arXiv:2611.32050v1 [cs.CY] 24 Nov 2026

临床医生评分示例能帮助大语言模型根据长达一小时的转录文本进行 ADOS-2 编码吗?

Zeteng Lin[1] Yikang Liu[2] Bolin Zeng[2] Mengyuan Ding[2] Xihui Wu[1]Kangyou Li[3] Chengzhe Fu[4] Xin Wang[3]

[1]香港科技大学（广州），中国广州[2]东北大学，美国[3]脑与认知科学及教育科学重点实验室、教育部脑研究与康复研究所、广东省心理健康与认知科学重点实验室，华南师范大学，中国广州[4]华南师范大学政治与公共管理学院，中国广州

摘要

大语言模型可以将书面评分标准应用于转录文本，使自闭症诊断观察量表（ADOS-2）的自动编码看似触手可及。使这类评定者与临床医生保持一致的一种常见方法，是向其展示专家评分示例，但正确标签是否能提高一致性却很少得到检验。我们针对48名儿童时长一小时的 ADOS-2模块3会话中的八个基于言语的条目进行检验，采用一个本地模型以及预注册的安慰剂对照设计。我们比较了少样本评分、锚定比较（模型只判断某名儿童相较于已评分锚点表现更好、相似或更差，再由固定规则给出分数）、事后校准、基于方案的证据路由，以及对完整转录文本的直接评分。在40名留出儿童上，与临床医生的一致性较为有限，二次加权卡帕系数介于0.22和0.30之间。将每个示例的标签替换为错误标签后，约三分之一的少样本预测结果发生变化，且更常朝向损坏标签，而非随机机会水平；但一致性并未降低，正确标签和错误标签下分别为0.25和0.23。第二个规模更大的模型也得出了相同结果。锚定比较并未优于少样本评分，其跨儿童判断的准确性并不高于绝对分数所体现的准确性，证据路由也没有帮助。在已注册但未经校准的各组中，对完整转录文本的直接评分表现最佳，一致性为0.30。探索性后续分析发现，模型的分级判断比其最终分数更能对儿童进行准确排序；现有的批次校准使同一批儿童的一致性提高到0.46；在受控的合成会话中，单独对片段评分并未优于校准后的整份转录评分。仅凭专家评分示例本身，并不能校准基于模型的临床

评定者；而关于校准或增加流程复杂度的主张，则需要标签损坏对照和强有力的基线方法。

1 引言

自闭症诊断观察量表第二版（ADOS-2）是一种半结构化标准化观察，由受训检查者引导儿童完成一系列固定活动，然后根据一组条目对儿童的行为进行编码 [15, 16]。每个条目都需要对整个会话作出判断，因此编码时，临床医生必须将长达一小时的互动过程牢记于心。大语言模型（LLM）能够读取转录文本、遵循书面评分标准并解释评分，这使计算机辅助条目编码看起来触手可及 [6, 8, 19]。我们针对 ADOS-2模块3 的八个基于言语的条目开展研究，这些条目依据一小时会话的自动转录文本进行评分。该任务是条目编码，旨在支持临床医生，而不是进行自闭症诊断。

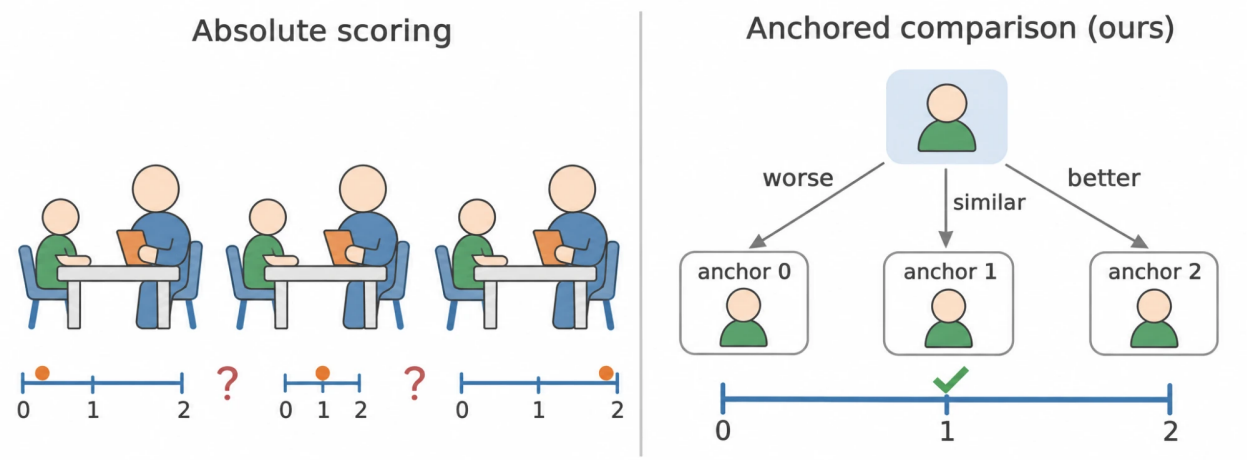

图 1：本研究测试的使用临床医生评分儿童的两种方式。左：模型分别为每名儿童评分；在我们的视频试点中，这种方式产生的分数无法在儿童之间进行比较。右：我们设计的锚定比较，模型只判断某名儿童相较于临床医生评分的锚定儿童表现更好、相似还是更差，再由固定规则根据锚定儿童的标签得出分数。本研究将这两种设计与对完整转录文本进行直接评分进行比较，也与锚点带有错误标签的安慰剂方案进行比较。在预注册评估中（第 7 节），比较并未提高一致性，正确的锚点标签也未比错误标签带来更好的结果。

使这类评定者与临床医生的评定保持一致的一种常见方法，是向其展示由专家评分的示例，将其作为少样本示例 [8]，或作为用于比较的锚点 [18]。其假设是，这些示例能将专家所使用的量尺传递给模型。这一假设很少得到直接检验。模型在看到已标注示例后可能会改变其预测结果，但这并不能说明正确标签使其评分与临床医生的一致性更高。如果没有标签错误的对照条件，就无法区分标签敏感性与校准收益。在我们的研究情境中，这一问题十分重要，因为在针对同一批48个会话开展的初步实验中（第4节），模型的判断能够在儿童内部对行为进行排序，却无法将不同儿童置于同一共同量尺上；而不足50个已标注会话太少，无法通过微调学会这一量尺。

因此，我们测试临床医生评分示例能否传递临床医生量表，以及示例的呈现形式是否会产生影响（图 1）。我们比较了展示示例分数的少样本评分、锚定比较（模型只需说明目标儿童相较于每个锚定儿童表现出的困难更少、相似还是更多，再由固定规则将这些关系转换为分数）、绝对分数的事后校准，以及对完整转录文本进行直接评分。每种方法都设有安慰剂对照：标签故意错误的锚点、来自随机窗口而非按协议路由窗口的证据，以及由模型自身绝对分数推导出的关系。所有实验组均使用同一个本地模型。在进行任何评分调用之前，我们已预注册假设、数据划分、评估指标和通过标准；之后的一项修订又使用第二个规模更大的模型重复了关键测试。

主要结果大多是负面的。在 40 名留出儿童中，所有组别与临床医生的一致性都不高（均值二次加权卡帕系数为 0.22–0.30）。将每个示例的标签都替换为错误标签后，约三分之一的少样本预测发生了变化，且变化方向更常指向损坏标签，而非随机机会水平所能解释的程度；但一致性并未降低（正确标签下为 0.25，错误标签下为 0.23）。第二个模型也表现出相同的情况。锚定比较并未改善少样本评分；在未经校准的预注册组别中，直接对完整转录文本评分与临床医生的一致性最高（0.30）。随后，注册方案之外的后续分析探究了对一小时会话评分时会在哪些环节出错。分析发现，模型的分级判断对儿童的排序优于其最终评分；现有的批次校准使同一批儿童的一致性提高到 0.46；而在受控的合成会话中，将会话拆分成若干片段分别评分，并未胜过校准后的整份转录评分。

因此，在这组数据中，示例改变了模型的评分，却没有使其更接近临床医生的评分；而我们尝试的额外处理步骤，反而丢失了通读整个会话所能保留的一些信息。

我们的贡献如下：

• **对专家评分示例进行受控评估**。 一项预注册、安慰剂对照测试，检验临床医生评分示例是否能提高大语言模型对长达一小时的 ADOS-2 转录文本与临床医生评分的一致性，并使用第二个模型进行复现。

• **将标签敏感性与校准收益区分开来**。 篡改专家标签改变了模型的预测结果，却没有降低其与临床医生评分的一致性；锚定比较也未能改善更简单的评分设计。

• **对长会话评分进行探索性分析**。 对概率读出的分析、采用混合证据的受控压力测试，以及情节级分解，揭示了在某些情况下，增加流程结构会比校准后的整份转录评分丢失更多信息。

