

机器犹豫的地方，正是你自由的地方

钢琴演奏转录的残余地图，与一个惩罚相似的质量指标

少 敏

SDE 学员专栏 · 音乐技术 · 发表于 2026 年 7 月 17 日

摘 要

一份音符起音 F1 达 0.98 的钢琴转录，可以让教师无话可说：它把学生弹得最像自己的地方，全部整理成了整齐的时值与干净的踏板。本文主张这不是转录不够好，而是质量的定义搞反了方向——这个领域的全部指标（F1、平均绝对误差、感知相似度）都单调地奖励“像”，而对一份要被人拿去用的转录，判准应当相反：**它的质量该用它能让别人生出一段多不像它的新演奏来衡量**。支撑这一判准的是本文的核心判断：**转录模型犹豫的地方，正是演奏者自由的地方**，而且同源——两者都是“音乐是一串离散可分离事件”这个假定失效的区域，它对机器是噪声，对人是全部的音乐。据此，残余地图不是失败的副产品，而是这段音乐允许选择的地区的地图，而现有系统正把它当垃圾扔掉。文章与西格（1958）的规定性／描述性之分划界——他的解药是纪律，本文的是架构；而机器带来了他不可能拥有的第三个对象：逐音置信度，它正是他无法测量的那个东西的测量。又与古德曼（1968）的合规类划界——MusicXML 是他的合规类的一个文件格式实现，F1 是他的完美要求的一个统计实现，而他判为非构成性的那些剩余，是本文的全部主题。工程上给出双轨表征、复合文档、六层保真阶梯与三道责任关口；评价上把可复生性定义为二维判断（相似度 × 可说明性）；最后给出四条证伪路径，其中核心检验可在 ASAP 公开数据集上一个下午跑完。

关键词：残余地图 离散性假定 可复生性 主体复生 双轨表征 自动音乐转录 MusicXML

引言

一个 F1 为 0.98 的失败

设想一份转录结果：音符起音 F1 达到 0.98，力度平均绝对误差极小，踏板事件识别准确，谱面整洁，可以直接进打谱软件。按这个领域的全部标准，它是一次成功。

把它交给一位钢琴教师，让他据此指导学生。他会发现自己无话可说。

不是因为它错了。是因为它对得没有留下任何余地——它把那位学生弹得最有意思的地方、弹得最像他自己的地方，全部整理成了整齐的时值和干净的踏板。谱面上什么都在，唯独没有那个人。

本文要论证的是：这不是转录做得还不够好，是**质量的定义搞反了方向**。这个领域的全部指标——F1、平均绝对误差、感知相似度——测的都是同一件事：产物有多像原演奏。而对于一份要被人拿去用的转录，本文主张一个相反的判准：**它的质量，该用它能让别人生出一段多不像它的新演奏来衡量**。

太像，是失败的一种。

支撑这个判准的，是本文的核心判断，它同样反直觉：**转录模型犹豫的地方，正是演奏者自由的地方——而且不是巧合，是同一个原因**。当前的工程实践把模型的置信度当作内部诊断数据，产出时扔掉，只交付最可能的那个答案。它扔掉的，恰恰是唯一有教学价值的东西。

全文八节。第一节说明相似度这把尺量错了什么。第二、三节与两个祖先划界——一九五八年的西格已经说了一半；一九六八年的古德曼则把本文的全部主题判为不属于作品，而这个领域在不知情的情况下把他的立场实现成了一个文件格式。第四节给出标本化悖论。第五节论证残余地图为何是自由的地图，这是本文的枢纽。第六节把它落成工程。第七节定义可复生性并给出那个惩罚相似的指标。第八节写明这套判断如何可能失败。

先声明一句，它决定了本文的分寸：**本文是一份研究框架与实验设计，不报告已完成的实验结果**。文中的指标、条件与假设属于待验证的设计，不能被表述为已有结论。这不削弱理论主张，只是防止用未来式设计冒充已有证据。第八节的检验一可以在一个公开数据集上、用一个下午跑完，它是本文最该被优先执行的部分。

