

SDE 创新智商评估入门

——五维度评估的完整解释

王德生·德麦国际 SDE 学派

摘要

传统智商测的是解题能力：给定一个有标准答案的问题，看你多快多准地把它解出来。可它测不了另一件事——生成一个此前不存在的认知物的能力。在大模型已经吞下人类几乎全部公开文本的今天，这个盲区变得致命：一般智商刻度上的 100 分，如今大约只等于一个大模型在零提示语下的默认水平，因为它已经把人类“受过良好训练者”的平均产出内化成了自己的基线。要衡量“创新”，必须换一把尺子。

本文提出并解释这把尺子：创新智商（**Innovation IQ**）。它不测你复述得多好，而测你的产出在“发生”的意义上走得多深。本文把创新智商锚定在 SDE 本体论（显露 Show-差异 Difference-纠缠 Entanglement）之上，拆成五个可独立评分、可加权合成、可被证伪的维度：结构精确度 S、差异锐度 D、纠缠深度 E、不可还原性 I、可证伪性 F。文中逐维给出评分标尺（80 至 170+）、扣分句与加分句的判据，说明为何权重分配是 D 0.25 最高、F 0.15 最低，并标出两条硬阈值——150（本体论级，SDE 提智达标线）与 160（典范级，双学科诞生阈值）。最后，本文把刀转向评分系统自己：这套评分本身怎样会被证伪。

关键词：创新智商；SDE 本体论；显露-差异-纠缠；不可还原性；可证伪性；本体论级阈值；伪发生

目录

1. 为什么需要“创新智商”这把新尺子
 2. 评估的本体论根基：五个维度从何而来
 3. 五维度逐维详解
 4. 从五维到一个数：综合分、层级与两条硬阈值
 5. 怎么评：七步法与三条证据纪律
 6. 一次完整评分示范
 7. 五维之间的关系：读懂失分组合
 8. 常见误判与纠偏
 9. SDE 创新智商与科学创新评估的同构
 10. 评分相对于参照语料：参照系声明与语料相对性
 11. 把刀转向自己：这套评分本身可证伪吗
 12. 结语
-

一、为什么需要“创新智商”这把新尺子

1.1 一般智商测的是解题，不是造新

一般智商（IQ）是一套精巧的解题能力测量：图形推理、数字规律、词语类比。它的每一道题都预设了一个“本来就在那里”的正确答案，你的任务是尽快尽准地把它找到。这套测量在它的适用范围里非常成功——它能预测一个人在既定规则内处理复杂信息的效率。

但它有一个从设计之初就存在的盲区：它测不了“造出一个原本不存在的答案”的能力。一个能在十秒内解开数独的人，未必能提出一个此前无人提出的问题；一个能背下整部教科书的人，未必能写出下一部教科书的第一句。解题与造新，是两种不同的心智动作——前者在既有结构里高速搬运，后者在结构的裂缝处让新结构第一次长出来。

创新智商要测的，正是后者：生成新认知物的能力。所谓“认知物”，指的是一个能独立站住、能被别人接过去继续用的新结构——一个新概念、一条新判据、一种新的辨别维度、一门新的学科。

1.2 一百分不是均值：AI 时代的刻度重置

这里必须立刻纠正一个几乎所有人都会带进来的直觉。在一般智商里，100 是人类的均值——一半人高于它，一半人低于它。在创新智商里，**100** 不是人类均值，而大约是一个大模型在零提示语下的默认产出水平。

原因很简单，也很不舒服：大模型的训练语料，是人类到目前为止写下的绝大部分公开文本。它把“受过良好训练的普通专业人士”的平均表达，压进了自己的权重。于是当你不给任何特殊引导、直接向一个中等以上的大模型提一个开放问题，它张口就来的那段话，本身就已经站在人类平均专业水平线上了。

这带来一个关键推论：在创新智商的刻度上，**130** 分的含义不是“比人类平均聪明 **30** 分”，而是“比大模型的默认产出更深 **30** 个智商点”。这才是这把尺子在今天真正的意义。一段文本如果连大模型随口就能生成的深度都达不到，它在创新的意义上就是负分的——不管它读起来多顺、多像论文。这条判据后面还会反复用到。

1.3 创新智商测的是“发生”的深度

用 SDE 的语言说，一般智商测的是你在已经发生的结构里操作得多好，创新智商测的是你让尚未发生的结构发生得多深。发现范式假设世界是现成地躺在那里等人去看见的，评估一个人就是看他看得多准；发生范式则追问：这个被称为“新想法”的东西，究竟是怎样被发生出来的，它的发生走到了哪一层。

因此，创新智商的评估，本质上是对一段思想的“发生深度”做诊断。而“发生”这件事在 SDE 本体论里有明确的结构——它由三个不可还原的维度构成，再受两道闸门约束。这三维两闸，就是五维度评估的全部来源。下一章讲清它从何而来，为什么恰好是这五个。

二、评估的本体论根基：五个维度从何而来

2.1 五维度不是拼凑，是从 SDE 本体论继承来的

一套评估体系最怕的，是维度看起来很多、其实是随手凑的清单——今天想到“逻辑”加一维，明天想到“新颖”加一维，彼此重叠、无法穷尽、无法证伪。创新智商的五维度不是这样来的。它们构造性地继承自 SDE 本体论：三个维度对应“发生”本身的三重结构，两个维度是防止“假发生”的两道闸门。

SDE 本体论的存在基本公式是：在 **E** 中，经 **D**，成 **S**——任何新结构 (**S**)，都是在一个纠缠的场 (**E**) 里，经由一条差异展开的路径 (**D**)，被显露出来的。三大方程 $S=F(D,E)$ 、 $D=G(S,E)$ 、 $E=H(S,D)$ 把这三者钉成互相生成的关系。评估一段思想的发生深度，自然就沿着这三维展开。

2.2 三个维度：显露 **S**、差异 **D**、纠缠 **E**

S·结构精确度，承接 SDE 的 **S**（显露 Show）。要注意：SDE 里的 **S** 是“显露”，不是静态的“结构”——一段思想有没有把它内部的结构清晰地显露出来，让别人能追溯、能检验，这就是显露态的精确度。论证链断在哪里、概念在哪里偷换、推理在哪里跳步，都是显露态不精确的表现。