2 相关工作

自动化自闭症评估。 计算机视觉在自闭症研究中的应用已得到系统性综述 [3], , 基于面部、语音和文本的基于人工智能的诊断也同样如此 [17]。与我们的研究设置最为接近的是，ADOS-Copilot 使用真实临床录音中的大语言模型对 ADOS-2 项目进行评分和解释 [8]。零样本大语言模型已被应用于 ADOS-2 第4模块的音频转录文本，以检测自闭症相关语言障碍 [6], ; 此外，一种多智能体对话系统会选择提问策略，以诱发第4模块中的社会语言特征 4 [7]。通用大语言模型根据文字稿、说话人分离转录文本或音频，预测 BOSCC 观察工具中与言语相关的编码 [19], ; 语音、视频和文本基础模型也已在包含儿童的自闭症诊断视频上进行基准测试：它们在活动识别等粗粒度任务上表现良好，但在细粒度任务上表现欠佳 [9]。CARE-VL 在专家标注的查询—响应区间上对视觉语言模型进行微调，并将片段级输出与大语言模型结合，得出自闭症/典型发育判定 [30]。这些系统会输出绝对分数或分类结果，其中一些会使用少样本示例。它们都没有通过比较，将儿童置于相对于临床医生评分的锚定儿童的位置上。

作为长时间会话评分标准评定者的大语言模型。 在自闭症领域之外，ChatGPT 已被用于零样本地依据观察工具对课堂文字稿片段进行评分 [27], , 语言模型也被应用于从冗长、嘈杂的课堂转录文本中评估需要高度推断的指令质量 [28]。在抑郁症评估中，大语言模型可根据访谈文字稿预测 PHQ-8 项目 [23]，并生成与 PHQ-8 对齐、可用于可解释识别的症状摘要 [32]。更广泛而言，关于大语言模型评判的研究已记录了立场偏差、冗长偏差和自我增强偏差 [26, 31], , 而 G-Eval 则结合思维链和表单填写对文本进行评分 [11]。

比较判断与锚点。 比较判断在心理测量学中有着悠久的历史 [1, 24]。对于大语言模型，研究表明，成对比较可以作为一种有效的零样本评估器 [13], , 并能高效地汇总比较结果 [14], ; 成对排序提示可用于检索 [20], , 还可将成对比较转换为作文分数 [22]。与本研究最相关的是，Poudel 等人将少样本直接评分与基于锚点的作文比较评分进行比较，使用人工评分锚点，发现比较方法更加稳定 [18]。我们采用锚点机制，并研究如何将其应用于包含众多活动、长达一小时的临床会话；这涉及协议路由、无评分证据，以及针对每个模块的安慰剂对照试验。

表1： 研究项目（$n = 48$ 名儿童）：各算法评分的数量、始终预测最常见评分时的准确率，以及每个项目读取的窗口。

Item	简称	0	1	2	主要输入	
A4	刻板语言	36	8	4	75%	all
A7	报告事件	24	14	10	50%	已路由
A8	对话	19	19	10	40%	已路由
B7	社交主动行为	10	28	10	58%	已路由
B9	社交回应	21	22	5	46%	all
B10	互惠性沟通	20	16	12	42%	all
B11	融洽关系	18	22	8	46%	all
D4	受限兴趣	32	10	6	67%	all

3 数据与设置

会话。 该数据集包含50个ADOS-2模块3会话，总时长为38.3小时，每个会话持续28至84分钟，由安装在房间角落高处的固定广角摄像机以 1920×1080、每秒25帧的速度录制。两个会话是字节完全相同的录制内容，却登记了不同的总分（2分和11分）。由于无法确定正确标签，二者均被排除，剩下$n = 48$ 名儿童。会话分两批获得，两批的ADOS总分均值存在差异（前16名儿童为6.06，后32名为10.50；曼-惠特尼检验$p = 0.017$），因此我们还分别报告每一批的数据结果。

项目。 临床医生提供了算法级项目评分（0、1 或 2；根据评分规则，原始分数为 3 时映射为 2）。我们研究了证据主要来自言语的八个项目（表1）：A4 刻板或特异性地使用词语或短语，A7 事件报告，A8 对话，B7 社交主动性的质量，B9 社交回应质量，B10 互惠性社会沟通量，B11 总体融洽关系质量，以及 D4 对不寻常或高度特定主题或物体的过度兴趣，或重复性行为。需要观察儿童的项目（眼神接触、指向性面部表情、手势、共同享受、感觉兴趣、手部习惯动作）不在研究范围内，因为儿童大多背对摄像头，人脸检测器仅在抽样帧的 1% 中检测到人脸，且头部关键点置信度为 0.005–0.28。

转录文本与方案时间线。每场会话均由自动语音识别器（Qwen3-ASR [21]）以连续的 16 秒窗口进行转录，且不含说话人标签。每场会话的转录文本平均约有 8,400 个汉字。大语言模型读取转录文本后，将每个窗口标记为 14 项模块3活动之一或“其他”（图 2）。与随机置换对照相比，连续活动区块的中位长度是随机置换标签时的 5.4 倍，且全部 50 场会话均超过预先设定的三倍阈值。

4 初步实验揭示的跨受试者不一致性

在主要研究开始前，我们进行了四轮初步实验，并为每轮预先设定了通过标准。这些实验使用视频片段和ADOS总分，而非条目级转录评分，因此它们为主要研究的问题提供了研究动机，但并未对其进行检验。图3对这些实验进行了总结，A 节则提供了详细信息。

在同一名儿童内部，模型判断包含有效信号。一个盲测的视觉语言评判模型对我们系统标记为最不典型的片段进行评分，其在0–10分量表上的评分比同一名儿童的随机片段高出0.64分（威尔科克森检验 $p = 0.0106$，$n = 45$）；另一种无关且规模小得多的评判模型也保留了这一效应（$p = 0.0044$）。但在儿童之间，同类判断并未实现迁移。对于ADOS总分，没有任何基于视频的预测器优于始终预测训练集均值的方法（平均绝对误差为4.73，而其他方法为4.96–5.63）；按儿童汇总的视觉语言模型逐片段评分也没有更好（4.83–4.85）。学习到的片段特征以79–85%的准确率识别出源录音，这表明跨儿童比较主要反映了

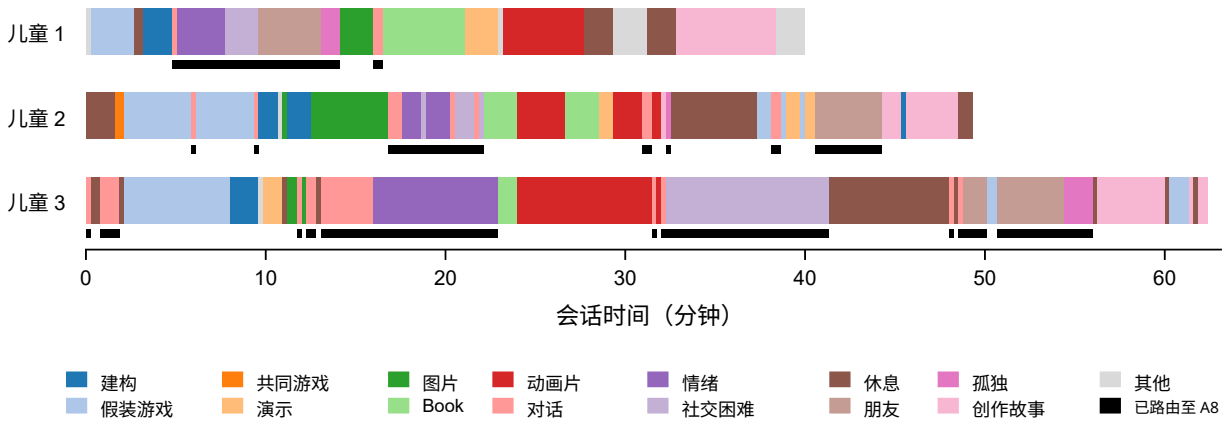

图2：三个儿童的方案时间线（每个16秒窗口的自动活动标签）。黑色刻度标记路由至项目 A8（对话）的窗口，即对话、情绪、社交困难、友谊和孤独活动。各会话的长度、活动顺序和持续时间各不相同，因此固定时间窗口无法代表整个方案。

由视觉语言模型按儿童汇总后，表现并未更好（4.83–4.85）。学习得到的片段特征以 79–85% 的准确率识别出源录音，这表明跨儿童比较主要反映了谁以及在哪里，而较少反映行为。

这些试点研究促使我们思考，临床医生评分的参考儿童能否提供片段级判断所缺少的跨儿童量表，并据此设计 第5节 中的锚定比较，作为一种实现方式。由于试点研究关注的是片段级判断和总分，因此应将其视为研究动机，而非条目级文字稿评分的证据。

5 锚定比较评分

设 s 为目标会话，i 为具有以下等级的项目 $\mathscr{L} = \{0, 1, 2\}$，并且 $\mathscr{P}$ 为其他已标注会话的集合。该方法包含四个模块（图4）。

M1：协议对齐。每个 16 秒的转录窗口都带有一个来自会话时间线的活动标签（第 3 节）。

M2：条目路由。每个条目都映射到预期会出现其证据的活动。对于 A7，评分说明明确列出了示范任务；对于 A8 和 B7，我们在进行任何评分之前，根据模块 3 协议的观察笔记推导出了相关活动。A7 被分配到示范任务、故事书、动画片和对话活动；A8 被分配到对话以及情绪、社交困难、友谊和孤独感访谈；B7 被分配到建构、假扮游戏和休息。其余五个条目均以整个会话为范围，因此涵盖所有窗口。