第一节

相似度是一把量错了东西的尺

把录音转成乐谱，通常被描述为"从音频中提取乐谱"。这个说法已经预设了一件事：乐谱本来就藏在音频里，算法只需把它找出来。若此说成立，高质量转化就只是更低的错误率、更细的力度、更准的踏板。

但演奏不是藏在声波里的乐谱。它是在具体时空里展开的身体事件：手指下沉的速度、腕部释放的时机、踏板半踩的深浅、前一个音的尾振、房间的反射、演奏者对下一句的期待。MusicXML 能记录其中一部分，不能把这个过程原封不动搬进文本。

这条链路上至少混着四个不同的问题，它们之间不存在自动等价。**识别问题**：弹了哪些音，何时开始，何时结束。**记谱问题**：这些事件该被组织成什么拍号、小节、声部、连音、装饰音、踏板符号。**诠释问题**：某个偏离标准谱的时值究竟是错误、身体限制，还是表达性选择。**再合成问题**：怎样用新的音源与声学空间重建一段可听的声音。音符识别准确，不等于记谱可读；谱面整洁，不等于表现忠实；表现参数相似，也不等于声音身份相同。

一个常见的反驳是：这些只是当前技术不够先进，未来模型可以记录更高分辨率的时间、更丰富的力度曲线、更细的踏板深度。这个反驳对工程误差有效，对表示层之间的结构限制无效。**只要目标格式仍要求把演奏表达为符号对象，系统就必须决定什么算一个可记录单位**。单位可以越划越细，不会因此消失。更要紧的是：提高精度提高的是记录的分辨率，不是记录的方向——**把一把量错了东西的尺刻得更细，量出来的还是错的那个东西**。

本文因此不追问"怎样无损"，而追问三件事：转化保留了演奏的哪些维度，遗漏了什么；系统通过哪些切分制造可计算对象，又如何把表达性变化写成错误；以及，产物进入教学与再演奏之后，它究竟让下一个人能做什么。

第二节

一九五八年已经说过一半

本文的第一个祖先，比这个领域的全部技术都早三十年。

一九五八年，查尔斯·西格在《规定性与描述性的音乐书写》中指出，音乐书写实践里潜藏着三重危害。第一重是：**假定音乐的完整听觉参数，可以被一个部分的视觉参数——一个只有二维的平面——所表示**。第三重是：未能区分音乐书写的规定性用法与描述性用法，即"一份关于某乐曲该如何被弹响的蓝图"与"一份关于某次演奏实际如何发声的报告"。

这两句话，几乎就是本文两个核心构造的内容。"双轨表征"——规范谱负责结构交流、演奏轨负责保存过程——正是西格那条分界的工程实现。任何把它当作新发现的说法都不成立；在民族音乐学里这条分界早已是常识，而且连本文所引的那篇转录综述的作者，也出现在西格那篇论文的引用网络里。

所以问题必须换成：**在西格已经说完的地方，还剩下什么？**

剩下两样，而它们都不是修辞。

其一，**西格的解药是纪律，本文的解药是架构**。西格说：分清你在用哪一种。这是对写谱的人的要求，靠自觉执行。可当写谱的是一台机器、而使用者是一个看不见机器内部的教师时，自觉无处安放——教师拿到的是一份已经完成的谱，它长得既像蓝图又像报告，没有任何东西告诉他哪一部分是哪一种。本文的主张因此不是"分清"，而是**别选：两个都产出，作为同一个版本化包里两份独立的、互不覆盖的产物**。规范谱里的整齐时值不许抹掉演奏轨里的 rubato，演奏轨里的细碎时间也不许强迫乐谱变得不可读。分清是一种态度，双轨是一个数据契约——后者可以被检查，前者不能。

其二，**也是要紧的那一样：机器带来了西格不可能拥有的第三个对象**。西格的描述性记谱，仍然是一个人写下的报告。写

的人当然知道自己在哪里拿不准，但那份拿不准随着报告的完成而蒸发了，纸上只剩结论。转录模型不同：它**逐音输出自己的置信度**。这是西格的世界里不存在的量——它回答的不是"这段音乐是什么"，而是"证据在哪里不唯一地决定一个答案"。