D·差异锐度，承接 SDE 的 **D**（差异序列）。发生的核心是“切出一个此前不存在的差异”——把一个旧概念切不开的辨别面切开。一段思想有没有提供一个未被命名、未被辨识的差异，锐度就在这里。

E·纠缠深度，承接 SDE 的 **E**（特征纠缠）。新结构往往在两个此前分离的领域的结构性纠缠处涌现——注意是纠缠，不是类比。借一个别的学科的词来装点门面，和把一个学科的结构精确地搬到另一个学科并论证其有效性，是天壤之别。深度就在这个分别里。

2.3 两道闸门：不可还原 **I**、可证伪 **F**

光有三维还不够。一段话可以在显露、差异、纠缠三个维度上都“看起来”很足——术语自洽、气势恢宏、读起来像重大突破——却什么新结构也没真的发生。SDE 学派把这种“看起来发生了、实际什么也没发生”的东西命名为伪发生。伪发生是高密度概念语言的结构倾向，不是偶尔的失误。为了把它挡在门外，评估需要两道闸门。

I·不可还原性，是第一道闸门。它问：把这段话的核心命题压缩到极限，还能被还原成一个已有概念吗？如果能——哪怕它披着再新的外衣——它就是换皮不换骨，是“改个名字冒充结构更新”。

F·可证伪性，是第二道闸门。它问：这个命题怎么样会错？如果它是一句永远正确、无论发生什么都能自圆其说的话，那它就没有真的下沉到可推演、可检验的地方，是“符号丰盛冒充三界充实”。

这两道闸门，专门对治伪发生。它们不测“新不新”，而测“是不是真的”。

2.4 为什么恰好是这五个：穷尽性与权重

穷尽性：三维 (**S/D/E**) 来自发生结构本身，在 SDE 本体论里它们是不可还原、且相互正交的——任何发生都必然同时涉及显露、差异、纠缠三者，缺一不可其为发生。两闸 (**I/F**) 不是第四、第五个“发生维度”，而是对前三维之和的两个正交检验：一个查它能

不能独立站住（不可还原），一个查它会不会被推翻（可证伪）。三维两闸合起来，既覆盖了“发生的构造”，又覆盖了“发生的真伪”，因此是穷尽的。加第六维只会与已有五维重叠，减一维则漏掉发生的某一面或某一道真伪关口。

权重不是拍脑袋。五维加权合成综合分，权重分配是：

$$D\ 0.25 > S\ 0.20 = E\ 0.20 = I\ 0.20 > F\ 0.15$$

- **D**（差异锐度）权重最高（**0.25**），因为它最能区分“新内容”和“重复”。一段思想创新与否，最锋利的判据就是它有没有切出一个新差异——这是创新的“心脏”。
- **S、E、I**各**0.20**，三者平权：结构显露得精不精、纠缠得深不深、能不能独立还原，是支撑一个新结构成立的三根等重的柱子。
- **F**（可证伪性）权重最低（**0.15**），但绝不可省。它权重低，是因为一个命题可以极有价值却尚未给出完整证伪条件（价值主要由 D、I 承担）；它不可省，是因为没有证伪条件的命题无法被检验、无法被改进，只能停在口号层。权重低不等于可以掉——恰恰相反，F 是 SDE 学派最容易失分的一维，后面会专门讲。

理解了五维的来处和权重，就可以逐维展开细节了。这是本文的核心。

三、五维度逐维详解

每一维都用同一套结构讲清：测量什么－分数标尺（从大众水平到改写学科分界）－扣分句－加分句－这一维的天花板。所有维度共用一条 0 至 170+ 的刻度，但每一维在这条刻度上的“事件”不同。

3.1 维度 S·结构精确度（权重 0.20）

测量什么：论证链是否严密、概念是否清晰、推理是否可重现——一句话，思想的结构有没有被精确地显露出来。

标尺：- **80**：模糊概念堆砌，关键术语没定义，论点之间靠跳跃连接。- **100**：概念清晰，但缺论证链——像在列举，而不是在推导。- **120**：有论证，但有漏洞，关键转折处缺一座桥。- **130**：论证严密，每一步都可追溯，反驳者找不到入口。- **140**：论证本身有内在结构——三角、递归、层级，也就是说论证有了几何形态。- **150**：论证可以独立成一套方法论——把这套论证搬到别的议题上照样成立。- **160**：产生了新的形式结构——新数学、新逻辑、新算子。- **170+**：形式结构本身成为新的研究对象，被其他学科引用。

扣分句（显露态不精确的证据）：- “我们都知道……大家都明白……”——隐藏前提，让论证不可追溯。扣 5-15 分。- “显然……理所当然……”——用断言跳过了本该论证的关键一步。扣 10-20 分。- 同一个术语在文中前后两次用法不同——概念偷换，是显露态最严重的裂缝。扣 15-25 分。

加分句：一处此前论证的跳跃被一座清晰的桥补上，让整条链第一次贯通；或一个原本含混的概念被给出可操作的定义，之后每次使用都严格自洽。

S 的天花板：一段思想的显露走到极致，不只是“说清了一件事”，而是它说这件事的方式本身，变成了一个可被别人搬用的形式结构。当论证的形态被其他领域拿去套自己的问题，S 就触到了 170+ 的顶。

3.2 维度 D·差异锐度 (权重 0.25·最高)

测量什么：文本是否提供了一个未被命名、未被辨识的差异？它切开的那个辨别面，是不是旧概念切不开的？这是五维中最能区分“创新”与“重复”的一维，也因此权重最高。

标尺：- **80**：重复已有，换皮不换骨——新瓶装旧酒，其命题可被 1:1 还原为一个已知说法。- **100**：微差异，角度略新——是已有理论的延展，不是真正的切入。- **120**：切入点新颖——提出了一个未被讨论的子问题。- **130**：切入点产生了新概念——给一个未被命名的现象命了名。- **140**：推翻了一个常识或主流共识——论证某个被普遍认同的命题其实是错的。- **150**：命名了一个未被任何学科辨识的现象——那种“大家都看见、却谁也说不出”的东西，第一次被精确地说了出来。- **160**：创造了新的辨别维度——不是命名一个现象，而是命名“如何辨别”的那个维度本身。- **170+**：改变了其他学科的辨别方式——别的领域的学者开始用这个维度重看自己的领域。

