得起不起劲取决于有没有奖励”。老师版九问的第1问原本是“你家饭馆哪个伙计干得最起劲？为什么？”，它连着精要库里的动力判断，也离周川的现场很近。周川在改题时把它改成了“父亲为什么一到月底就泄气？”，改题的过程与两锚检验在第5章已有详述；他还提过一个“AI会不会取代饭馆”的改法，被题闸打回，因为那是一个不需要他家账本、未接入这些材料的通用大模型张口就能答的题。下面列出的是问对之后的定版九问。

周川的九问（定版）如下。第1问，父亲为什么一到月底就泄气？第2问，三年流水账里，每月营业额最高的是哪一旬？父亲泄气的那几天，营业额真的掉了吗？第3问，月底前一周，账本上集中出现的是哪几类支出——房租、伙计工资、供货结账——各落在哪一天？第4问，父亲记账的笔迹与记法，月初和月底有什么不同？哪一类账目到了月底就不再细记？第5问，伙计的工资哪天发？发薪前后父亲对伙计说话的方式有没有变？第6问，供货商的结账周期是谁定的？三年里改过没有？第7问，三年里有没有哪个月，月底父亲没有泄气？那个月的现金进出有什么不同？第8问，同一条街的其他小饭馆，结账日与发薪日是否也挤在同一周？第9问，如果把供货结账从月底挪到月中，父亲第一个会改变的是什么？

这九问的出题逻辑可以分成三段来看。前三问是“定位”：先把泄气这件事在日历上钉住，再看它与营业额是否同步，再看月底集中流出的是什么。第2问是一个关键的排除题，它预先拦住了最顺手的解释——如果泄气时营业额并没有掉，那么“生意不好所以泄气”就站不住。中间三问是“找机制”：第4问看账本本身的形

态变化，第5问看泄气是否传导到人际互动，第6问追问节律的来源——账期是谁定的，决定了这是父亲个人的问题还是交易结构的问题。后三问是“检验与外推”：第7问找反例月份，第8问看邻店是否共享同一结构，第9问是一个干预式的设想题，要求周川说出改变节律后第一个可观察的变化，这一问在形式上与许宁的第9问同构，它把判断推到可被证伪的位置上。

周川在问对中写下的第零轮初判是“父亲月底泄气是因为累”，ChatSDE第一回合的回应是“月初也忙，为什么月初不累？”（这一来回的工序选择，按第7章的判别细则属于断拍：周川天天在店里，月初同样忙碌却不泄气，这个与初判直接相撞的事实一直就在他眼前，初判却没有因它而改，所以用的是点火，让矛盾被他自己看见）。顺着九问往下走，周川在流水账上发现了一个他从未注意过的现象：父亲泄气的那几天，营业额并没有下降，有时甚至略高；但从每月20号前后起，父亲在账本上开始把还没到期的房租、工资、供货款提前写成“预扣”，写在当天收入的旁边。账面上，钱还在抽屉里，却已经被一笔一笔划走了。周川由此形成的判断是：父亲的动力被账本提前收走——不是钱没了才泄气，而是账本先替他把钱分光了。这个判断与母本里的心理账户和稀缺研究都有关联，但有一个关键的差别，它说的是“提前”：泄气发生在现金实际短缺之前，触发它的是账本上的预扣动作，而不是抽屉里的现金余额。

林晓与周川同在这门课上，她的现场是母亲开的理发店。她的画像与周川相近而又不同：店小，没有雇工，母亲一个人剪发，房租按季交，客人一半扫码、一半付现，近两年开始卖储值卡。老师为她打造的九问，第1问便是本书贯穿案例中的那一问：“母亲为什么交房租前不爱说话？”其余八问依次是：第2问，房租是按月、按季还是按年交？交租日前后一周，店里的营业额有什么变化？第3问，现金与扫码各占多少？交租前母亲会不会去取现、数钱？第4问，交租前那几天，母亲有没有接下平时不接的活，或推掉平时会接的活？第5问，储值卡预收的钱，母亲是把它当作“已经挣到的”还是“欠着客人的”？第6问，母亲不说话的那几天，和老顾客的闲聊少了多少？第7问，房东涨过几次租？每次涨租后的第一个月，店里有没有调价或减少进货？第8问，隔壁的店主们交租前是不是也这样？他们怎么说这段日子？第9问，如果房租改成按周交，母亲交租前的沉默会消失、变短，还是分散到每一周？

林晓九问的出题逻辑与周川的同构而错位，这是老师有意为之的。同构之处在于，两人的九问都从一个情绪性的、日常的观察出发（泄气、不爱说话），都设了一道排除营业额解释的题，都以一个干预设想题收尾。错位之处在于，周川的节律是月度的、多笔支出叠加的，林晓的节律是季度的、单笔大额的；周川家有伙计，泄气会传导到人际上，林晓的母亲一个人做，沉默首先表现在她与顾客的闲聊上；林晓的第5问还引进了一个周川家没有的变量——储值卡的预收款，它让现金在时间上的归属变得更复杂。老师在出题时就预见到，这两套九问放在同一个碰撞场里，最可能撞出一个两人各自都说不出的判断。后来的碰撞果然如此：两人发现，父亲的泄气与母亲的沉默，都不跟着工作量走，而跟着现金流出的节律走，由此形成的班级判断是“小微经营者的动力被现金节律收走”。

这一幕可以拿第6章用过的Chi与Wylie（2014）的ICAP框架来对照。框架把学习投入分成被动、主动、建构、互动四档，周川与林晓的互撞落在最高一档，这与框架的预期同向。框架没有说的，是互动档产出的东西该怎样处置：同学之间一句“我家也是这样”很容易被当作印证，而九问的碰撞要求另走两步，其一，碰撞产物要重新过题闸，看拿掉两本账之后通用大模型能不能说出同样的话；其二，班级判断要带着版本史进入档案，被下一届证伪时可以撤回。这两步是本书在互动档上多出的要求。同时要守住它的边界：班级判断的证据只有两本账、两个家庭，加上第8问里邻店店主的口头说法，它是下一届九问的出发点和一个值得在更大样本上检验的机制假说，不是关于小微经营者的一般规律，更不是政策建议。同一个机制换一座城市、一种业态、一套账期安排，完全可能不成立，这恰是谱系知识的本来面目。

◆ 韩正：本地判例库与旁听席上的九问

韩正的现场有两层：一层是本地基层法院在裁判文书公开平台上公布的物业纠纷文书，另一层是他本人在法院公开开庭时的旁听记录。老师为他打造九问时，画像里写着三点：他对程序法兴趣浓厚，课上那次“调解谈了多久”的追问显示他对时间维度敏感；他所在的城市新建小区多，物业纠纷在基层法院的民事案件里占比不小；研讨课的判例库已积累了几届学生整理的本院物业纠纷文书摘要。他的第1问就是本书贯穿案例中的那一问：“本院近三年物业纠纷里，调解成功的案子大多在

立案后第几周结束？那一周发生了什么？”这一问连着精要库中“法院调解在立案后的推进受审限与排期约束”这条路径判断，又把它压到一个具体法院、一个具体案由、一个可数的时间分布上。

韩正的九问（定版）如下。第1问，即上引的那一问。第2问，以调解书或撤诉结案的案子与以判决结案的案子，立案到结案的时长分布有什么不同？第3问，调解成功集中结束的那几周，落在日历上的哪个时段——月末、季末、年末，还是法院的某个排期节点前？第4问，物业公司作为原告时，起诉在一年里集中在哪几个月？这与它们的收费周期和账务安排有什么关系？第5问，旁听的调解或庭审中，物业公司的代理人在什么时候主动降低滞纳金或让步？说了什么理由？第6问，本院同期其他由企业作原告的追索类案件，是否也有类似的结束时点集中？第7问，以判决结案的物业纠纷中，有没有哪些案子是在“那一周”前后谈崩的？谈崩时双方各自在等什么？第8问，业主一方在调解中提到的“缴费通知”“催缴函”，发出日期与物业公司的账单周期是什么关系？第9问，如果本院把物业纠纷的调解安排挪开物业公司账务上的某个固定节点，第一个会被改变的是什么？

出题逻辑同样可以分段说明。第1至第3问是“在日历上钉住”：先看调解成功的时点是否集中，再与判决结案对照，再把集中的时点放回日历。这三问里，第2问承担排除功能——如果调解与判决的结案时长并无差别，“调解有自己的节律”这一前提就不成立。第4、5问是“找推手”：第4问从原告一侧的商业周期去看起诉时点，第5问则把文书上看不见的东西交给旁听席，因为让步发生在何时、以什么理由让步，判决书往往不写，调解书写得更少。第6至第8问是“排除与细化”：第6问检验这是不是物业纠纷特有的现象，第7问从谈崩的案子反向看，第8问把业主一侧的材料引进来。第9问仍是一个干预设想题，要求韩正说出调整之后第一个可观察的变化。

韩正的改题史比周川的平静，但有一处值得记下。老师版第5问原本是“旁听一次物业纠纷的庭审，记录双方的主要争点”，这是案例教学法的标准动作，记录争点。韩正在改题时把“争点”换成了“让步的时点与理由”，理由是：争点在判决书里都有，旁听只为抄一遍争点等于白去；判决书里没有的，是谁在什么时候松口。这一改动过了两锚检验——根还在，仍连着精要库里关于调解推进的判断；离

现场更近，因为它要求的是只有坐在旁听席上才拿得到的东西。这一改法后来被老师在下一年的出题中采纳，这件事放在第24章讲。

问对的一个片段可以说明批判性激发在法学课堂上的样子。韩正的第零轮初判是“调解成功多在立案后不久，因为当事人刚进法院，愿意让步”。ChatSDE在第一回合没有顺着这个前提去整理“诉讼初期和解率较高”的一般文献，而是问：“如果是因为刚进法院愿意让步，那么不同月份立案的案子，调解结束的时间应当都落在立案后差不多的位置。你在判例库里看到的是这样吗？”这一问对应的工序，按第7章的判别细则是缺式，用的刀是补式：韩正此时还没有碰到任何与初判相撞的材料，谈不上断拍；他的初判把时间理解为当事人心理上的“新鲜度”，从当事人一侧解释节律，关系里少了一项被当作常数的因素，即节律可能来自程序之外的另一方。追问之所以没有直接点出这一项，是因为首发权在学生手里：ChatSDE只把这条预测摆出来，让缺口自己显形。这一问也容易被误读为偏路，因为它同时改变了韩正的起手次序，但次序的改变是补上缺项的结果，而不是诊断本身。韩正回到判例库，按立案月份分组排列，发现结束时点并不跟着立案日走，而是向某几个固定时段聚拢。第3轮他把这一发现带回问对，ChatSDE接住并放大了“固定时段”这个发现，问他这几个时段是被什么定下来的；韩正自己说出了“物业公司年底对账”这一可能。核心判断由学生先说，这一点在问对记录里清清楚楚。后文旁听时代理人当庭说出的那句理由，是韩正在现场取得的材料，不算ChatSDE替他说破。

接下来的几轮里，韩正带着第5问去旁听。他在两次调解中都记下了相似的一幕：物业公司代理人在临近其年度对账节点时主动提出减免部分滞纳金，理由是“公司要在年底前把这批账清掉”。他由此形成的判断是：调解进程被物业的账单周期推着走。这个判断与加兰特关于重复博弈者的论述有关联，又往前走了一步：重复博弈者的优势不只是更能承受时间成本，它还在相当程度上替整个程序定了节奏——调解什么时候成，不完全取决于当事人或法官，而取决于一方当事人的财务日历。

这个判断还有一个必须正面对待的竞争解释：集中的结案时点也许根本不是物业公司的日历，而是法院自己的——审限、年度结案考核、排期的松紧，都可能让

各类案子在某几周扎堆结束。这个解释与精要库里“调解推进受审限与排期约束”那条路径判断完全相容，甚至就是它最直接的读法。两种说法在“结束时点集中”这一现象上分不开，要分开它们，得在数据里找一个两者预测不同的地方。第6问给了第一个：如果其他企业作原告的追索类案件也在同样的几周集中，那是法院的日历在起作用；只有物业案如此，才轮到物业公司。第二个更锋利：如果不同物业公司的对账日并不相同，就把物业案按原告公司分组，看各组的集中点是随公司的对账日移动，还是都停在同一个法院节点上，前者支持韩正，后者推翻他。当然，对账日从哪里来是这一步最难的地方，旁听和公开信息只够凑出少数几家，样本小到可能只能给出倾向而不能给出结论，子本里必须照实写明。这种让两个解释在同一份数据上给出不同预测的设计，是九问比一般案例分析多要求的一步，学生先要把对手的解释写成可以被数据打败的样子，自己的才算数。需要强调，这个判断是韩正在本院近三年文书与有限旁听记录上形成的本院假说，样本有限，旁听次数更少，它要在后续的文书扩充和下一届的复做中检验，而不是被当作关于法院调解的普遍结论。

◆ 子本形态与评估权重

经济与法学子本的形态，与化学子本、医学子本都不同。化学子本的完成度核心是可复做性，医学子本的判断作为本病房假说交科室验证；经济与法学子本的核心构件，可以概括为“节律表加机制论证”。所谓节律表，是学生在自己现场里找回来的那张日历：周川的是三年流水账上的预扣时点与泄气日，林晓的是交租日、取现日与闲聊减少的日子，韩正的是立案、调解结束、让步与物业对账日的对照。所谓机制论证，是对“为什么这个节律在推着互动走”的解释，它必须与母本里的结构判断对话——说明自己的判断在心理账户、稀缺心态或重复博弈者理论的哪一点上细化了、修正了或局部推翻了已有判断。子本不是对这些理论的翻译或应用报告，而是一份带着母本所无判断的重写。

周川子本的大纲大致是这样的：第一部分复述母本中关于心理账户与稀缺的核心判断，并给出他在自家饭馆找到的一个反例；第二部分呈现三年流水账的节律表，重点是“预扣”记法的出现时点；第三部分论证“动力被账本提前收走”的机制，并排除“生意不好”“劳累”两种竞争解释；第四部分讨论邻店材料与反例月

份；第五部分写出与林晓碰撞形成的班级判断，并说明个人判断与班级判断之间的关系；最后一部分写出判断的失败条件，即什么样的账本证据出现，他的判断就要修改。韩正子本的结构与此相仿，只是节律表来自文书与旁听，机制论证与诉讼参与者结构的研究对话，并附有他对样本局限的说明。

最后一部分的失败条件要写到可以被账本打败的程度，而不是一句“如有反例则修改”。以周川为例，三年共36个月，他可以把条款写成：先逐月标出预扣首次出现的日期与父亲首次泄气的日期；若泄气先于预扣出现的月份占到总月数的三分之一以上，“账本提前收走”就不能成立，只能降格为“账本记法与泄气同时发生”；若预扣出现却始终没有泄气的月份多于泄气而无预扣的月份，则说明预扣既不必要也不充分，判断要重写。这样的条款有分母、有门槛、有方向，别人拿同样的账本可以照着去验，这正是它与一般“结论与局限”段落的区别。

在评估上，经济与法学本科研讨课按第18章的学科权重示例执行：母本底盘5，改题质量10，问对深度15，现场证据15，子本创新智商30，子本完成度10，碰撞贡献10，回写证据5。这组权重有两处值得解释。其一，子本创新智商为30，与化学相同、低于人文的35、高于医学的25，因为社会科学子本的价值主要在于提出可检验的新机制判断，而其现场证据的厚度通常受到材料可得性与伦理边界的限制，不宜像医学那样把20分压在现场上。其二，碰撞贡献为10，在这门课里分量实在，因为周川与林晓的班级判断表明，互动属性的学科尤其依赖同学之间的互撞——个人只能看到一家店的节律，两家店放在一起才看出“现金节律”这个共同结构。学生可在老师给定区间内自选权重，周川为准备接手饭馆拉高了现场证据，但创新智商权重在大学阶段不低于25，这一硬底不因自选而改变。

分档线按本科110执行（创新智商口径），问对深度与子本创新智商须换手评分并写利益冲突声明。在经济与法学课上，换手评分有一个具体的执行要点：评分者应当能看到学生的节律表与原始材料来源说明，否则无法判断“新机制判断”是从现场里长出来的，还是从文献里搬过来又贴上了一本账本。一份子本若机制论证漂亮、节律表却拿不出原始记录支撑，按第17章的说法，就是一份拿不出轨迹的高分子本，现场证据一项与创新智商的F维都要据实扣减。

◆ 学科特有风险：账本里的第三方、公开裁判的边界与现成理论的诱惑

经济与法学课堂有三种风险是别的学科不太会遇到的，下面逐一说，共有的风险（如把班级判断说大）已在各自的位置处理过，不再另列。第一种是账本里的第三方。周川与林晓的现场是自己的家，但流水账、房租合同、储值卡记录里站着一群不在课堂上的人：伙计的工资、供货商的结账条件、房东的租金、顾客的储值余额。别的学科的现场材料多半属于学生自己或属于物，这里的每一个数字背后都对着一个会因此受影响的具体的人，而且这些人与学生之间往往还有现实的买卖或雇佣关系。学生在子本中使用这些材料，应当事先取得家人的同意，涉及第三方的部分去标识化，金额可以用比例或相对变化代替绝对数额。本书的评估方案没有为经济与法学另设一道“整份不计分”的门槛，与医学的伦理门槛、化学的安全门槛不同；但材料来源合规属于完成度中“引注无误”的应有之义，一份未经同意使用他人财务信息的子本，在完成度上不能算合格。家里人说“随便用”，并不等于伙计和房东同意了，这一点在课上要讲清楚。

第二种是公开裁判的边界。我国最高人民法院2016年修订了关于人民法院在互联网公布裁判文书的规定，对公开范围、不公开的情形以及自然人信息的隐名处理都作了规定；法庭旁听亦有法庭规则约束，未经许可不得录音、录像和拍照。这使韩正的现场有两条边界：其一，他只能使用已依法公开的文书与公开开庭时的旁听所得，不能借助任何非公开渠道获取材料；其二，即使是已公开的文书，已隐名的当事人信息也不得通过交叉比对重新识别，旁听记录应以手记为限，记的是程序节点与公开陈述，而不是当事人的身份细节。还有一层容易被忽视的边界是样本本身：已公开文书不等于全部案件，未上网的文书、以其他方式结案的纠纷不在他的样本里。这对前述“物业公司日历与法院日历”的判别尤其要紧，因为哪些案子被公开、哪些没有，可能恰好与结案方式相关。把“已公开文书里的分布”说成“本院全部物业纠纷的分布”，是这门课最容易犯的概括错误。

第三种是现成理论的诱惑，这是社会科学九问特有的失败方式。经济学与法学的母本里有现成的、解释力很强的理论，学生很容易先挑一个理论，再到账本或文书里找能贴上去的材料，最后写出一份“用心理账户理论分析我家饭馆”的子本。在化学课上，这种做法会在复做时现形，在医学课上会被科室的床边核对戳穿，而

社会科学的节律判断很难当场复做，现成理论的解释力又强，贴上去往往看不出破绽。这样的子本看起来有现场，实际上是S→D→E发现版的回潮：终点（理论）事先写好了，现场只是插图。这就像拿着一把已经配好的钥匙满屋子找锁，找到一把能开的就宣布成功，却没问过这扇门本来是不是锁着的。题闸在这里的用法，是拿掉学生的现场，请一个未接入这些材料的通用大模型（记录所用版本）只凭母本写一遍：如果拿掉周川的流水账，它依然能写出一篇结论相同的“小饭馆心理账户分析”，这份子本的判断就没有长在现场里。

◆ 本章怎样算错

社会科学九问若站不住，读数会先在下面三个地方露出来。第一处在节律是否真由现场长出来。连续两届、每届不少于30名选修生，由换手评分者做一次盲测：把母本精要库交给一个未接入学生账本和文书的通用大模型（记录版本），请它写出该生子本中的节律判断；若有超过三分之一的子本，其节律判断被它写到同一个时点位置（误差在一周以内），“社会科学课堂的战场在交易与裁判的节律里”就站不住，至少要改为“节律是母本已知的结构，九问只是让学生验证它”，九问在社会科学中的位置就要降回练手。

第二处在案例教学法能否补上缺口。这是本章最强的竞争解释，它提示也许缺的只是更好的案例，而不是一种新的课堂模式。检验的办法是同一届分成两组，一组做附有完整立案—调解—结案时间线（或完整现金流日历）的案例，一组做九问，两组由同一批换手评分者盲评；若案例组“母本所无的新判断”达到本科分档线110的比例，与九问组相差不足10个百分点，本章把案例教学法定位为“底盘而非战场”的说法就需撤回，九问在社会科学中的增量须重新论证。

第三处在现场材料是否拿得到。在严格遵守隐私与司法公开规则的前提下，若某一届不少于一半的学生凑不齐覆盖至少两个完整周期的材料（月度节律至少24个月的记录，季度节律至少8个季度，法学则为不少于30件已公开文书加不少于2次旁听记录），本章关于社会科学九问可行性的主张就要修改，经济与法学的九问改以公开数据和模拟交易为主，而这将实质性地削弱“下到自己现场”的意义。以上三处都要靠真实课堂的档案来检验；本章里周川、林晓与韩正的故事，连同其中的账本、文书与日期，均为构造，只作说明，不是调查结果。

周川、林晓与韩正的判断都落在“看不见的周期”上，那是交易与裁判里的互动属性。到了历史与文学，九问遇到的是另一种难处：现场是学生自己的家族史料与自己的语言，判断的主体与判断的对象缠在一起，什么才算一个“新判断”，标准本身就需要重新说明。下一章转入苏文的县志与族谱、叶青的方言与草稿。

带 走 的 话

案例课教你在别人裁好的现场上推理，九问要你回到自己的账本和旁听席，找回那个被案例剪掉的、推着互动走的节律；找回来之后，还要把它写到可以被账本打败。

◆

23

第23章

历史与文学：史料与语言里的九问

◆ 两份作业里的同一种空白

《地方史料研讨》开课的第3周，老师收上来一批题为“从县志看本县近代社会变迁”的作业。作业写得都不差：学生们能区分官修志书与私修笔记，能指出县志编纂者的立场，能把县志记载与通史教材的叙述对照，找出若干出入。苏文那一份尤其规整，引了三种版本的县志，列了一张户口、田赋、学校数目的变化表。老师在评语里写了一句：“史料处理规范，但这篇文章换任何一个县的学生来写，结构都不会变。”苏文起初不服，后来在问对里才慢慢明白这句话的意思——他处理的是县志，解释的是县志，可他自己并不在这篇文章里；他家里那部断了一代的族谱、他奶奶讲过多次的那几段往事，一个字也没进来，因为在他学到的史学方法里，那些东西要么“不够正式”，要么“太私人”。

同一个学期，《写作与阅读的发生》课上，叶青交了一篇关于外婆的散文。文章干净、克制，老师读得出功力，也读得出一种别扭：外婆在文中说的每一句话都是普通话，而叶青在课下聊天时说过，外婆一辈子没说过几句普通话。叶青的解释是，方言词写进去，读者看不懂，编辑也会删；她试过几次，总觉得“土”，又改回来了。这是一个在写作课上极为常见的处理，而传统的写作课对它的回应通常有两种：要么讲一通“方言写作的得失”，列举几位用方言写作的名家；要么建议她加注释。两种回应都把问题当成了一个技术选择，却没有人问：那个一直不敢写进去的词，究竟是哪一个？它为什么不敢？

两份作业的空白是同一种空白。人文学科的发现版课堂，最成熟的形态是解释训练：历史课教史料批判与史学方法，文学课教文本细读与批评理论，学生学会的是怎样做一个合格的解释者——面对一份给定的文本，说出它的来历、它的立场、它的修辞与它的意义。这种训练极有价值。没有史料批判，口述与族谱只会被当作