这个差别的分量，第五节才会完全显出来，但可以先说一半：**西格说记谱假定了一个它兑现不了的东西，他没办法测量那个假定在哪里失效。转录模型的置信度，就是那个测量**。六十八年之后，一句哲学上的抱怨第一次变成了一张逐音的、可计算的地图。

第三节

一九六八年把本文的主题判为不属于作品

第二个祖先更危险，因为这个领域已经把他的立场实现成了产品，而且不知道自己选了边。

一九六八年，纳尔逊·古德曼在《艺术的语言》中给出了音乐作品同一性的经典分析：乐谱定义了一个合规类，一次演奏是否为该作品的实例，取决于它是否完全符合乐谱的规定。他把后果说得毫不含糊：**"最悲惨的、没有实际错误的演奏"是该作品的一个实例，"而最辉煌的、有一个错音的演奏不是"**。他的理由也不是随口：若允许最小的偏离，作品保存的全部保证都将丧失——经由一连串的单音增删改，我们可以从贝多芬第五一路走到《三只瞎老鼠》。

现在把这段话与一个技术事实并排看。

MusicXML 精确地编码了古德曼所说的那些构成性属性：音高、时值、记谱力度、声部。它把其余一切——触键质地、微时序、踏板的浑浊、声场的粘性——留作"未指定或在限度内可变"。而这个领域的评价体系，是逐音计数、一个错音即一次错误。

MusicXML 是古德曼的合规类的一个文件格式实现；F1 是他的完美要求的一个统计实现。这个领域把一九六八年的一个哲学立场，实现成了一种交换格式和一套指标，并在此后的半个世纪里把它当作中立的工程事实使用。

本文与古德曼的分界，不在于说他错了。就他自己的问题——"这还是不是同一部作品"——他的答案有力，那个贝多芬到瞎老鼠的论证站得住。分界在于：**他的问题不是唯一的问题，而这个领域把他对甲问题的答案，当成了对乙问题的答案**。

甲问题是作品同一性。乙问题是演奏的传承：一个后来者能否凭这份材料，理解前一次演奏在做什么，并据此弹出一段属于他自己的音乐。

对甲问题，古德曼把演奏细部判为非构成性的——它们不属于作品。**对乙问题，那些非构成性的剩余，是全部的主题**。一个学生要从老师的录音里学到的，恰恰是古德曼定义上不属于作品的那一堆东西。

所以本文不推翻古德曼，本文指出一处错位：**这个领域用一把为"是不是同一部作品"打造的尺，去量"能不能传下去"**。量出来的东西不是错的，只是与要问的事无关。而这把尺量得越准，那处错位就越难被看见——这正是下一节。

第四节

标本化：越成功越致命

西格把第一重危害描述为一个假定，一个可以被指出、被纠正的错误。本文的判断是：它不是一个静态的错误，而是一个**随技术进步而加剧的回路**。

这个回路有四步。第一，为提高演奏的可保存性、可传播性与可比较性，系统必须增强结构化、离散化与标准化。第二，结构化越成功，输出越像一个完整、自足、确定的对象。第三，输出越像完整对象，使用者越容易忘记它的条件性、推断过程与排除项。第四，系统越成功地履行"保存整体"的表面承诺，越彻底地破坏使用者对演奏整体性的认识。

压缩成一句：**保存的技术成功，制造了"整体已经被保存"的认识失败**。

这就是本文与西格的第二处分别。西格诊断的是一个假定；本文诊断的是这个假定的**斜率**——转录做得越好、谱面越干净，那个假定越有说服力，越少有人去查它。一份 F1 为 0.6 的转录骗不了人，它满纸都是可疑之处；一份 F1 为 0.98 的转录

几乎无法不骗人，因为它剩下的 0.02 恰恰藏在最需要被看见的地方。**技术进步不是这个问题的解药，是它的燃料。**

必须立刻区分两件事，否则本文会被读成反结构化。**结构化记录本身是必要的**：没有它，音乐经验无法跨越身体与时间；标准乐谱之所以能传几个世纪，正因为它舍弃了具体声场，保留了可在不同身体中重建的骨架。本文反对的不是整理，是**把整理的结果包装成完整替身**。