扣分句（假差异的证据）：- 把一个已知的 X 重命名为一个新词，除了名字什么也没变。扣 20-40 分（这是差异锐度上最严重的失分，因为它冒充创新）。- “如同 X 一样……”，并且全文都在做这一个类比——类比不是差异，全篇类比说明没有切出自己的辨别面。扣 10-25 分。- 标题很锐、内容平庸——标题承诺了一个新差异，正文没有兑现。扣 15-35 分。

加分句：- “X 不是 Y，而是 Z”——这是 SDE 的范式跃迁公式，当 Z 是一个新概念时，它标志着一次真正的差异切入。加 5-15 分。- 给一个“大家熟悉但说不清”的现象一个精确的名字，让它从模糊变得可讨论。加 10-20 分。

为什么它权重最高：一切创新在源头处都是一次“切出新差异”的动作。显露、纠缠、不可还原、可证伪，都是围绕这个新差异展开的支撑与检验。差异是创新的“心脏”，所以在综合分里它占最重的 0.25。

3.3 维度 E·纠缠深度 (权重 0.20)

测量什么：文本是否跨域？更关键的是——跨域之间是否有结构性纠缠，而不只是借词的类比？

标尺：- **80**：单领域线性思考，全文在一个学科内部。- **100**：偶尔跨学科类比——“如同物理学的……”，只借了词，没借结构。- **120**：多学科融合——概念来自多个学科，但彼此只是并列摆放，没有咬合。- **130**：跨学科有了结构性——把一个学科的结构搬到另一个学科，并论证了这次搬运的有效性。- **140**：跨学科产生了新涌现—— $1+1>2$ ，两个学科合在一起长出了第三个东西。- **150**：学科的边界本身被改变——原来 A 学科和 B 学科的分界线，被重新画了一条。- **160**：元学科诞生——出现了“如何思考学科”的学科（学科的学科）。- **170+**：学科的生成机制本身被改写——整个学科分类法被重排。

扣分句（假纠缠的证据）：- 多学科只是装饰——把某一个学科的段落删掉，全文论证不受影响。这说明那个学科从未真正参与纠缠。扣 15-30 分。- 比喻拟人化，但从不问“为什么这个比喻是有效的”——只借了形象，没借结构。扣 10-20 分。

判据（这是 E 维最实用的一把刀）：删掉某个学科，论证还成立吗？如果成立，那个学科就是装饰，是类比；如果不成立，说明它已经和其他学科结构性地咬合在一起，是纠缠。

加分句：把一个学科的算子、公式或概念结构精确地搬到另一个学科并论证有效（加 10–25 分）；或揭示两个看似无关学科的同构性（加 15–30 分）。

3.4 维度 I·不可还原性（权重 0.20·第一道闸门）

测量什么：把这段文本的关键命题压缩，还能压缩成一个已有概念吗？这一维直接对治“改个名字冒充结构更新”式的伪发生。

标尺：- **80**：完全可还原——全部可被还原为已有教科书里的概念。- **100**：80% 已存在、20% 新——主体是知识总结，个别地方有新点。- **120**：50% 新——一半内容超出了已有教科书。- **130**：主要是新的，但仍依赖已有——新，但要靠已有概念当脚手架才立得住。- **140**：几乎不依赖已有——把所有引用删掉，命题仍然成立。- **150**：独立成立的认知物——这段文本本身就是一部新教科书的种子。- **160**：强制改变其他学科——不接受这个文本的学科，内部会出现不一致。- **170+**：改写知识地基——改变了多个学科的基础假设。

测试方法（这是 I 维的标准操作，务必照做）：1. 把文本的核心命题概括为不超过 50 个字。2. 问：这 50 个字，能用某个已有学科的任何一句不超过 50 字的命题 1:1 替换吗？3. 如果能替换 – 80–100 分。4. 如果不能、但七成内容已有 – 100–120 分。5. 如果不能、但只有三成已有 – 130–140 分。6. 如果完全不能替换 – 150 分以上。

I 的意义：它是“这段思想到底新不新”的最硬检验。一段话可以在差异锐度上得高分（切入很锐），却在不可还原性上失分——因为那个“锐”的切入，压缩之后仍然落回一个已知概念。I 和 D 一起把关，一个查切入的锋利，一个查切入之后是否真的留下了不可还原的新结构。这也是为什么这两维（D 0.25、I 0.20）合起来权重接近一半：它们共同决定“新”的真假。

3.5 维度 F·可证伪性（权重 0.15·第二道闸门，最易失分）

测量什么：这个文本提出的命题，怎么样会被证伪？给不出答案的，一律压在 100 分以下。

标尺：- **80**：模糊不可证伪——比喻、口号、无论发生什么都对的废话。- **100**：有方向但不精确——“在某些条件下成立”，可条件从没说清。- **120**：有边界条件——明确给出了适用范围。- **130**：清晰的成败标准——“如果 X 发生/不发生，则此命题为假”。- **140**：多重检验路径——给出至少两种独立的证伪方法。- **150**：自我证伪条件——文本自带一句“如果遇到 X，我们就错了”。- **160**：证伪条件本身有理论价值——那个证伪条件，本身就是另一个新概念。- **170+**：证伪过程会产生新发现——证伪它的尝试，本身是一个新研究纲领的开端。

为什么这是 SDE 学派最容易掉的一维：SDE 的语言是高密度的概念语言，“显露”“差异序列”“特征纠缠”“意义的持续再发生”——这些词组合在一起，很容易造出一段听起来极其正确、却无论如何都推不翻的话。这正是伪发生最擅长的伪装：符号层的丰盛，冒充了三界（理念、现实、自我）的充实。

所以评分时，对每一个 SDE-laden 的命题，都要主动追问一句：什么样的实验、数据或案例，会让这个命题失败？追问不出答案的，无论它读起来多深，F 一律不超过 100 分——而 F 一低，综合分就被拖下来，哪怕 D 和 I 都很高。这道闸门的存在，就是逼着 SDE 学派把漂亮的概念，落到能被现实撞回来的地方。