M3：无评分证据。对于每个会话和项目，大语言模型会读取路由后的窗口，列出最多八条有关儿童的证据，每条证据均包含时间戳、简短的逐字引文和客观记录，随后给出由两到三句话组成的总结。提示词禁止使用分数和评价性词语，并要求模型根据内容推断哪些话语是儿童说的，因为文字稿不带说话人标签。

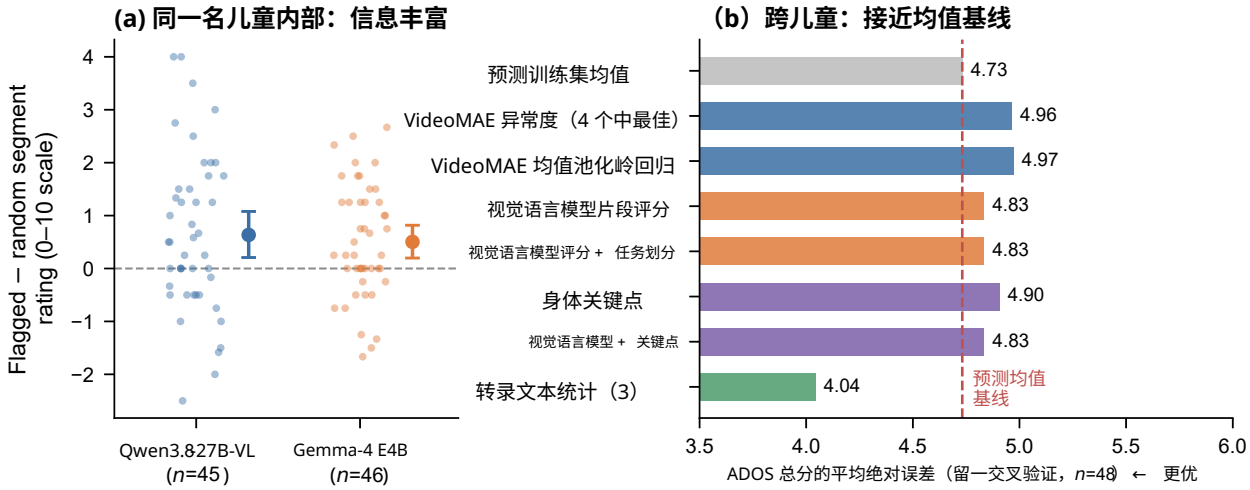

图3：试点证据。(a) 对同一名儿童，盲评者对系统标记的片段进行评分，结果高于对该儿童随机片段的评分（每个点代表一名儿童，误差线表示均值的自助法95%置信区间）。这一效应在两个互不相关的评判模型中均成立。(b) 在不同儿童之间，对于ADOS总分的所有视频或视觉衍生预测因子，其误差均不低于始终预测训练集均值的误差（虚线）。只有简单的转录文本统计能够改善预测效果。

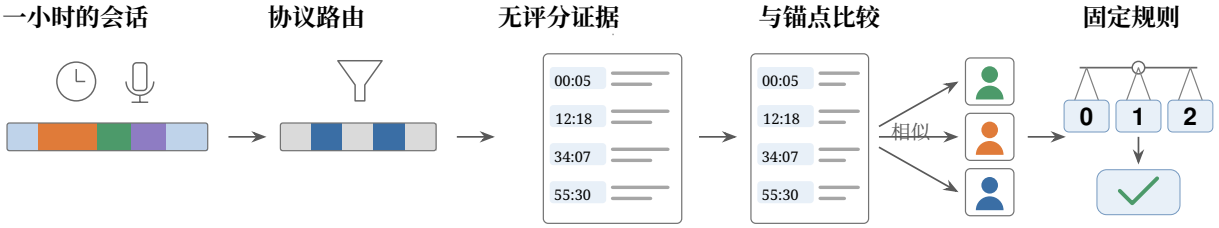

图4：锚定比较评分。对于某个项目，大语言模型会从协议活动中提取该项目可被观察到的无评分证据，然后在看不到锚定儿童评分的情况下，判断目标儿童相较于每名锚定儿童是更好、相似还是更差。评分根据这些关系以及锚定儿童的临床医生标签，按照固定规则计算得出（公式（1））。

M4：锚定比较。 对于每个等级 $\ell \in \mathscr{L}$，以确定性方式抽取一名锚定儿童$a_\ell \in \mathscr{P}$，其在项目 ℓ 上的得分为i（通过目标、项目和等级的哈希值确定）。大语言模型会看到该项目的完整评分标准、三名锚定儿童的 M3 证据（不含其分数）（按固定的随机顺序排列），以及目标儿童的证据；随后仅针对每个锚点输出一种关系$r_\ell \in \{-1, 0, +1\}$：目标儿童在该项目上表现出的困难少于、约等于或多于该锚定儿童。得分为

$$\hat{y}(s,i) = \arg\max_{y \in \mathscr{L}} \prod_{\ell \in \mathscr{L}} \begin{cases} 0.7 & r_\ell = \operatorname{sign}(y - \ell) \\ 0.15 & \text{otherwise,} \end{cases} \tag{1}$$

平局时倾向于采用 $\mathscr{P}$ 的中位标签。该规则预先固定，且不包含任何拟合参数。临床医生量表仅通过锚定儿童的标签体现。

模型在 M4 中从未给出分数。设计依据是将绝对评分混合在一起的两种角色分开。模型评审者判断两个儿童中谁表现出更多某种行为，而锚点的标签则提供评分尺度。第一种角色是否确实比绝对评分更可靠，将通过 H6 直接检验。

6 实验设计

模型。 所有实验组均使用同一款本地部署的大语言模型，即 Qwen3 系列的 Qwen3.8-27B [29], , 由 vLLM [10], 提供服务，温度设为 0；评分和比较调用启用推理，描述性提取步骤则禁用推理。录制内容或文字稿均不会离开研究团队的服务器。

组别。 我们将方法（A）与基线方法和安慰剂进行比较；这些方法分别移除或破坏一个要素：

- **D** （直接）：完整转录文本和项目评分标准，由模型输出原始分数。
- **Z** （基于证据的绝对值）：先执行M1–M3，然后由模型根据评分标准对证据进行评分。
- **Zcal**：使用相同标签通过留一校准对Z进行映射（对于每个原始分数，取其他儿童中具有该原始分数者的最常见真实分数）。
- **F** （少样本）：展示相同的三个锚点及其分数，模型输出该分数。这是 Poudel 等人的直接条件 [18] , 也是 Jiang 等人所采用的少样本设置 [8]。
- **PF**：将每个锚点标签循环移位后的 F（$0 \to 1 \to 2 \to 0$）。
- **PA**：将 A 的关系与移位后的标签进行聚合。
- **R**：对于三个路由项目，从相同数量的随机选取窗口中提取 A 以及支持目标和锚点的证据。

数据划分和评估指标。 在进行任何评分调用之前，先根据总分进行分层，并选取八名儿童，仅用于检查输出格式。结果报告基于其余40名儿童，并将全部48名儿童的结果作为次要分析。目标对象的锚点始终来自其他儿童。每个项目的一致性采用二次加权卡帕系数（QWK） [2] 进行测量，主要终点是八个项目的平均QWK。各组之间的配对差异采用受试者层级自助法（10,000次重抽样）和单侧95%界限 [4]。对 A 的绝对检验采用置换零假设，即在各项目之间联合打乱儿童的标签向量。无法解析的输出将替换为留一多数预测的多数分数，并予以计数。

假设。 H1：A 与临床医生的一致性高于随机机会水平，且优于直接评分 D。H2：A 优于使用相同锚点的少样本评分 F。H3：锚点标签具有因果作用（A >PA 和 F >PF）。H4：协议路由具有因果作用（A >R 用于路由项目）。H5：A 优于使用相同标签的事后校准 Zcal。H2–H5 作为一组进行 Holm 校正 [5]。提示词未根据任何儿童的一致性进行调优。

机制检验。 H1–H5 检验该方法是否有效，H6 则检验其原因。对于每个（目标、项目、锚点）三元组，根据临床医生标签可知其真实关系，$r^* = \text{sign}(y_{\text{target}} - \ell)$。我们比较模型在 A 中陈述的关系准确率与其自身绝对分数在 Z 中所推断出的关系准确率，即 $\text{sign}(\hat{z}^{\text{target}} - \hat{z}_{\text{anchor}})$；换言之，先按模型的量表为每名儿童评分，再进行比较。H6a：A 的关系准确率高于随机机会水平（对儿童标签进行置换检验）。H6b：该准确率高于 Z推断的准确率（受试者层级自助法）。若 H6b 未通过，则任何端到端增益都不能归因于相对判断更可靠。H6 在进行任何评估之前已加入预注册，并与 H2–H5 这一组假设分开报告。

解码与模型规模。 为了探究忽略锚点标签是否是 270亿参数模型的局限，在主要结果得出后、调用第二个模型前，团队进行了另一项预注册修订：使用 DeepSeek-V4.1-Flash 重新运行 Z、F、PF 和 A 组，该模型在团队的 GPU集群上本地运行。由于该硬件只能以不进行推理的模式运行该模型，因此修订采用强制选项 解码。沿用相同的提示词，并在其后添加固定的答案前缀；答案则取允许选项中概率最高的一项（分数数字，或针对每个锚点的更好／相似／更差）。通义千问Qwen3.8-27B也以完全相同的模式重新运行，作为匹配对照，从而避免模型与解码方式相互混杂。针对较大模型的预注册检验包括 F