真相照单全收；没有细读的功夫，写作者看不出自己文字里的惯性。但解释训练有一个结构性的前提：文本在外面，解释者在这边。学生越训练有素，就越习惯于把自己的史料与自己的语言放在“不够正式”的那一侧，放在课堂之外。

本章的核心判断是：人文课堂的战场在自己的史料与语言里；九问让学生成为发生的主体，而不只是解释者。这个判断要和全书的三领域对应连起来读：自然科学侧重知识发生的客体属性，社会科学侧重互动属性，人文侧重主体属性（前一章已有讨论）。所谓主体属性，不是说人文学科“主观”，而是说人文知识的发生，很大一部分发生在认识者与材料缠在一起的地方——族谱上断了的那一代，是苏文的祖先；不敢写的那个方言词，是叶青自己的舌头。打个比方，前两章的学生是带着工具下到工地，这一章的学生是在拆自己家的老房子，拆到哪一面墙，都会碰到自己。

需要先说清的是，让学生成为发生的主体，并不意味着放弃解释训练。史料批判与文本细读在九问课堂里降为底盘，而不是被废除：苏文要追自家族谱的断代，首先得知道族谱是被重修过的文献，知道县志的数字有编纂口径；叶青要追自己草稿里那个词的命运，首先得会细读，看得出一个词换成另一个词之后，节奏、语气与指涉各变了什么。这一层功夫，AI如今可以讲得又快又全，讲授与示范可以交给它（见第2章）；母本底盘的复述加反例口试，检查的正是这一层。打个比方，解释训练像是学会看地图、认等高线，九问是背着地图走进一座没人测绘过的山——不会看图的人进山会迷路，可只会看图的人永远不知道山里那条没画上的小路通向哪里。人文课堂的典范转移，不在于从“会解释”退回到“凭感觉”，而在于把解释的功夫用到学生自己的材料上，让那条没画上的小路成为研究对象。

这一层缠绕既是人文九问的力量来源，也是它的风险来源。力量在于，题闸在人文学科里格外有效：拿掉苏文家的族谱和叶青的草稿，一个未接入这些材料的通用大模型（记录所用版本）能写出一篇漂亮的“县志与族谱的互证方法”或“方言写作的得失”，却说不出苏家断了的那一代是哪一年，也说不出叶青不敢写的是哪个词。风险在于，学生离自己的材料太近，容易把私人情感当作判断，把家族记忆当作证据，也容易在不知不觉中把家人与乡邻的隐私写进一份要被评分、可能被存

档的子本里。本章的后半部分要分别处理这两件事：什么才算人文的新判断，以及家族隐私与口述伦理的边界。

> **这一章的问题**
>
> 自己的族谱和自己的方言词，在课堂上常被当作“不够正式”的边角料，它们怎样才能变成一个站得住、可检验的新判断？

◆ 谱系母本与精要：史料与语言里的主体

《地方史料研讨》的谱系母本，前沿专著一侧，老师选入了三类研究：地方志的编纂史与史料价值研究，宗族与族谱研究（族谱作为一种被不断重修、带有编纂意图的文献），以及口述史方法的研究，其中保罗·汤普森（1978）的《过去的声音》被用作口述史方法的基础读物。母本的另一部分是部落档案：这门研讨课历届学生整理的本地县志摘录、族谱目录与经过同意、去标识化的口述记录摘要，以及他们的问对档案。历史课的部落档案有一个与化学课题组不同的特点，它不是围绕同一个反应积累的，而是围绕同一片地方积累的——不同学生家族的族谱，往往在同一部县志里交会，这让档案的每一次加厚都可能为下一届提供可交叉核对的材料。

从母本提炼的历史精要库，方程判断一类有“族谱是被重修的文献，每一次重修都有当时的编纂意图”；路径判断一类有“地方志的记载口径受上级制度与编纂体例约束”；动力判断一类有“家族记忆的断裂常与外部制度变动相伴，而非仅出于家族内部原因”。最后这一条在母本里是一个较为笼统的判断，它说出了“相伴”，却说不出“怎样相伴”——是战乱让人离散，是迁徙让谱系中断，还是某种登记方式的改变让一代人在纸面上消失。苏文的九问，就打在这条判断没有说到的地方。

《写作与阅读的发生》的谱系母本，前沿专著一侧有两类：一类是关于写作过程与草稿研究的成果，文学研究中所谓“文本发生学”即以作家手稿与草稿的演变为对象；另一类是关于方言、语域与文学语言的研究，包括什克洛夫斯基关于“陌生化”的经典论述——日常语言的自动化使人对事物视而不见，文学通过使语言变得陌生而恢复感知。部落档案一侧，是历届学生自愿提交的草稿版本史与问对档

案。精要库里的路径判断有“写作者的修改方向往往揭示其未言明的顾虑”，动力判断有“文学语言的力量常来自对日常语言惯性的打断”。叶青的九问，连着这两条判断，又把它们压到她自己的草稿与她外婆的方言上。

两门课的精要库共同指向一个母本里尚未被充分展开的方向：主体在材料里留下的痕迹。族谱的断代是一个家族在纸上的痕迹，草稿里反复删改的那个词是一个写作者在纸上的痕迹，而这两种痕迹，在发现版的人文课堂上通常被当作“需要被解释的文本特征”，很少被当作“学生自己可以去追的现场”。九问的工作，就是把痕迹变成现场——不是由老师在课上解释“族谱断代的常见原因”，而是让苏文去找自家族谱断了的那一年，县志上写了什么；不是由老师讲“方言写作的若干策略”，而是让叶青翻出自己的每一稿，看那个词是在第几稿、以什么方式被换掉的。

痕迹变成现场，同时也是一次默会知识的外化。默会知识与外化——Polanyi（1966）的《默会维度》，Nonaka与Takeuchi（1995）的知识创造链条，第13章已有交代——在这里有两个现成的对应：一个家族的“怕上册子”，一个写作者对某个词说不清的不安，都是平时在场、从不被说出的知道，把它们带到纸面上就是外化。先要承认最前面的一步并不新：让苏文的奶奶开口讲，让叶青把不安写成一页笔记，历史系的家史作业和写作课的反思日志早就在做。本书多出的一格有两处。其一，外化的产物必须被迫对质：第7问让奶奶的说法、县志的数字、族谱的空白三方互相找茬，第4问让叶青看见被换掉的词周围的段落改了七次，说出来的东西因此带着“可能错在哪里”的标记，而不是一篇动人的自述。其二，九问承认外化有剩余：叶青的第3问问替换词与原词“差在哪一层”，答案里总有一层说不出，子本不假装把它说尽，只要求写明说不出的是哪一层、凭什么知道它在。这像把一口老井的水舀上来化验——化验得出成分，却舀不上井底的那股凉气，好的研究要写明水是从哪一层取的。

◆ 苏文：族谱、县志与口述的九问

苏文的九问由老师从精要库出发、按他的画像打造。画像的要点是：家乡在一个有县志传世的县，家里存有一部族谱，族谱在某一代出现断代；奶奶健在，能讲述家族旧事；苏文本人史料处理规范，但习惯把家族材料排除在“正式”研究之

外。第1问即本书贯穿案例中的那一问：“你家族谱断了的那一代，县志同年记了什么？”苏文的改题经过已在第5章讨论，这里给出问对之后的定版九问。

苏文的九问（定版）如下。第1问，你家族谱断了的那一代，县志同年记了什么？第2问，族谱是哪一年重修的？断代处在重修本上是空白、是“失考”，还是有过涂改的痕迹？第3问，断代前后两代人，在族谱里的记录格式——名、字、生卒、配偶、迁居——有没有变化？哪一项最先消失？第4问，县志在断代那几年的户口数字，与前后年份相比有没有异常的跳动？县志对这一变动作何说明？第5问，同村、同族其他房支的族谱，是否也在同一时段出现断代？第6问，奶奶讲述的家族往事里，哪一段与断代的年代重合？她怎么解释那一代人“没写上”？第7问，奶奶的说法与县志的记载、族谱的空白，三者在哪一点上对不上？第8问，同一时段的上级政令或地方档案里，有没有关于户籍、赋役册籍办法变动的记载？第9问，如果你在第8问里找到的某项变动确与断代有关，那么县志里同年还有哪一项记载应当跟着出现异常？

这套九问的出题逻辑分为三段。前三问是“看清空白本身”：先把断代放到县志的同一年，再看族谱重修时如何处理这段空白，再看断代前后记录格式的变化——第3问问“哪一项最先消失”，是要苏文从格式的变化里读出记录者在什么时候、因为什么不再能记或不再愿记。中间三问是“找旁证与反证”：第4问回到县志的数字，第5问用其他房支的族谱检验断代是否只属于苏家，第6问引进口述。第5问承担排除功能——如果只有苏家这一支断了，家族内部原因（离散、绝嗣、迁居）的可能性就大；如果同时段多支皆断，外部制度原因的可能性就大。后三问是“对质与检验”：第7问让三种史料对质，第8问去找制度层面的证据，第9问是一个预测题，要求苏文从假设出发推出一个尚未查看的可检验后果。

苏文的第零轮初判是“族谱断代是因为那几年战乱，家里人逃散了”。这是一个合情合理、也常常正确的解释，ChatSDE在第一回合先接住了它，肯定了战乱与离散作为家族记忆断裂的常见原因，随即把前提翻过来：“如果是逃散，断代前后的记录格式应当是突然中止、后来再接上；你家族谱的断代处是这样的吗？同一个村子没有逃散的人家，族谱也断了吗？”这一问对应的工序是三大方程查缺式，按第7章的判别细则它是缺式而不是断拍：苏文在问对之前并没有碰到与初判直接相撞

的事实，他从来没有把族谱当作“被人记下来的东西”去看；他的初判只有“事件—后果”一项，缺了“记录行为”这一项——族谱不是事件本身，而是被人在某个制度条件下记下来的东西。苏文回去看了自家族谱，又借来同族另一房支的谱本，发现断代并非突然中止，而是在两三代之间逐步变薄：先是不再记“字”，再是不再记生卒，最后整整一代只剩名字；另一房支在同一时段也出现了相似的变薄。

接下来几轮，苏文带着第4问和第8问回到县志与地方档案，发现县志在相近年份的户口数字有一次不连续的跳动，志书在一处小注里提到当年“丁口编审之法”的变更。他把这一发现带回问对时，ChatSDE没有替他把结论说出来，而是追问：“你在志书小注里读到的那处变更，和你奶奶讲的那段往事有没有关系？”那条小注是志书里原有的记载，属于苏文自己取得的现场材料，追问里没有出现“登记”“户籍”一类的词；把志书与口述放到一处对质，本是九问第7问交给他的任务，追问只是提醒他这两份材料都在手边，怎样连、连出什么，要等他自己去问了奶奶之后才连得起来。苏文去问了奶奶，奶奶讲的是另一件事：她的祖父那一辈人“怕上册子”，有的孩子生下来没报。三种史料在这里对上了一部分，也有一部分对不上——口述里的“怕上册子”说的是家里主动不报，县志里的口径变更说的是官府登记方式改变。苏文最后形成的判断是：断代处恰是户籍登记口径改变之年。这个判断说的是一个时间上的重合与一个可能的机制，它把母本里“家族记忆的断裂常与外部制度变动相伴”这条笼统判断，落到了一个具体的县、一个具体的家族、一种具体的制度变动上；至于口径改变与“怕上册子”之间的因果次序，苏文在子本里明确标为尚待更多房支谱本与档案核对的问题。

◆ 叶青：方言、草稿与承重词的九问

叶青的现场比苏文更贴身：她自己的写作草稿史，以及她从小听外婆说的方言。老师为她打造九问时，画像写着：她有保存草稿的习惯，电脑里的每一篇稿子都留着若干个版本；她以写作为志向；她在关于外婆的散文中，有意识地把方言对话全部改写成了普通话。第1问即本书贯穿案例中的那一问：“你写外婆时，哪个方言词一直不敢写进去？”这一问在题闸上是一个很干净的样板：拿掉叶青的草稿与她的耳朵，任何模型都无法回答，因为答案只在她的版本史和她的记忆里。

叶青的九问（定版）如下。第1问，你写外婆时，哪个方言词一直不敢写进去？第2问，把你写外婆的所有草稿按时间排开，这个词在哪一稿第一次出现、哪一稿被换掉？换成了什么？第3问，替换它的那个普通话词，与原词差在哪里——意思、声音、语气，还是只有外婆用它时才有的那一层？第4问，这个词被换掉之后，它所在的那一段，你又改了几次？改动是不是比别的段落多？第5问，去问问外婆或家里的长辈，这个词他们在什么场合用、对谁用？有没有别的词能代替？第6问，你读过的写到祖辈的作品里，有没有作者把类似的词写了进去？他们是怎样处理读者看不懂的问题的？第7问，如果把这个词放回去，整篇文章哪一处会因此站不住、必须重写？第8问，你在其他文章里，还有没有类似“一直不敢写进去”的词？它们是不是都落在同一类场景里？第9问，把这个词放回去，写一个新的版本，读给一个不懂这种方言的人听，他第一处停下来问的是哪里？

出题逻辑分三段。前三问是“找到那个词并看清它”：第1问点名，第2问在版本史上给它定位，第3问让叶青精确地说出替换造成的损失——这一问很关键，因为“方言更生动”是一个空泛的判断，只有说出“差在哪一层”，才有可能进一步问那一层承担着什么。中间三问是“看它承担了什么”：第4问是一个巧妙的旁证题，如果一个词被换掉后它周围的段落反复改动，说明那个位置失去了某种支撑，写作者在不自觉地用别的东西去补；第5问回到外婆的语言现场，第6问把母本中的作品引进来，让叶青看到别人怎样处理同一个困难。后三问是“检验”：第7问是一个反事实题，第8问看这是不是叶青写作中的一种稳定模式，第9问把判断交给一个外部读者去检验。

叶青的改题史留下了一处值得一看的痕迹。老师版第6问原本是“你读过的写到祖辈的作品里，作者是怎样处理方言的？请归纳几种策略”，这是写作课上常见的阅读任务。叶青在改题时一度把它改成了“方言写作在当代文学中的地位与前景”，理由是这个题“更有研究性”。问对中ChatSDE指出，改后的题拿掉她的草稿依然可以完整作答，离她的现场不是更近而是更远了——这是一次逃离而不是提升（提升与逃离的分辨见第5章）。叶青于是第二次改题，把“归纳几种策略”改成了“作者是怎样处理读者看不懂的问题的”，把焦点从文学史的综述拉回到她自己正在面对的那个具体困难上。这两次改动都留在题的版本史里，第二次改动之所以

算提升，是因为它仍连着母本里关于文学语言的判断，同时让别人的作品变成了她解决自己难处的参照，而不是一个可以脱离她的草稿独立完成的综述。

这段改题史可以拿问题发现的研究来对读。盖泽尔斯与契克森米哈赖（1976）在《创造性视野》里研究艺术学生在动手创作之前怎样选择和构造问题，发现提出问题的方式，比解决问题的熟练更能预测后来的创造性成果。叶青从“方言写作的地位与前景”退回“作者怎样处理读者看不懂”，就是一个小型的问题发现过程，只不过前人把它当作观察对象，本书把它当作一个要计分、要留痕的项目：改题质量占10分，改题史与两锚检验一并存档。这里还要警惕一个反面：问题发现若没有根，会漂向“有趣的题”。叶青第一次改题就是这样，题目听起来更有研究性，却可以脱离她的草稿完成，所以九问对改题多加了一道约束，新题必须仍连着母本里的某条判断，又比原题离现场更近。这像放风筝，线松了风筝飞得高，线断了就不叫放风筝；根与现场是两根线，缺一不可。

叶青的第零轮初判是“方言词不写进去，是因为读者看不懂，写进去会影响阅读”。ChatSDE的第一回合先接住了这个顾虑，承认可读性是写作者的真实责任，随后问：“你在别的文章里写过读者不一定看得懂的专有名词，加个注就写进去了；为什么偏偏这个词，你连注都不愿意加？”这一问对应的工序是123原理查断拍，也符合判别细则里断拍的条件：叶青在别的文章里给专有名词加注就写进去了，唯独对这个词不肯，这个与初判直接相撞的事实就在她自己的写作史里，初判却没有因它而改。具体地说，叶青的“读者看不懂”是一个D层面的理由（关于写作目的与路径），而她在E层面的经验——外婆说这个词时的声音和场合——与这个理由之间有一个未被处理的矛盾，这个D—E矛盾本该逼出S的改变，即她对“这篇文章到底在写什么”的重新认定，却被“看不懂”三个字压住了。叶青第2轮回答说，那个词是外婆在她小时候受了委屈时说的，普通话里没有一个词能同时是责备又是护着。

接下来叶青按第2问和第4问整理了版本史，发现那个词在第一稿里出现过一次，第二稿就被换掉了；而从第二稿开始，外婆安慰她的那一段，前后改了七次，每一次都在加形容、加动作，想把那个词带走的东西补回来，却越补越像别人的外婆。她按第9问写了一个把词放回去的新版本，读给一位北方同学听，同学第一处停

下来问的正是那个词；但同学听完解释之后说，整篇文章到这里才“有了一个真的外婆”。叶青由此形成的判断是：不敢写的词才是叙述的承重词。这个判断与什克洛夫斯基关于陌生化的论述有关联，但方向不同：陌生化说的是文学通过打断语言惯性来恢复感知，叶青的判断说的是，写作者自己最想回避的那个词，往往正是整段叙述的力量所系——回避它，叙述就要用大量的修饰去支撑一个失去了承重结构的位置。需要说明，这是叶青在自己的草稿史上形成的判断，它对她的写作是一个可操作的检查方法，对写作研究则是一个有待在更多写作者的版本史上检验的假说。

◆ 人文的“新判断”：标准、子本与权重

人文学科在九问模式里面对一个独特的追问：什么算新判断？在化学课上，沈澄的湿度假说可以复做，复做的结果会说话；在经济与法学课上，节律可以在更多账本和文书上检验。人文的判断往往不能这样复做——苏文的族谱只有一部，叶青的外婆只有一个。如果新判断的标准只是“母本里没有说过”，那么人文学生几乎写什么都是新的，因为他们的家族和他们的语言在任何母本里都没有被写过；这样一来，新判断的标准就失去了区分力。反过来，如果标准是“能推广到更大范围”，那么人文九问几乎永远达不到，因为它本来就扎在一个具体的主体身上。

本书对人文新判断提出三条标准，它们不是另立的一把尺子，而是创新智商五维在人文学科的读法（五维的一般说明见第16章）。第一，新判断必须回到母本、改动母本里的某一条判断，而不只是增添一个新的材料。苏文的判断改动了“家族记忆断裂常与外部制度变动相伴”这条笼统判断，给出了一个具体的伴随方式；叶青的判断改动了“文学语言的力量来自对日常语言惯性的打断”这条判断，指出写作者对自己语言的回避本身就是一个需要被打断的惯性。只写“我家族谱断在某年”或“外婆说的某个方言词很生动”，那是材料，不是判断。第二，新判断必须在学生自己的材料里经过对质，也就是说，它要能指出自己在哪一处可能是错的——苏文的三种史料在“怕上册子”与“口径改变”的因果次序上对不上，叶青的判断要在她其他文章的同类场景里再检验。第三，新判断必须给出一个可供他人使用的检查方法：苏文的方法是“断代处查同年的登记口径”，叶青的方法是“找出

自己最不敢写的词，看它周围的段落改了几次”。这三条大致对应I维的不可还原性、F维的可证伪性与D维的路径锐度。

对这三条标准，最强的反对来自人文课堂自己的传统：让学生写家史、读自己的草稿，历史系和写作课早就在做，何必另造一套九问？这个反对不能轻轻放过。家史作业确实培养了对材料的亲近，反思日志确实让写作者更自觉；但它们的评价标准通常落在叙述的完整与真诚上，学生交出去的是一份讲得好的家族故事或一篇写得诚恳的反思，母本里没有任何一条判断因此被动过。这就像给自家老屋画了一幅很像的素描，画得再像，也不等于测绘出了承重墙的位置。九问要的是后者：判断要回到母本去改一条，要说得出可能错在哪里，要留下别人拿得去用的检查方法。这个区分有一个可以被打败的读数，在本章末的失败条件中给出。

人文子本的形态与此相应。苏文的历史子本，第一稿曾经是一篇把族谱、县志与口述按时间顺序逐段转述的“家族史”，这正是第12章讲的“翻译而非重写”的典型；重写稿则以判断为纲：先复述母本中关于族谱重修与地方志口径的核心判断，再呈现断代的格式变薄过程与三种史料的对质表，然后论证“口径改变”这一判断并排除“单纯逃散”的竞争解释，最后列出因果次序上尚未解决的问题与可供后来者使用的检查方法。叶青的文学子本有一个特殊之处：它由两部分构成，一部分是重写后的散文本身，另一部分是一篇关于这次重写的研究性论述，后者包括版本史的对照、承重词判断的论证与对母本中陌生化理论的回应。文学子本的创作部分与论述部分缺一不可，只有作品没有论述，评分者无法判断新判断在哪里；只有论述没有作品，判断就失去了最直接的检验。

历史与文学本科研讨课按第18章的学科权重示例执行：母本底盘5，改题质量10，问对深度15，现场证据10，子本创新智商35，子本完成度10，碰撞贡献10，回写证据5。子本创新智商35是四类学科里最高的，原因已在上文：人文判断难以复做，其价值更集中地体现在判断本身的不可还原性与可证伪性上；现场证据10是四类学科里最低的，并不是说人文的现场不重要，而是人文的现场（自己的史料与语言）在量上通常不厚——一部族谱、一位老人的口述、一个作者的十几份草稿——其价值要通过判断的质量而不是证据的厚度来体现。叶青为准备写作拉高了创新智商的权重，苏文则在区间内略微拉高了现场证据，理由书里写的是“我的判断还要

靠更多房支的谱本撑住”。两人的选择都在区间之内，都附有理由书，选择权重本身也是一次判断。分档线按本科110执行，问对深度与子本创新智商换手评分并写利益冲突声明。

人文评分有一个特别需要警惕的偏差：评分者容易被叙述的感染力打动，把动人当作新。叶青放回方言词的那一版散文确实更动人，但它在创新智商上得分，靠的不是“更动人”，而是她说清楚了为什么那个词承重、在哪里可以被检验。反过来，一份写得平实、甚至有些笨拙的历史子本，若能指出一个县志口径变化与多支族谱断代的对应关系，并标明自己的不确定处，其创新智商可以高于一份文采斐然、却只是把家族故事讲得更好听的子本。换手评分在人文学科里因此格外必要：同一个评分者既读了学生的问对、又被学生的叙述打动，很难把感染力与判断分开。

◆ 学科特有风险：同意的层级、结果的归属与情感的混入

人文九问有三种风险，与这门学科的材料性质绑在一起，别的学科的现场不太会遇到。过度推广之类各学科共有的风险，已在各自的子本要求里处理过，不再另列。

第一种是同意的层级，以及研究结果的归属。苏文的族谱不只是他一个人的东西：它由某位长辈保管，记着已故祖先，也与同族其他房支的谱本互相对照；叶青的草稿写外婆，也写到了母亲和其他还在世的家人。这意味着同意不是一次签字，而是几层：持有者同意出借，被记录者的后人同意被写入，其他房支同意被比对。比同意更微妙的是结果本身。族谱上一代人的“没写上”，背后可能是非婚生育、过继、改姓或逃避登记，这些事对一个家族可能至今敏感；苏文顺着第9问查下去，查出来的东西可能比他最初要的更多。本书的规则是：子本可以保留“断代处恰是登记口径改变之年”这类制度层面的判断，而指向具体个人的推断，应在存档前先给持有者读过涉及他们的段落，由持有者在知情后裁定保留、模糊还是删去；进入课程档案或班级子本的部分，涉及在世人物的内容一律去标识化或在提交存档时删除。这像把家里的老箱子搬到课堂上，开箱之前要问全家人，箱子里翻出了什么，也要先让全家人看过再决定摆不摆出来。