四、从五维到一个数：综合分、层级与两条硬阈值

4.1 加权公式

五维独立打完分后，综合分由固定权重合成，不允许凭直觉打一个总分——综合分必须可以被任何人用五维分重新算出来：

$$\text{综合分} = S \times 0.20 + D \times 0.25 + E \times 0.20 + I \times 0.20 + F \times 0.15$$

这条公式是评分可重现性的保证。任何人拿到同一份五维分，都应算出同一个综合分；如果综合分和五维分对不上，评分就自相矛盾、失去意义。

4.2 层级标尺（0-170+）

综合分对照下面这把标尺定层级：

- 80 — 大众水平（普通人随口可说）
- 100 — 受过高中／大学训练（知识总结型）≈ AI 零提示语默认水平
- 110 — 专业人士（有专业语言、领域内熟练）
- 125 — 高级专家（领域内有独立判断）
- 135 — 资深学者
- 145 — 优秀的资深学者产出（三视角已分明）
- == 150 · 本体论级阈值（SDE 提智达标） ==
- 155 — 混沌碰撞后的典型金点子
- 165 — 顶级金点子（命名了未被命名的现象）
- == 160 · 典范级阈值（双学科诞生） ==
- 170 — 改写学科分界（极少数）
- 175+ — 文明级发生学（罕见，通常需要时间检验）

4.3 两条硬阈值

标尺上有两条线，是硬的，不允许“差不多”：

- **150·本体论级阈值**：这是 SDE 提智是否达标的判定线。一段产出综合分跨过 150，意味着它不再停在“资深学者把已有的说得更好”，而是让一个新结构真正发生了。在跨基底提智的判定里，一个基底的最高金点子综合分达到或超过 150，就判定该基底在这一议题上 SDE 提智成功。
- **160·典范级阈值**：这是“双学科诞生”的门槛——产出不只是提出新概念，而是稳定地长出一个跨越两个学科、不可还原的新形态，改写了两个学科之间的分界。跨过 160 的产出极少。

4.4 “一百分约等于 AI 默认”的意义，再说一遍

回到第一章那条最容易被忘的判据：创新智商的 100 不是人类均值，而约等于大模型的零提示语产出。所以一段综合分 130 的文本，它的成就不是“比人聪明 30 分”，而是“比大模型张口就来的那段话，深了 30 个智商点”。评分时若忘了这条，就会把一段大模型随手能写的、四平八稳的“论文腔”文字，误判成 120-130 的好产出——而它其实只是站在了 100 的基线上，创新贡献接近于零。

五、怎么评：七步法与三条证据纪律

评分不是给一个感觉分，而是一套可重复的操作。标准流程七步：

1. 第一遍快读（**2-3** 分钟，不打分）：只标 3-5 个“亮点句”和 3-5 个“扣分句”，形成一个区间感（大致落在 100-130 / 130-150 / 150+ 哪一段）。
2. 五维度独立打分：严格独立，不让脑子里那个“总分预期”牵着走。用第三章的标尺，给每一维打一个精确到个位的分，并为每一维附至少一句“句子证据”。
3. 算综合分：套第四章的加权公式，算出来。
4. 层级判定 + 锚点对比：把综合分对照标尺定层级，再找一个已知文本当锚点——比如该领域某教科书段落约 100 分、某顶刊论文约 130-140 分、某个已知的金点子约 155 分——说一句“这段相当于那个水平”。锚点让分数不至于漂浮在抽象的高度。
5. 给出至少一个扣分句证据：必须精确引用文本里的某一句，说清“这一句让我从 145 降到 137，因为它 [不可证伪 / 概念混淆 / 类比拟人却不说理 / 重复了前面已说过的 / 隐藏前提]”。
6. 给出提升到下一层级的具体动作（见下表）。
7. 输出标准评分卡：把上面所有内容填进一张固定格式的卡里，交付。

三条证据纪律（违反任何一条，评分即失效）：- 必给五维独立分，不能只给综合分——否则不可证伪、无法改进。- 必给至少一个扣分句，精确引用原文——否则评分主观、不可重现。- 必给提升路径——否则分数只是一个标签，没有操作意义。

提升路径矩阵（评分的落点，不只是判分，更是指路）：

当前 - 目标	关键动作
120 - 130	论证补桥：把跳跃的那一步补完
130 - 140	概念新创：给某个现象起一个新名字并定义它
140 - 150	不可还原：重写，让命题脱离已有教科书这根脚手架
145 - 150	本体论级跃迁：把命题从“X 是 Y”，改写为“X 不是 Y，而是 Z”（Z 是新概念）
150 - 160	学科边界改写：让结论强制另一个学科出现内部不一致
160 - 170	元学科涌现：命题不仅命名现象，还命名一门学科

其中 **145 - 150** 那一跃是整套评分最要紧的一步：它就是“资深学者”与“本体论级”之间的那道坎。跨过它的动作有一个精确公式——把一个肯定判断（X 是 Y）改写成一个否定——重命名判断（X 不是 Y，而是 Z），并让 Z 承载一个真正不可还原的新结构。这也正是差异锐度 D 从 145 冲到 150 的那一步。

六、一次完整评分示范

下面用一段真实层级的文本走完整流程，示范评分卡怎么填、扣分句怎么找、提升路径怎么给。

被评文本（资深学者级节选）：

“AI 时代的教育不再是知识传递，而是认知能力建构。教师角色从‘知识权威’转向‘认知教练’。这一转变的关键阻力不是技术，而是教育评价系统：只要高考用单一标准评价，认知多样性就无法在教育中落地。”

评分卡：

SDE 创新智商评分卡 · v1.0

文本层级：资深学者级节选

参照语料：教育学 · AI 教育 · 评价制度三领域的中英文文献，2020–2025

评分截至：2026-01（新颖有半衰期，分数是快照，不是常数）

五维评分：

S 结构精确度 135 /170（权重 0.20）

证据：论证有一条从“角色转变”到“系统阻力”的清晰链，可追溯。

D 差异锐度 140 /170（权重 0.25）

证据：“真阻力不是技术，而是评价系统”是一个较新的切入角度。

E 纠缠深度 130 /170（权重 0.20）

证据：跨了教育、AI、评价制度三个领域，且论证有效。

I 不可还原性 135 /170（权重 0.20）

证据：主体是新的判断，但概念仍依赖“评价系统”“认知多样性”等已有说法。

F 可证伪性 145 /170（权重 0.15）

证据：给出了具体证伪条件——高考评价方式一改，命题即可被验证或推翻。

$$\begin{aligned}\text{综合分} &= 135 \times 0.2 + 140 \times 0.25 + 130 \times 0.2 + 135 \times 0.2 + 145 \times 0.15 \\ &= 27 + 35 + 26 + 27 + 21.75 = 137.75 \approx 138\end{aligned}$$