标准化的压力来自三处，都不是恶意。**不出错**：工程必须控制错误，于是把微时间拉回节拍格，把不稳定音头归入最近音符，把轻微装饰滤掉——这种"不出错"本身制造了一种错误，它把演奏者用来表达意义的变化修正掉了。**不浪费**：算力与存储有限，于是"对目标任务无用"的信息被判为冗余——而琴槌噪声、制音器动作、泛音云与声场粘性，正是音色与质感之所在。**不亏本**：产品必须存活，于是路线指向主流场景——古典钢琴、标准调律、清晰录音；边缘音乐、自由速度、当代技法不是无法处理，是不值得优先处理。更隐蔽的是，商业逻辑会反过来定义技术指标：评价范围之外的变化不再算失败，因为它们没有被纳入问题。

三者叠加，输入的是一段带着个人痕迹的声音，输出的是一个干净、合格、可比较、可合成的对象。它并非全无用处，只是不再是那次演奏的延续，而是一具可流通的标本。

第五节

残余地图：机器的犹豫是自由的地图

现在到本文的枢纽。

先问一个纯技术的问题：**一个转录模型在哪里失去置信度**？答案是已知的，任何做过转录的人都能列出来：密集和弦里被遮蔽的内声部；接近听阈的极弱音；踏板重叠处前音尾振与新音的混叠；自由速度段落里失效的拍点网格；时值不唯一的装饰音。

再问一个纯音乐的问题：**一个钢琴演奏者在哪里拥有选择**？密集和弦里让哪个声部浮出来；极弱音藏多深；踏板留下多少前一个和声的影子；句法的呼吸在哪里松、在哪里紧；装饰音落在拍上还是拍前。

同一批区域。

这不是巧合，而它的原因也不是什么神秘的对称。两者共享一个成因：**这些正是"音乐是一串离散、可分离事件"这个假定失效的地方**。

在这个假定成立的地方——中音区一个响亮、干净、孤立的四分音符——模型几乎不会出错，而演奏者也几乎没有余地：那个音就是那个音，弹法的空间被物理与记谱共同压得很窄。在这个假定失效的地方，声音不再干净地分解为独立可控的事件：和弦里的音互相融合，弱音沉进噪声，踏板把边界抹掉，rubato 让网格失去参照。模型在这里犹豫，是因为声学证据不唯一地决定一个记谱；演奏者在这里自由，是因为演奏的艺术**恰恰就是管理这种不可分离性**——配置发声的重量而不是逐音敲击，在可听与不可听的边缘上工作，控制浑浊的程度而不是消灭浑浊。

换句话说：**离散性假定失效的区域，同时是机器的困难区和人的表达区。它对机器是噪声，对人是全部的音乐。**

由此得到本文对残余地图的重新定义。残余地图记录低置信度音符、多个候选解释、密集和弦、弱音、踏板重叠与无法稳定量化的区域，并把每一项链接回原始音频的时间码。传统上，这类数据是内部诊断信息，产出时被丢弃，系统只交付最可能的那个答案。本文主张：

残余地图不是失败的副产品，是这段音乐允许选择的地方的地图；而那是下一个演奏者需要知道的、几乎唯一的一件事。

这也补上了第二节留下的那半句。西格说记谱假定了一个它兑现不了的东西，他无法测量那个假定在哪里失效。**转录模型的置信度就是那个测量，而残余地图就是那张地图**。六十八年之后，第一重危害第一次从一句哲学抱怨，变成了一份逐音的、可计算的、可以印在谱子旁边的坐标。

一个直接的工程推论：现有系统正把这张地图当垃圾扔掉。它们优化的方向是让地图变小——置信度越高越好，犹豫越少越好。按本文的判断，**它们在系统性地删除唯一有教学价值的那一层**。一个把所有音符都以 0.99 置信度输出的模型，如果它诚