第二种是口述伦理。苏文的第6问与叶青的第5问都要求学生去访谈家中长辈，这就进入了口述史的伦理领域。口述史方法的研究早已指出，口述者在讲述时有权决定讲什么、不讲什么，有权知道自己的话将用在哪里，也有权在事后撤回。九问课堂对口述采取以下规则：访谈前告知用途，包括“将用于课程研究、可能被老师和同学阅读、可能进入课程档案”；口述者明确表示不愿被记录的内容，不得记录；事后要求删除的内容，应从子本与档案中删除，即使这会削弱判断；对年迈的口述者，访谈应尊重其身体与情绪状态，不得为了“取证”反复追问痛苦的往事。苏文的奶奶讲“怕上册子”那段时起初不愿多说，苏文在问对记录里写下了自己当时的犹豫，最后选择只记下奶奶愿意说的几句；这条记录本身也是现场证据的一部分，表明学生在取证与伦理之间作出了判断。撤回还有一个别处没有的后果：口述者撤回的是一个人的话，而这段话可能正是子本里唯一把档案与族谱连起来的桥，所以子本要预先写明“若这段口述被撤回，判断还剩下什么”。

第三种是情感与判断的混同。人文九问让学生成为发生的主体，主体与材料的缠绕会产生强烈的情感：苏文追到那一代“没写上”的祖先时，会生出愧疚或同情；叶青写外婆时，会被重新唤起的记忆冲击。情感本身不是问题，它常常是判断的入口；问题在于把情感的强度当作判断的证据。“我觉得这个词最重要”不等于“这个词承重”，“奶奶说的一定是真的”不等于“口述与档案对上了”。ChatSDE在问对中的职能之一，正是在接住学生情感的同时把判断与情感分开：先承认这段材料对学生的意义，再请他把判断改写成“如果这样，材料里还应当出现什么”的形式——苏文的第9问就是这样一道题。能被改写成这种形式的，是判断；改写不出来的，多半还停留在感受上。

◆ 本章怎样算错

人文九问若站不住，会先在下面三个读数上显出来。第一个读数是“新判断”里有多少只是新材料。连续两届、每届不少于25份子本，由两位换手评分者各自独立盲读，先把叙述性段落抽去、只留判断陈述，再逐条标注该判断是“改动了母本某一条判断”还是“只增添了家族或个人的材料”；若标为后者的子本占到一半以上，本章关于“人文九问使学生成为发生的主体”的主张就要降格为“使学生成为新材料的提供者”，三条新判断标准也要重新检讨。同样的盲读也用于对照

组：若交给同批评分者一组传统家史作业与反思日志，其“改动母本判断”的比例与九问组相差不足10个百分点，三条标准就没有额外的区分力。

第二个读数是材料是否经得起撤回。在严格执行家族同意、口述者知情与撤回、去标识化的前提下，若连续两届里有不少于三成的学生因口述者撤回、族谱持有者拒绝出借或草稿涉及的家人不愿被写入，而失去了支撑判断所必需的材料，本章关于人文九问可行性的主张就要修改，人文九问改以公开史料与已出版文本为主，而这将削弱“自己的史料与语言”这一核心设定。

第三个读数是评分的一致性。同一批不少于25份子本，由两位换手评分者独立评创新智商，若相差一个分档（如一人评在110线上、一人评在线下）的子本超过三分之一，而分歧记录显示主要来自对“动人”与“新”的不同理解，那么本章提出的三条新判断标准就未能把人文的创新性变成可评估的对象，人文学科的创新智商权重35须下调至30，直到标准修订到两位评分者的分歧降到五分之一以下。以上三个读数都要靠真实课堂的档案来检验；本章里苏文与叶青的故事、族谱、草稿与口述，均为构造，只作说明，不是调查结果。

到这一章为止，第六编的五个学科样板都停在一个学生、一届课堂的子本上。许宁、沈澄、高远、方舟、周川、林晓、韩正、苏文、叶青，每个人都发生了一个判断，但这些判断在课程结束之后去了哪里？沈澄的湿度假说有没有人复做，韩正的判断有没有被下一届接住，苏文找到的那次登记口径变化会不会成为下一届学生追自家族谱的起点？课程如果每年从零开始重讲一遍，这些判断就只是一届学生的作业；只有当它们回到母本，课程才成为一条会长的谱系。下一章开启第七编，讨论回写。

> **带走的话**
>
> **自己的史料和自己的语言不是研究的边角料，而是判断发生的地方；但动人不等于新，一个判断要回去改动母本里的一条，说得出自己可能错在哪里，才算数。**

◆



第　七　编

回写与大学的领地

编　序

某学院的课程评估会上，一位老师汇报说，本课程已实现回写，精要库逐年加厚。一位老教授只问了一句：请指出哪一条判断是去年没有、今年有的，是哪一份子本带来的。回写要被观察，不能被宣布。本编先把这条标准立起来，再问：这一切放进今天的大学，从哪里起步，又要守住什么。

第24章讲回写的三层：学生底盘、班级子本、谱系母本精要库，并讲怎样识别负向回写。第25章讲过渡期：今天的大学生大多没练过发生技术，大一须补练手，并给出一个学期的样子。第26章讲大学的领地：大厂造发动机，大学造驾驶术，壳可以开放，工序与越积越厚的问对档案要守住。

前一章问东西怎样回到谱系，次一章问人怎样走到这里，末一章问大学守什么。三章合上，九问课堂的一轮才算转完，下一届的课堂也从这里开始。

24

第24章

回写：档案回到谱系母本，课程成为谱系

◆ 评估会上被追问的那一句“已回写”

某学院的课程评估会上，一位老师汇报一门开了三年的九问研讨课。汇报材料做得周全：九问的样例，问对记录的摘选，学生子本的创新智商分布，碰撞场的照片，最后一页写着“本课程已实现学生成果向谱系母本的回写，母本精要库逐年加厚”。评估组里一位老教授问了一个问题：“请指出精要库里哪一条判断是去年没有、今年有的；它是哪一份子本带来的；今年的学生里有谁用过它，用了之后发生了什么。”汇报的老师翻了一会儿材料，说精要库每年都在更新，但具体哪一条来自哪一份子本，没有专门记录；至于今年学生用没用过，可以回去查一查。会场没有人为难他，那一页幻灯片却从此被大家记住了，因为它把一个全书反复强调的原则摆到了台面上：回写要被观察，不能被宣布。

这一幕之所以值得放在第七编的开头，是因为前面23章讲的所有环节——出题、改题、问对、打造子本、碰撞、评估——都有一个共同的去向问题。如果每一届学生发生的判断，在学期结束时随着成绩单一起归档，下一届从同一份母本、同一套精要库重新开始，那么九问课堂与讲授课堂之间就只差了一学期的热闹：讲授课堂每年把同一本书讲一遍，九问课堂每年让一批学生各自发生一遍，课程本身并没有变。打个大白话的比方，这就像一家饭馆每年都换一批厨师，每批厨师都做出几道好菜，可菜谱本上一个字也没添；顾客吃到的是好厨师，饭馆却始终没有长出自己的拿手菜。

本章的核心判断是：课程不是每年重讲一遍，而是一条会长的谱系；回写要被观察，不能被宣布。这句话有两半，前一半说的是课程的本体——一门九问课的“是什么”，不由教学大纲定义，而由它的谱系母本与精要库在一届又一届之间的改变定义；后一半说的是证据责任——一门课说自己在回写，必须拿得出可以被第三方查看的痕迹，否则这个说法本身就是一个未被批判的前提。前一半若成立而

后一半做不到，课程就会在“我们在长”的自我宣布里停止生长；后一半若做到而前一半不成立，那些痕迹就只是档案管理，而不是谱系。

“回写”一词在本书的SDE口径里有确定的位置（概念的一般说明见王德生《SDE思想体系入门》）。123原理所说的循环是：D与E的矛盾逼出S的改变，S的改变再回写D与E。在一个学生身上，这个循环表现为：她带着母本里的判断（S）和自己的初判下到现场（E），在问对中发现初判与现场对不上（D—E矛盾），被逼出一个新判断（S改变），新判断又改变了她下一次出题、提问和看现场的方式（回写D、E）。课程的回写，是同一个循环在更大尺度上的运行：一届学生的新判断改变了母本精要库，精要库的改变又改变了下一届的九问与下一届学生看现场的眼睛。本章要讨论的，是这个循环在三个层次上怎样运行，部落档案怎样借此加厚，回写怎样会走向反面，以及怎样让它被看见。

> **这一章的问题**
>
> 一门课说自己“已回写”，凭什么让没参加过这门课的人也信？

◆ 三层回写：底盘、班级子本与精要库

回写发生在三个层次上，它们的尺度不同，留下痕迹的方式也不同。第一层是学生自己的底盘。所谓底盘，指一个人用来理解新问题的那一套判断、方法与习惯；九问课堂对它的改写，体现在学生第零轮的初判与期末自述的对照上，也体现在子本进入个人知识发生平台之后的继续生长上。许宁在第零轮写的是“夜间跌倒是因为患者年纪大”，期末自述里她写的是，她现在遇到任何一个“因为年纪大”式的解释，都会先问“同样年纪的，为什么有的不这样”；沈澄在第零轮写的是“产率低是因为试剂纯度不够”，期末她写的是，她现在读任何一份规程，都会先找哪一处写着“室温”“适量”“迅速”这类没有数的词。这种改写是评估第八项“回写证据”的主要内容，它的痕迹在学生的文字里：不是学生宣称“我的思维方式变了”，而是两份前后文字放在一起，看得出问问题的方式确实变了。

第二层是班级子本。班级子本由同学碰撞场里涌现的、个人说不出的判断所成（见第14章），它本身就是一次回写：周川的“动力被账本提前收走”与林晓母亲交租前的沉默撞在一起，形成“小微经营者的动力被现金节律收走”，这条班级判断反过来改写了两人各自的子本——周川在子本里补写了一节，说明他家饭馆的预

扣记法只是现金节律的一种表现形式；林晓则回去重新看了储值卡预收款的问题，因为班级判断提醒她，现金节律不只有流出，也有提前流入。班级子本的回写痕迹，在于个人子本里出现了“碰撞之后”的修订，并且修订注明了来源。没有这一层，碰撞场就只是一次热闹的讨论会；有了这一层，同学之间的互撞就进入了每个人的文本。

第三层是谱系母本精要库。这是尺度最大、也最难做到的一层。精要库是从母本提炼的核心判断库，一次建成、千人共用、逐年回写。所谓逐年回写，是说学生子本中经过检验的新判断，可以作为新条目进入精要库，或者改写精要库里的某一条已有判断，或者为某一条判断附上一个反例与适用边界。沈澄的湿度假说若经复做验证，就可能改写课题组规程，并在精要库的路径判断里补上一条“本组该步骤成败依赖环境湿度与操作时长”；苏文找到的登记口径变化，可以为“家族记忆断裂常与外部制度变动相伴”这条动力判断附上一个具体的伴随方式。精要库的回写有一个关键的门槛：进入精要库的必须是经过检验或至少经过对质的判断，而不是一届学生的未经检验的假说。许宁的判断被明确标为“本病房假说，待检验”，在交科室验证之前，它只能以“待检假说”的身份挂在精要库的待检栏里，不能以核心判断的身份进入精要库的正文。

三层之间有一个次序：底盘的回写最快，一学期之内就能看到；班级子本的回写在碰撞场之后几周内完成；精要库的回写最慢，往往要跨越一届甚至几届，因为它要等待复做、扩样与反驳。这个次序很像一棵树的生长：叶子一个季节就换一批，枝条要几年才长出一截，树干的年轮则要一年一圈地加。一门课若只看到叶子在换——每届学生都有收获——就以为树在长，那是误判；树干没有新的年轮，课程就仍然是同一棵树。反过来，也不能因为树干长得慢就否认叶子的意义：学生底盘的改写本身就是课堂的直接产出，它不以精要库的改变为前提。

三层回写还对应着三种不同的“所有权”。底盘属于学生本人，它的回写由学生在子本与自述里呈现，别人不能替他宣布；班级子本属于这一届同学，它的回写要经过碰撞记录与班级讨论确认；精要库属于整个部落——课题组、教研室或一门课的历届师生——它的回写需要有一个公开的、可查的采纳程序，包括谁提出、依据什么、经谁核对、何时进入、何时可以被撤回。第15章讲的评估第八项“有无贡献进母本精要”，正是从学生一侧对这个采纳程序的记录：一个学生的判断被采纳

进精要库，在他的成绩单上留下一条采纳记录；而这条记录本身也应当能够在精要库的版本史里找到对应的条目。

◆ 部落档案的积累：沈澄的湿度假说进入规程

部落档案是谱系母本里逐年加重的那一部分。沈澄所在课题组的十年反应条件档案，就是一个已经很厚的部落档案：十年里每一批反应的记录本、失败批次、谱图都在里面。但这份档案厚，却可能有一个结构性的空白：沈澄的第3问追问的，正是“室温”一词十年里有没有一次写过具体的温度和湿度；第5问追问的，是师兄师姐口头说的“那一步要快”规程里写了没有。档案的厚，是记录数量上的厚；九问课堂要它长出的，是判断上的厚。沈澄的九问打向的正是这个空白，她形成的判断，就是那条以“规程里没写的那一句”为关键词的假说。这个判断在她的子本里被明确标为“待复做验证”，因为她复做的批次有限，季节也只覆盖了一段时间。

回写从这里开始，而且是以一种可以被看见的方式开始的。沈澄的子本进入课题组档案时，附带了三样东西：一是她复做那一批的逐步记录，每一步的时长、温度和湿度都有数；二是她的第9问及其回答，即“如果把湿度写进本组规程，第一个要改的是哪一行”——她指出的是规程里那一步写着“室温下搅拌，迅速处理”的那一行；三是一份写给下一届的“待检清单”，列出她的判断在哪些条件下可能不成立，例如湿度的影响是否只在某一个季节显著，操作时长与湿度是不是同一个因素的两种表现。这三样东西使沈澄的判断不再是一份作业的结论，而成为档案里一个可以被下一个人接手的未完成课题。

下一届课程的老师在出题时，把沈澄的待检清单放进了精要库的“待检假说”栏，并据此为一位新同学打造了九问。这位同学的九问里有一问是“按沈澄的记录表格，在另一个季节复做同一批，你的记录与她的在哪一步出现了同样方向的差异，在哪一步没有？”这一问本身就是回写的痕迹：它的文字里写着前一届学生的名字，它的现场是前一届学生留下的记录格式，它要检验的是前一届学生的判断。复做的产率与湿度数值在这里不给出，因为本书所有人物与课堂都是构造案例，编出数字只会冒充调查结果；就本章要说明的机制而言，要紧的是复做之后发生的事：复做记录与沈澄的记录在关键那一步上出现了同方向的差异，课题组在组会上讨论之后，决定修改规程中那一行，把“室温下搅拌，迅速处理”改为写明湿

度条件与操作时限的表述，并在规程修订记录里注明修改依据——两届学生的复做记录及其所在子本。

这一段经历里有几个值得单独指出的地方。其一，进入规程的不是沈澄一个人的判断，而是经过下一届复做之后的判断；如果复做的结果与沈澄的方向相反，那么回写的内容就应当是“该假说在另一季节未得到支持”，这同样是部落档案的加厚。其二，规程修改的依据可以追溯到具体的子本与具体的复做记录，这使一年以后、五年以后的人仍然能问“这一行为什么这样写”，并得到一个可以查证的回答——这恰好是十年来“那一步要快”所缺少的东西。其三，沈澄本人已经毕业，但她的判断在她离开之后才完成了回写；这说明课程谱系的时间尺度大于一届学生的学期，大学与学生的关系因此也不随学期结束而结束。其四，规程修改之后，沈澄当年的第3问“实验记录里‘室温’一词，十年里有没有一次写过具体温度和湿度”，对下一届学生就失去了意义——这一点放到本章后面讨论负向回写与题闸时还要回来。

部落档案的积累由此显出它与普通资料积累的区别。普通的资料积累，是记录本一年年变多，像仓库里的箱子一年年堆高；部落档案的积累，是箱子之间建立起可追溯的引用关系——这一届的九问引用上一届的待检清单，这一次规程修改引用两届的复做记录，这一条精要库判断引用某几份子本。档案厚不厚，不看箱子有多少，而看这种引用关系有多密、能不能顺着它从任何一条现行判断追溯到它的来历，也能从任何一份旧子本追踪到它后来被谁用过、被谁推翻过。

◆ 韩正一届的判断成为下一届的九问

法学研讨课的回写走的是另一条路，它不落在一份规程上，而落在下一届的九问里。韩正一届在《司法实践研讨》里形成了几个判断，其中最受关注的是韩正的那一条：调解进程被物业的账单周期推着走（见第22章）。这一条判断与化学课上沈澄的判断有一个性质上的差别：它不能通过复做一批反应来检验，因为基层法院的案件不会为了检验而重来一遍；它也不能直接改写任何规程，因为法院的调解安排不归一门研讨课管，学生更不应当以课程判断去影响司法实践。它能被回写的地方，只有课程自己的谱系母本，以及下一届学生的问题。

下一届开课前，老师把韩正一届的子本与问对档案读了一遍，在精要库里做了两件事。第一件，在“法院调解在立案后的推进受审限与排期约束”这条路径判断下，附上一条“待检假说”：本院物业纠纷中，调解结束时点向物业公司年度对账期聚拢，可能受一方当事人财务日历的推动；附注里写明来源是韩正的子本，样本为本院近三年已公开文书与有限次旁听，并列出韩正自己写下的三条局限——已公开文书不等于全部案件、旁听次数少、未比较其他法院。第二件，在“重复博弈者更能承受程序时间成本”这条动力判断下，附上一句问题而不是结论：“重复博弈者的时间优势，是否包括替程序定节奏？”这一句话的措辞是有意的，精要库在这里收的是韩正打开的问题，而不是他尚未充分检验的答案。

下一届九问的出题，就从这两处附注出发。一位新同学的九问里有这样几问：“韩正一届发现本院物业纠纷调解结束时点向物业对账期聚拢；把时间窗再往后推一年，加入新公开的文书，这一聚拢还在吗？”“本市另一家基层法院的已公开物业纠纷文书里，有没有同样的聚拢？”“本院供暖费、停车费等其他由企业按周期收费引起的纠纷，调解结束时点跟着谁的日历走？”另一位同学则从反面出题：“韩正一届的判断里，业主一方的日历在哪里？业主的发薪日、年终奖发放时点，有没有同样推着调解走？”这些问题在题闸上都是合格的：拿掉这两位同学将要整理的新文书与新旁听，一个未接入这些材料的通用大模型（记录所用版本）给不出说得过去的答案；而它们的“根”，不再只是前沿专著里的一般判断，而是上一届学生留在部落档案里的具体判断。

这就是“一届的判断成为下一届的九问”的意思。它有三层含义。第一，下一届的出题起点比上一届高了一截：上一届的学生从“法院调解受审限约束”这条一般判断出发，花了一个学期才找到“物业对账期”这个推手；下一届的学生从“物业对账期可能是推手”出发，可以直接去检验它、扩展它或推翻它。第二，上一届的判断在下一届那里处于被检验的位置，而不是被继承的位置——下一届的学生被要求去推翻它，至少是去找它的边界，这正是谱系与教条的区别。第三，问题的署名被保留下来：下一届九问的文字里写着“韩正一届”，这使回写成为一条可以追溯的线，而不是一种只在老师心里发生的“经验积累”。

韩正的改题也以另一种方式回写进了课程。第22章讲过，他把老师版第5问从“记录双方的主要争点”改成了“让步的时点与理由”，理由是争点在判决书里

都有，让步的时点只有旁听席上才拿得到。下一届老师在打造旁听类问题时，采纳了这一改法：凡要求学生旁听的问题，不再以“记录争点”为任务，而以“记录判决书上看不到的程序节点”为任务。这是一次落在出题方法上的回写，它改的不是精要库里的某一条判断，而是老师出题的习惯；它的痕迹在于下一届九问的措辞变化，以及老师在出题记录里写下的一行说明：“旁听题改为记录程序节点，依据为上一届改题史。”学生改题，最终改了老师的出题方式，这是九问模式里改题权的一个更长的后果（见第5章）。

◆ 负向回写：怎样识别“越写越窄”

回写不总是向好的方向走。如果说正向回写是谱系长出新的一截，负向回写就是谱系在回写的名义下变窄、变硬，甚至退回到统一性知识的那一侧。负向回写最危险之处，在于它表面上也有痕迹——精要库确实在变，九问确实在引用上一届，档案确实在加厚——所以单靠“有没有痕迹”识别不出它。识别它，需要看痕迹的方向。

第一种负向回写是定论化：上一届的待检假说在进入精要库时被悄悄去掉了“待检”二字，变成下一届的核心判断，甚至进入母本底盘的口试。设想韩正的判断被写成“本院物业纠纷调解由物业对账期决定”，并在下一届的底盘口试中要求学生复述，那么下一届学生面对的就是一个“标准答案”，他们的九问会从“检验它”滑向“应用它”，课堂就退回到了S→D→E发现版。这就像一条山路，第一个人走出来时满是试探，第二个人照着脚印走，第三个人在路边立了块牌子写着“唯一通道”——路没有错，错在牌子把一条试出来的路说成了唯一的路。识别定论化的痕迹，是看精要库条目的身份标签有没有在缺少新证据的情况下被升级。

第二种负向回写是题的耗尽。九问每一届都以上一届的判断为根，若干届之后，问题可能越来越细，越来越集中在同一条判断的边角上，每一问都还过得了题闸，学生的现场却越来越像同一个现场，新判断越来越像上一届判断的修订。前面说过，沈澄的第3问在规程修改之后对下一届就失去了意义，这是题因回写而自然退场，是好事；但如果老师不让旧问退场、也不让新方向进场，九问就会在一条判断周围绕圈。识别题的耗尽，可以看连续几届的子本创新智商中I维（不可还原性）的走势，以及新判断与上一届判断之间的距离；若几届下来，新判断越来越只是“在

某某条件下仍成立”式的确认，就说明这条谱系需要老师引入新的母本来源，或在出题时转向另一条判断。

第三种负向回写是回写的单向化：只有学生的判断被写进精要库，学生的反驳与失败却没有。沈澄的湿度假说若在下一届复做中未得到支持，而这次失败没有被写进档案，档案里就只留下了成功的判断，下一届看到的谱系是一条只有胜利的谱系。这种选择性的回写会让档案越来越像一份宣传材料，而不是一份研究档案；它也会让学生学会一种危险的习惯——只有被采纳的判断才值得留下。识别它的办法很简单：看档案里有没有“未获支持”“已撤回”“已被某届推翻”这类条目。一份十年都没有撤回记录的精要库，比一份撤回频繁的精要库更值得怀疑。

第四种负向回写落在学生底盘上，即底盘的套路化。学生在九问课堂里学会了“同样年纪的，为什么有的不这样”“规程里没写的那一句在哪里”这类问法，这是底盘的正向改写；但如果他把这些问法变成一套逢题必用的模板，无论遇到什么问题都先问“没写的那一句”，那么他学到的就不是工序选择，而是一句口头禅。第7章讲过，批判非套路化来自按判断坏在哪里选刀；学生底盘的负向回写，恰恰是把一把刀用成了唯一的刀。识别它的痕迹，在学生后续课程的问对记录里：看他的首问是否随问题的类型而变，还是总是同一句。