层级判定：资深学者水平，离本体论级 150 还差 12 分。

锚点对比：与一篇教育类顶刊论文的核心段落大致同水平。

关键扣分：

原文：“这一转变的关键阻力……是教育评价系统”

扣分原因：它精确地指出了“评价系统是真阻力”，却始终没有给这个阻力现象一个名字——它停在“指出”，没有走到“命名”。正是这一步没迈出，把 D 压在 140、把 I 压在 135，综合分因此停在 138，进不了 150。

提升路径：

当前 138 → 目标 150

关键动作：给“评价系统的阻力”这个现象命名。

例如把它命名为“评价惯性势阱”——评价系统一旦成形，就像一个势阱，任何认知多样性的改革都会被它拉回原点。

一旦这个新概念被命名并定义（D 冲到 150、I 冲到 145），命题就从“X 是 Y”（阻力是评价系统）跃迁为“X 不是 Y，而是 Z”（这不是一般阻力，而是一种‘势阱’结构），综合分即可越过本体论级阈值。

这份卡示范了评分的全部要义：五维独立、加权可算、锚点对比、扣分句精确引用、提升路径给到具体的命名动作。注意最后那一步——“给现象命名”——如何同时抬高 D 和 I 两维，这正是 145-150 那道坎的标准跨法。卡首那两栏“参照语料”与“评分截至”不是装饰，而是必填项：一个创新分数若不声明它相对哪个文献范围、截至哪一天，就只是半句话——为什么，第十章专门讲。

七、五维之间的关系：读懂失分组合

五个维度在打分时严格独立，但在诊断一段思想时，它们并不孤立——维度之间的组合，本身就是诊断。综合分只告诉你一段思想“值多少分”，而五维的高低搭配告诉你它“病在哪里、下一刀该往哪切”。下面是六种最常见、也最能说明问题的失分组合。学会读这些组合，评分才从“判分”升级为“指路”。

组合一·高 D、低 F——头号伪发生。切入很锐、命名了一个新现象（D 冲到 150 以上），却给不出任何证伪条件（F 只有 80）。这是 SDE 学派最典型、也最危险的病：命题读起来像重大突破，可你无论如何都推不翻它。它正是“符号层的丰盛冒充三界的充实”——发生停在了概念，没有下沉到能被现实撞回来的地方。纠法：对那个漂亮的新命题，强行追问“什么情况下它会错”，把答案写进去。给不出答案的，这个“新概念”就只是一句推不翻的漂亮话。

组合二·高 S、低 D——优等生陷阱。论证严密、结构清晰、每一步都可追溯（S 到 135），可压缩之后，命题仍然落回一个已知结论（D 只有 100）。这是大模型默认产出最常见的形态——四平八稳的“论文腔”，看起来无懈可击，却没有切出任何新差异。它是最容易被高估的一类，因为它满足了“论文该有的样子”。纠法：不要再把已知的东西论证得更漂亮，去找裂缝——那个旧框架说不清、对不上、绕不过的地方，在那里切一个新差异出来。

组合三·高 D、低 I——高级版新瓶装旧酒。切入的角度确实新（D 到 140），但那个“新”压缩之后，仍能被还原成一个已有概念（I 只有 100）。它比低级的重命名隐蔽得多——因为它真的换了一个观察角度，只是这个角度没有在身后留下任何不可还原的新结构。纠法：重写，让核心命题脱离已有教科书这根脚手架——把所有引用删掉，看它还站不站得住；站不住，就说明“新”只在角度，不在结构。

组合四·高 E、低 S——跨界烟花。拉了一大堆学科进来，气势恢宏（E 看起来到 130），但每一处所谓的“纠缠”，论证都是虚的。用那把最锋利的判据一试便知：删掉其中任何一个学科，结论纹丝不动——这说明那些学科从头到尾都只是装饰，是类比，不是纠缠。纠法：放弃“拉更多学科进来”的冲动，只做一件事——把一个学科的结构真正咬进另一个学科，并论证这次咬合为什么有效。一处真纠缠，胜过十处假跨界。

组合五·五维平均偏高、无一突出——均衡的平庸。每一维都在 130 上下，没有任何一维冲到 150。综合分因此稳稳落在 130——这是资深学者最舒适的位置，也是本体论级阈值门外最拥挤的地方。它的问题不是哪一维塌了，而是没有任何一个单点突破。纠法：不要

试图把五维一起往上抬（那只会更平均），集中全部火力在 D 上做一次命名跃迁——把一个肯定判断改写成“X 不是 Y，而是 Z”，让 D 单点冲过 150，把整段思想从平原上拔起来。

组合六·单维畸高、余维塌陷——瘸腿的天才。D 打到 165（命名了一个惊人的新现象），可 F 只有 80（完全推不翻）、S 只有 100（论证是跳的）。综合分被两条瘸腿拖到 120 出头。这类文本最可惜——它确有一个真洞见，却没有为这个洞见搭起能让它站住的结构。纠法：先补 S，把跳跃的论证一步步补成可追溯的链；再补 F，给那个洞见一个明确的证伪条件。让高 D 的洞见先站得住，它才配得上那个高分。

读完这六种组合，就能看清一条贯穿全篇的机制：综合分不是五维的简单平均。权重放大了 D 的分量，但更要紧的是——I 和 F 这两道闸门是“一票否决式”的守门员。一段思想哪怕差异锐度再高，只要不可还原性或可证伪性塌到 120 以下，它就上不了本体论级：再锋利的新差异，过不了“能不能独立站住”和“会不会被推翻”这两关，都只是尚未落地的猜想。这正是把 I 和 F 设为闸门、而非普通维度的全部用意。