实，那说明这段演奏没有自由；如果这段演奏有自由，那它就是在撒谎。

这个主张有一个清楚的失败条件，写在第八节的检验一：若模型的低置信度区与不同演奏者的分歧区在真实数据上不相关，本节整节垮掉，残余地图就只是一张信噪比地图，与音乐的自由无关。

第六节

工程：双轨、复合文档、六层阶梯

把上面的判断落成可实施的东西，需要放弃一个习惯：让单个 .musicxml 文件承担它无法承担的全部责任。

最低工程单元是一个版本化转化包，不是一个文件。它至少包含：原始媒体（固存并计算哈希）、未经破坏性处理的分析音频、未量化的演奏事件、经人工校订的规范 MusicXML、残余地图、标准谱对齐报告、标准回放、表现回放、人工判读记录。每一项都记录输入、模型、参数、版本与哈希，使任何变化都能被追溯到演奏、算法、人工编辑还是渲染。

双轨不许互相覆盖。这是数据契约层面的，不是态度层面的：量化后的音符位置不得写回并替换未量化时间；教师确认不得删除模型原候选；新音源的渲染不得覆盖旧版本而继续沿用同一份听感报告。只要这些对象各有标识与来源链，系统就能回答"这个变化来自演奏、模型、人工还是渲染"，并在发现错误时局部回退。

踏板必须三分。演奏轨里的"连续踏板"只表示踏板控制轨保真，不自动等同于原钢琴的阻尼器运动，更不等于最终声学效果。MIDI CC64 即使保存 0—127 的连续值，不同钢琴、传感系统与音源的响应曲线也不同。控制轨保真、乐器响应模型保真、声学效果保真是三个量，报告里不许合并。

质量是一把阶梯，不是一个分数。本文固定采用六层，每层各自报告目标、证据与限度。**源完整性**：哈希一致、解码错误、削波比例、有效信噪比、同步误差。**事件准确性**：起音 F1、带终止的音符 F1、力度平均绝对误差、踏板事件 F1、置信度校准误差——并须注明时间容忍窗，慢速抒情作品用固定毫秒阈值与按拍长比例阈值可能得出相反结论，两者都要报。**记谱可用性**：小节结构正确率、拍号调号错误、声部一致性、节奏拼写错误、教师校订操作数与校订时间——校订时间比符号距离更接近真实成本，两个错误数量相同的结果可能因错误集中而产生完全不同的修复负担。**表现保真**：局部速度曲线、相对拍点偏移、和弦内部起音散布、力度轮廓相关、踏板曲线距离——全部须在结构对齐完成之后才能比较，否则位置错配会制造虚假变化。**感知相似**：分开询问自然度、结构正确感、与原演奏相似度、个体特征可辨性、方向一致性，不合并成"像不像"。**可复生性**：下一节定义。

六层不可互相替代：音频哈希一致不代表转录正确；音符 F1 高不代表谱面可用；谱面可用不代表演奏个性仍在；盲听相似不代表对齐准确；再演奏受到启发也不证明声音被保存。

三道关口，各管一件事。可记谱关口：模型结果能否形成结构自治、可被打谱软件解析的候选——通过只意味着格式成立。可评价关口：是否存在版本一致、教师确认的参考，是否完成结构对齐——只有通过之后，才允许报告教育意义上的准确率。可使用关口：教师、演奏者或盲听者是否确认产物满足具体目的。三者分别对应机器可运行、知识可判断、实践可采用，前一道不能替代后一道。

没有教师确认且版本对齐的参考谱时，评价状态必须显式标记为 **NOT_EVALUABLE_NO_REFERENCE**。解析通过、渲染通过、下载成功，只证明管线在跑，不证明转录准确，更不等于教师认可。用可运行性遮蔽未知，是这条链路上最容易犯、也最难被发现的错误。

第七节

可复生性：一个惩罚相似的指标

现在可以给出那个反方向的判准。

可复生性：转化产物在不要求声音身份复制的前提下，使后来者能够恢复原演奏中具有教学或艺术意义的结构关系、表现方向与判断依据，并据此生成具有自身完整性的新演奏的能力。