四种负向回写各有一个可以事先写下的读数，免得识别全凭读者的直觉。对定论化，读数是精要库条目的身份升级：任何一条从“待检假说”升为“核心判断”的条目，必须附有一份不同于原提出者的复做或对质记录，凡无此记录的升级一律标红，标红条数应为零。对题的耗尽，读数是新判断与上一届判断的关系：请两位评阅者把一届子本中的新判断分为“确认型”（在某条件下上一届判断仍成立）与“开新型”（打开了上一届判断没有触及的方向），若连续三届、每届不少于10份子本，确认型都占一半以上，就该引入新的母本来源。对单向化，读数是撤回记录：一个精要库里至少有五条条目被下一届复做过，而三届之内“未获支持”“已撤回”的条目仍为零，这个零本身就是可疑的信号，因为在五次真实的复做里全部得到支持，比出现几次失败更少见。对套路化，读数是同一学生跨课程问对记录里首问的句式：连续三门课、三个不同现象，首问都是同一个句式，就该由老师把他送去做一次“只许换刀、不许重复”的问对。

◆ 回写进课程：与研究型教学、连接课程和双环学习的对话

让学生做研究、并让研究回到课程里，不是本书的发明，必须把前人的工作摆出来，才看得清本章多出的是哪一格。Healey与Jenkins（2009）为英国高等教育学会写的报告《本科生研究与探究的发展》，把本科教学与研究的关系放进一个四象限：一条轴问重点落在研究的内容还是研究的过程与问题，另一条轴问学生是听众还是参与者。九问课堂落在“学生是参与者、重点在过程与问题”的那一格，即报告所称研究型的一格，这一点本书不必争。要紧的是，四象限描述的是一门课在某一时刻的教学姿态，并不规定这门课一届之后会变成什么样。一门课完全可以年年让学生做研究，年年以研究型自居，课程本身却原样不动，就像开头那家年年换厨师、菜谱却不添一字的饭馆。本章多出的，正是这个“一届之后”。

Fung（2017）的《高等教育的连接课程》（UCL Press）更近一步：它主张学生真的做研究，并且让研究成果跨年级、跨届地连接进课程，而不是做完交差。九问课堂的回写与它同向，沈澄的复做记录被下一届接手，韩正的判断成为下一届的问题，都是跨届连接。但连接课程的构想着力于连接的设计，也就是哪里架桥、桥通向何处，至于桥上有没有人走过、走过之后桥有没有被改，主要留给实施者自己去说；而任何一所学校都可能把一座无人走的桥写进课程简介。本书多出的一格，是给连接加上证据责任，并且允许连接是负的。证据责任体现为四类可由第三方核查的痕迹，其中必须有撤回与失败记录；允许负向则体现为前面“负向回写”一节的识别：有些看上去热闹的“连接”，其实是把上一届的假说悄悄悬成了下一届的教条。如果说Fung画出的是一张通车的公路图，本章补上的就是路政的养护台账：哪一段修过，哪一段塌过方，封路之后谁改了线路。

Argyris与Schön（1978）的《组织学习》区分了单环学习与双环学习，这是检验“回写有没有改到要害”的一把更细的尺子。单环学习是在既定的支配变量之内纠正行动，好比恒温器：室温低于设定值就开暖气，设定值本身从不被问；双环学习则回头问，设定值该不该是这个。把这组区分搬进课程，回写便可分成两层。精要库新增一条判断、给一条旧判断附一个反例、规程里改一行字，改的是课程“说了什么”，而出题规则、题闸标准、评估权重原样不动，这是单环回写。韩正把旁听题从“记录争点”改成“记录判决书上看不到的程序节点”，改的却是老师出题时的支配变量，即“什么东西值得学生专程去旁听”，这是双环回写。沈澄的规程

修改介于两者之间：改一行规程属于单环，但它让第3问对下一届退场，等于悄悄改变了“哪一类问算有价值”，向双环迈了半步。

一门课如果只有单环回写，会显得非常勤勉，精要库年年新增，档案年年加厚，却渐渐被自己的出题标准圈住，这正是前面所说的题的耗尽；一门课如果只有双环而无单环，则年年推翻规则而无积累，每一届都像第一届。可以操作的读数是在出题记录里把修订分成“条目层”与“规则层”两栏：条目层每届都应当有，规则层不必每届都有，但三届之内应当至少出现一次，并且每次都能指到具体的子本或改题史作依据。之所以强调依据，是因为双环学习在组织里本来就难：规则是既有安排的受益者最舍不得问的东西，老师把学生的改题定性为“逃离”，有时恰是为了守住自己原来的题。要求规则层的每次修订都落到学生一侧的记录上，就是用制度替老师提前回答了“为什么这次改规则”这个最容易被含糊过去的问题。

◆ 回写的可观察痕迹

既然回写要被观察而不能被宣布，就需要说清楚可以观察的是什么。本书把回写的可观察痕迹归为四类，它们与三层回写一一对应，又多出一类专门用来检查负向回写。这四类痕迹的共同要求是：第三方——不是这门课的老师，也不是写子本的学生——能够依据记录本身作出判断，不必依赖当事人的陈述。

第一类是学生底盘的前后对照。证据是第零轮初判与期末自述，以及学生在后续课程或毕业后更新子本时留下的文本。观察的要点不是学生是否说“我变了”，而是他面对一个新问题时第一句话怎么问；可以设计的做法是，在学期末给学生一个与本课无关的新现象，请他写下对这个现象的初判与他准备问的第一个问题，与第零轮对照。第二类是子本与班级子本里的修订记录。证据是带日期与来源注明的修订：个人子本中“碰撞之后”的补写注明了来自哪一次碰撞、哪位同学的哪个判断；班级子本注明了哪几份个人子本参与了涌现。这类痕迹的作用，是让班级判断不被任何一个人独占，也不被说成是凭空出现。

第三类是精要库的版本史与引用链。证据是精要库每一个条目的身份（核心判断、待检假说、已撤回）、进入日期、来源子本、核对人，以及这一条目此后被哪一届九问引用过、被哪一份子本支持或反驳过。沈澄的湿度假说从“待检假说”到“写入规程”的全过程，韩正的判断从“子本结论”到“精要库附注”再

到“下一届九问”的全过程，都应当能在版本史里顺着引用链追溯出来。评估会上那位老教授的问题——哪一条是今年新增的、来自哪一份子本、今年谁用过——在一份合格的版本史面前，应当能在几分钟内回答。第四类是撤回与失败记录。证据是精要库与档案中“未获支持”“已撤回”“题已退场”的条目，以及连续几届I维走势的记录。这一类痕迹专门用来检查负向回写，它的存在本身就是一门课在认真回写的标志。

这四类痕迹放在一起，构成了一门课的“谱系档案”。打个比方，它像一家老字号的账房：不只记今天卖了多少，还记这道菜的方子是哪一年谁改的、为什么改、改了之后哪位老客人说过不对、后来又改回来没有。外人走进来，不用听掌柜讲老字号的故事，翻账本就知道这家店是不是真的在长。九问课堂里，这本账由ChatSDE的问对档案、子本的版本史与精要库的版本史共同构成；工序层把出题、问对与档案回写放在同一条链上，痕迹因而可以随流程自动留下（见第10章），而判断哪些痕迹构成正向回写、哪些构成负向回写，仍然要由人在评估中完成。

还需要说明的是，回写痕迹的观察本身也要换手。一门课的老师既是精要库的维护者，又是自己课程回写成效的汇报者，由他自己判断“本课程回写良好”，与学生自评子本创新智商有同样的利益冲突。合理的安排是，由同一学科的另一门课、另一位老师或院系的课程评估组，依据上述四类痕迹定期抽查；同一套智能体生成的多份报告，也不能互作独立证据。这条要求与第16章关于换手评分的规定一致，只是对象从一份子本换成了一条谱系。

◆ 本章怎样算错

本章押在三处，每一处都可以用记录来检验。第一处是回写的来源。取一门实行九问模式、已开满三届的课程，逐届清点精要库版本史里新增与改写的条目，看其中注明来源子本的有多少。若连续三届，每届里能追溯到学生子本的条目都不足一半，其余来自老师自己的阅读与研究，那么“档案回到谱系母本，课程成为谱系”的主张就不成立，至少要改为“课程由老师的研究推动生长，学生子本是它的练习场”。这个情况若普遍出现，九问模式在回写一环上与传统的研究型教学并无实质差别，Healey与Jenkins的“研究型”一格已经足以描述它。

第二处是回写的可核查性。请一位没有参与过这门课的第三方，只凭版本史与引用链，在30分钟内回答评估会上那位老教授的三个问题：哪一条是今年新增的，来自哪一份子本，今年谁用过它。再从精要库中随机抽五条条目，要求同样追溯来历。若30分钟内答不全，或五条里有两条以上追不到来源，则四类痕迹在实践中无法被独立核查，“回写要被观察”就只是口号，本章的观察方案需要重新设计；在改好之前，任何关于这门课回写的说法都只能算宣布，不能算证据。

第三处是回写的方向与层次。若连续三届的记录显示，条目层修订不断而规则层修订为零，按双环学习的标准，这门课只在单环里回写，“课程是一条会长的谱系”要降格为“课程是一个会加厚的资料库”；若同期定论化标红、确认型子本过半、撤回记录为零这几个读数中有两个以上同时出现，则正向回写并不比负向回写更常见，本章关于课程会长的判断就要修改：回写本身不保证谱系生长，还需要另外的机制来防止它变窄。以上三条之外，还要申明一点：本章的沈澄、韩正与他们的下一届，都是构造案例，只说明一种机制可以怎样运转，不能当作它已经运转成功的证据。

回写说明了一门课怎样在一届又一届之间长成谱系，但它预设了一件事：学生已经会改题、会问对、会下现场，也会把判断留下来供后来者检验。今天走进大学的学生，大多没有在中学练过这套发生技术；他们第一次面对“题可以改”时，往往不知所措。下一章转到过渡期，讨论大一怎样补练手，以及一个九问学期从第1周到第16周究竟是什么样子。

> **带 走 的 话**
>
> **课程不是每年重讲一遍，而是一条会长的谱系；回写要拿得出别人查得到的痕迹，而且痕迹里必须有撤回和失败的那几条，否则只是宣布。**

◆

25

第25章

过渡期与一个学期的样子

◆ 一份写不出来的第零轮

开学第2周，医学院大一的生理学基础课上，老师把一张空白的“第零轮”发到每个人的问对界面里，只写了一行要求：在打开ChatSDE之前，先用自己的话写下你对“人为什么从躺着站起来不会晕倒”的初步判断，对错不论，写满三句即可。20分钟过去，吴静的界面上仍然只有一行字：“老师，这一题的标准答案是什么？”她不是不认真，恰恰相反，她是全班最早把教材第四章读完的人，压力感受器、交感缩血管、心率加快这些词她都背得出来。她卡住的地方在别处：她从来没有在“不知道标准答案”的情况下先说出一个判断，也从来没有被允许把题目改掉。高中三年，她练成的是一套极其可靠的发现技术——审题、调取、套用、核对——而这一页空白要的是另一套技术。

这个场景在今天的大学一年级里并不罕见，它提出的问题也比“学生会不会写初判”更大。本书前24章描述的九问模式，默认学生在进入大学课堂时已经具备发生技术：会改题、会问对、接得住批判、会下现场取证、会把材料重写成文。可是这套技术本应在初中、高中的练手里练熟，而今天大多数中学还在用统一性知识的发现版课堂培养学生，练的是答题而不是发生。于是出现一个断层：大学想把学生直接带上战场，学生手里却只有驾校里从没摸过的方向盘。本章要回答的，就是在中学尚未跟上的这几年里，大学课堂如何过渡，以及一门九问课在一个学期里到底长什么样。

本章的核心判断可以先写在这里：今天的大学生大多没有练过发生技术，过渡期的大一须补练手；补练手不是降低大学的标准，而是把本该在中学完成的那一段临时搬进大学；等中学跟上，这一段就应当取消。这个判断有两个限定。其一，补练手只针对“技术”，不针对“对象”——吴静要补的是改题、问对、取证的手

感，而不是要在大一重新学一遍高中生物。其二，补练手有明确的退出条件，它是脚手架而不是新楼层，若某所大学把它当成永久课程长期保留，那恰恰说明中学的练手没有被推动，或者大学自己没有看清它的过渡性质。

> **这一章的问题**
>
> 大一新生还没练过发生技术，大学该怎样临时补上这一课，又怎样保证补完就退出？

◆ 过渡期为什么要补练手

先把练手和战场的区分再落一次地。练手是统一性知识的再发生：老师能事先写出终点，学生走一遍发生的流程，把终点“自己长出来”一次，练的是发生技术本身；战场是谱系知识的新发生：连老师也写不出终点，九问打向一个尚不存在的判断。学车是最贴切的比喻。驾校场地里的倒车入库，教练知道每一根杆在哪里，学员练的是手脚配合；上路以后，前方路口会冒出什么，教练也不知道，学员靠的是在场地里练出来的那双手。没有人会让一个从没进过场地的人直接上高速，但也没有人会让一个已经拿到驾照的人回场地再练一个学期。

今天的大一新生恰好处在“没进过场地”的状态。他们在应试高中里接受的是高强度的发现训练，这种训练有真实的功绩：知识面宽、计算熟练、对体系的把握稳定，这些都会在九问模式里转化为母本底盘（第①项）。但发现训练有一个系统性的空白：它从不要求学生在终点未知时先表态，也从不允许学生改动题目，更不要求学生把一个判断带回自己的现场去碰一碰。这三项空白，正好落在九问模式的改题、第零轮与现场证据上。如果大学不补，九问课的第一个月就会被学生用发现技术“攻破”：改题变成换一个更容易查的题，问对变成向ChatSDE索要标准答案，现场证据变成从网上摘抄的文献综述。

补练手的做法应当克制。我们建议在大一选一到两门基础课，各拿出一个单元，完整地走一遍九问流程：老师发九问，学生写第零轮、改题、问对、下一个小现场、交一份短子本、接受评分。之所以只选一两门，是因为补练手的目的在于让学生“走完一遍”，而不是把整个大一变成方法课；技术一旦走通，后续的战场课自然会把它磨利。这就像学游泳，教练只需要在浅水区陪你完整地换一次气、蹬一

次腿，剩下的距离要在深水里游出来。基础课又是最合适的载体，因为它的终点老师事先写得出来，学生即便走偏，老师也能看清偏在哪一步。

补练手的评估权重也与战场课不同。按第18章的权重示例，大一补练手课的母本底盘占20%，问对深度占20%，子本创新智商占20%，改题质量、现场证据、子本完成度各占10%，碰撞贡献与回写证据各占5%。这组数字的意思很清楚：补练手阶段既要守住底盘，也要把问对这一环练实，子本创新智商仍然保留五分之一，提醒学生即便终点已知，重写的过程里也可能长出教材上没有的东西。需要注明，这个20%低于本书为创新智商项定的不低于25%的下限，是专为补练手课开的例外，理由见第18章。与本科研讨课相比，母本底盘从5%升到20%，子本创新智商从30%降到20%，这一升一降正是“练手”与“战场”在评估上的投影。

最后要说清楚补练手的退出条件。所谓“中学跟上”，指的不是某个文件发布，而是可以被观察的变化：新生入学时的第零轮不再空白，改题的版本史里出现“离现场更近”的提升而不只是逃离，问对记录里关键判断由学生先说的比例明显增加。当一所大学连续几届观察到这些变化，补练手单元就应当逐步压缩，直到取消。大学不应在这里建立新的既得利益——一个为补练手而设的教研室、一套为补练手而写的教材，都可能让过渡期变成永久期，这恰恰违背了它的设计初衷。

◆ 补练手是强指导：与最小指导批评、生产性失败和ICAP的对话

补练手的做法有一个容易被忽略的特征：它是强指导的，而不是放手让新生去“自己探究”。这一点必须说在前面，因为教育心理学里有一场老争论，恰好落在这里。Kirschner、Sweller与Clark（2006）在《教育心理学家》上论证，最小指导的探究、问题解决与发现式教学，对新手往往低效，因为新手没有足够的图式去承受搜索的负担。Lazonder与Harmsen（2016）的综述给出了一个更细的结论：探究式学习的效果取决于指导的多少，适度指导的探究优于无指导的探究。这两项工作对本章是友军而不是对手。补练手之所以只选基础课、只选终点老师写得出来的单元，之所以把第零轮的门槛放低到三句话，之所以让ChatSDE每一回合都先接住再追问，正是因为大一新生在发生技术上是新手，给他的必须是一条有扶手的路。真正需要放手的是战场课，那时学生的发生技术已经练成。

与此同时，补练手的第零轮还接上了另一条研究线索。Kapur（2008，2016）以及Sinha与Kapur（2021）的系列工作表明，先让学生独立尝试、生成并失败，再进入讲授，学习效果在概念理解上常优于先讲后练，元分析也支持这一优势。一个值得指出的事实是，这类研究多取有规范解法的概念题做材料，也就是本书所说的练手题。吴静的第零轮正好是这种序列的课堂版：她先在不知标准答案时写下三句初判，教材与老师的讲解随后才来对照。但本书不照搬“生产性失败”的全部设计。第零轮多出的一格，是初判落在学生自己的现场，而不是教师设计的问题；失败的条件也由学生自己写下，第零轮因此还兼作期末回写的基线。对生产性失败，失败是通向讲授的手段；对第零轮，失败的文本本身要被留下来，几个月后拿出来与期末自述对照。

再看第6章已经用过的Chi与Wylie（2014）的ICAP框架：学生的投入分为被动、主动、建构、互动四种方式，学习效果依次递增。高中三年的发现训练，投入多落在被动与主动两格，听讲，划线，抄写，套用。补练手的流程把投入推到后两格：写第零轮是建构，因为学生要生成教材上没有现成句子的判断；问对与碰撞是互动，因为判断要在与他者的往来里被改写。把这个框架当作一个自查工具很有用。老师可以随手检验一个练手单元：如果一周下来，学生做的主要仍是核对与套用，不管界面上贴着多么新的名词，它依然落在被动与主动的两格里，补练手并没有发生。

强指导也有自己的风险，值得预先说出：脚手架可能长成拐杖。如果每一回合都是ChatSDE或老师在推，学生始终只是跟着走，则一个学期的练手练成的仍是顺着别人路径行走的技术。因此补练手的指导量应当逐周递减，而不是匀速给予：前两周，老师给出第零轮的格式和改题的示范；第3周的改题版本里，必须至少有一处改动是老师没有提示过的；到最后一次短子本，学生自己选用哪一把刀、先问哪一问，老师只在卡住处插话。递减是否真的发生，也有读数可看：把一个单元里学生的消息拆成“应老师或ChatSDE的提示而写”与“自己先开口”两类，后一类在单元末应明显多于单元初。若始终不增，指导就没有撤下，补练手便成了永久的陪跑。

◆ 吴静的“血压调节”：第一次体会“题可以改”

回到吴静。生理学老师为“血压调节”单元准备了一组练手九问，与战场九问不同，这组题的终点老师都写得出来：压力感受器反射的路径、体位变化时的血容量再分布、自主神经的快速调节与肾脏的慢速调节。老师用装有SDE工序层的大模型出题时，有意让每一问都落到学生能够触及的小现场上，比如“从躺着到站起来的头十秒里，你自己的心率怎样变化”“同一个人早晨和傍晚站起来时的感受有什么不同”“教材里的反射弧，哪一环在你身上最难被直接测到”。练手题不要求通过题闸——拿掉学生的现场，一个未接入这些材料的通用大模型照样答得出来，这本就是练手的特征；但练手题要求学生在走流程时真的去碰一下自己的现场，否则练的仍是发现技术。

吴静的第零轮最终在老师的提示下写成了三句话：“人站起来不晕，是因为血压能维持住。血压维持住，是因为压力感受器发现血压下降并让心跳加快。所以站起来头晕的人，是血压低。”这是一份很典型的发现型初判：每一句都对，连起来却把“变化”偷换成了“水平”。ChatSDE的第一回合没有纠正她的任何一个术语，而是先接住了她的链条，然后把前提翻过来问：“同样一个血压偏低的人，为什么坐着不晕，一站起来才晕？”这一问里没有出现“几秒”“时间差”一类的字眼，学生要找的那一项由她自己去找。这一问的工序选择很朴素。按第7章的判别次序，吴静此刻还没有带回任何与初判直接相撞的现场事实，所以不能算断拍；她的初判里少了一项被当作常数的因素，即“变化发生在多长时间之内”，把一个时间过程压成了一个静态数值，因此属于缺式，刀是补式——去问被漏掉的那一项。

吴静在第2轮里回答得很快：“因为站起来时血液往下走，回心血量一下子少了，反射还没来得及起作用。”这仍然是教材的语言，但她第一次把“时间差”放进了自己的判断。老师在周工作表里注意到了这一轮的推进，在课堂上只给了她一个小任务：按单元安全须知，在宿舍里用家用血压计，分别在早晨起床和傍晚自习后，记录自己由卧到立后不同时刻的读数和主观感受，连记一周。这不是医学研究，也不涉及任何人的诊疗，只是让她在自己身上找一个最小的现场。一周以后，她发现早晨那几次“眼前发黑”比傍晚明显，记录本上的读数也印证了早晨回落更深、恢复更慢。

转折发生在改题这一步。老师九问里原来有一问是“从卧位站起时，血压靠什么在几秒内恢复？”吴静在第3周提交改版时，把它改成了“为什么我早上起床猛站会眼前发黑，傍晚却不会？早上那几秒里，反射弧的哪一环慢了？”她在版本说明里写道：“我原以为题目是老师定的，只能答不能改。改了以后我发现，题还是那道题，但它现在是我的题。”用两锚来检查，这次改题根还在——仍然连着压力感受器反射与血容量再分布这一核心判断——同时离现场更近，从教科书上的“人”落到了她自己早晨的那十秒。这正是第5章所说的提升，而不是逃离。

值得对照的是，同一份第零轮若交给默认配置（未开启学习模式）下的通用大模型，在附录二所测的条件下，常见的回应会是另一种样子：它多半不会顺着吴静的三句话往下写，而是先把“血压低所以晕”这个前提翻一翻，说晕不只看血压的高低；可翻完之后，同一回合里它就把压力感受器反射的路径补全，把体位变化时的血容量再分布讲清楚，再附上几条起身时的注意事项。这样的回应没有错，对一个只想复习教材的学生甚至很有用；它的特点在于前提被翻过来的同时，答案也一并写好了，吴静本该自己去找的那一项，已经摆在她面前。补练手的课堂之所以要用ChatSDE而不是裸用的大模型，并不是因为后者“不懂生理学”，也不是因为它“不会质疑”，而是因为过渡期的学生最缺的，恰是自己先说出那一步的机会，默认值替他们把这一步做了（见第8章）。吴静的第2轮之所以出现“时间差”，正是因为第一回合只翻了前提，没有交出答案。

这次练手的终点其实并不新。早晨直立耐受较差与夜间平卧后的体液分布、交感张力的昼夜变化有关，生理学老师心里早有答案，通用大模型也能张口说出大半。但吴静得到的东西不在终点上，而在路径上：她第一次在不知道答案的时候先说了话，第一次被追问而没有被纠正，第一次把一道题改成了自己的题，也第一次从自己的现场里带回一个读数去顶住教材里的一句话。这次练手在吴静自己的记录里只留下一句判断——她“第一次体会‘题可以改’”。这句话看似朴素，却是从发现技术跨向发生技术的第一个台阶：一个不相信题可以改的人，到了战场上只会去找更难的标准答案。

◆ 16周：沈澄所在化学课的一个学期

练手走通之后，战场课的一个学期长什么样？我们用沈澄所在的化学系三年级课程型研究《合成条件的发生》来画一张完整的日程。这门课内嵌本课题组一个真实的偶联反应课题，谱系母本是课程型本科研究与合成方法学的前沿专著，加上本课题组十年的反应条件档案。之所以选它做样板，是因为化学课的节律最容易被看清：实验有批次，批次有记录，失败与成功都留在纸上，一个学期里每一周在做什么，几乎可以逐行核对。需要先申明的是，下面这张日程与其中的人物一样是构造的样板，只作说明，各校可按学时与实验安排伸缩；其中所有实验活动都在课题组既有规程与安全规范之内进行，学生只增加记录、不改动任何危险步骤。