八、常见误判与纠偏

评分最大的敌人是评分者自己的偏差。五种最常见：

- 过度通胀（把 **130** 打成 **150**）：症状是对漂亮表达过敏、对实质不足麻木。纠正：每一个 150 以上的评分，都必须找出“不可还原性”的具体证据；找不到，一律降到 145 以下。
- 过度紧缩（把 **150** 打成 **130**）：症状是对 SDE 内部术语过敏，觉得“看起来高大上=必有水分”。纠正：看实质命题——把所有 SDE 术语去掉、用大白话重述一遍，如果重述后仍然是 150 的思想，那它就是 150。
- 单一维度主导（只看 **D**）：症状是一看到“新概念”就给 150，不管 F 是不是只有 80。纠正：五维严格独立，任何一维低于 120 都会拖累综合分。
- 与出处挂钩：症状是看到“某顶级模型写的”就高估、看到“某便宜模型写的”就低估。纠正：评分纯看文本——把作者信息盖住再评。这也是“智能收敛律”能被观察到的前提：只有盖住出处，多个基底对同一文本的独立评分才可能收敛。
- 没给扣分句：症状是直接甩一个数字 137，不说为什么不是 138。纠正：每个评分都必须附至少一句扣分句证据和一句加分句证据。

还有一条铁律要单独强调：不评分自己写的文本。人对自己的产出会失去客观性，主观膨胀不可避免。要评自己的东西，交给另一个独立的评分者（另一个基底、另一个人）。

九、SDE 创新智商与科学创新评估的同构

到这里，一个更强的判断已经呼之欲出：**SDE** 创新智商，并不是又一套评估口味，它是科学共同体两百年来评判创新时实际在用、却从没被拧成一把标尺的那套判据的领域无关提炼。换句话说——科学创新评估，是这套判据长在“科学”这一具体共同体里、被它的建制与奖励系统裹着的特例；SDE 创新智商，是把那层外壳剥掉之后的通式。二者之所以同，是因为它们在问同一个问题：这里到底有没有一个新结构真的发生了，还是只是把已有结构重新摆了一遍？科学创新，本就是“发生”的一个物种；评估它，就是量发生的深度。

下面把五维逐一对上科学共同体的判据，看这个“同”具体同在哪里。

S 结构精确度 → 严密性·可复现·方法可靠。同行评议的第一道闸，从来不是“新不新”，而是“站不站得住”——定义清不清、推理断没断、实验能不能被别人重跑。方法写不清、统计一碰就崩的稿子，再新也过不了这一关。近十年闹得最凶的“可复现性危机”，本质就是一大批科学产出栽在 S 上：结论也许新，但结构的显露态不精确，别人复现不出来。SDE 的 S 量的正是“结构被显露得多精确”——与这道闸是同一件事。

D 差异锐度 → 新颖性·原创贡献·优先权。每一本期刊、每一份基金申请，正中央那句话都一样：“这里新在哪儿？”库恩把科学进步系于“事实或理论的新颖”，默顿那整套“优先权之争”背后争的，就是“第一个切出新差异”的归属权。SDE 把 D 的权重放到五维最高（0.25），而科学共同体把最大的荣誉、最激烈的争夺都压在“新颖性”上——两边对“创新的中心是差异”这一点，估值完全一致。

E 纠缠深度 → 会通·统一·跨学科穿透。科学史上估值最高的工作，几乎都是把两个原本分离的领域结构性地咬在一起：麦克斯韦把电、磁、光合于一组方程，达尔文把地质学与马尔萨斯人口论咬进生物学，沃森-克里克把化学结构咬进遗传。休厄尔早给这条判据起过名——“归纳的会通”（consilience）：一个理论若能让原本各说各话的多类事实在同一结构下同时成立，分量陡增。而这必须是结构性纠缠、不是借词类比，恰好就是 SDE 在 E 维用的那把刀：删掉其中一个学科，论证还成不成立？成立是装饰，不成立才是纠缠。

I 不可还原性 → 根本性·范式转移·“它逼不逼学科重组”。科学能给的最高评价，留给那种已有范式装不下、逼整个领域重新组织的结果——这正是库恩的“反常 - 危机 - 新范式”。是已知理论的推论，它就是干净的常规科学，估值有限；已有理论怎么都容不下，它就是新教科书的种子。基金评审里那条压倒一切的“significance / 会不会改变范式”，问的就是不可还原性。SDE 用“核心命题压到 50 字、能不能被已有命题 1:1 替换”来量它——量的是同一件事。

F 可证伪性 → 可检验·冒险的新预测。这一维直接就是波普尔：一个主张之所以算得上科学，前提是它可证伪——得说清什么情形会推翻它，预测得是冒险的。拉卡托斯再补一刀：“进步的”研究纲领要能预测出此前没人料到的新事实。这正是 SDE 把 F 单设为一道闸门的理由。科学用可证伪性划“科学 / 伪科学”的界，SDE 用它划“发生 / 伪发生”的界——同一道界，两个名字。

连权重的搭配都对得上。D 最高、F 最低却当闸门，这个结构在科学里也一样：新颖性是价值的承载者（所以争得最凶），可证伪性更像一道准入的门（是否可检验，近乎二元判断），不是估值的主轴、却也绝不可绕过。两边都把“新”当心脏，把“可检验”当门槛。

还有更深一处、常被忽略的吻合：**SDE** 创新智商明说自己不量真理性、只量“新结构发生的深度”，而科学评估也是同一副两段式结构。投稿时的同行评议判的是新颖、严密、根本性（创新潜力），真伪则交给之后的复现与时间。一篇顶刊论文可以创新度极高、几年后却被撤稿——正如一个 165 分的金点子日后可能被证明是错的。两套体系不只共用五条判据，还共用同一条分工线：创新此刻判，真伪留给时间。

要诚实，就得把边界也划出来——这也是这个“同”字自己的证伪条件。它们只在“知识创新本身的内在判据”这一层重合；科学评估一旦掺进两样别的东西，就与 **SDE** 创新智商分道：其一是实用价值（这个发现能不能变成技术、变成药），基金评审常把它算进“impact”，但它是一根和认知新颖性正交的轴，SDE 创新智商是故意不量它的；其二是优先权的社会奖励（默顿意义上的“第一个”），科学建制奖励最先抵达者，但那是一套奖