表 2: 在测试集上与临床医生的一致性（$n = 40$ 名儿童）：每个项目的二次加权卡帕系数（QWK）及各项目均值、平均完全一致率，以及无法评分的输出占比（以留一多数预测替代）。†R 仅涵盖三个路由项目；其均值按这些项目计算。多数投票（留一法）对每名儿童进行预测时，取其他儿童中最常见的评分。剔除该儿童后，只要类别分布接近平衡，多数类别就可能与该儿童自己的评分不同，因此这一基线的一致性可能低于随机机会水平（QWK 为负值）。

Arm	Mean	A4	A7	A8	B7	B9	B10	B11	D4	精确匹配	NA
多数投票（留一法）	**−0.14**	0.00	0.00	−0.63	0.00	−0.46	0.00	0.00	0.00	42%	0.0%
D：直接、完整转录文本	**0.30**	0.21	0.25	0.21	0.11	0.08	0.48	0.51	0.57	45%	1.2%
Z：基于证据的绝对值	**0.22**	0.06	0.13	0.48	0.08	0.17	0.36	0.19	0.24	37%	0.6%
Zcal：Z +留一法校准	**0.06**	0.00	0.00	−0.06	0.00	0.17	0.18	0.17	0.00	49%	0.9%
F：带锚点的少样本	**0.25**	0.10	0.28	0.46	0.17	0.10	0.47	0.19	0.21	38%	0.6%
PF：F，错误标签	**0.23**	0.07	0.25	0.36	0.23	0.26	0.30	0.26	0.14	40%	0.6%
A：锚定比较	**0.22**	0.35	0.06	0.33	0.12	0.11	0.32	0.22	0.22	45%	0.3%
PA：A，错误标签	**0.15**	0.12	0.24	−0.14	0.11	0.31	0.19	0.17	0.18	43%	0.3%
R: A, 随机窗口†	**0.22**	–	−0.05	0.42	0.28	–	–	–	–	43%	0.0%

表 3: 测试集上的预注册配对检验：平均QWK之差（H6b：关系准确率之差）、单侧95%自助法下界、单侧p值及结论。H2–H5 采用 Holm 校正；H3 需同时满足 H3a 和 H3b。差值根据未舍入的估计值计算，显示值经过舍入。A 与随机机会水平的置换检验：平均QWK为 0.22，$p = 0.0002$。H6a：A的关系高于随机机会水平，置换检验$p = 0.0049$。

	比较	Δ	单侧95%下界	p	Pass
H1	A与D比较	−0.08	−0.19	0.898	no
H2	A与F比较	−0.03	−0.10	0.774	no
H3a	A 与 PA	0.07	−0.06	0.222	no
H3b	F 与 PF	0.01	−0.08	0.419	no
H4	A 与 R（路由项目）	−0.05	−0.20	0.684	no
H5	A 与 Z + 留一法校准	0.16	0.06	0.004	yes
H6b	关系准确性：A 与 Z推断	0.01	−0.02	0.242	no

> PF（使用的标签，H7）和 A > F（H8），经 Holm 校正。对于不含通过/不通过标记的通义千问对照组，也报告了相同的差异。

7 结果

主要比较。 在 40 名测试儿童中，锚定比较（A）的一致性高于随机机会水平（平均QWK 0.22，置换检验 $p = 0.0002$，表 2）。完整转录文本的直接评分（D）达到 0.30，是预注册组别中最高的；这些组别均未进行校准。H1 要求 A 超过 D，但未通过（$\Delta = -0.08$，单侧检验 $p = 0.90$，表 3）。提取的证据的绝对评分（Z）达到 0.22，使用锚点的少样本评分（F）达到 0.25。H2，即 A 与 F 的预注册检验，也未通过（$\Delta = -0.03$，$p = 0.77$）。

正确标签与损坏标签。 所有锚点标签都被损坏时，少样本一致性与使用正确标签时几乎相同（F 0.25，对比 PF 0.23）。使用损坏标签汇总 A的关系后，一致性降低了 0.07，但差异不显著（PA 0.15）。H3 要求两种效应都成立，因此未通过检验。使用正确的临床医生标签时，模型的得分并未显示出与临床医生更高的一致性。差异不显著并不能证明没有效应，但观察到的差异很小（F − PF $= +0.01$，单侧 95% 下限 -0.08）。

表 4： 机制检验（H6）：测试集上 960 个（目标、项目、锚点）三元组中跨儿童关系（更好／相似／更差）的准确率，以及真实关系并非“相似”的三元组上的准确率。

关系来源	准确率	仅排序
A：明示关系	46.0%	52.0%
Z推断：sign($\hat{z}_t - \hat{z}_a$)	44.7%	40.9%
Z与锚点标签：sign($\hat{z}_t - \ell$)	52.9%	60.9%
F 与锚点标签：符号($\hat{f}_t - \ell$)	55.8%	64.7%

表 5： 次要分析：全部48名儿童的平均QWK；测试集中各录音批次内的平均QWK（批次1：较早的会话，批次2：较晚的会话）；测试集中预测评分总和与临床医生评分总和在八个条目上的斯皮尔曼相关系数。

Arm	全部48名儿童	批次1	批次2	总和 ρ
多数投票（留一法）	−0.14	−0.18	−0.09	−0.69
D：直接、完整转录文本	0.34	0.22	0.27	0.60
Z：基于证据的绝对值	0.24	0.11	0.22	0.61
Zcal：Z +留一法校准	0.08	−0.01	0.06	0.31
F：带锚点的少样本	0.27	0.07	0.26	0.57
PF：F，错误标签	0.27	0.09	0.22	0.53
A：锚定比较	0.24	0.09	0.23	0.53
PA：A，错误标签	0.17	0.07	0.17	0.25

对标签的敏感性（事后分析）。 没有获益并不意味着模型忽略了标签。因此，我们使用已保存的预测结果，不再进行模型调用，逐个儿童比较 F 和 PF。损坏标签后，318 个（儿童、项目）单元中有 107 个单元的预测等级发生变化，占 34%。作为比较，对于同一项目，两名互不相关儿童的少样本预测在 58% 的单元中存在差异。在发生变化的单元中，61% 的预测恰好朝循环标签移位所指示的等级移动；若方向随机，这一比例应为 50%（二项检验 $p = 0.017$）。在强制选择运行中，由于仅使用单次前向传播，因此不存在推理变异性；其中 14% 的预测发生变化，而这些变化中有 67% 朝移位所指示的方向移动（$p = 0.013$）。因此，模型确实对标签作出了反应，并在一定程度上遵循了标签，但变化相互抵消了。将正确标签换成损坏标签后，45 个预测由错误变为正确，38 个由正确变为错误。推理运行的结果有一项注意事项：我们没有测量在标签相同的情况下重复运行时预测发生变化的频率，因此 34% 的变化中可能有一部分源自运行间变异。强制选择的结果不存在这个问题。

协议路由。 在三个路由项目上，从数量相同的随机选取窗口中获取的证据，其平均一致性略高于协议路由（这些项目上 R 为 0.22，A 为 0.17）。H4 未通过。

校准基线。 A 优于对 Z 进行的留一校准（H5 通过，$\Delta = +0.16$，$p = 0.004$）。然而，校准组的得分低于未校准的 Z（0.06 对 0.22）。由于每个原始分数对应的儿童数量较少，逐值多数映射经常会塌缩为单一标签，导致四个项目的 QWK 为 0.00。因此，H5 表明，这一校准规则是一个较弱的基线，但并未说明锚定能够增加信息。

机制检验。 在 960 多组三元组（目标、项目、锚点）中，模型在 A 中陈述的关系有 46% 是正确的。这高于随机机会水平（置换检验 $p = 0.005$），但并不优于其自身绝对分数所隐含的关系（45%，表 4），因此 H6b 未通过。将模型的绝对分数与锚点的真实标签进行比较，准确率更高（Z 为 53%，F 为 56%）；对于真实关系并非“相似”的三元组，准确率更高（A 为 52%，F 为 65%）。