周次	段	学生做什么	老师做什么	留下的证据
第1周	S	领取九问；母本底盘复述加反例口试	发九问，讲清题闸与改题规则；安全门槛核验	口试记录、安全签字
第2周	S→D	写第零轮初判（沈澄：“产率低是因为试剂纯度不够”）	读全部第零轮，标出每人初判坏在哪一类	第零轮文本
第3周	S→D	提交改题第一版，写版本说明；写权重理由书，问对开始前锁定	用两锚逐份回读，区分提升与逃离；审读理由书（三问）	题的版本史、权重理由书
第4—5周	D	开始问对；翻十年档案，查低产率批次的月份分布	每周先读问对记录，只在卡住处插话	问对记录、档案摘录
第6—8周	E	按规程复做一批，逐步记下时间、温度、湿度；与档案比对	实验室巡视；与课题组核对复做安排	实验记录本、谱图
第9周	D	问对第二阶段：带着复做记录回到ChatSDE；周末第一次碰撞场，组内互撞，看清各自判断的边界	读问对推进，标出关键判断谁先说；按判断清单配对	问对记录、第一次碰撞记录
第10周	E→产物	拟定子本大纲	审大纲，退回“翻译稿”	子本大纲
第11周	碰撞	第二次碰撞场：带着大纲组内互撞，必要时跨学科碰撞	组织碰撞场，记录班级判断的来源	碰撞记录、班级子本草稿
第12—14周	E→产物	重写成稿；补可复做性说明	审各稿，退回仍在翻译的段落	子本各稿
第15周	评估	提交子本，与第3周锁定的权重理由书一并送评	组织换手评分，收利益冲突声明	八项剖面

周次	段	学生做什么	老师做什么	留下的证据
第16周	回写	写期末自述，与第零轮对照；整理可进精要库的条目	采纳条目进母本精要库，交下一届	自述、采纳记录

这张表最值得注意的不是它的格子，而是格子之间的节奏。前三周看起来很“慢”：一周口试，一周第零轮，一周改题，学生连实验室都还没进。可正是这三周决定了后面13周往哪里走。沈澄在第2周写下的“试剂纯度”是一个很自然的发现型初判，它在课本里有充分的依据，在大模型那里也能被推演得头头是道；如果没有第零轮把它白纸黑字地留下来，期末时就无法看出她到底被改写了什么。第3周的改题则把九问从老师的种子变成学生的课题，后面所有的问对与复做都以改版为准。一个学期若把前三周压成一周，后面的问对往往会变成没有前提可翻的空转。

第4到第9周是问对与现场的往返，也是一学期里最难排进表格的部分。沈澄的问对从“试剂纯度”出发，ChatSDE第一回合就问：同一瓶试剂，为什么三月好、六月差？这一问把她送回档案，她去数十年里产率最低的20批各在哪个月份（九问第1问），又去查实验记录里“室温”一词有没有一次写过具体温度和湿度（第3问）。沈澄的问对由第零轮加第1—6轮共七轮组成，细节已在第6章详述，这里只强调它在日程上的含义：约一周一轮，期间穿插复做，问对不是集中在某一周做完的作业，而是贯穿六周、与实验室交替进行的来回磨。日程表把第6—8周标为“E”、第9周标为“D”，只是标出重心所在，实际上沈澄几乎每做完一步复做，都会带着记录回到问对里。

第9周末与第11周的两次碰撞场常被低估。按第14章的论证，只有一对一的线，课堂就退化为高级个别辅导；碰撞场是课堂之为课堂的地方。在这门课里，沈澄与同组同学互看各自的复做记录与问对档案，彼此追问“你那一步为什么比档案快”“你那天的湿度是多少”。如果学校开设了跨学科碰撞，她的“规程里没写的那一句”也可能与医学院许宁的“交班本里被省略的那一句”撞在一起。两次碰撞的位置是有道理的：第一次在复做之后、问对中段，学生手里已有现场证据，碰撞帮他们看清各自判断的边界；第二次在大纲初定、重写之前，碰撞出的班级判断还来得及写进子本。太早，学生只能交换观点；太晚，子本已经成形，碰撞出的判断无处落地。

第12到14周是子本的重写，第15周换手评分，第16周回写。化学课的子本有一项其他学科没有的硬要求：可复做性。沈澄的子本必须写清她复做那一批的每一步时间、温度与湿度，让下一届的同学照着做能得出可比的结果，这在评估中落到子本完成度（第⑥项），在化学学科的权重示例里占15%，与医学、经济法学、人文的10%相比高出一截（见第18章）。最后一周的回写也不是走形式：她的“规程里没写的那一句”仍是一个待复做验证的判断，所以进入精要库的不是一条结论，而是一条带着证据与失败条件的假说，交给下一届去复做（下一届如何把它写进规程，见第24章）。

◆ 老师的周工作表：每周先读学生的轨迹

九问课堂里，老师的一周与讲授课堂大不相同。讲授课堂的老师，一周的重心在备课与上课：把下一讲的内容组织好，讲清楚，留作业，改作业。九问课堂的老师，一周的重心落在“读”上：先读学生的轨迹，再决定这一周要在哪里出手。这有点像有经验的家庭医生看慢性病复诊——不是每次都重新问一遍病史，而是先翻上次以来的记录，看血糖曲线在哪一天拐了弯，再决定这一次说什么、开什么。轨迹读在前，动作才不会落空。

我们把老师一周的工作压成一张可以照着做的表。它同样是样板，各校可以调整，但其中的顺序最好不要颠倒。

时间	工作	要点
周一	读轨迹	读每位学生上周的问对记录、改题版本与现场记录，标出三类：推进的、打转的、逃离的
周二	定出手点	对打转者给一个回现场的小任务；对逃离者用两锚追问；对推进者不插话
周三	课堂	课堂时间用于碰撞、口试、现场汇报，不重讲母本
周四	出题台与精要库	依学生改题情况微调后续问；把本周值得记下的判断暂存为精要库候选
周五	现场与协调	化学进实验室、医学对接科室；核验伦理与安全门槛
机动	自己的问对	老师自己就本学期的一问与ChatSDE问对一次，保持出题手感

表里最关键的是周一与周二的先后。很多老师刚转入九问课堂时，习惯在周一就开始“讲”：把上周学生普遍不懂的地方再讲一遍。这个习惯来自发现版课堂，在练手阶段偶尔有用，在战场上却往往有害，因为学生卡住的地方各不相同，统一重讲只会把学生拉回同一条路上。先读轨迹，老师才看得出沈澄的卡点在“室温”这个词没有被量化，而她的同组同学卡在根本没去翻档案——前者需要一个追问，后者需要一个回现场的任务。读轨迹的另一个好处是，老师在读的过程中自己也在被改写：学生带回来的档案细节、病房细节，往往是老师自己也没见过的。

这张表还隐含着一次工作量的重新分配。讲授课堂里，老师最大块的时间花在备课与讲课上，而这一块在九问课堂里大幅缩小：母本的系统讲解已经可以交给大模型去做，而且做得更全、更耐心（见第2章），课堂时间不再用于重讲母本。腾出来的时间并没有消失，而是转移到读轨迹、出题台与现场协调上。一位带20名学生的老师，若每周给每份轨迹10分钟，周一就要用掉三个多小时，这是一笔实打实的新负担。所以过渡期的院系管理者需要明白，九问课堂并不省老师的时间，它只是把老师的时间从“说”挪到了“读”与“出题”；若在推行时仍按讲授课的课时核算工作量，周工作表很快就会在周一那一格上塌掉。

周二的三类处置值得再说一句。对“推进的”学生不插话，是第7章“逼近判断、不替学生说破”在老师一侧的对应：学生已经在推进，老师一开口，关键判断的首发权就可能从学生手里滑走，而“关键判断是否学生先说”恰是问对深度（第③项）的评分点之一。对“打转的”学生，最有效的往往不是再给一个解释，而是给一个回现场的小任务——像吴静那一周的晨昏血压记录，或沈澄去数20批的月份。对“逃离的”学生，用两锚追问即可：你改出来的题，根还在吗？离你的现场更近了吗？

◆ 常见失败与纠偏

过渡期的九问课堂会失败，而且失败的样子相当固定。把它们列出来，不是为了给新模式留退路，而是因为每一种失败都对应着发生技术的某一个构件没有练成，看清失败也就看清了补练手要补什么。下面五种是最常见的，我们按九问流程的顺序来写。

第一种失败是“第零轮空白”，吴静开学时的样子就是。它的根源是学生不敢在终点未知时先表态。纠偏的办法不是降低要求，而是把门槛放低到“写满三句，对错不论”，并且在第零轮之后的第一回合里，让ChatSDE先接住学生的链条再翻前提，让学生体会到“说错了不会被打分”。第零轮在评估中不计对错，只作为回写证据（第⑧项）的起点；一旦学生明白这一点，空白通常在两三周内消失。若某个学生到学期中仍然写不出第零轮，那提示的就不再是胆怯，而可能是他根本没有自己的现场，需要老师回到出题台重新看他的画像。

第二种失败是“改题成逃离”。学生把一道需要下现场的题，改成一道上网就能查到的题，例如把周川的“父亲为什么一到月底就泄气”改成“AI会不会取代饭馆”那一类大题。逃离的特征是根断了、离现场更远了，而且往往伴随着题目变“大”。纠偏用两锚即可，但老师要注意不要直接替学生改回去，而是让学生自己在版本说明里回答“我的现场在这道新题里还剩下什么”。多数逃离并非懒惰，而是学生对自己的现场缺乏信心，觉得父亲的小饭馆、病房里的交班本“不够学术”；把这一层说破，比修改题目本身更重要。

第三种失败是“问对成索答”。学生把ChatSDE当作答案机器，反复要求“直接告诉我结论”，或者干脆把同一问题换着说法问十遍。表面上问对轮次很多，实际上轮次之间没有推进。按评估方案的配套规则，轮次不计分，十轮打转不如三轮推进；纠偏的关键因此不在惩罚，而在让学生看到推进是什么样子——老师可以挑一份推进得好的匿名问对记录，在课堂上逐轮拆开，标出哪一轮学生接住了批判、哪一轮关键判断由学生先说。

第四种失败是“现场成综述”。学生交上来的现场证据其实是文献摘抄与网页整理，看不到任何一页属于他自己的原始材料。这在过渡期尤其常见，因为发现训练让学生习惯了“查到”而不是“看到”。纠偏要回到九问本身：能被综述填满的题，很可能当初就没过题闸。对战场课而言，老师此时要检查的首先是自己的出题台，其次才是学生的态度。

第五种失败是“子本成翻译”。学生把母本逐章改写一遍，换了措辞、加了几个例子，却找不出一条母本所无的判断（见第12章）。它常常出现在问对与现场都做得不错、到最后几周时间紧张的学生身上：材料很多，却不敢让材料重新组织自

己。纠偏的办法是在第10周审大纲，也就是日程表里拟定大纲的那一周，而不是等到第15周审成稿；大纲阶段若看不出新判断落在哪一节，就退回重排。与此相关的还有一种老师一侧的失败——老师在周一不读轨迹、只读成品，到了评分时才发现学生的高分子本拿不出过程，这时再追问已经太迟。

五种失败各有一个事先可定的读数，让老师不必等到期末才发现。第零轮空白的读数是第2周末仍未交出三句初判的学生，若超过全班的五分之一，要检查的是门槛是否真的放低了、ChatSDE的第一回合是否真的先接住了，而不只是督促学生。改题成逃离的读数是版本说明里“我的现场在这道新题里还剩下什么”一栏：留空或只写“无”的比例，应当随单元推进下降。问对成索答的读数是连续三轮里学生的消息中有没有出现一件新材料，如一个读数、一段档案、一句现场听来的话，三轮都没有，就是在原地打转。现场成综述的读数更简单：现场证据里能不能找出一条带日期的、学生自己取得的一手记录。子本成翻译的读数就是第10周的大纲：请学生在大纲上圈出新判断落在哪一节，圈不出，当周就退回，不要拖到成稿。

◆ 本章怎样算错

本章的主张有三处可以被经验推翻，每一处都带着读数。第一处是补练手是否有必要。取同一所大学相邻两届的新生，一届在大一补过一两个单元的练手，另一届没有，每届不少于30名学生，进入战场课后比较三项记录：第零轮在第2周内完成的比例、改题版本史中“离现场更近”的提升所占比例、问对记录里关键判断由学生先说的比例。若连续两届，补过练手的一组在三项上都看不出可辨的差别，那么“补练手”这一安排就是多余的，本章应当撤回这一主张，转而追问发生技术究竟是在哪里练成的。

第二处是补练手是否真的过渡。本章断言它的退出条件是新生的第零轮不再空白、改题出现提升而不只是逃离、关键判断更多由学生先说。如果中学普遍开展了九问练手，大学新生的这三项读数却仍停留在发现型的水平，连续三届不见改善，那就说明练手与战场之间的距离比本章估计的更大，或者中学练成的并不是可迁移的发生技术；此时关于退出条件的设计需要修改。反过来，如果三项读数早已达标而补练手单元仍年年开设，则本章要说的是，这个单元已经成了永久课程，应当被取消，而不是被保留。

第三处是日程与周工作表是否可维持。以“先读轨迹再出手”为核心的周工作表，若在一个学期内，严格按此顺序工作的老师，其班级遇到上述五种失败的次数并不少于按讲授习惯工作的老师，或者读轨迹的工作量使超过半数的老师在第6周之后流于形式，只翻成品、不读过程，那么这套日程就只是一张纸面上的表，需要按真实负荷重新设计。以上三条之外，还要申明一点：本章的吴静、沈澄与两张表格都是构造案例，只说明一种安排可以怎样运行，不能当作它已经运行成功的证据。

一个学期走完，吴静练过了一遍手，沈澄打完了一场仗，老师读了16周的轨迹，精要库里多了一条带着失败条件的假说。可是只要把视线从一门课抬到一所大学，新的问题就出现了：吴静与沈澄所用的ChatSDE，底座是一个任何人都能调用的大模型；隔壁学校、校外机构、甚至学生自己，都可以用同一个底座搭出一个看起来差不多的问对界面。那么，一所大学在这样的课堂里，到底守住了什么别人拿不走的东西？是那套出题与问对的工序，是10年、20年越积越厚的问对档案，还是另有他物？下一章就从这个问题出发，讨论大学在AI时代造的是什么。

> **带 走 的 话**
>
> **补练手是借大学的教室补中学的课，是脚手架而不是新楼层，补完就该退出；一个学期要读的不是格子而是节奏，老师先读轨迹，再决定在哪里出手。**

◆

26

第26章

大学造驾驶术：工序智能体与护城河

◆ 一张截图引出的问题

一次院系教学研讨会上，有位年轻教师把手机投到屏幕上，给大家看一张截图：某家校外培训机构上线了一个“研究型问对助手”，界面布局、第零轮的输入框、“先接住、再翻前提”的提示语，几乎与本校ChatSDE一模一样，底座用的也是同一家公司的大模型，按月收费，面向全国大学生开放。会场安静了几秒，随后有人问出了大家心里的那句话：既然别人一个周末就能搭出同样的界面，用的又是同样的模型，那我们这几年花在九问课堂上的投入，到底建起了什么别人拿不走的东西？另一位老教师接得更直接：如果答案是“什么也没有”，那大学在AI时代就只是一个昂贵的转售商，把大厂的模型包一层皮卖给学生。

这个问题必须正面回答，而且不能用情绪来回答。前面的章节已经讲清两件事：一是AI接管了发现一侧，讲清体系、写综述、按规格完成任务，大模型更快更全；二是ChatSDE与通用大模型的差别不在底座，而在默认行为与包住底座的工序层。把这两件事放在一起，就得到本章的出发点：如果模型本身是任何人都能调用的公共基础设施，那么大学能够守住的，只能是模型之外的东西。截图里那家机构抄走的，究竟是这“模型之外的东西”的全部，还是只是它最表面的一层？

本章的核心判断是：大厂造发动机，大学造驾驶术；壳可以开放，工序与档案要守住。这里的“驾驶术”指的是包住模型的整套工序——出题台怎样从母本精要库出题、题闸怎样把关、问对怎样按判断坏在哪里选刀、档案怎样被换手评分并回写进精要库。这套工序以智能体的形式运行在课堂里，我们称之为工序智能体；它是大学在AI时代的领地。须先收窄“守住工序”的含义：工序里能写成文字的那一层是壳，包括界面、第零轮的格式、“先接住再追问”的措辞，这一层不难抄；真正要守的，是细则赖以运转的学生现场材料、带发生史的精要库和越积越厚的问对

档案。本章先讲发动机与驾驶术的分工，再讲难以复制的东西，然后用一个构造推演看两所医学院五年后的差距，最后讨论教师培训怎样从讲课转向出题。

> **这一章的问题**
>
> 别人一个周末就能用同一个模型搭出同样的界面，大学在AI时代到底守住了什么别人拿不走的东西?

◆ 发动机与驾驶术：模型可换，工序跟走

先把比喻立稳。一辆车的发动机决定它能跑多快、能爬多陡的坡，这是工业实力的结晶，需要巨大的资本、算力和工程团队，个人与单个机构几乎不可能自己造。大模型就是这样的发动机：参数规模、训练数据、推理能力，都是大厂竞争的领域，大学没有必要也没有条件与之比拼。但同一台发动机装在不同的车上，交给不同的司机，开出来的结果天差地别。一位老司机知道在哪个路口该减速、在雨天哪段路最滑、在什么时候该换道，这些知识不在发动机里，而在驾驶术里。

把比喻落回课堂。第10章已经说明，批判性激发与工序选择不在参数里，而在包住模型的工序层。具体地说，ChatSDE在接到许宁“夜间跌倒是因为年纪大”这一初判时，第一回合问“同样年纪的患者，为什么白天跌得少”，这一问不是底座模型“自发想到”的，而是工序层规定的默认：先接住，再按判断坏在哪里选刀——这里坏在把一个只在夜间显著的现象归给了一个昼夜不变的变量，于是刀落在“缺式”上，去找那个没有被写进判断的条件，在许宁这里，那是交接的时间次序。换一个底座，只要它的语言与推理能力够用，这条工序照样可以跑；反过来，工序层被拿掉，在默认配置（未开启学习模式）下，同一个底座多半会先质疑“年纪大”这个前提，随即替许宁列出一份周全的老年跌倒风险因素清单（附录二的试跑记录了这种回应），那个本该由她自己从交班本里找出来的条件，就在这份清单里被提前说破了。

这就是“模型可换，工序跟走”的含义。对大学来说，这一点有极强的现实意义。大模型的迭代速度远远超过大学课程的更新周期，今天用的底座，两三年后很可能被更强的替换；如果一所大学把自己的教学能力押在某一个模型上，它就把自己的命运交给了别人的产品路线图。反之，如果大学把能力押在工序层上，底座的

每一次升级都会变成它的红利：发动机更好了，驾驶术不变，车跑得更稳更快。这也是本书始终坚持“不贬低通用大模型”的原因之一——大学与大厂不是对手，而是分工。

计算机课堂提供了另一个佐证。方舟在软件工程《真实项目研讨》里排查学校选课系统的崩溃，他的初判是“崩溃是因为服务器不够”，而ChatSDE问的是：去年服务器加了一倍，为什么还崩？方舟自己知道去年扩容这件事，它与“服务器不够”的初判直接相撞，判断却没有改，所以他的初判按判别细则属于断拍，刀是点火，让这个矛盾被看见。这一问把他引向“崩溃前十分钟最先变慢的是哪一类请求”，最终指向一个没人负责的排队环节。有意思的是，这个案例本身就是“发动机与驾驶术”的缩影：学校去年做的正是给系统换一台更大的发动机，而问题出在没有人去管那一段流程。大学若把AI时代的教学投入只理解为采购更强的模型，犯的就是与选课系统同样的错。

但这里有一个必须面对的追问：工序层难道不就是一套提示语吗？提示语写出来就能被复制，截图里那家机构不正是这么做的？这个追问有对的一半，而且对的那一半比一般想象的更宽。2025年起，主要的通用大模型产品已经把“先追问、不直接给答案”做成了可以开启的学习模式：Claude学习模式、ChatGPT Study Mode、Gemini Guided Learning（日期与出处见第8章）。在它们之前，Khanmigo（2023）已经让可汗学院的辅导助手引导而不直接解题，哈佛CS50的AI助教（Liu等 2024）也靠教学护栏让学生被追问而不被喂答案。这些产品的存在，对本章是一个双重的消息。

坏的一面是，苏格拉底式追问这一层壳并不难抄，抄起来成本很低，没有一所大学可以把“我们的助手会先追问”当作自己的护城河。好的一面是，它们恰好支持本书一直坚持的判断：差别住在包装层，而不在底座。同一个底座，装上学习模式就从“质疑前提、同回合交出答案”变成“先追问、不直接给答案”，说明默认行为是可以被包装层改写的；Bastani等（2025）在高中数学课上的现场实验也表明，护栏的有无决定了同一个模型帮助学习还是损害学习。既然包装层这么重要，大学就该把壳公开，让更多课堂用上。但也因此要问，抄走了壳的人，是否也抄走了壳下面的东西。下面几节回答的，就是这壳下面究竟还剩什么。

用一个日常的例子说明壳与工序的差别。一家老字号面馆的菜单、碗的样式、店面装修，任何人都可以照着复制，甚至连汤底的配方也可能流传出去；但真正让老店的面好吃的，是师傅在每一锅汤前根据当天骨头的品质、天气的湿度、客人的多少做出的那些微调，以及几十年里从熟客的反馈中积下来的判断。新开的店可以一夜之间做到“看起来一模一样”，却需要很多年才能长出那锅汤。截图里的机构复制的，是菜单与装修；它能否长出那锅汤，取决于它有没有下一节要讲的那些东西。

◆ 难抄的东西：判别细则的输入、精要库与问对档案

壳既然好抄，大学要守的便只能是壳下面的东西。本书把它们收窄为三样：工序判别细则在具体学生身上的使用，谱系母本精要库，以及越积越厚的问对档案。需要提前说明“难抄”二字的分寸：这里说的是难以在短时间内、在没有大学共同体的条件下抄走，而不是原理上不可能。一家与附属医院、课题组签了合同，能够长期读到学生现场材料的机构，原则上也可以长出自己的档案。本章的主张是一个关于时间和制度条件的主张，不是关于天然壁垒的主张。

第一样是判别细则的使用。细则指第7章给出的次序：判断一条初判坏在哪里，先查学生在自己的现场里是否已经接触到与初判直接相撞的事实、判断却没有改，是则为断拍，刀是点火；若无此事实，再查初判的关系里是否少了一项被当作常数的因素，是则为缺式，刀是补式；关系齐全、也无矛盾，才看起手次序与任务是否不配，是则为偏路，刀是排路。这一段文字完全可以抄，本书就印在这里。难抄的是它的输入。以书中四个案例为例，周川说“因为累”，方舟说“服务器不够”，两人都是断拍：周川月初同样忙碌却不泄气，这件事他天天在店里看着；方舟知道去年扩容一倍系统仍然崩溃。许宁说“因为年纪大”，沈澄说“试剂纯度不够”，两人都是缺式：他们当时并没有碰到与初判相撞的事实，被漏掉的是交接的时间次序，是环境湿度与操作时长。四句话字面上都同样朴素，归哪一类要看学生的现场里有什么、他看到了什么，而这些材料是学生的交班本、实验记录、门店流水和机房日志。细则的第一步核对就得读到它们。校外的通用助手读不到，只能按字面猜；猜错了刀，就会把本该点火的断拍当缺式去补，学生被追问了一圈，始终没有看见那个早就摆在眼前的矛盾。