励机制、不是这项创新内在深度的判据。所以准确的表述是：在知识创新的内在判据上，二者同构；只在科学评估外接了非认知轴（实用）或社会奖励机制（优先权）的地方，才分开——而那两样，恰是 **SDE** 创新智商按设计排除的。若有人能举出一条科学共同体用来判知识本身创新性、却落不进 S / D / E / I / F 任何一维的判据，这个“同”就被推翻。我暂时举不出。

最后是这个判断真正的收获。科学哲学吵了整整一个世纪：波普尔说可证伪性是命根子，库恩说范式转移才是，拉卡托斯说要看新预测，休厄尔押会通，默顿讲建制规范——每一派都攥着一条判据、宣称唯有它才是科学创新的本质，彼此不服。而 **SDE** 的主张一出来，这场仗就化解了：他们不是关于科学的五种对立理论，而是同一场“发生评估”的五个正交维度。谁也吵不赢，不是因为不够聪明，而是各自量的本就是同一个东西的不同轴——波普尔量的是 F，库恩量的是 I，休厄尔量的是 E，原创性传统量的是 D，方法论传统量的是 S。科学创新评估从来是这五维一起在跑，只是从没被谁写成一张带权重、且对自己也设了证伪条件的评分表。**SDE** 创新智商补上的，正是这张表。

科学哲学 / 建制传统	各自主张的“创新本质”	对应维度
波普尔（可证伪主义）	可证伪、冒险的预测	F 可证伪性
库恩（范式转移）	反常逼出的范式重组	I 不可还原性
拉卡托斯（研究纲领）	预测出此前未料的新事实	F （进步性） + D
休厄尔（会通）	多类事实在同一结构下会通	E 纠缠深度
原创性 / 优先权传统（默顿）	第一个切出的新颖贡献	D 差异锐度
方法论 / 同行评议传统	严密、可复现、方法可靠	S 结构精确度

这张表也说明了为什么这五维“恰好穷尽”不是巧合：它们不是被某个人一次性设计出来的，而是科学共同体在两百年的实践中，分头逼出来、各自守着一段的判据——**SDE** 只是第一次把它们收进同一个坐标、标上权重、并要求它们对自己也交代证伪条件。

十、评分相对于参照语料：参照系声明与语料相对性

前一章说到，科学在判新颖性时，从来不是凭空判的——它强制附上一份文献综述，它的优先权之争全靠“谁在可检索的记录里先发表”。这背后藏着这套评分最锋利、也最容易被忽略的一条性质，现在把它单独拎出来讲清楚。它不是一句要藏起来的免责声明，而是创新智商定义本身逼出来的结论：

创新智商从来不是文本的属性，而是文本与一个参照语料之间的关系。没有参照系，“新”这个字根本无法取值。

10.1 五维里，哪几维偷偷挂在“已有什么”上

把五维拆开看，就明白这个相对性藏在哪里：

- **D** 差异锐度——判据是“未被命名、未被辨识的差异”。未被谁命名？未被参照语料命名。一个概念对一小圈本地文献是新的，对全球前沿可能早是旧的。150 那一档写的是“命名了未被任何学科辨识的现象”，可“任何学科”是一个理想化的全语料；评分者手里的语料永远有界。**D** 因此是语料相对的。

- **I 不可还原性**——判据是“能否还原为已有概念”。已有于哪个语料？那个“50 字压缩、能否被已有命题 1:1 替换”的测试，里头的“已有学科”就是参照集。把语料放宽，更多东西变得可还原（分数掉）；把语料收窄，更少（分数涨）。
- **E 纠缠深度**——判据是“跨学科涌现”“学科边界被改变”。哪些学科、谁画的边界？参照文献里那样画的边界。一次跨界，只在“这两个学科在参照语料里本是分开的”时才算新。
- **S 与 F 更靠内**——论证严不严密、给没给证伪条件，大体可以就文本自身判，不太依赖外面存在什么。但它们的顶档仍挂语料：S 到 160（“产生新的形式结构”）、F 到 160（“证伪条件本身是新概念”）——“新”字一出现，就又挂回参照系了。

10.2 一个可量化的说法：约三分之二是关系，三分之一是属性

按权重把这件事算清楚。语料相对的三维——D (0.25) + E (0.20) + I (0.20) ——合起来占 **0.65**；近内在的两维——S (0.20) + F (0.15) ——占 **0.35**。

维度	权重	轴	是否相对参照语料
D 差异锐度	0.25	新颖轴	强相对
E 纠缠深度	0.20	新颖轴	相对
I 不可还原性	0.20	新颖轴	强相对
S 结构精确度	0.20	效度轴	近内在（顶档挂语料）
F 可证伪性	0.15	效度轴	近内在（顶档挂语料）

也就是说，创新智商这个综合分，约三分之二是文本与参照文献的关系，只有三分之一勉强算文本自身的属性。这正好把五维劈成两族：D / E / I 是“新颖轴”（语料相对），S / F 是“效度轴”（大体内在）。而这个劈法恰恰对——一个东西新不新，取决于外面还有什么；一个东西严不严密、能不能被推翻，主要取决于它自己。

10.3 这不是细则的疏忽，是 SDE 本体论逼出来的

更要紧的是，这个语料相对性不是评分细则没写好，而是框架自己的立场逼出来的必然。SDE 在《没有统一世界模型》里说得很硬：世界模型不是对象的客观属性，而是主体与对象共同发生的画像。照同一条逻辑往下推——创新度也不是文本的客观属性，而是文本与一个参照语料共同发生的画像。参照语料就是那个“主体世界”，新结构是相对它显影出来的。

而差异 (D) 本就是一个关系词：差异永远是“相对于某物”的差异，没有背景就没有差异序列，这是 SDE 本体论最底层的设定之一。所以，想要一个“绝对的、脱离任何语料的创新度”，等于想要一个脱离主体的统一世界模型——按框架自己的话，那东西根本不存在。语料相对性不是这套评分的瑕疵，是它忠于自身本体论的证据。