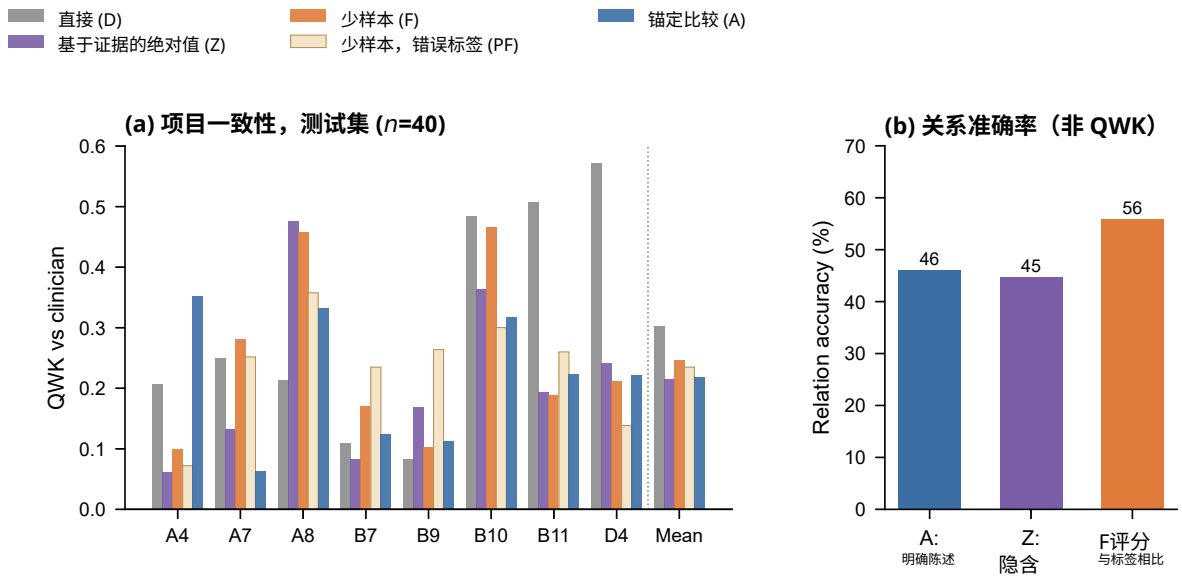

图5：测试集上的预注册结果。 （a）每个项目的二次加权卡帕系数。完整转录文本的直接评分（D）具有最高的平均一致性，主要得益于全会话项目 B10、B11 和 D4。使用有意错误的锚点标签进行少样本评分（PF）并不比使用正确标签（F）差。（b）另一项指标：其更好/相似/更差关系正确的（测试儿童、锚点）对所占的百分比。“A：明示”是模型在锚定比较中直接陈述的关系。“Z：隐含”比较模型对两个儿童的绝对分数。“F 分数与标签比较”将模型对测试儿童的少样本评分与锚点的临床医生标签进行比较。明示关系并不比隐含关系更准确，且不如分数与标签比较准确。

直接阅读何时更胜一筹（探索性分析）。 这些比较未经预注册。在相同评分标准下，直接评分优于对提取证据进行绝对评分（D $-$ Z $=+0.09$，自助法 95% 置信区间下限 $+0.01$，$p=0.03$）。差异完全来自 5 个按整个会话定义的项目（A4、B9、B10、B11、D4），其中 D 达到 0.37，Z 为 0.21（$+0.16$，$p=0.0005$）。在 3 个路由项目上，两者没有差异（D 为 0.19，Z 为 0.23）。将一小时的互动压缩成少量证据片段，似乎会丢失会话整体判断所需的信息；而对于与特定活动相关的项目，这样做则毫无损失。

强制选择解码与模型规模。 不进行推理时，通义千问对照组呈现出相同的模式（表6）。使用错误锚点标签的少样本一致性与使用正确标签时相当（PF 0.24 对 F 0.23，$\Delta=-0.01$，$p=0.64$），且锚定比较并未优于少样本评分（A 0.24，$\Delta=+0.01$，$p=0.47$）。其明示关系再次不如将其绝对分数与锚点标签进行比较准确（48% 对 55%）。因此，缺乏校准收益并非由推理模式造成。对于规模更大的 DeepSeek-V4.1-Flash，两个预注册检验均未通过。正确标签并不优于被篡改的标签（F 0.26 对 PF 0.26，$\Delta=+0.00$，$p=0.45$，H7），且锚定比较并不优于少样本评分（A 0.24，$\Delta=-0.02$，$p=0.67$，H8）。其明示关系同样不如将其绝对分数与锚点标签进行比较准确（45% 对 55%）。DeepSeek 对标签的反应不如通义千问强烈。篡改标签使其 10% 的预测发生变化，而变化后的预测遵循偏移后标签的频率并不高于随机机会水平（48%，$p=0.64$）。因此，第二个规模更大的模型同样不存在校准收益。

直接评分的错误（事后）。为了解直接评分在哪些方面会失败，一项未预注册的单独分析考察了其320个测试（儿童、项目）评分。D在其中175项（55%）上出错。它过度评分的次数远多于评分不足（112次对63次），并将大多数儿童归为中间水平。对于刻板

表6：强制选择解码（无推理；答案是在固定答案前缀之后概率最高的允许选项），提示词和锚点与主要实验组相同，测试集（$n = 40$）。均值和逐项目QWK。

模型	Arm	Mean	A4	A7	A8	B7	B9	B10	B11	D4
Qwen3.8-27B	Z	**0.26**	0.09	0.29	0.42	0.20	0.22	0.39	0.31	0.18
	F	**0.23**	0.06	0.20	0.42	0.17	0.25	0.46	0.31	−0.01
	PF（错误标签）	**0.24**	0.06	0.17	0.40	0.11	0.40	0.27	0.25	0.25
	A	**0.24**	0.29	0.25	0.20	0.06	0.14	0.19	0.38	0.37
	PA（错误标签）	**0.12**	0.13	0.23	−0.06	0.30	−0.03	0.11	0.01	0.29
DeepSeek-V4.1-Flash	Z	**0.22**	0.04	0.13	0.43	0.16	0.30	0.31	0.24	0.15
	F	**0.26**	0.09	0.23	0.34	0.17	0.47	0.35	0.27	0.15
	PF（错误标签）	**0.26**	0.07	0.14	0.40	0.17	0.51	0.30	0.22	0.26
	A	**0.24**	0.19	0.11	0.37	0.30	0.26	0.37	0.26	0.05
	PA（错误标签）	**0.10**	0.16	0.08	−0.10	0.07	0.03	0.27	0.07	0.19

语言（A4），它将40名儿童中的31名评为1分，尽管临床医生将其中31名评为0分；对于社交回应（B9），它将40名儿童中的37名评为1分。我们还要求模型在评分后引用其评分所依据的转录文本行，并将这些内容与说话人分离结果进行核对。它归属于儿童的文本行，往往来自由检查者主导的语段。不过，这一标记在错误评分和正确评分中出现的频率大致相同（分别为46%和47%）。模型用一句话给出的理由，只有在2%的情况下描述了与其评分不同的水平。这项核查依赖于并不完美的说话人分离，以及评分后给出的理由；这些理由展示的是模型所提供的支持依据，不一定是它实际关注的内容。分析未发现被标记的说话人归属与评分错误之间存在关联，但无法排除真实的归属错误会影响评分。所述理由与评分之间的一致性也仅表明内部一致性，并不能证明正确运用了临床标准。错误中最明显的模式是量表被压缩，导致难以有效区分儿童。

次要分析。对全部48名儿童的结果相似（表5）。对于每个主要实验组，第一录音批次的11名测试儿童之间的一致性低于第二批次的29名测试儿童，尽管第一批次的估计基于较少的儿童。对八个条目进行汇总后，对于主要实验组，预测评分与临床医生评分呈中等程度的相关（表5）。

8 后续分析

主要研究发现，临床医生评分示例并未提高与临床医生之间的一致性；在未经过校准的预注册组别中，直接对完整转录文本进行评分的表现最佳。由于增加参考案例和处理流程步骤并无帮助，我们进一步探究一小时的会话在评分过程中为何会出错。三项后续分析对此进行了探讨。这些分析均不属于主要注册，且都是在已知主要研究结果后设计的。它们的证据等级各不相同，每个小节都会说明这一点。

8.1 概率读出与序数校准

问题与证据层级。 完整转录文本模型对儿童的排序是否优于其最终评分所显示的效果？本分析再次使用了40名测试儿童，因此这是对我们已经分析过的一个队列进行的额外评估，并非验证性研究。其校准基线是在我们看到通义千问的概率之后添加的。

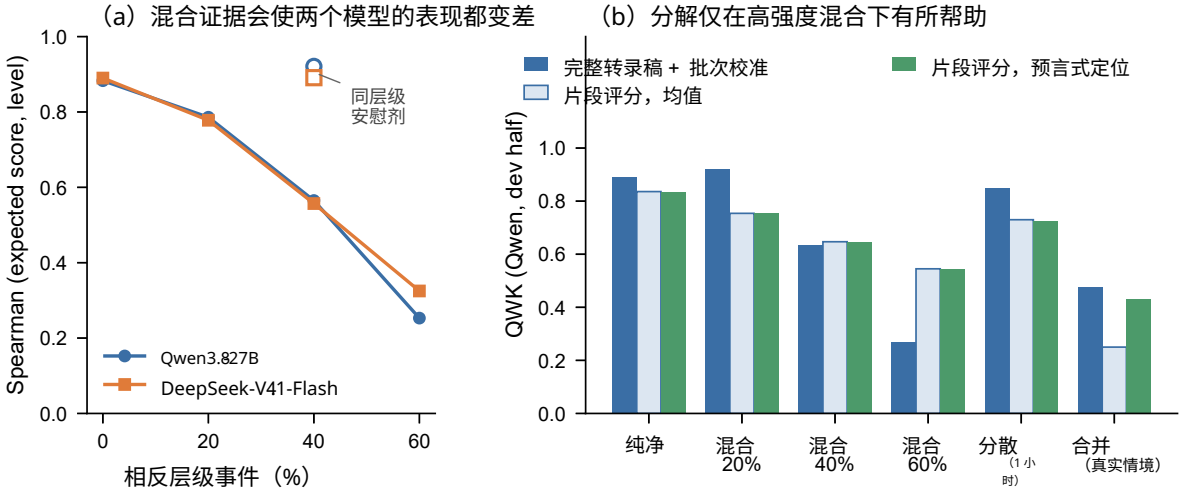

图 6：对合成会话的后续分析（第 8.2 节和 8.3）。 (a) 在病例短文中加入相反等级的片段，会降低两个模型对病例短文等级的排序能力，且加入的片段越多，影响越大。加入相同数量的同层级文本（空心标记）则没有这种影响。(b) 合成数据集的开发集部分，每种条件下有 48 个会话。仅在高强度混合条件下，单独为片段评分并取均值（浅色柱）才优于校准后的整份转录评分（蓝色），而在分散和组合条件下则逊于后者。绿色柱仅保留生成器记录显示包含病例短文文本的片段。这个理想定位器仅适用于所测试的事件处理流程，并不代表所有可能的会话拆分方法都能达到的上限。