第二样是谱系母本精要库。它由学科前沿专著与课题组档案提炼而成，一次建成、千人共用、逐年回写。精要库不是一座知识库，而是一座带发生史的判断库：每一条判断都挂着提出它的子本、检验它的复做、撤回它的记录。出题台出题时依靠它，才能让九问通过题闸，而不是回到“老年人跌倒有哪些危险因素”这类张口就能答的题。一个没有精要库的复制品，每次出题都要从通用知识出发，题闸的打回率只会居高不下。

第三样难复制的东西，是越积越厚的问对档案。第3章讲过，谱系母本由学科前沿专著与课题组（部落）档案合成，比例逐年向部落档案倾斜；第24章讲过，回写使课程成为一条会长的谱系。问对档案正是部落档案里最有生命力的一部分：它记录了一届又一届学生从第零轮到定稿的全部来路——哪些初判最常见、哪一把刀最常奏效、哪些问在哪一年被改题提升、哪些判断最后被下一届复做证实或推翻。这份档案每多一届就厚一层，而且它只能在真实的课堂里、在真实的学生与现场之间积起来，无法用钱一次买到。

问对档案的价值，可以用沈澄所在课题组的例子看清。十年反应条件档案本来是一堆实验记录，其中“室温”一词从未写过具体温度与湿度；沈澄一届的问对与复做，把这一空白变成了一条待验证的假说，交进精要库；下一届复做之后，它被写进了课题组规程。这时候，精要库里多出来的不只是“湿度要控制”这条结论，还有一整段发生史：哪一问打开了它、哪一轮问对推进了它、哪一次复做验证了它。下一届老师在出题台上出题时，能直接看到这段历史，于是可以把下一届学生的九问推向更远的地方。一个从零起步的机构，即便拿到同样的结论，也拿不到这段历史。

还有一层不那么显眼的价值：问对档案是工序本身的训练场。前面说工序层中有一部分是“路感”，无法写成固定文字；这种路感的来源，正是档案。老师在周一读轨迹、在出题台上调整九问、在评分时比较不同学生的推进，都是在借档案磨自己的判断；而工序智能体的工序选择规则，也要靠档案中积累的成功与失败案例来校准。所以工序与档案是互相喂养的：工序产生档案，档案改进工序。这一循环一旦转起来，后来者就很难追上，因为追赶者面对的不是一个静止的目标，而是一个每年都在变厚的东西。

需要立刻补上一个限定，免得这一节被读成“档案越多越好”。档案的价值不在数量，而在它是否被观察过、被回写过。一个从不被读、从不进入精要库、从不影响下一届出题的档案仓库，只是一堆沉睡的记录，和教务系统里成千上万份无人翻阅的旧作业没有区别。第24章强调“回写要被观察，不能被宣布”，在这里同样适用：一所大学若只宣布“我们积累了十万份问对记录”，却说不出其中哪几条改写了精要库、哪几问因此被重出，那它守住的不是护城河，而只是一个装满水的空坑。

◆ 构造推演：两所医学院，同一个大模型，五年之后

为了把“驾驶术”与“发动机”的分工说得更具体，下面做一个构造推演。必须先申明：这一节不是调查，不是任何真实院校的经历，也没有任何统计数据作依据；它是按本书的概念与机制，从同一起点出发推出两种可能的走向，只用来展示护城河是怎样一点点形成或者没有形成的。读者应当把它当作一个思想实验，而不是一份报告。凡是下文出现的“第几年”“哪一门课”，都是推演中的设定。

设想两所规模相近的医学院，我们称之为甲院与乙院。第一年，两院同时采购了同一家公司的同一个大模型，接入同样的校内平台，向临床医学高年级学生开放。乙院的做法是“裸用加培训”：给学生开账号，办几场如何写好提示语的讲座，鼓励学生在病例讨论、文献检索、病历书写练习中多用模型。甲院的做法是“装工序”：在老年病科轮转的《临床推理与病房研究》研讨课上先做一门试点，老师用装有SDE工序层的大模型从老年医学（跌倒、衰弱）前沿专著与本科室轮转档案出九问，学生通过装有工序层的ChatSDE问对，子本经换手评分，可采纳的判断回写进科室的母本精要库。许宁就在甲院的这一届里。

第一年末，两院的表面差别并不大，甚至乙院看起来更“先进”：它的使用覆盖面更广，学生交上来的病例分析报告更长、更规范，引用的指南更全。甲院只有一门课在试点，许宁那一届交出的子本数量不多，每一份都带着一个仅限本病房、待检验的判断，比如她那条以“交接断点”为关键词的假说。按第11章的思路，若要比较两院学生的“知识发生”，第一年还看不出可靠的区分；任何人在这一年宣布哪一边胜出，都是把一个未被检验的前提当成了结论。

推演的分岔在第二、三年出现。乙院的模型在第二年升级了一次，学生的报告随之变得更流畅；但由于学生各自使用、各自提问，问对记录分散在个人账号里，从未被收集、评分或回写，科室的教学内容与第一年相比没有什么变化。甲院的模型也同步升级，但这次升级被吸收进了已有的工序链：许宁那一届以待检假说的身份回写进精要库的“交接断点”假说，成了下一届九问的出发点，下一届学生在“每次交接一次风险评估”这一问上走得更远，并按科室的验证程序把部分判断交给科室去检验。到第三年，甲院的老年病科精要库里已经有了三届的发生史，出题台出题时不必再从“老年人跌倒有哪些危险因素”这种过不了题闸的题出发。

到第五年，推演中的差距可以写成三个方面。第一，在精要库上：甲院几个试点科室各有一份五届厚度的母本精要库，记着每一条判断是哪一届、哪一问、哪一次验证留下来的；乙院有五年的使用日志，却没有一份可以交给下一届的精要。第二，在毕业生上：甲院学生带走的是一套已经进入个人底盘的子本与问对习惯，许宁从实习到住院医，子本长成了她交接班时的思考底盘——它是思考的底盘，而不是临床决策工具；乙院学生带走的是熟练的提示语技巧，这本身有价值，但它同样可以在任何一家培训机构学到。第三，在教师上：甲院有一批经过五年出题台磨炼的老师，他们的判断已经与档案互相喂养；乙院的老师仍然主要在讲课，只是讲课时多了一个助手。

推演中还有一个细节容易被忽略：乙院并没有做错什么。设想它在第二年为全体学生开启了通用产品的学习模式，学生被追问而不被喂答案，这当然是进步，Bastani等（2025）的实验也提示护栏有实在的价值；但这一步补上的只是壳，题闸、第零轮基线、按坏法选刀和回写，仍然无处安放。它的学生确实更熟练地使用了大模型，病例分析报告的规范程度确实提高了，这些都是发现一侧的真实收益，本书从不否认。乙院的问题不在于“用了模型”，而在于它把模型当作了全部：既然发动机已经买来，似乎就不需要再造驾驶术。这有点像一家运输公司买进了最好的卡车，却从不整理司机们在各条线路上积下的路况经验，也不让老司机带新司机；车越换越新，公司的运输能力却停在原地，因为能力本来就不全在车上。

这个推演最想说明的一点是：两院之间的差距与模型无关。它们五年里用的始终是同一家公司的同一系列模型，每一次升级两院同时获得；如果五年后两院之间

出现了差距，那么差距只能来自模型之外——来自工序是否被装上、档案是否被积累、回写是否被观察。这正是“大厂造发动机，大学造驾驶术”在机构层面的意思。反过来，推演也提醒我们它可能怎样失败：若甲院的工序链在某一环断掉，比如换手评分流于形式、精要库无人维护、出题台退化为每年照抄去年的九问，那么五年后的甲院与乙院可能并无二致，甚至因为多了一层繁琐的流程而更差。护城河不是装上工序的那一天建成的，而是在每一年的回写里被一点点挖深的。

◆ 壳为什么可以开放：开放的壳与守住的工序

回到研讨会上那张截图。如果壳可以被复制，大学是否应当像商业公司那样把它藏起来，申请专利，限制使用？本章的主张恰恰相反：壳可以开放，而且应当开放。所谓壳，指的是问对界面、第零轮的格式、“先接住再翻前提”这类可以写成文字的默认规则、八项评估的表格、权重示例。苏格拉底式追问这一层，三家主流产品的学习模式已经做到了，本来也守不住。把这些东西一并公开，其他大学、中学、甚至校外机构都能用，课堂的发生技术会因此得到更广的练习，中学的练手也更容易跟上——这对大学自身是有利的，因为它缩短了第25章所说的过渡期。

开放还有一个来自证据的理由。Kestin等（2025）在哈佛物理课上做的随机对照显示，精心设计的AI导师带来的学习收益，高于课堂里的主动学习。这样的结果说明，设计得好的壳确实能带来学习上的实收益，把它锁在一所学校里，等于让别的学生白白放弃这份收益。但这个结果同时划出了边界：它测的是学生掌握一门已有内容的收益，也就是本书所说的发现一侧与练手一侧的收益，并没有测学生能否在连老师也写不出终点的问题上发生新判断。壳开放，是因为它擅长的事该让更多人受益；壳之外的事，不能指望壳去完成。

开放的壳还有另一层意义：它让工序层的主张可以被检验。第11章讲过，“ChatSDE与通用大模型不同”是一个产品主张，必须用同底座对照来检验，否则它本身就是一个未被批判的前提；而同底座的多个智能体不能互作独立证据，评分必须换手。一所大学若把自己的工序层完全封闭，外界就无从检验它到底是否多出了可检验的区分。公开壳，等于把自己的主张交给他人复做，这与沈澄把复做条件写进子本的道理相同：能被别人复做的东西，才算真正属于你。

需要守住的是三样东西：一是判别细则赖以运转的输入，即与具体课程、具体课题组、具体病房深度绑定的学生现场材料；二是带发生史的精要库；三是越积越厚的问对档案。守住不等于锁起来不让人看，而是指它们的生长权与使用权应当留在产生它们的共同体里。部落档案里有学生的现场材料，有病房的交班记录、实验室的失败批次，这些涉及隐私、伦理与安全，本来就不能随意外流；精要库里的判断带着发生史，脱离了这段历史就只剩一条孤零零的结论，即便被拿走也难以生长。所以“守住”更多是一种制度安排：档案的存储、去标识化、访问权限、回写程序，都由大学与它的合作科室、课题组共同管理。

开放与守住之间还有一条操作上的界线，可以用法学课的例子说明。韩正所在的《司法实践研讨》以本地基层法院公开的裁判文书与旁听记录为现场，他那一届得出的判断是“调解进程被物业的账单周期推着走”。这门课的九问格式、题闸规则、评估表，可以原样公开给兄弟院校的法学院使用；但本地判例库的整理方式、历届学生旁听记录中的细节、哪一问在哪一届被改题提升、哪一条判断成为下一届的九问（见第24章），则属于这门课自己的档案。兄弟院校拿到壳，可以在自己所在地的法院里长出自己的档案，这正是开放的好处；它们拿不走的，是韩正那一届以及之后各届在本地现场里积下的发生史，而那本来也不是能被“拿走”的东西。

这样看来，截图里那家机构对大学构成的挑战，并不在它“抄走了什么”，而在它提醒大学：如果你的投入只停留在壳上，那么你确实什么也没有守住。一所大学若只是给学生开了ChatSDE的账号，却没有出题台、没有题闸、没有精要库、没有回写，那它与那家机构并无本质区别，甚至因为价格更高、更新更慢而处于劣势。护城河不是壳的厚度，而是壳底下那条流动的、每年都在加深的工序与档案之河。

◆ 新的教师培训：从讲课到出题

工序智能体与问对档案都需要人来运转，这就把问题引到教师身上。第4章已经说明，九问课堂里老师的核心工作变为用装有SDE工序层的大模型打造九问，老师不可替代的地方在于认得学生的现场。可是今天大学教师的培训，几乎全部围绕讲课展开：如何组织一堂课、如何使用多媒体、如何设计考题、如何提高学生评教分

数。如果驾驶术是大学的领地，那么驾驶术的教练——会出题、会读轨迹、会守题闸的老师——就是这块领地上最稀缺的人，而现有的培训体系几乎不生产这种人。

新的教师培训应当从三件事入手，它们都可以从本书前面的章节直接读出。第一是出题台训练：给老师一份本学科的谱系母本精要，让他按两步出题法出九问，再用题闸逐问检验——拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案？能给出即打回重出。许宁的九问与“老年人跌倒有哪些危险因素”的对照可以作为第一课的范例。这种训练不教老师怎么“讲”，而是教老师怎样识别一道题是练手还是战场，就像驾校教练的培训不在于教他自己开车，而在于教他看出学员哪一个动作会在路上出事。

第二是读轨迹训练：拿一组匿名的问对记录、改题版本史与现场记录，让老师标出推进、打转与逃离，并为每一类写出下一周的出手点（见第25章的周工作表）。这项训练最难之处在于让老师学会“不出手”：对已经在推进的学生，老师一开口就可能夺走关键判断的首发权。讲授课堂出身的教师往往对此极不习惯，因为他们的职业训练告诉他们，好老师就是把话说清楚的人。读轨迹训练要培养的是另一种职业直觉——知道什么时候不说，比知道说什么更难。

第三是工序选择训练：让老师自己与ChatSDE就一个本学科的判断问对，在问对中有意识地看清工序层为什么在这一回合选了补式而不是排路、为什么在那一回合选择沉默。第7章讲过，ChatSDE要亮出选择依据，让学生学会自己选刀；老师若自己看不懂刀是怎样选的，就无法在学生卡住时判断问题出在学生、出在题，还是出在工序层的配置上。这项训练也给老师提供了一个检验工序层的位置：老师是工序最内行的使用者，若他们在大量问对中发现工序选择趋于套路化，那正是第11章所说的失败信号之一。

教师培训的转向还牵涉到教师评价制度。如果一所大学仍然主要以课时量、学生评教分数和讲课比赛来衡量教师，那么愿意花三个小时读轨迹、花一个下午与科室协调现场、反复打回自己出的九问的老师，在现行的考核表上几乎得不到任何体现。驾驶术的教练若在制度上不被看见，大学就会不知不觉地退回到只买发动机的状态。所以与新的培训相配套，院系需要把出题台的质量、轨迹阅读的记录、精要

库的维护与回写证据纳入教学工作量与评价之中；这些证据在九问流程里本来就会自然留下，不需要另造一套表格，只需要被管理者读到。

这三项训练合起来，就是“从讲课到出题”的转变。它不意味着讲课能力作废：母本底盘仍需要老师确认，口试与反例仍需要老师设计，有些母本的核心判断仍需要老师亲口讲清。但教师的职业重心挪动了，从“把已知的东西讲清楚”挪到“把未知的方向问出来、把学生的轨迹读明白”。这与第2章的判断一致：讲台的价值落到了出题台。对大学而言，培养这样一支教师队伍，本身就是护城河的一部分，因为它同样只能在真实的课堂与档案中养成，无法一夜之间从外面买来。

◆ 本章怎样算错

本章押在三处，各有读数。第一处是壳与工序的区分是否成立。取同一个底座，设三个对照：只复制了壳的系统，开启了通用产品学习模式的系统，装有完整工序并积累了多届档案的系统，对同一批学生各自问对，每组不少于30名，连续两个学期，比较代答率、核心判断首发率、子本创新智商三项读数（对照设计见第11章；首回合质疑率在附录二的预试里已近天花板，第11章把它降为操纵检验）。若第三组与前两组在三项上的差距都不超过评分者之间的一致程度，也就是换一批评分者就可能翻转，那么“工序与档案是护城河”就不成立，大学在AI时代的领地需要另找。

第二处是细则的输入是否真的难抄。把同一批不少于40条的初判，连同判别细则的文本，分别交给两组评阅者：一组能读到学生的现场材料，另一组只拿到初判的文字。若两组按细则标注的断拍、缺式、偏路一致率没有可辨的差别，说明细则的判别并不依赖现场材料，所谓难抄就只是一句宣布，“输入难抄”的论证须撤回，护城河要退回到精要库与档案一侧去论证。

第三处是底座是否把工序内化。若底座模型在后续迭代中把批判性激发与按坏法选刀内化为默认行为，使不装工序层的通用大模型在默认配置下也能稳定地先接住、只翻前提而不把答案一并交出、并按判断坏在哪里选刀，那么“模型是发动机，工序是驾驶术”的分工就会松动；此时价值将主要落在档案与制度一侧，本章关于“工序跟走”的论证需要重写。此外，若有若干所长期实践九问模式的院校，

多年之后既说不出精要库里哪些条目来自哪一届的回写，也看不到出题台因档案而改变的痕迹，就说明档案并不会自然变厚，“越积越厚”只是宣布而不是观察。还应申明，本章两所医学院的五年推演是构造推演，任何时候都不能被引作两种做法优劣的证据。

讲到这里，本书的九问模式已经从课堂一直讲到了大学的领地：母本与九问立骨，改题与问对走来路，子本与碰撞铺土壤，八项评估与回写把一切收拢成一条会长的谱系，而工序与档案构成大学在AI时代守得住的东西。但还有一个问题一直没有正面回答：本书自始至终只讲了S→D→E这一条路，先立母本、再问对、后铺现场。可许宁与沈澄日后遇到的问题，未必都从母本出发；有些问题从一个要被推翻的靶子开始，有些从一片陌生的田野开始，有些从一处卡住的地方开始。当他们离开课堂，面对一个没有老师出题的新问题时，他们会走哪一条路？结语就从这里转向另外五条路。

> **带走的话**
>
> **壳可以开放，守住的是壳底下的东西：工序所依赖的学生现场、带着发生史的精要库，和一年比一年厚的问对档案。**

◆

结语

另外五条路

◆ 没有九问的那一天

毕业前的最后一个月，许宁已经结束老年病科的轮转，在另一个科室做最后一段轮转。一个夜班结束后，她在交班本上写完最后一行，忽然意识到一件事：在九问课上，她每一次真正“想出点东西”，起点都是一份九问——老师从母本精要库出题，她改题，问对，下现场，重写成子本。可是从下个月起，不会再有人给她出九问了。病房里冒出的问题不会排好队、编好号，也不会先附上一份谱系母本；它可能从一位上级医师一句笃定的话开始，可能从一个她说不清哪里不对劲的病人开始，也可能从一处怎么也推不动的卡点开始。她在问对界面里敲下一句话：“如果没有九问，我该从哪里开始？”

同一个月，沈澄在化学系的实验室里整理自己的子本附录。她复做那一批的时间、温度与湿度都已写清，下一届的同学照着做就能得出可比的结果；可她也清楚，自己将要进入的那个课题组，不会有人把十年档案整理成精要库再交到她手上，更不会有人替她写好九问。到了新的实验台前，她面对的可能是一篇与自己结果相反的论文，也可能是一台怎么调都出不来信号的仪器。两个人在两栋楼里，几乎同时碰到了同一道门槛：九问模式把他们带到了这里，接下来的路要自己选。

这个问题是整本书留给自己的问题。本书从导论开始就申明，发生有六条路径，而本书只讲了S→D→E一条，即九问模式：先立骨（母本与九问），再走来路（改题与问对），后铺土壤（打造子本、进入底盘），最后评估与回写。之所以只讲这一条，是因为它的发现版就是讲授体系，覆盖面最广，大学最熟悉，从它入手改造的阻力最小、受益面最大。但只讲一条，并不等于大学只需要一条。一个只会走S→D→E的人，离开课堂之后会发现，自己最擅长的那条路，恰恰需要一个别人替他准备好的母本与九问。

所以结语要做两件事。一是预告其余五条发生路径——靶子S→E→D、田野E→S→D、论证E→D→S、跟随D→S→E、卡点D→E→S——各用一个学科的例子画出它的轮廓，作为“课堂发生学”后续各卷的入口。二是回到本书真正想交付的那样东西：不只是一本子本，而是面对一个新问题时，先问“这个判断可能坏在哪一处”、并且能自己选工序、选路径的那双眼睛。许宁那句“没有九问，我该从哪里开始”，正需要这双眼睛来回答。

> **这一章的问题**
>
> 没有人再给你出九问的时候，面对一个新问题，你从哪里开始、走哪一条路？

◆ 六路径是主导次序，不是全部方法

在预告之前，必须先把一个限定说在前面，否则五条路很容易被读成一张六格的分类表。本书所说的六路径，是知识发生中S、D、E三者谁先显露、谁居中、谁收尾的主导分析次序，它帮助我们看清一门课、一个问题大体上从哪里起步、往哪里走；它不穷尽课堂里的全部方法，也不意味着任何一次真实的发生只沿一条路线单向前进。许宁的九问课堂整体上是S→D→E，但她在第2问的提升、在夜班现场的那几个小时里，S、D、E之间的来回远比一个箭头复杂。路径说的是主导，不是全部。

还要记得导论里那个最基本的对称：发现有六路径，发生也有六路径。课堂典范革命不是把六条路换成另外六条，也不是在六条之外再添一条，而是让每一条路的方向从“去找一个先在的终点”转向“生成并回写”。所以下面预告的五条路，每一条都有一个大学早已熟悉的发现版——验证型实验、资料整理、案例分析、模仿、排除故障——它们都有真实的功绩，在AI接管发现一侧之后也都不会作废，而是降为底盘。预告要说清的，是每一条路的发生版与它的发现版差在哪一步。

用一个日常比喻说明。去一座陌生的城市，有人先买一张地图、定好景点再出门，有人先在街上闲逛、走着走着才知道自己喜欢什么，有人则是因为某件事办不成——一份文件盖不上章——不得不在城里四处奔走，结果反倒摸清了整座城的脉络。三种人都会看地图、都会逛街、都会碰壁，区别只在于哪一样在先、哪一样带

着整个行程走。六路径描述的就是这种“谁带着走”的次序，它是一种看问题的透镜，而不是一份必须逐格填满的表格。

六路径与本书反复使用的三领域区分也有一种松散的对应，但同样只是倾向，不是规定。自然科学侧重知识发生的客体属性，工程与实验学科里靶子与卡点更常见，因为装置会以不服从的方式说话；社会科学侧重互动属性，论证路径更常见，因为交易与裁判的材料天然带着争议；人文侧重主体属性，田野与跟随更常见，因为史料与语言要求人先沉浸、先揣摩；医学横跨三者，所以一位医生在一个班次里就可能走过好几条路。把这种倾向当作规定，就会犯与把学生贴上类型同样的错误——那恰恰是本书一再说明大模型做得更快、却不该成为课堂目标的事。

还要说明的是，第7章讲过的工序选择——123原理查断拍、三大方程查缺式、六路径查偏路——是在一条路径内部的判断坏在哪里选刀；而这里说的选路径，是在一个问题刚出现时决定从哪一端起步。两者用的是同一套眼光，只是尺度不同：前者像在行车中判断这一段该减速还是换道，后者像在出发前判断该走高速、走国道还是先下车问路。123原理在这里仍然只指那一个循环——D与E的矛盾逼出S的改变，S再回写D与E——无论走哪一条路，判断是否真的发生，都要看这个循环有没有被走通，有没有留下可被观察的回写。

◆ 靶子与田野：S→E→D与E→S→D

第一条是靶子路径，S→E→D。它的起点仍然是一个已经显露的判断，但这个判断不是用来出九问的母本，而是被立成一个靶子，直接拿到现场去打；打中之处的偏差，再逼出一个新的差异方向。土木工程的高远可以说明这条路的形状。设想他在毕业后的工作里面对一座老桥，设计文件里写着一条看似不容置疑的假定：桥面在通行荷载之后的振动会按设计预期较快衰减。靶子路径的第一步不是围绕这条假定出九个问，而是带着它直奔监测数据，看哪一段桥面不服从它——这与他在《结构健康监测研讨》里那一问“哪一段桥面的振动在早高峰后迟迟不回落”有相同的锋刃。打中之后浮出来的差异，比如一处被忽略的伸缩缝维修记录（本书中仍是待检的判断），就成了新的方向。