它分三维，彼此不可替代。**结构复生**：作品骨架是否被正确恢复——证据是音高、节奏、声部、段落结构与教师校订。**方向复生**：原演奏的句法与表现牵引是否被后来者接住——证据是速度轮廓、力度方向、踏板控制、专家判断与演奏者陈述。**主体复生**：新演奏是否形成了自身的完整性。

第三维是本文与整个指标传统分道的地方，值得把话说死。在主体复生上，**"太像"是一个失败信号**。

这不是修辞。这个领域的每一个数字——F1、平均绝对误差、感知相似度——都单调地奖励相似：产物越接近原演奏，分越高。主体复生反过来：一段与原演奏高度相似、却说不出为什么这样弹的新演奏，是机械复制，它证明材料只传递了表面，没有传递依据。

但这个判准需要一个对称的护栏，否则它会奖励任意的不像。一段与原演奏很不同、却无法证明与所提供材料有关的优秀新演奏，也不能算在复合文档的功劳里——那可能只是这位演奏者本来就很好。

所以主体复生是一个二维判断，不是一个标量：**相似度 × 可说明性**。目标区域是低相似度加高可说明性——他弹得不像，而且他说得出他读懂了什么、并据此选择了别的。高相似度加低可说明性是机械复制；低相似度加低可说明性是无关；高相似度加高可说明性是理解之后选择了忠实，这也合格，但它是特例而非常态。

这里也必须限制"意愿"这个词。系统可以保存**意愿标注及其来源**——演奏者事后陈述、多次演奏的稳定模式、教师的结构解释——但不能宣称已经保存了演奏者的意愿本身。意愿只能作为多源支持的假说存在，不是能从单次声音里直接读出的实体。

第八节

这套判断如何可能失败

一套关于"看不见的东西"的理论，最方便的活法是永远正确。以下写明本文在什么情形下应当被判为错。

检验一 · 犹豫与自由的相关（核心检验，可在公开数据上立刻做）。第五节的核心主张是：模型的低置信度区与演奏者的自由区是同一批区域。这可以直接测，而且材料现成——ASAP 这类数据集把同一乐谱的多次演奏与音频、MIDI 对齐在一起。设计：其一，在每次演奏的音频上跑转录模型，取逐音置信度；其二，在对齐后的乐谱位置上，计算不同演奏者之间的分歧（微时序方差、力度方差、踏板轮廓距离）；其三，做相关。**预测：模型置信度与演奏者间分歧显著负相关——机器越犹豫的位置，人越分歧。**

若不相关，第五节整节垮掉，残余地图退化为一张信噪比地图，与音乐的自由无关，本文的枢纽随之失效。这个检验不需要教师、不需要再演奏者、不需要伦理审查，一个下午即可完成，因此它应当被最先执行。它也可能给出一个中间结果：相关只存在于某些成因（比如踏板重叠成立、极弱音不成立），那将把残余地图从一个整体概念拆成需要分别对待的若干类——这同样是有信息量的失败。

检验二 · 聚集而非均匀。预测：残余的收益集中在低置信度、密集和弦、弱音与复杂踏板区，而非均匀分布全曲。若低置信度音符在全曲上大致均匀散布，残余地图就没有定位价值。

检验三 · 分层增量。四条件逐层增加信息：A 规范谱；B 加人工校订；C 再加演奏轨；D 再加残余地图与证据入口。所有条件使用同一渲染器、音源、混响与响度标准，并控制总学习时间、材料浏览时间与教师介入时间，防止"材料更多"被误判为"结构设计更优"。预测：B 相比 A 降低结构错误与校订成本；C 相比 B 提高方向一致性；D 相比 C 缩短教师定位疑难片段的时间、提高教师间"错误—诠释—不可判定"的分类一致性。**若 C 不提高方向一致性，演奏轨的必要性受到削弱；若 D 不降低定位时间或不提高教师一致性，残余地图的实践主张失败**。本文刻意不采用"合成音频与原录音不可区分"这类几乎不可能触发的失败条件——那种条件混杂了音色、房间与噪声，永远不会开火。