设计。 我们使用强制选择解码重新运行了D的提示，该方法会为每个等级给出一个概率，并将这些概率下的期望评分与概率最高的等级以及现有的读出方法进行了比较。

发现。 期望评分在各条目内对测试儿童的排序效果良好（平均斯皮尔曼相关系数为 0.54，置换检验 $p = 0.0001$），而 D 生成的评分为 0.36。在临床医生评分相差两个等级的儿童配对中，期望评分有 91% 的概率能正确排序。相比之下，最可能等级的预测大量集中在中间水平。对于 B9，95% 的儿童都被预测为 1 级，尽管期望评分对他们的排序仍达到 0.57。在我们按照 D 的格式撰写的明确病例短文上，同一提示的 QWK 达到 0.90，且没有出现预测集中于中间水平的情况，因此，当证据明确时，模型能够正确应用评分标准。现有的无标签校准方法应用于这些概率后，将平均 QWK 从 D 的 0.30 提高到批次校准的 0.46 和先验匹配的 0.45。基于期望评分的逐条目切点，使用其余 47 名儿童的临床医生标签进行留一法拟合后，QWK 达到 0.50。提升最大的是预测集中在中间水平的条目（B9 从 0.08 提高到 0.53，A8 从 0.21 提高到 0.66）。

局限性。 批次校准和先验匹配具有传导式特征。它们不使用任何临床医生标签，但会利用整个测试儿童批次的模型输出调整每名儿童的读出结果，因此没有儿童是在孤立状态下获得评分的。0.50 的结果使用了标签，属于不同的条件。在 DeepSeek-V4.1-Flash 中，期望评分再次比最大值索引更好地对儿童进行排序（0.46 对 0.38）；但 47 名儿童的校准结果（0.38）并未显著优于 D（$\Delta = +0.08$，95% 置信区间下限为 -0.06）。因此，校准增益仅展示了两个模型中的一个，且所基于的队列并不独立。

8.2 混合行为证据的受控压力测试

问题与证据层级。 一小时的会话具有哪种特性，会削弱模型的分级判断？本分析仅使用合成会话，不包含儿童数据，并在首次调用模型之前完成了预注册（设计、评估指标

和决策规则）。

设计。 我们构建了 96 个基础合成短案例：针对八个条目中的每一个，设计四组最小对集，每组包含三个简短案例，分别对应 0、1 和 2 级。我们使用大语言模型助手，根据评分标准的等级描述撰写这些案例和常规检查者引导填充内容。没有复制任何真实转录文本，仅使用真实转录文本的汇总统计量来检查内容是否逼真。每个基础案例都在 33 种条件下呈现，每种条件以受控剂量加入一个因素。这些因素包括出现相反等级的片段（混合证据）、减少儿童自身的片段数量（稀疏性）、将内容嵌入长达一小时的填充内容中（分散性）、字符级识别噪声、不考虑话语边界截取窗口、交错的检查者话语，以及同时加入上述所有因素。每个因素都设有安慰剂条件：加入相同数量的文本，但不包含该因素，例如加入与案例本身等级相同的片段。通义千问 Qwen3.8-27B 和 DeepSeek-V4.1-Flash 均通过强制选择为每种呈现形式评分。分析单位是基础案例。主要指标是每种条件下 96 个案例的期望评分与设计等级之间的斯皮尔曼相关系数。只有当该相关系数经安慰剂条件校正后的下降值具有经 Holm 校正的 $p < 0.05$ 且其 95 % 置信区间下限大于零时，才认定某个因素的影响已得到证实。

发现。 混合证据在两个模型中都产生了最大的影响，安慰剂校正后的降幅为 +0.36(95% 置信区间 +0.22 至 +0.49)，通义千问为 +0.34 (+0.20 至 +0.47)，DeepSeek 为，且随着剂量增加而扩大（图 6a）。两个模型中也都证实了稀疏性（+0.07 和 +0.13），而分散性仅在 DeepSeek 中得到证实（+0.22）。噪声、窗口截断和交错在两个模型中都未得到证实。两个模型赋予相反的少数部分的权重都超过了其占比。对于评分标准按某种行为出现频率计分的条目，四个片段中出现一个非典型片段，就会使期望评分向全为非典型片段时的取值移动 42–45%。对于 A7，其评分标准认可最佳示例，评分仍接近最佳示例所对应的评分，这正是评分标准要求的结果。综合所有因素后，相关性降至 0.22（通义千问）和 0.11（DeepSeek），低于真实转录文本上的相关性（0.54 和 0.46），因此合成条件涵盖了真实情况。

局限性。 按建构方式，混合情境案例的设计层级就是其多数等级。真实的 ADOS-2 条目并非都按多数等级评分。有些条目侧重特定行为或其频率，有些则侧重整体互动质量，因此测得的下降幅度中有一部分可能符合临床判断。研究发现，在这项合成任务中，混合不同层级的行为会导致排名大幅下降，且下降幅度随混合程度增加；这一结果在两个模型中都成立。这并不能证明真实 ADOS-2 评分失败主要是因为证据混合。要得出这一结论，需要临床医生编码的混合案例。

8.3 为什么情节级分解效果不佳

问题与证据层级。 如果模型对相反的少数部分赋予过高权重，那么将每个情节分别评分，再按照明确规则合并分数，是否有帮助？本分析使用第 8.2 节中的合成会话。该方法及其决策规则均已预注册。对其失败原因的分析属于探索性分析。

设计。 我们分别对每个片段进行评分，并使用条目特定规则合并情节评分：对于统计某种行为发生频率的条目，采用均值；对于涉及典型行为的条目，采用截尾均值；对于 A7，采用最佳示例。片段是同一方案活动的连续运行，每个片段又被划分为最多包含 12 个窗口的块。我们在合成数据集的开发集的一半上应用无效性规则，以决定是否运行留出的一半。该方法必须在混合证据条件和综合现实条件下都优于校准后的整份转录评分。第二种预注册变体会先移除模型探针判定为无关的片段。

发现。 在高强度混合条件下，该方法表现更好。当相反片段占 60% 时，其 QWK 达到 0.59，而采用批次校准或先验匹配的整份转录评分为 0.27–0.32；在三种混合比例下，其得分为 0.71，而后者为 0.66。在综合条件下，八种配置均未接近目标（最佳为 0.31，而目标范围为 0.48–0.52），因此研究在运行留出的一半之前停止。第二种变体在其定位门槛处停止，因为探针将带有证据的双窗口单元与填充内容区分开的 AUROC 仅为 0.74，低于要求的 0.80。生成器记录了哪些转录窗口包含病例短文文本，这显示了性能损失发生的位置（探索性分析）。包含病例短文文本的片段仍能在每个条目内区分等级 2 和等级 0；在综合条件下 AUROC 为 0.78，仅使用分散性时为 0.90，而仅含填充内容的事件无法做到这一点（0.51）。然而，有用的片段很少，而且大多是填充内容。在综合条件下，每次会话的 19 个片段中，平均只有 1.4 个包含任何病例短文文本；在这些片段中，病例短文文本占文本的中位数为 11%。仅对这些片段取均值——这是任何实用的定位器都无法做到的——可将 QWK 从 0.25 提高到 0.43，但仍低于校准后的整份转录评分（0.48）。仅使用分散性时，这种理想定位器并未获得任何提升（0.73 对 0.85，图 6b）。在两种长条件下，与均值相比，使用上分位数或最大值进行汇总的表现更差。

局限性。 这些结果仅适用于我们测试的事件处理流程、该流程所采用的粒度，以及合成会话。在这一流程中，将会话拆分会使稀疏证据分散得更薄，而不是将其隔离出来；即便定位完全准确，也未能达到整份转录基线。采用其他方式拆分会话，结果可能有所不同。

9 讨论与局限性

敏感性与收益。 模型是否会对 示例标签产生反应，以及正确标签能否带来校准收益，是两个不同的问题，而且在本研究中答案各不相同。模型确实对标签作出了反应。标签损坏后，约三分之一的少样本预测发生了变化，且变化方向指向损坏标签的频率高于随机机会水平。然而，使用正确标签时，一致性并不高于使用偏移标签或完全不提供示例时的水平。这说明未检测到改善，并不能证明效果为零；但对于40名儿童而言，任何大到足以产生实际影响的收益都未能显现。标签改变了个别预测，但总体而言，并未使预测更趋近临床医生的判断。不具备推理能力的通义千问和规模更大的 DeepSeek 模型表现相同，而 DeepSeek 对标签的反应甚至更弱。因此，凡是通过向大语言模型展示评分示例来使其与专家意见保持一致的系统，在声称存在收益之前，都应加入标签损坏对照，因为模型可能会对标签作出反应，却并未采用专家的评分标准。