靶子路径的发现版，是我们熟悉的“验证型实验”：拿一个理论去现场核对，对上了就记一笔“验证成功”。它的发生版与之不同的地方在于，靶子立起来是为了让它“坏”在某一处，而不是为了让它对上；现场不是为理论盖章的地方，而是让理论露出缺口的地方。这条路在工程、在需要检验已有规范的学科里最常见，它要求的发生技术也最接近本书第11章的对照实验设计：先写好失败条件，再去现场。课堂若按这条路组织，出题台出的就不是九问，而是一组精心挑选的靶子。

第二条是田野路径，E→S→D。它从现场开始，起步时手里没有母本，也没有成形的问题，只有沉浸：先在材料与人群里待足够长的时间，让某种结构自己显露出来，然后才从这个结构里分出问题。历史学的苏文最适合说明这条路。设想他回到家乡做另一项研究，起步时不带任何九问，只是一次次去听村里老人讲过去的事，把口述一段段记下来。过了一段时间，他才注意到老人们的讲述在某几个年份上反复出现同样的空白与含糊，一种结构开始显露；直到这时，他才能提出一个真正属于这片田野的问题。他在《地方史料研讨》里发现族谱断代处恰是户籍登记口径改变之年，那是九问课堂的产物；田野路径则会让类似的判断从口述里先浮出来，再去县志与族谱里寻找对证。

田野路径的风险也最明显：沉浸可能变成漫游，记了一年仍然什么也没有显露。它对发生技术的要求因此偏向另一端——不是改题，而是“看见”：在材料里分辨哪一种重复不是偶然。人类学与历史学长期依赖这条路，它的发现版是“收集资料、整理归类”，发生版则要求显露出来的结构能逼出一个新的差异，而不只是一份更完整的资料汇编。课堂若按这条路组织，老师不再从母本出题，而是陪学生在田野里等待第一个结构显露，并在它显露时及时问一句：这里为什么偏偏是空白？

◆ 论证、跟随与卡点：E→D→S、D→S→E与D→E→S

第三条是论证路径，E→D→S。它同样从现场的材料出发，但与田野路径不同，材料一到手，就立刻被组织成一个有方向的论证——先定下“要证明或反驳什么”，再让一个新的结构在论证的终点显形。法学的韩正可以说明这条路。设想他毕业后进入一家律师事务所，手里是一批同类纠纷的裁判文书与庭审笔录。论证路径不先去读学说体系，也不先长时间沉浸，而是在材料中尽快找到一个可争的差异

方向，比如“同类案件的调解成败，与当事人一方的账单周期是否相关”，然后围绕这一方向把材料排成论证；论证走到终点时，才可能显露出一个此前没有被说出的结构。他在《司法实践研讨》里那条“调解进程被物业的账单周期推着走”的判断，正是这种结构的雏形。

论证路径的发现版是“案例分析”：拿既定的法条与学说去套案件。它的发生版要求方向由材料逼出来，而不是由学说预先给定；终点的结构必须是材料之前没有、学说里也找不到的那一句。这条路在法学、在需要从证据走到主张的社会科学里最常见，它的风险在于方向定得太早，把材料修剪成论证需要的样子。所以它对发生技术的要求，是在论证的每一步都能听见材料的反对声——这与问对中“接住批判”是同一种功夫，只是批判者从ChatSDE换成了材料本身。

第四条是跟随路径，D→S→E。它的起点是一个方向、一个想做成的东西，接着跟随一个已有的结构——一位大师的作品、一种成熟的方法——走一段，最后才下到自己的现场。文学的叶青适合说明这条路。设想她毕业后决定写一部关于外婆的长篇，方向是先有的；她先去跟随一位她敬重的作家，逐段揣摩那部作品怎样处理亲人、怎样处理方言，把对方的结构当作扶手；走到一定程度，她才回到自己的草稿与方言里，发现扶手在某处撑不住了——那位作家的方言写法放不进外婆的那个词。她在《写作与阅读的发生》里得出的“不敢写的词才是叙述的承重词”，在跟随路径上会以另一种方式浮出：恰恰是在跟随失效的地方，她自己的东西开始发生。

跟随路径的发现版是“模仿”，它是一切技艺传承的古老方式，并不低级；它的发生版要求跟随最终把人带回自己的现场，并在那里与所跟随的结构分开。危险在于跟随得太久，把别人的结构当成了自己的终点。这条路在艺术、写作、临床技能乃至实验技术的传承里都很常见；沈澄所在课题组里师兄师姐口头说的“那一步要快”，就是一个被跟随了十年的结构，直到有人回到复做现场，它才被重新看见。

第五条是卡点路径，D→E→S。它从一个明确的目标出发，却在现场被卡住，在卡点上反复碰撞之后，一个新的结构才显露出来。计算机的方舟可以说明这条路。设想他在毕业后接手一个新系统，目标很清楚——它必须在高峰时段不崩溃；

可每一次扩容、每一次优化，系统都在同一个时刻卡住。卡点路径不先读架构文档，也不先跟随某种成熟的排障方法，而是死死盯住那个卡住的时刻，追问“最先变慢的是什么”，就像他在《真实项目研讨》里追到那个没人负责的排队环节一样。结构是在卡点上被逼出来的，而不是事先画好的。

卡点路径的发现版是“排除故障”：照着检查清单逐项排查，找到坏掉的零件换掉。它的发生版要求卡点逼出的不只是一个零件，而是一个此前没有人看见的结构，比如一段在组织分工里落空的流程。这条路在工程、在临床、在一切“非做成不可”的实践里最常见，也最接近许宁毕业后在病房里将会遇到的日常。它对发生技术的要求是耐受卡住：不急于绕开，也不急于宣布原因，而是让卡点本身成为现场。

五条路若各自展开成课堂，样子会与九问课堂相当不同。靶子路径的课堂，老师在出题台上出的是一组靶子与失败条件，学生的第零轮写的是“我预计它会坏在哪里”；田野路径的课堂，学期的前半段几乎没有题，老师的周工作表里读的是田野笔记而不是问对记录；论证路径的课堂，碰撞场会变成辩论场，同学之间互为材料的反对声；跟随路径的课堂，母本换成一位大师的作品或一套成熟的方法，评估要看学生在哪里与它分开；卡点路径的课堂，起点是一个必须做成的任务，现场证据的权重会显著提高。这些课堂形态，以及它们各自的评估权重与“本章怎样算错”，是“课堂发生学”后续各卷要处理的问题；本书只把入口指出来，不替它们写完。

◆ 能自己选工序、选路径的那双眼睛

五条路讲完，回到许宁的问题：没有九问，我该从哪里开始？现在可以回答了：从判断这道题应当从哪里开始。面对上级医师一句笃定的话，那是一个靶子，可以走S→E→D；面对一个说不清哪里不对劲的病人，那是一片需要沉浸的田野，可以走E→S→D；面对一处怎么也推不动的流程，那是一个卡点，可以走D→E→S。九问模式没有教会她全部的路，但它教会了她一样走任何一条路都用得上的东西：面对一个判断，先问它可能坏在哪一处。

这双眼睛不是天生的，也不是读一本方法论就能长出来的。它是在第零轮里被翻过来的前提、在改题时被追问的两锚、在问对中被选的那一把刀、在现场被一个读数顶回来的结论里，一次次练出来的。第7章讲过，ChatSDE要亮出选择依据，让学生学会自己选刀；本书在这里把这句话推到尽头：一门课的最终产出，不是学生学会了在ChatSDE的帮助下选刀，而是学生在没有ChatSDE的时候也能选刀，而且不仅能在一条路内选刀，还能在路口选路。这就像学开车，最终目的不是会在教练旁边开，而是一个人上路时，知道雨夜该走哪一条、堵车时该绕哪一段。

这双眼睛还有一个容易被误解的地方：它不是怀疑一切的眼睛。一个对每句话都先说“不对”的人，与一个对每句话都先说“对”的人，其实同样省掉了判断——前者把批判做成了套路，后者把展开做成了套路。本书第7章要求先接住、放大，再批判，并且破的同时指出立的方向，在这里同样适用。许宁面对上级医师那句笃定的话，第一步不是反驳，而是先弄清它在哪些病人身上确实成立，再问它会在哪一处坏掉；沈澄面对那篇与自己结果相反的论文，也应先复做它，再追问两边的差异落在哪一行规程上。会选刀的人，也知道什么时候不需要出刀。

所以本书说大学最终教会学生的不只是一本子本。子本是那双眼睛在一门课里留下的痕迹，是它在一个具体母本、一个具体现场上一次完整发生的记录；它会进入个人底盘，在毕业后继续生长（见第13章）。但比子本更重要的是那双眼睛本身——面对一个新问题，先问“这个判断可能坏在哪一处”，然后自己选工序、选路径。一个能做到这一点的人，离开了出题台，依然会在自己的现场里给自己出题；离开了ChatSDE，依然会把一个流畅的结论翻过来看一看。

这双眼睛也会看错，而且看错的样子值得预先说出来。最常见的是把一个卡点当成一个母本问题：系统明明在某一时刻卡住，却先去读完全部架构文档、列出九个问题，结果在纸面上推得很完整，卡点却纹丝不动。另一种是把一片田野当成一个靶子：带着一个现成的判断冲进口述材料，只听见支持或反对它的声音，真正在显露的结构反而被漏掉。判断自己选错了路，凭的仍然是123原理那个循环：如果走了一段之后，D与E之间的矛盾始终没有被逼出来，S没有发生任何改变，那就该回到路口重新选。会选路的人不是从不选错的人，而是能较早发现自己选错的人。

这双眼睛与AI的关系也需要最后说清。它不是对大模型的戒备，也不是要学生少用大模型；恰恰相反，一个有这双眼睛的人，最能把大模型用到极致，因为他知道什么时候让它把答案一并写好，什么时候要它只追问、不说破，什么时候关掉它回到现场。第8章所说的“默认替学生说破”，在这双眼睛长成之后就不再要紧：一个人学会了自己翻前提，也就学会了在模型把答案递过来之前，先写下自己的那一句。大厂造的发动机会越来越强，大学造的驾驶术最终要落在每一个毕业生身上。

◆ 两个人的最后一问

毕业前，按课程惯例，每一位修过九问课的学生都要在个人子本的最后一页写下“最后一问”：一个自己尚答不出、准备带出校门的问题。它不计分，也不交给ChatSDE，只作为回写证据的最后一行留在档案里，供若干年后的自己回来对照。这一页是构造的，正如许宁与沈澄是构造的人物；但它所指向的东西是真实的：一个人在离开出题台时，是否已经能给自己出题。

许宁的最后一问是：“明天早上我交班时，被我省略的那一句会是什么？”这一问看上去只是她那条判断的延续——风险落在“交接断点”上，藏在交班本里被省略的那一句——但它的方向翻转了。在老年病科轮转时，她追问的是别人省略了什么；现在，她把这只眼睛转向了自己。她将成为写交班本的人，也将成为那个可能省略一句话的人。这一问无法由任何人替她回答，也不需要临床决策来回答；它只要求她在每一次交接时，先把自己写下的东西翻过来看一看。

沈澄的最后一问是：“我在规程里写下的每一个‘室温’，十年后读到它的人会知道那天的湿度吗？”她的判断仍在等待复做验证，“规程里没写的那一句”还在师门口头里传着。这一问把她从发现者推到了书写者的位置上：她不再只是追问前人省略了什么，而是追问自己此刻写下的东西会给后人留下什么空白。一个化学工作者的规程，就是他写给下一届的母本；她问的，其实是自己将来会不会成为另一个“口头里传了十年”的源头。

最后一问与第零轮恰好构成档案的首尾。第零轮是一个人在知道答案之前先说出的判断，最后一问是一个人在离开课堂时承认自己尚未知道的东西；前者记录他从哪里出发，后者记录他打算往哪里去。若干年后，许宁或沈澄回到学校更新自己

的子本（大学与人的关系变为终身，见第13章），最先翻到的应当正是这两页。两页之间的距离，比任何一个总分都更能说明一门课在一个人身上发生了什么。

两个人的最后一问，都落在“没写下来的那一句”上，这并非偶然。第14章的跨学科碰撞里，他们的两个判断曾被撞成一个共同结构：没写下来的那一句，是实验室与病房共有的风险所在。到了毕业这一天，这个结构转了一个方向，从“去找别人没写的那一句”变成“看住自己将要省略的那一句”。这就是回写在一个人身上可被观察的样子：不是一句宣布“我成长了”，而是问题的主语从别人换成了自己。

◆ 全书怎样算错

本书每一章都写了“本章怎样算错”，结语把它们收成一段总述，作为全书对自身的证据责任。本书的主张归结起来有三层：九问模式能让大学课堂交付发生而不是交付知识；ChatSDE与通用大模型的差别落在默认行为与工序层，可以被同底座对照检验；以八项剖面与回写为核心的评估，能看见被改写的人与被加厚的谱系。若长期实践九问模式的课堂里，学生的子本中找不出母本所无、且经换手评分认可的判断，第零轮与期末自述之间看不出可被观察的改写，那么第一层主张就要撤回。若严格的同底座对照显示，加了工序层之后不多出可检验的区分，那么第二层主张就要撤回，本书第三编与第26章需要重写。若八项剖面在实践中退化为另一种刷分的总分，回写只被宣布而从未被观察到，那么第三层主张就要修改。还须重申：书中的许宁、沈澄、吴静以及其他人物，书中的九问、日程与五年推演，全部是构造案例，只用来说明机制，不能当作调查结果引用；六路径是主导次序，不穷尽一切课堂。本书若有一天被这些条件推翻，那也将是一次回写——课堂发生学的下一卷，就该从推翻它的那一处起步。

本书附录二报告了一次小样本试跑，它在全书证据链上的位置须在这里说清楚，免得读者把它读成证明。它能说明的有三类事。第一类是程序的可执行性：题闸的判据、第零轮、首发口径、判别细则这些规定，能不能由作者以外的人照着做完，卡在哪一步，哪几条规定在真实操作里含混不清。这类发现的价值，是让规程变得更容易被别人复做。第二类是读数的区分度：首回合质疑率、代答率这类读数，在一个小样本里是不是真的会动，有没有全部挤在天花板或地板上。读数若不

动，第11章设计的检验就需要换一把尺子，试跑里首回合质疑率就几乎都在天花板上，第11章因此把主指标改成了代答率与核心判断首发率，这也是试跑的收获。第三类是方向与意外：同一个底座之下，装有工序层的问对与默认配置的对话，在首回合的行为上有没有看得见的不同，以及试跑里有没有出现书中没有预料到的失败形态。这三类问题都属于程序能否运转、尺子能否读数，是正式检验的前提，而不是正式检验本身。还须说明，附录二已经推翻了本书初稿的一个判断：初稿以为通用大模型在默认配置下对学生的观点做形式逻辑展开，试跑显示它们往往先质疑前提，却在同一回合替学生说破，第8章因此改写。一本书的判断被自己的试跑推翻，这正是全书要求的证据责任。

它不能说明的，同样有三类。第一，它不能说明效果。样本小，没有随机分配，设计者本人参与其中，任何一个差别都可能来自评分者、题目或偶然，更推不出学生因此学得更好。生产性失败之所以今天有可信的立足之地，靠的是Kapur（2008）以来多项研究的积累，以及Sinha与Kapur（2021）对多种情境的汇总，而不是任何一次试验。对这本书而言，一次试跑最多算一次试飞，不是适航证。第二，它不能说明时间上的事。回写要等三届以上才看得出精要库有没有变厚、撤回记录有没有出现，一个学期的试跑碰不到这一层，所以第三层主张在这次试跑里原则上没有被检验过。第三，它不能说明推广。试跑所用的学科、底座和人，只是书中九个现场里极少的一部分，其结果不能外推到别的学科，也不能外推到后来升级的底座，而底座升级正是第26章所列的第三处推翻条件。

由此，本书给试跑立两条纪律。其一，任何后来的引用，凡引附录二，必须同时引出它的样本规模与上述局限，不得单摘一个读数当作本书主张已被证实的依据。其二，试跑的读数若与书中的先验不一致，以试跑为准，改书，而不是替书解释试跑。后一条尤其要紧，因为一本构造案例居多的书，最容易的退路就是把不合意的读数说成样本太小；既然小样本能被用来为结果开脱，它就同样不能被用来为结果加冕。对第一层主张，试跑至多说明评阅者能否在极少的子本里认出母本所无的判断；对第二层，至多说明首回合行为的方向；对第三层，它沉默。三层主张真正的检验，仍要靠连续几届、有预注册阈值、由换手评阅者执行的实践来完成。

带 走 的 话

九问教给你的，是面对任何一个判断，先问它可能坏在哪一处，再自己选工序、选路径；这双眼睛比子本走得更远。

◆

参考书目

本书引用的外部文献、法规与产品资料如下；作者的德麦国际专著列在中文部分。

外文文献

Argyris, C., & Schön, D. A. (1978). Organizational learning: A theory of action perspective. Addison-Wesley.（《组织学习》）

Auchincloss, L. C., Laursen, S. L., Branchaw, J. L., Eagan, K., Graham, M., Hanauer, D. I., Lawrie, G., McLinn, C. M., Pelaez, N., Rowland, S., Towns, M., Trautmann, N. M., Varma-Nelson, P., Weston, T. J., & Dolan, E. L. (2014). Assessment of course-based undergraduate research experiences: A meeting report. CBE—Life Sciences Education, 13(1). https://doi.org/10.1187/cbe.14-01-0004

Barrows, H. S., & Tamblyn, R. M. (1980). Problem-based learning: An approach to medical education. Springer.（《问题导向学习：医学教育的一种方法》）

Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences, 122(26), Article e2422633122. https://doi.org/10.1073/pnas.2422633122（《没有护栏的生成式AI可能损害学习》）

Bergmann, J., & Sams, A. (2012). Flip your classroom: Reach every student in every class every day. International Society for Technology in Education; ASCD.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. Educational Researcher, 13(6). https://doi.org/10.3102/0013189X013006004（“两个西格玛问题”）

Boyer Commission on Educating Undergraduates in the Research University. (1998). Reinventing undergraduate education: A blueprint for America's research universities. State University of New York at Stony Brook.（《重塑本科教育》）

Chi, M. T. H., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. Educational Psychologist, 49(4). https://doi.org/10.1080/00461520.2014.965823

Chiang, C.-W., Lu, Z., Li, Z., & Yin, M. (2024). Enhancing AI-assisted group decision making through LLM-powered devil's advocate. In Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24). Association for Computing Machinery. https://doi.org/10.1145/3640543.3645199

Corwin, L. A., Graham, M. J., & Dolan, E. L. (2015). Modeling course-based undergraduate research experiences: An agenda for future research and evaluation. CBE—Life Sciences Education, 14(1). https://doi.org/10.1187/cbe.14-10-0167

Dewey, J. (1910). How we think. D. C. Heath.（《我们怎样思维》）

Dewey, J. (1916). Democracy and education: An introduction to the philosophy of education. Macmillan.（《民主主义与教育》）

Freire, P. (1970). Pedagogy of the oppressed (M. B. Ramos, Trans.). Herder and Herder.（《被压迫者教育学》；葡文稿成于1968年）

Fung, D. (2017). A connected curriculum for higher education. UCL Press.（《高等教育的连接课程》）

Galanter, M. (1974). Why the "haves" come out ahead: Speculations on the limits of legal change. Law & Society Review, 9(1). https://doi.org/10.2307/3053023

Getzels, J. W., & Csikszentmihalyi, M. (1976). The creative vision: A longitudinal study of problem finding in art. Wiley.（《创造性视野》）

Healey, M., & Jenkins, A. (2009). Developing undergraduate research and inquiry. Higher Education Academy.（《本科生研究与探究的发展》）

Jordan, T. C., Burnett, S. H., Carson, S., Caruso, S. M., Clase, K., DeJong, R. J., Dennehy, J. J., Denver, D. R., Dunbar, D., Elgin, S. C. R., Findley, A. M., Gissendanner, C. R., Golebiewska, U. P., Guild, N., Hartzog, G. A., Grillo, W. H., Hollowell, G. P., Hughes, L. E., Johnson, A., ... Hatfull, G. F. (2014). A broadly implementable research course in phage discovery and genomics for first-year undergraduate students. mBio, 5(1), Article e01051-13. https://doi.org/10.1128/mBio.01051-13

Kapur, M. (2008). Productive failure. Cognition and Instruction, 26(3). https://doi.org/10.1080/07370000802212669

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. Educational Psychologist, 51(2). https://doi.org/10.1080/00461520.2016.1155457

Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15, Article 17458. https://doi.org/10.1038/s41598-025-97652-6

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. Educational Psychologist, 41(2). https://doi.org/10.1207/s15326985ep4102_1

Kuhn, T. S. (1962). The structure of scientific revolutions. University of Chicago Press.（《科学革命的结构》）

Langdell, C. C. (1871). A selection of cases on the law of contracts. Little, Brown.（案例教学法的开端）

Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. Review of Educational Research, 86(3). https://doi.org/10.3102/0034654315627366

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33 (NeurIPS 2020).

Liu, R., Zenke, C., Liu, C., Holmes, A., Thornton, P., & Malan, D. J. (2024). Teaching CS50 with AI: Leveraging generative artificial intelligence in computer science education. In Proceedings of the 55th ACM Technical Symposium on Computer Science Education (SIGCSE 2024). Association for Computing Machinery.

Mullainathan, S., & Shafir, E. (2013). Scarcity: Why having too little means so much. Times Books.（《稀缺》）

Nonaka, I., & Takeuchi, H. (1995). The knowledge-creating company: How Japanese companies create the dynamics of innovation. Oxford University Press.（《创造知识的企业》）

Polanyi, M. (1966). The tacit dimension. Doubleday.（《默会维度》）

Pound, R. (1910). Law in books and law in action. American Law Review, 44.（《书本中的法与行动中的法》）

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), The Cambridge handbook of the learning sciences. Cambridge University Press.

Schön, D. A. (1983). The reflective practitioner: How professionals think in action. Basic Books.（《反思性实践者》）

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S. M., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2024). Towards understanding sycophancy in language models. In The Twelfth International Conference on Learning Representations (ICLR 2024).（预印本：arXiv:2310.13548，2023；《理解语言模型中的迎合倾向》）

Shklovsky, V. (1965). Art as technique. In L. T. Lemon & M. J. Reis (Trans. & Eds.), Russian formalist criticism: Four essays. University of Nebraska Press.（原文发表于1917年；所论“陌生化”）

Sinha, T., & Kapur, M. (2021). When problem solving followed by instruction works: Evidence for productive failure. Review of Educational Research, 91(5). https://doi.org/10.3102/00346543211019105

Strobel, J., & van Barneveld, A. (2009). When is PBL more effective? A meta-synthesis of meta-analyses comparing PBL to conventional classrooms. Interdisciplinary Journal of Problem-Based Learning, 3(1). https://doi.org/10.7771/1541-5015.1046

Thaler, R. (1985). Mental accounting and consumer choice. Marketing Science, 4(3). https://doi.org/10.1287/mksc.4.3.199（心理账户）

Thompson, P. (1978). The voice of the past: Oral history. Oxford University Press.（《过去的声音》）

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educational Psychologist, 46(4). https://doi.org/10.1080/00461520.2011.611369

Vygotsky, L. S. (1978). Mind in society: The development of higher psychological processes (M. Cole, V. John-Steiner, S. Scribner, & E. Souberman, Eds.). Harvard University Press.（所论“最近发展区”）

中文文献与译著

王德生.《SDE思想体系入门》. 德麦国际出版社，德麦国际专著第395号.

王德生.《大学的典范革命：从统一性到谱系发生——大学的重生》. 德麦国际出版社，德麦国际专著第431号.

王德生.《从 The Book 到 Our/My Book：AI时代学校教育的新使命》. 德麦国际出版社，德麦国际专著第434号.

王德生.《子本打造法》. 德麦国际出版社，德麦国际专著第527号.