检验四 · 相似度的方向。这是对第七节的直接反驳条件。预测：拿到 D 组材料的再演奏者，其新演奏对原演奏方向的理解应优于 A 组，**而与原演奏的表面相似度不应更高——若更高，可复生性的定义方向就是错的**，材料传递的只是模仿的便利，不是理解的依据。

边界声明，四条。其一，**本文是设计不是结果**：以上全部属于待执行的检验，本文没有跑过其中任何一个。其二，**适用域有限**：本文针对以音高明确的钢琴独奏为中心的链路；室内乐、声乐、预制钢琴与自由即兴涉及更复杂的音源分离与非标准音

色，不能套用同一阈值。其三，**残余地图也会成为新的权威**：一张标着"此处自由"的地图，可能反过来暗示"别处不许自由"，因此它必须允许修订与删除错误推断。其四，**主体复生的"自身完整性"最终仍靠人判断**，两位教师可以就此长期分歧；本文能做的只是要求分歧被记录、被报告一致性统计，并允许"不可判定"成为合法类别，而不是把它交给模型伪装成客观。

结语

把犹豫交出去

回到那份 F1 为 0.98 的转录。

它之所以让教师无话可说，不是因为它错了那 0.02，而是因为它把那 0.02 扔了。那 0.02 不是残余的错误，是残余的音乐——它落在密集和弦、弱音、踏板与呼吸上，也就是落在这段演奏之所以是这个人弹的那些地方。

现有的系统把置信度当作内部数据，产出时丢弃，只交付最可能的答案。这个习惯有它的道理：用户要一个答案，不要一堆犹豫。但对于一份要被人拿去理解、拿去教、拿去再弹一次的转录，这个习惯扔掉的正是全部价值。

高质量不是没有损失，是不让损失隐身。MusicXML 的最高价值也不在于成为过去声音的密封容器，而在于成为一颗开放的种子：它保存足以交流的骨架，连接尚未被符号吞没的声音证据，并标出那些它自己也无法决定的地方——因为那些地方，正是下一个人可以进去的地方。

所以，下一次交付一份转录，值得多做一件事：

把机器犹豫的地方，一起交出去。

参考文献

- Seeger, Charles. "Prescriptive and Descriptive Music-Writing." *The Musical Quarterly*, 1958, 44(2): 184–195. (规定性／描述性之分与三重危害的提出；重印于 *Studies in Musicology 1935–1975*, University of California Press, 1977)
- Goodman, Nelson. *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis: Bobbs-Merrill, 1968 (2nd ed., Hackett, 1976). (合规类与完美要求；"最悲惨的无错演奏／最辉煌的错一个音"见 1976 年版第 186 页)
- W3C Music Notation Community Group. *MusicXML 4.0. W3C Community Group Final Report*, 2021.
- Benetos, E., Dixon, S., Duan, Z., & Ewert, S. "Automatic Music Transcription: An Overview." *IEEE Signal Processing Magazine*, 2019, 36(1): 20–30.
- Hawthorne, C., Elsen, E., Song, J., et al. "Onsets and Frames: Dual-Objective Piano Transcription." *Proceedings of ISMIR*, 2018.
- Hawthorne, C., Stasyuk, A., Roberts, A., et al. "Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset." *Proceedings of ICLR*, 2019.
- Foscarin, F., McLeod, A., Rigaux, P., Jacquemard, F., & Sakai, M. "ASAP: A Dataset of Aligned Scores and Performances for Piano Transcription." *Proceedings of ISMIR*, 2020. (本文检验一所用材料)
- Cancino-Chacón, C. E., Grachten, M., Goebel, W., & Widmer, G. "Computational Models of Expressive Music Performance: A Comprehensive and Critical Review." *Frontiers in Digital Humanities*, 2018, 5.
- Holzapfel, A., Stylianou, Y., Gedik, A. C., & Bozkurt, B. "Three Dimensions of Pitched Instrument Onset Detection." *IEEE Transactions on Audio, Speech, and Language Processing*, 2010, 18(6): 1517–1527.
- Repp, B. H. "Diversity and Commonality in Music Performance: An Analysis of Timing Microstructure in Schumann's 'Träumerei'." *Journal of the Acoustical Society of America*, 1992, 92(5): 2546–2568.