比较为何没有帮助。 我们基于视频试点得出的前提构建了锚定比较，即相对判断会比绝对判断更可靠。对于转录文本，机制检验并不支持这一点。模型在两个儿童之间陈述的关系，其正确率并不高于其绝对分数所暗示的关系，而且低于将其绝对分数与锚点的真实标签进行比较所得的结果。在这些数据中，模型依据评分标准对儿童进行判断的效果优于依据另一个儿童的证据进行判断。我们没有检验其中的原因。一种可能性是，不同儿童的无分数证据摘要读起来很相似。另一种可能性是，同时判断两组证据比依据书面标准对其中一组进行判断更困难。这两点仍然是假设。

信号所在之处。在所有组别中，互惠性沟通（B10）都是评分最高的条目之一。对话（A8）在基于证据的组别中表现良好，而融洽关系（B11）和受限兴趣（D4）则在直接评分中表现良好。八个条目的总分与临床医生的总分呈中度相关。第一批录音的一致性要低得多，而该批次中的测试儿童较少。总体一致性

较为有限；由于没有第二位临床医生对相同会话进行评分，我们无法判断其与人工评定者之间的一致性相比差距有多大。

结构更少，一致性更高。 在这些预注册设计中，没有任何一项经过校准，而对完整转录文本进行直接评分的准确性最高。在探索性比较中，它相较于对提取的证据进行评分所具备的优势，来自于那五个基于整个会话定义的条目。对于这些条目而言，少量证据片段无法充分概括长达一小时的互动，而长上下文模型通过阅读全部内容取得了更好的表现。对于与特定活动相关的条目，提取不会带来任何成本。对于长会话，应逐条将先进行压缩再予以评判的处理流程与直接阅读进行比较。后续分析（第 8 节）得出了相同结论。在合成会话中，只要一小时内的证据较为稀疏，即使采用预言式定位，将片段分别评分后再汇总，其表现也低于校准后的整份转录评分。在我们测试的通义千问输出中，最佳方案将整份转录评分与现有的批次校准相结合，不过该结果来自一个并非独立的队列，并且未能在DeepSeek上复现。在合成压力测试中，混合不同程度的行为是对两个模型影响最大的条件。这如何对应到真实的ADOS-2评分仍有待研究，因为其中并非所有条目都采用多数决策。新的改进方案应与校准后的整份转录评分进行比较，而不应与未经校准的直接评分比较。

结论。 在本研究中，专家评分示例未能校准对 ADOS-2 转录文本进行评分的大语言模型评分者，而更复杂的证据处理丢失了会话层面信息。关于示例能够校准此类评分者，以及增加处理流程步骤会有所帮助的说法，应通过标签污染对照和校准后的整份转录评分进行检验。

范围与局限。 本研究评估的是八个基于语音的ADOS-2条目的条目编码，而非自闭症诊断，也未评估临床实用性。没有临床医生使用模型的输出结果，并且本文报告的一致性水平均未在独立队列中得到验证。本研究使用一个数据集，其中包含48名儿童；这些儿童均接受了中文ADOS-2模块3评估，并在两个评分分布不同的批次中录制。每名儿童只有一组临床医生评分，因此我们无法估计该样本中人工评分者之间的一致性，这限制了任何自动评分者所能达到的水平。我们仅研究能够通过语音进行判断的八个条目。摄像机位置排除了视觉条目。转录文本没有说话人标签，因此模型必须推断哪些话语属于儿童，并继承语音识别器的错误。该方法需要一个由使用相同方案评估、且有临床医生评分的儿童组成的样本池，以便从中提取锚点。我们按照固定规则为每个评分等级选取一个锚点，并未研究一致性如何取决于锚点的选择或数量。主要实验组使用一个大语言模型，而强制选择复现增加了另一个模型。第4节中的试点证据第4节 依赖模型评审者和预注册检验，而不是临床医生对标记片段进行审查。第 第 8 节中的校准结果重复使用了40名测试儿童，并且只对两个模型中的一个进行了复现，因此需要独立队列。最后，合成会话由大语言模型助手起草，其内容比真实行为更为纯净；对于混合情境案例，其设计等级通过多数决定正确答案，而对于将偶发行为计入评分的条目，临床医生可能不会接受这种做法。

数据与隐私。录音和文字稿属于未成年人的临床数据，不予公开。所有针对儿童数据的模型评分均在研究团队控制的服务器上运行，未使用第三方 API。第 8 节中的合成会话不包含儿童数据。本文未复现任何转录文本摘录。

伦理。 华南师范大学政治与公共管理学院科学研究伦理委员会伦理分委员会将该研究审查为涉及人类受试者的研究，

认定其内容和方案符合研究伦理要求，并于2026年11月12日批准其继续开展（批准编号：SCNU-PAP-2026-006）。

参考文献

[1] Ralph Allan Bradley 和 Milton E. Terry。不完全区组设计的秩分析：I. 成对比较方法。*Biometrika*，39(3/4):324–345，1952。doi: 10.2307/2334029。[2] Jacob Cohen。加权卡帕系数：名义量表一致性，可用于处理等级差异或部分评分。*PsychologicalBulletin*，70(4):213–220，1968。doi: 10.1037/h0026256。[3] Ryan Anthony J. de Belen、Tomasz Bednarz、Arcot Sowmya 和 Dennis Del Favero。孤独症谱系障碍研究中的计算机视觉：对2009至2019年发表研究的系统综述。*TranslationalPsychiatry*，10(1)，2020。doi: 10.1038/s41398-020-01015-w。[4] B. Efron。自助法：重新审视刀切法。*The Annals of Statistics*，7(1):1–26，1979。doi: 10.1214/aos/1176344552。[5] Sture Holm。一种简单的序贯拒绝多重检验程序。*Scandinavian Journal of Statistics*，6:65–70，1979。[6] Chuanbo Hu、Wenqi Li、Mindi Ruan、Xiangxu Yu、Shalaka Deshpande、Lynn K. Paul、Shuo Wang 和 Xin Li。利用大语言模型诊断孤独症相关语言障碍并识别其不同特征。*npj Digital Medicine*，8(1)，2025。doi: 10.1038/s41746-025-02133-9。[7] Chuanbo Hu、Minglei Yin、Bin Liu、Wenqi Li、Lynn K. Paul、Shuo Wang 和 Xin Li。用于评估孤独症社会语言障碍特征的主动式多智能体对话框架，2026。[8] Yi Jiang、Qingyang Shen、Shuzhong Lai、Shunyu Qi、Qian Zheng、Lin Yao、Yue ming Wang 和 Gang Pan。通过大语言模型在真实临床场景中辅助孤独症诊断，2024。[9] Aditya Kommineni、Digbalay Bose、Tiantian Feng、So Hyun Kim、Helen Tager-Flusberg、Somer Bishop、Catherine Lord、Sudarsana Kadiri 和 Shrikanth Narayanan。多模态基础模型能否帮助分析包含儿童的孤独症诊断视频？载于*Interspeech 2025*，第3050–3054页。ISCA，2025。doi: 10.21437/Interspeech.2025-150。[10] Woosuk Kwon、Zhuohan Li、Siyuan Zhuang、Ying Sheng、Lianmin Zheng、Cody Hao Yu、Joseph E. Gonzalez、Hao Zhang 和 Ion Stoica。使用 PagedAttention 提升大语言模型服务的内存管理效率。载于第*29*届操作系统原理研讨会论文集，第611–626页。ACM，2023。doi: 10.1145/3600006.3613165。[11] Yang Liu、Dan Iter、Yichong Xu、Shuohang Wang、Ruochen Xu 和 Chenguang Zhu。G-Eval：使用 GPT-4 进行自然语言生成评测，实现更好的人类对齐。载于*2023*年自然语言处理经验方法会议论文集，第2511–2522页。计算语言学协会，2023。doi: 10. 18653/v1/2023.emnlp-main.153。[12] Adian Liusie、Potsawee Manakul 和 Mark Gales。缓解零样本提示分类器中的词语偏差。载于计算语言学协会论文集：*IJCNLP-AACL 2023* 精选论文，2023。[13] Adian Liusie、Potsawee Manakul 和 Mark Gales。大语言模型比较评测：使用大语言模型通过成对比较进行零样本自然语言生成评测。载于第*18*届欧洲计算语言学协会分会会议论文集（第*1*卷：长篇论文），第139–151页。计算语言学协会，2024。doi: 10.18653/v1/2024.eacl-long.8。[14] Adian Liusie、Vatsal Raina、Yassir Fathullah 和 Mark Gales。高效的大语言模型比较评测：用于成对比较的专家混合框架。载于*2024*年自然语言处理经验方法会议论文集，第6835–6855页。计算语言学协会，2024。doi: 10. 18653/v1/2024.emnlp-main.389。

[15] Catherine Lord、Susan Risi、Linda Lambrecht、Edwin H. Cook、Bennett L. Leventhal、Pamela C. DiLavore、Andrew Pickles 和 Michael Rutter。自闭症诊断观察量表通用版：用于衡量与自闭症谱系相关的社交与沟通障碍的标准工具。《自闭症与发育障碍杂志》，30(3):205–223，2000。doi: 10.1023/A:1005592401947。