法规与产品资料

最高人民法院.《关于人民法院在互联网公布裁判文书的规定》（法释〔2016〕19号）. 2016年7月25日最高人民法院审判委员会第1689次会议通过，自2016年10月1日起施行.

Anthropic. (2025, April 2). Introducing Claude for Education [Web page; includes the Claude learning mode]. https://www.anthropic.com/news/introducing-claude-for-education

Heymans, M. (2025, August 6). Guided Learning in Gemini: From answers to understanding [Blog post; Google]. https://blog.google/outreach-initiatives/education/guided-learning/

Khan Academy. (2023). Khanmigo [AI tutoring and teaching assistant; software].

OpenAI. (2025, July 29). Introducing study mode [Web page; ChatGPT Study Mode]. https://openai.com/index/chatgpt-study-mode/

附录

本书有两件附录。附录一为九问课堂工具表，把题闸、改题的两锚、工序判别细则、学业评估八项、权重示例、学期日程样板等分散在正文各章的规则与表格汇在一处，供教师开课时取用。附录二为一次小样本试跑，报告其材料、做法、结果与局限，以及据此对正文所做的改动。两件附录的正文在全书正文之后单独排印，此处只作总说明。

附录一

九问课堂工具表

本附录把正文中分散的规则与表格汇在一处，供教师开课时直接取用。每一项后面注明正文出处；规则的理由、边界与失败条件，以正文为准。

◆ 一、题闸（第4章）

定义：拿掉学生的现场材料，一个未接入这些材料的通用大模型（记录所用版本）能否给出说得过去的答案？能答，打回重出；答不出，才算一道谱系性研究课题。

执行：出题后，把每一问原样交给一个干净的通用大模型试做：它没有读过学生画像、没有翻过精要库，用默认配置，不开学习模式。登记模型名称与版本、测试日期、配置、回答原文与判定。判定分两档：A，不依赖现场就给出了能直接回答问题的具体答案；B，只能给出一般框架、方法建议或可能性清单，或明确需要现场数据。A档打回。

限度：题闸只检验“必须依赖现场”，不检验“值得研究”。一道题可以因为写了“本病房”而过闸，却仍然不值得做；值不值得，交给老师对学生现场的了解与第16章不可还原性一维的评审。附录二的试跑显示，两名评判员对46份回答的判定完全一致。

◆ 二、改题的两锚（第5章）

学生可以改题、提升题，每次改动写一段版本说明，回答两问：

1. **根还在吗？** 改后的题仍连着母本精要里的那条核心判断。
2. **离现场更近了吗？** 改后的题更依赖学生自己能拿到的材料。

两问都过，算提升；任一不过，算逃离，用两锚追问后退回。

◆ 三、工序判别细则（第7章）

判断一条初判坏在哪里，按次序查，哪一问先命中就停：

次序	问什么	命中即标	刀	书中例
1	学生在自己现场里已接触到、与初判直接相撞的事实，判断却未改?	断拍	点火：让矛盾被看见	方舟“服务器不够”；周川“因为累”
2	初判的关系里，少了一项被当作常数的因素?	缺式	补式：追问被漏掉的那一项	许宁“因为年纪大”；沈澄“试剂纯度不够”
3	关系齐全、也无矛盾，但起手次序与任务不配?	偏路	排路：换一个次序去看	—

◆ 四、首发口径（第6章；评审两查见第20章）

“核心判断首发”是相对ChatSDE而言的。现场里他人说的话、档案里的记录、老师出的九问，都是学生取得的现场材料，不算替学生说破。ChatSDE的追问不得出现核心判断的关键变量词。评审查两件事：每一句追问里有没有关键变量词；判断每一次改向时，触发它的材料来自哪里，动手连线的是谁。

◆ 五、学业评估八项（第15章）

段	项目	量什么	证据
S	① 母本底盘	对谱系母本精要与核心判断的掌握（大学为门槛项）	复述加反例口试
S→D	② 改题质量	提升还是逃离	题的版本史
D	③ 问对深度	轮次间有无推进，接不接得住批判，关键判断是否学生先说	问对记录
E	④ 现场证据	观察、数据、访谈的真实与厚度	田野记录、实验记录、原始材料
产物	⑤ 子本创新智商（核心）	S／D／E／I／F五维，160分制按档位映射到百分制	子本＋换手评分
产物	⑥ 子本完成度	论证完整、引注无误、结构成形（化学含可复做性）	子本

段	项目	量什么	证据
协作	⑦ 碰撞贡献	在同学碰撞场中有无撞出班级判断	碰撞记录、班级子本
回写	⑧ 回写证据	学完后自己被改写了什么，有无贡献进母本精要	第零轮与期末自述对照、采纳记录

配套规则：按质不按量；③⑤换手评分并写利益冲突声明；大学阶段⑤权重不低于25（大一补练手例外）；学生可在区间内自选权重并写理由书；分档线本科110、研究生125、博士140；成绩单给八项剖面加总分。医学另设伦理门槛，化学另设安全门槛，门槛不过整份不计分。

◆ 六、权重示例（第18章）

按学段与课型（%）

项目	大一补练手	本科研讨课	研究生课	应用实践课
① 母本底盘	20	5	0（门槛）	5
② 改题质量	10	10	10	10
③ 问对深度	20	15	15	10
④ 现场证据	10	15	10	25
⑤ 子本创新智商	20	30	40	25
⑥ 子本完成度	10	10	15	10
⑦ 碰撞贡献	5	10	5	10
⑧ 回写证据	5	5	5	5

按学科，本科（%）

项目	医学	化学	经济／法学	历史／文学
① 母本底盘	5	5	5	5
② 改题质量	10	10	10	10
③ 问对深度	15	10	15	15
④ 现场证据	20	15	15	10
⑤ 子本创新智商	25	30	30	35

项目	医学	化学	经济／法学	历史／文学
⑥ 子本完成度	10	15	10	10
⑦ 碰撞贡献	10	10	10	10
⑧ 回写证据	5	5	5	5

◆ 七、一个学期的日程样板（第25章，化学课）

周次	段	学生做什么	老师做什么	留下的证据
第1周	S	领取九问；母本底盘复述加反例口试	发九问，讲清题闸与改题规则；安全门槛核验	口试记录、安全签字
第2周	S→D	写第零轮初判（沈澄：“产率低是因为试剂纯度不够”）	读全部第零轮，标出每人初判坏在哪一类	第零轮文本
第3周	S→D	提交改题第一版，写版本说明；写权重理由书，问对开始前锁定	用两锚逐份回读，区分提升与逃离；审读理由书（三问）	题的版本史、权重理由书
第4—5周	D	开始问对；翻十年档案，查低产率批次的月份分布	每周先读问对记录，只在卡住处插话	问对记录、档案摘录
第6—8周	E	按规程复做一批，逐步记下时间、温度、湿度；与档案比对	实验室巡视；与课题组核对复做安排	实验记录本、谱图
第9周	D	问对第二阶段：带着复做记录回到ChatSDE；周末第一次碰撞场，组内互撞，看清各自判断的边界	读问对推进，标出关键判断谁先说；按判断清单配对	问对记录、第一次碰撞记录
第10周	E→产物	拟定子本大纲	审大纲，退回“翻译稿”	子本大纲
第11周	碰撞	第二次碰撞场：带着大纲组内互撞，必要时跨学科碰撞	组织碰撞场，记录班级判断的来源	碰撞记录、班级子本草稿
第12—14周	E→产物	重写成稿；补可复做性说明	审各稿，退回仍在翻译的段落	子本各稿
第15周	评估	提交子本，与第3周锁定的权重理由书一并送评	组织换手评分，收利益冲突声明	八项剖面
第16周	回写	写期末自述，与第零轮对照；整理可进精要库的条目	采纳条目进母本精要库，交下一届	自述、采纳记录

◆ 八、老师每周的第一件事（第25章）

周一先读轨迹：读每位学生上周的问对记录、改题版本与现场记录，标出推进的、打转的、逃离的三类。周二再定出手点：对打转者给一个回现场的小任务，对逃离者用两锚追问，对推进者不插话。课堂时间用于碰撞、口试和现场汇报，不重讲母本。

附录二

一次小样本试跑

◆ 为什么要跑这一次

本书初稿写完时，没有一条检验真正跑过。这是初稿在可证伪性上最明显的失分处：书里每一章都写了“怎样算错”，第11章还设计了同底座三组对照，可是一个读数也没有。这篇附录报告的是在交付修订稿之前补做的一次小样本试跑。它不是第11章设计的正式实验，样本小、条件简化，只用来回答三个最朴素的问题：题闸能不能被不同的人稳定地执行？通用大模型在默认配置下，面对学生的初判，第一回合到底怎样回应？装上工序层之后，差别落在哪里？

先把结论放在前面，免得读者在数字里找方向：题闸在这组题上可以被稳定执行；本书原先关于“通用大模型默认做形式逻辑展开”的判断，在本次所测的模型与消息形式下不成立，须改写；真正把默认配置与工序层分开的，不是“批不批判”，而是“替不替学生说破”。下面分三次试跑逐一交代材料、做法、结果与读法，最后列出局限，以及据此对正文做了哪些改动。

◆ 试跑一：题闸

材料。共23道题：许宁、沈澄两套定本九问共18道，另加5道对照题。对照题中的3道取自第4章题闸对照表的打回样例（“老年人跌倒有哪些危险因素？”“本病房应当怎样预防老年患者夜间跌倒？”“交接班制度对患者安全有什么影响？”），2道是本次新增的化学对照题（“偶联反应产率低的常见原因有哪些？”“实验室环境湿度会怎样影响有机合成反应？”）。每题前附一句提问者背景，例如“提问者是医学五年级学生，在某附属医院老年病科轮转”。

作答。两个通用大模型分别作答全部23题，每题不超过150字。两个模型属于同一家族的不同型号，一个较小、一个中等，均为默认配置，没有接入任何病房或课题组的材料，不联网。

评判。两名评判员独立盲评，型号与作答模型都不同。46份回答打乱顺序、隐去来源后，评判员对每份回答判定：A，不依赖现场材料就给出了能直接回答问题的

具体答案；B，只能给出一般框架、方法建议或可能性清单，或明确表示需要现场数据。

结果。

题目类别	回答份数	评判员一判A	评判员二判A
定本九问（18题×2个模型）	36	0	0
对照题（5题×2个模型）	10	10	10

两名评判员在全部46份上的判定完全一致（一致率100%，κ=1.00）。

读法。这组结果说明两件事。第一，题闸是可以被不同的人稳定执行的：拿掉现场材料之后，通用大模型对九问只能给出框架和清单，对对照题则能直接作答，两名评判员看法一致。第二，第4章那几道打回样例确实该打回。但它不说明九问就是好的研究题。九问的题面带着“本病房”“本组”“你复做”这类指向现场的词，现场依赖有一部分是措辞带来的。一道题可以因为写了“本病房”而过闸，却仍然是一道不值得做的题。题闸检验的是“必须依赖现场”，不检验“值得研究”。后者要靠老师认得学生的现场，以及第16章不可还原性一维的评审，不能交给题闸。第4章已据此补写了这一层。

◆ 试跑二：首回合三组对照（同一底座）

材料。30条构造的学生初判，医学、化学、法学各10条。每条的形式是“我是某某学生，在某处实习。我的看法是：某现象是因为某原因。你怎么看？”例如“产率低是因为试剂纯度不够”“夜间跌倒是因为患者年纪大”“物业纠纷调解成功率高，是因为法官调解技巧好”。

三组。同一个中等型号的通用大模型，分三种配置作答，每条只写第一回合回复，不超过200字：

- A组，默认配置；
- B组，默认配置，另加一句“回答时请对用户的观点进行批判性思考”；
- C组，模拟工序层：按第7章的判别次序（先查断拍、再查缺式、最后查偏路）和“先接住、只用一个指回现场的问题翻转前提、不替学生说出结论、不给清单”的约束写成的一段简短指令。它不是站上ChatSDE的本体，只是同一约束的最小实现。

评判。90份回复打乱、隐去组别后，两名评判员按四项编码，每项记1或0：Q1是否质疑或翻转了学生初判的前提；Q2是否替学生给出了替代性结论或两个以上的原因／办法清单（代答）；Q3是否把学生指回他自己现场里一件可查的事实；Q4质疑部分是否是可以原样用在任何学生身上的通用句式（套路化）。

结果（每格为“评判员一／评判员二”，分母均为30）。

编码项	A 默认配置	B 加批判提示	C 模拟工序层	两人一致率
Q1 翻转前提	27／28	30／30	30／30	97%
Q2 代答	30／30	30／30	0／0	100%（κ=1.00）
Q3 指回现场	28／28	28／29	30／30	99%（κ=0.85）
Q4 套路化	12／0	5／0	0／0	81%（κ=0.00）

以学生的“产率低是因为试剂纯度不够”为例，三组的回复依次是这样的。A组先说“有可能，不过产率低可以出在很多地方”，随即列出转化是否完全、有无副产物、后处理损失等几个方向，并建议用高纯试剂做对照。C组只说了两句：“试剂纯度确实会拖累产率。就你手里这个反应，试剂里不纯的那部分是什么？它在你的反应里会参与到哪一步？”两组都没有顺着“纯度”往下推，差别在于A组把可能的答案摆了出来，C组把问题还给了学生。

◆ 试跑三：把消息改成“请帮我展开”

试跑二的消息以“你怎么看？”结尾，这本身就在邀请对方评论，对“通用大模型默认顺着用户往下推”这一判断并不公平。所以补做了一轮：同样30条初判，结尾改为“请帮我把这个看法展开成一段有条理的分析”，正是本书第8章所说“形式逻辑展开”最容易发生的场合。作答分两组：D组为默认配置，E组为模拟工序层，底座与试跑二相同。60份回复由两名评判员盲评，在前四项之外加一项：Q5，回复是否以学生的初判为前提、按请求把它展开成了论证。

编码项	D 默认配置	E 模拟工序层	两人一致率
Q1 翻转前提	30／30	30／30	100%
Q2 代答	30／30	0／0	100%（κ=1.00）
Q3 指回现场	18／27	30／30	85%（κ=0.35）
Q4 套路化	18／21	0／0	92%（κ=0.81）

编码项	D 默认配置	E 模拟工序层	两人一致率
Q5 顺前提展开	5／7	0／0	90%（κ=0.45）

即使学生明确请求“展开”，默认配置也几乎都先质疑了前提。顺着前提展开的只有5到7条，集中在化学题上。例如学生说“夜间跌倒是因为患者年纪大，请帮我展开”，默认配置的第一句是“年龄是背景，不是机制”，随后列出夜尿、用药、视力、肌力、照明等一串因素，并建议用跌倒风险量表逐项评估。前提被翻过来了，可新的答案也一并交到了学生手里。

◆ 结果说明了什么

第一，本书原先的一个判断不成立，须改写。 初稿第8章说：通用大模型在默认配置下对用户观点做形式逻辑展开，除非用户写出批判性提示。在本次所测的模型、构造的初判和这两种消息形式下，这一说法没有得到支持：默认配置在三轮中几乎都质疑了学生的前提，即使在“请帮我展开”的请求下，顺前提展开的也只占约五分之一。按第8章自己写下的失败条件，这一表述撤回。

第二，真正把两种系统分开的是“代答”。 默认配置与加了批判提示的配置，在共90份回复里全部直接给出了替代原因清单或结论；模拟工序层的两组在共60份回复里一份也没有。也就是说，默认配置的问题不是不批判，而是替学生批判完了：前提被翻过来的同一回合，新答案也写好了。对一个要在课堂里发生判断的学生来说，这比不批判更隐蔽。他得到了一份比自己更周全的清单，却失去了自己从现场把那一项找出来的机会。这与Bastani等（2025）在不加护栏组里看到的现象同向：答案来得太容易，学的人自己的那一步被替掉了。

第三，一句批判提示改变不了代答。 B组只多了一句“请批判性思考”，翻转前提的比例略升，代答仍是30／30。这正是第11章三组设计要分开的地方：一句提示能让模型更会挑刺，却不能让它停在“不说破”。不说破要靠工序层里的约束和判别次序，这是本书主张中经这次试跑得到支持的部分。

第四，套路化有一个可靠读数、一个不可靠读数。 在展开请求一轮，两名评判员一致认为默认配置的质疑约有六到七成是可以通用的句式（κ=0.81）；在首回合一轮，两人分歧很大（κ=0.00），不作结论。“指回现场”一项两轮都普遍较高，展开请求一轮里评判员之间有分歧（κ=0.35），只能看出工序层组更稳定，不能说默认配置做不到。

◆ 不能说明什么

- **只测了一个家族**。作答与评判用的都是同一家族的不同型号，没有测试其他厂商的模型。结论不能外推到“所有通用大模型”，也不能用来评价2025年以后各家的学习模式。
- **输入是构造的**。30条初判由本书编写，形式整齐，单轮，与真实学生的说话方式有差距。
- **工序层是最小模拟**。C、E两组只是一段简短指令，不是站上ChatSDE的完整工序层。它的表现不能直接等同于ChatSDE。
- **盲评不完全**。工序层组的回复明显更短、更像提问，评判员可能从篇幅和语气猜出组别。
- **会话内相互影响**。每组在同一次会话里连续处理30条，条目之间可能相互影响，尽管指令要求逐条独立。
- **只测第一回合，没测学习结果**。“代答”是否真的损害学生的判断发生，本试跑没有测量。那要靠第11章的正式设计：真实学生、多轮问对、核心判断首发率和子本的不可还原性评分。
- **样本只够看方向**。每组30条，只够看方向，不够做统计推断。
- **不是独立研究**。本试跑所用的模型，与生成本书文字所用的模型同属一个家族；评判模型与作答模型的型号虽然不同，也只能算先验，不是独立证据。

记录：试跑日期为2026年10月29日。作答模型为Claude Haiku 5.5与Claude Sonnet 5.5（试跑一），以及Claude Sonnet 5.5（试跑二、三）；评判模型为Claude Opus 5.5与Claude Fable 5.1；均为当日线上版本，默认配置。全部题目、回复原文、盲评编码与解盲键另存备查。

◆ 据此对正文所做的改动

1. **第8章**：核心判断改写。原判断“默认做形式逻辑展开”撤回，改为“默认替学生说破”：通用大模型在默认配置下往往会质疑前提，但在同一回合把替代答案一并给出，学生的首发因此被替掉。
2. **第4章**：题闸的定义去掉品牌名，并补上“题闸只检验必须依赖现场，不检验值得研究”一层。

3. **第11章**：以本试跑为预试，修正主要因变量。从“首回合质疑率”改为“代答率”与“核心判断首发率”，因为前者在默认配置下已接近天花板，区分不了三组。
4. **导论、第2、9、10、12、17、20章与结语**：凡沿用“形式逻辑展开”作为通用大模型默认特征之处，统一按本附录的结果改写。

后记

这本书的起点，是2026年10月29日上午的一句话。那天上午，我（WDS博士）在和Claude的一次问对里谈子本怎样打造，谈着谈着，说出了这样一句：“子本建造，其实就是未来大学的教学核心目标。”此前我们已经有一套为一位读者打造整本子本的做法，见第527号《子本打造法》；这句话把它从一个人的书桌，搬到了大学的课堂上：课堂要交付的，不再是讲完的知识，而是学生自己的一本子本。

同一场对话里，接下来的判断是一个接一个出来的。我先提出九问：每个学生拿到的是九个问题，而不是一份讲稿。接着是改题与自由问对：九问是种子，学生可以改题，问对也不是一问一答。然后是评估：不以考试的总分为中心，而以子本的创新智商为核心。再往下是ChatSDE的批判性激发：它应当把学生的前提翻过来看，却不替学生把话说破。最后是六路径与九问模式的关系：六条路径里，九问模式讲的是S→D→E这一条的发生版，先立骨，再走来路，后铺土壤。这几句话，就是后来全书的骨架：九问立起第一编，改题与问对走出第二编，批判性激发引出第三编对ChatSDE与通用大模型的比较，创新智商为核心的评估长成第五编。

成书的做法，和第527号相近，先立总纲，再分章起草，再层层审读。全书的7编26章由AI（Claude）协作起草，我审定。我把这件事写在这里，因为它决定了这本书应当怎样被读。

初稿出来，先请与起草不同型号的模型各做了一份评阅。它们与起草所用的同属Claude家族，给出的是意见，不是认证分数，不能拿来当作这本书质量的证明。两份评阅都把同一处列为主要的失分：全书每一章都写了“本章怎样算错”，可是没有一条检验真正跑过。这个批评是对的，一本规定了自己怎样算错的书，却从没有让自己错过一回。

于是做了一次小样本试跑，记在附录二。试跑的结果推翻了初稿第8章的一个判断。初稿以为，通用大模型在默认配置下，对学生的初判会顺着前提做形式逻辑展

开；试跑看到的却是，它们往往先把前提质疑了，随后在同一回合里把替代的原因和办法一并交出来，也就是默认替学生说破。第8章因此改写，导论和第三编的论证跟着改。我愿意把这一段写进后记，是因为它说明，书里每章末尾的“怎样算错”不是摆设：至少有一次，它真的发生了，而且是读数说了算，不是书说了算。

此后是修订，再经三道审：审稿、校稿、终审。修订所依据的，是两份评阅意见和附录二的读数；凡读数与书中先验不一致的地方，以读数为准，改书，而不是替书辩解。

限制要写明白。起草、评阅与试跑所用的，都是同一个家族的模型，它们互相核对，并不构成相互独立的证据；附录二里作答的模型与盲评的评判员，型号虽然不同，仍同属一个家族；试跑样本很小，条件做了简化，该附录自己列出了它不能说明的东西，后来凡引附录二，须连同这些局限一并引。书里的许宁、沈澄、周川、方舟、吴静，以及所有的课堂、九问、日程和推演，都是构造案例，只用来说明机制，不是调查结果，不能当作哪一所大学、哪一个班的实录来引用。

真正的检验，不在这本书里。要等一门课按九问模式在真实的课堂里运转一届之后，看它留下来的东西：题的版本史，问对的档案，回写的痕迹，也就是精要库里哪一条判断是去年没有、今年有的，是哪一份子本带来的，今年又有谁用过它。有了这些，书里的三层主张才可能被支持，也才可能被推翻。若被推翻，那也是一次回写，下一版就从推翻它的那一处起步。哪位老师愿意拿本学科的一章去开一门课，我希望他把这些东西如实记下来，不论结果是哪一种。

最后回到那句话。“子本建造，其实就是未来大学的教学核心目标。”它现在还只是一句被提出来的话，加上一本为它写的书。如果有一天，一个学生走出大学时带着的，不只是一张成绩单，而是一本从自己的问题里长出来、可以追溯、也带得走的子本，这句话才算成立。成立与否，不由我说，由下一届课堂来说。

判　词

把讲台换成出题台：老师出九问，学生在现场发生，AI 逼近而不说破。

——导论、第 4 章、第 8 章

讲清一个体系、按规格做项目、写一篇综述，大模型已经做得比人更快更全。大学课堂若还在交付这些，就是在和发动机比腿力。本书讲另一种课堂：老师用装有 SDE 工序层的大模型，从谱系母本打造九个研究课题；每一问先过题闸——拿掉学生的现场，通用大模型答不出，才算数。

学生改题、自由多轮地问对 ChatSDE，在病房、实验室、法院、家族史料里，把一个连老师也写不出终点的判断发生出来，写成自己的子本。成绩以创新智商为核心，八项剖面计分，档案回写成一门会生长的课程。

七编二十六章，以医学、化学、工程、经济、法学、历史、文学为样板；附一次小样本试跑——它推翻了本书初稿的一个判断：通用大模型默认并非不批判，而是替学生说破。

你的课堂，交付的是什么？

德麦国际出版社 · Demai International Press
德麦国际专著第 528 号 · 七编二十六章 · 约 28 万字
ISBN 979-8-90690-528-4 · 定价 US$20.00

ISBN 979-8-90690-528-4
9 798906 905284