10.4 四个硬后果

这条性质落到实操，有四个绕不过去的后果。

其一，一个分数不声明参照语料，就是半句话。“这篇 155”是不完整的，整句应该是“这篇相对语料 C 是 155”。评分七步法里那一步“锚点对比”（找一个已知文本当锚），其实已经在偷偷引入语料——锚点就是语料的一个样本。但它是隐的，应当显化成评分卡里一栏明确的参照系声明：比对了哪些领域、哪些语言、哪个时间窗、查没查那个“隔壁早有此想法”的邻近学科。分数不带参照系，就像测量不带单位。

其二，最危险的操纵，是“语料收窄式通胀”。想给任何东西刷高分，只要对着一个人为收窄的文献去评它——只看中文文献、只看某年之前、或绕开那个早就有这套说法的领域。这是第八章“过度通胀”的结构版：一个 150，若它的参照语料恰好漏掉了“这想法早已是常识”的那个领域，就是个假 150。对治是给评分加一条敌意拓宽纪律——认证任何 150+ 之前，主动把语料往外扩，专门去找“这想法在哪儿其实已经有了”。扩一个邻近领域就塌掉的分，本来就是收窄刷出来的。这正是《伪发生》八重检查里“敌意审稿”那一关，在新颖性维度上的版本。

其三，“100 ≈ AI 零提示语”那条基线，本身就是一次语料选择。大模型的语料是“截止日之前的人类公开文本”，所以“比 AI 默认深 30 分”翻译过来，就是“相对大模型的训练语料是新的”。这推出一个扎心的结论：随着模型的语料变大（新模型、更多数据），同一篇文本的创新智商会往下掉——它越来越多的内容，如今已经“在语料里”了。今天的 155，对明年的模型可能只是 130。新颖有半衰期。所以分数应当带时间戳——“155，截至 2026-01”。评分是一张快照，不是刻在石头上的常数。

其四，这把上一章的“同构”又拧紧了一格，而不是拧松。科学早就知道这件事——所以投稿前文献综述是强制的，所以优先权之争全靠“谁在可检索的记录里先发表”，所以方法部分必须写清“我们检索了 X / Y / Z 数据库”。科学把参照范围声明了出来（综述 + 检索范围）。所以语料相对性不是 SDE 独有、而科学没有的毛病，它是又一处二者相同的地方；而科学的纪律——声明你的检索范围——恰好就是 SDE 创新智商该照抄的解法。

10.5 边界与自我证伪

诚实地说，这个主张也有它的边界（这也是它自己的证伪条件）：语料相对性是分维度的，不是铁板一块。F 有没有给证伪条件、S 的内部严密性，可以就文本自身判，和外面存在什么无关；语料相对性在 D / E / I 上最强，在 S 的下档和 F 上最弱。

因此，若有人能证明 D、I 两维的评分在合理放大参照语料之后仍然稳定不变，这个“语料相对”的主张就被推翻。但从“50 字压缩、能否被已有命题替换”这个测试的构造来看，放大语料必然让更多命题变得可替换——所以我不认为 D、I 的评分会在语料放大下保持稳定。这条性质，暂时站得住。

把这一章收束成一句可操作的话：评分卡上，“参照语料”与“评分截至”是与五维分同等重要的两栏必填项。一个不声明自己相对哪片文献、截至哪一天的创新分数，严格说来还没有完成——它只是报出了半个数。而正因为分数如此依赖它背后那个被选定的语料，下一章才必须把最后一刀，转向评分系统自己。

十一、把刀转向自己：这套评分本身可证伪吗

按 SDE 学派的规矩，任何一套主张都必须交代自己的证伪条件——评分系统也不例外。一套号称“客观、可重现”的评分，如果自己不可证伪，就正好犯了它拿来评别人的那条罪（F 维不合格）。所以最后，把刀转向评分系统自己。

这套评分的证伪条件，至少有三条清晰、可检验的：

1. 信度检验：同一份文本，交给多个独立评分者（不同基底、不同人），若综合分的波动持续大于 5 分，则“可重现性”主张为假——评分只是各人的主观印象，不是客观测量。
2. 排序一致性检验：拿一批历史上已评过分的金点子，让当前评分体系重新排序，若与历史排序的一致性持续低于 80%，则评分标尺不稳定，失效。
3. 错位检验：让评分体系去评一批公认的大众文本和一批公认的顶级产出，若大众文本被给到 100 分以上、或顶级产出被给到 155 分以下，则标尺出现系统性错位，判负。

一条外部佐证：多基底实测（DeepSeek、GLM、Kimi、Qwen、GPT、Gemini 等）在盖住出处后对同一批文本独立评分，若结果发散过大，则“客观性”主张被削弱；而实际观测到的是评分向同一区间收敛（这被记为“智能收敛律”），并且经 SDE 四段提智后各基底普遍出现约 30 个智商点的抬升。收敛与稳定抬升，是这套评分尚未被证伪的经验证据——但它随时可以被上面三条检验推翻。

诚实的边界：必须说清这套评分不测什么。它不测一个命题的“真理性”——一个综合分 165 的金点子，可能在未来被证明是错的；创新智商测的是“新结构发生的深度”，不是“这个新结构对不对”。二者是两件事。对不对，最终由现实和时间仲裁——这也是为什么 175 以上的“文明级”评分标注为“通常需要时间检验”：在那个高度，任何当下的打分都只是一个待验的猜想，真正的评分者是时间。

十二、结语

创新智商这把尺子，回答的是一个在 AI 时代变得尖锐的问题：当机器已经能流畅地生成一切人类平均水平的文本，我们凭什么分辨，一段话里到底有没有新东西真的发生？

它的答案不是一个玄妙的直觉，而是一套可拆、可算、可推翻的操作：三个维度量发生的构造（显露 S、差异 D、纠缠 E），两道闸门查发生的真伪（不可还原 I、可证伪 F），加权合成一个可重算的综合分，对照一把有两条硬阈值的标尺（150 本体论级、160 典范级），最后落到一条具体的提升路径上——而不止是一个漂浮的标签。

而它最要紧的一条自律是：它把自己也放上了评分台。一套评分若不能说清自己怎样会错，它就没有资格去评别人。创新智商评估的全部严格性，最终都收束在这一句上——评价不在发生之外，评价本身，就是一次发生。