[16] Catherine Lord、Michael Rutter、Pamela C. DiLavore、Susan Risi、Katherine Gotham 和 Somer L. Bishop。自闭症诊断观察量表第二版（*ADOS-2*）。Western Psychological Services，2012 年。

[17] Fatemeh Mohammadi、Hassan Shahrokhi、Afsoon Asadzadeh、Saeed Pirmoradi、Ali Moghtader 和 Peyman Rezaei-Hachesu。人工智能在孤独症谱系障碍诊断中的应用：面部、语音和文本分析方法的范围综述。《健康科学报告》，8(11)，2025 年。doi：10.1002/hsr2.71476。

[18] Prashreet Poudel、Huayue Gu、Collin Lynch 和 Zhikai Gao。人工评分锚点的两种用途：少样本直接评分和基于锚点的比较评分。载于 Joshua Wilson、Christopher Ormerod 和 Magdalen Beiting-Parrish 编，人工智能测量与教育会议（*AIME-Con*）论文集：全文论文，第 458–465 页，匹兹堡，宾夕法尼亚州，美国，2026 年 10 月。国家教育测量委员会（NCME）。ISBN 979-8-9983004-0-0。网址：https://aclanthology.org/2026.aimecon-main.51/。

[19] Jordan Prescott、Aditya Kommineni、Tiantian Feng、Megan Micheletti、Alyssa Viggiano、Luis Angeles、Anfeng Xu、Lynn Perry、Catherine Lord、Daniel Messinger 和 Shrikanth Narayanan。《迈向基于大语言模型的自动化临床行为编码：使用儿童 BOSCC录音的案例研究》，2026。

[20] Zhen Qin、Rolf Jagerman、Kai Hui、Honglei Zhuang、Junru Wu、Le Yan、Jiaming Shen、Tianqi Liu、Jialu Liu、Donald Metzler、Xuanhui Wang 和 Michael Bendersky。《借助成对排序提示，大语言模型可有效充当文本排序器》。载于计算语言学协会论文集：北美计算语言学协会 *2024* 年会论文集，第 1504–1518 页。计算语言学协会，2024。doi：10.18653/v1/2024.findings-naacl.97。

[21] Xian Shi、Xiong Wang、Zhifang Guo、Yongqi Wang、Pei Zhang、Xinyu Zhang、Zishan Guo、Hongkun Hao、Yu Xi、Baosong Yang、Jin Xu、Jingren Zhou 和 Junyang Lin。《通义千问3-ASR 技术报告》，2026。

[22] Takumi Shibata 和 Yuichi Miyamura。LCES：利用大语言模型通过成对比较进行零样本自动作文评分。载于 *2025*年自然语言处理实证方法会议论文集，第29976–29989页。计算语言学协会，2025年。doi：10.18653/v1/ 2025.emnlp-main.1523。

[23] Bazen Gashaw Teferra、Argyrios Perivolaris、Wei-Ni Hsiang、Christian Kevin Sidharta、Alice Rueda、Karisa Parkington、Yuqi Wu、Achint Soni、Reza Samavi、Rakesh Jetly、Yanbo Zhang、Bo Cao、Sirisha Rambhatla、Sri Krishnan 和 Venkat Bhat。利用大语言模型进行自动抑郁症筛查。*PLOS*数字健康，4(7)：e0000943，2025年。doi：10.1371/journal.pdig.0000943。

[24] L. L. Thurstone. 比较判断定律。《心理学评论》，34(4):273–286，1927。doi: 10.1037/ h0070288。

[25] Limin Wang、Bingkun Huang、Zhiyu Zhao、Zhan Tong、Yinan He、Yi Wang、Yali Wang 和 Yu Qiao。VideoMAE V2：通过双重掩码扩展视频掩码自编码器。载于*2023*年*IEEE/CVF*计算机视觉与模式识别会议（*CVPR*），第14549–14560页。IEEE，2023。doi: 10.1109/CVPR52729.2023.01398。

[26] Peiyi Wang、Lei Li、Liang Chen、Zefan Cai、Dawei Zhu、Binghuai Lin、Yunbo Cao、Lingpeng Kong、Qi Liu、Tianyu Liu 和 Zhifang Sui。大语言模型并非公平评估者。载于第*62*届计算语言学协会年会论文集（第*1*卷：长篇论文），第9440–9450页。计算语言学协会，2024。doi: 10.18653/v1/2024.acl-long.511。

[27] Rose Wang 和 Dorottya Demszky。ChatGPT 是一位优秀的教师教练吗？衡量其在零样本条件下对课堂教学进行评分并提供可操作性见解的表现。载于第*18*届*NLP*构建教育应用创新用法研讨会（*BEA 2023*）论文集，第626–667页。计算语言学协会，2023年。doi：10.18653/v1/2023.bea-1.53。

[28] Paiheng Xu、Jing Liu、Nathan Jones、Julie Cohen 和 Wei Ai。使用语言模型衡量教育中的指令质量：前景与陷阱。载于*2024*年北美计算语言学协会分会会议论文集（*NAACL 2024*）。计算语言学协会，2024年。doi：10.18653/v1/2024.naacl-long.246。

[29] An Yang、Anfeng Li、Baosong Yang、Beichen Zhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Dayiheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang、Jian Yang、Jianhong Tu、Jianwei Zhang、Jianxin Yang、Jiaxi Yang、Jing Zhou、Jingren Zhou、Junyang Lin、Kai Dang、Keqin Bao、Kexin Yang、Le Yu、Lianghao Deng、Mei Li、Mingfeng Xue、Mingze Li、Pei Zhang、Peng Wang、Qin Zhu、Rui Men、Ruize Gao、Shixuan Liu、ShuangLuo、Tianhao Li、Tianyi Tang、Wenbiao Yin、Xingzhang Ren、Xinyu Wang、Xinyu Zhang、Xuancheng Ren、Yang Fan、Yang Su、Yichang Zhang、Yinger Zhang、Yu Wan、Yuqiong Liu、Zekun Wang、Zeyu Cui、Zhenru Zhang、Zhipeng Zhou 和 Zihan Qiu。《Qwen3 技术报告》，2025 年。

[30] Cheol-Hwan Yoo、Jang-Hee Yoo 和 Jaeyoon Jang。CARE-VL：用于早期ASD筛查的领域专用视觉语言模型。载于 医学图像计算与计算机辅助干预 – *MICCAI 2025*，《计算机科学讲义》，第57–66页。Springer，Cham，2025年。doi：10.1007/978-3-032-04971-1_6。

[31] Lianmin Zheng、Wei-LinChiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。使用 MT-Bench 和聊天机器人竞技场评判大语言模型裁判。载于 神经信息处理系统进展（数据集与基准赛道），2023年。

[32] Wenjie Zheng、Qiming Xie、Jianfei Yu、Yang Wang、Lei Cao、Fei Wang、Shijin Wang、Rui Xia 和 Chengqing Zong。通过与 PHQ 对齐的症状摘要实现临床访谈中的可解释多模态抑郁识别，2025年。

[33] Han Zhou、Xingchen Wan、Lev Proleev、Diana Mincu、Jilin Chen、Katherine Heller 和 Subhrajit Roy。批次校准：重新思考上下文学习和提示工程中的校准。载于第十二届国际学习表征会议（*ICLR*），2024 年。

试点实验详述

根据视频预测总分。由视频编码器提取的片段特征（VideoMAE V2 [25]），转换为每名儿童的异常分数，并通过岭回归进行汇总，未能通过预测 ADOS总分的四项预注册检验（最佳置换检验$z = +1.06$，$p \geq 0.167$）。始终预测训练集均值的平均绝对误差为 4.73，低于所有尝试过的基于视频的方法（4.96–5.63）。视觉语言模型对每个 16 秒片段给出的绝对 0–4 分行为评分，按儿童汇总后，并不优于均值（平均绝对误差 4.83–4.85）。

同一儿童内的信号。 在一次盲测中，视觉语言模型以随机顺序查看了系统评定为最非典型的四个片段，以及来自同一儿童的四个随机片段，并对每个片段的非典型社会行为按0–10分进行评分。标记片段的得分高出0.64分（威尔科克森检验 $p = 0.0106$，$n = 45$名儿童）；在移除每次录制内容开始后的前两分钟或前五分钟后，该效应仍然存在。将评分者替换为一个无关且规模小得多的模型（Gemma-4-E4B）后，该效应依然存在（$+0.50$，$p = 0.0044$，$n = 46$）。

跨儿童比较。 这些结果结合起来表明，片段级模型判断能够对同一儿童内的片段进行排序，但无法将不同儿童置于共同量尺上。一项诊断性分析也指向同一结论。学习到的片段特征以79–85%的准确率识别出源录音，因此跨儿童比较主要反映了儿童及录制环境的差异，而较少反映行为本身。

添加特征。 三项与说话者无关的文字稿统计量能够预测总分（平均绝对误差 4.04，$z = +3.03$，$p = 0.004$；经九个特征类别校正后）。在此基础上加入视觉、姿态或按说话者分离的互动特征，会削弱或消除这一信号；按说话者分离的互动特征在全部十二项预注册的条目级检验中均未通过。在 $n = 48$ 的样本量下，在特征与得分之间学习更丰富的映射，在这些试点中并未成功。