使用多模态大语言模型进行手写工程考试评分

Janez Perˇs、Jon Muhoviˇc、Andrej Koˇsir 和 Boˇstjan Murovec 卢布尔雅那大学电气工程学院{janez.pers},{jon.muhovic},{andrej.kosir},{bostjan.murovec}@fe.uni-lj.si

arXiv:2601.00730v1 [cs.CV] 2 Jan 2026

摘要

手写*STEM*试卷能够展现开放式推理和图表，但人工评分耗时且难以扩展。我们提出了一套端到端工作流，用于对扫描后的手写工程测验进行评分，并采用多模态大语言模型（*LLM*），同时保留标准考试过程（*A4* 纸、不限制学生的书写方式）。讲师只需提供手写参考解答（*100%*）和一组简短的评分规则；系统会将参考解答转换为纯文本摘要，以此指导评分，同时不暴露参考答案扫描件。我们采用多阶段设计来保证可靠性，包括防止对空白答案评分的格式／内容存在性检查、独立评分器集成、主管汇总，以及配合确定性验证的固定模板，从而生成可审计、机器可解析的报告。我们在洁净室方案下，使用斯洛文尼亚语对一份留出的真实课程测验评估冻结后的处理流程，其中包含手绘电路示意图。使用先进的后端（*GPT-5.2* 和 *Gemini-3 Pro*）时，完整流程与讲师评分的平均绝对差为 ≈ 8 分，偏差较低，估计人工复核触发率为 $\approx 17\%$（置信度$D_{\max} = 40$）。消融实验表明，简单提示以及移除参考解答都会显著降低准确率，并导致系统性过度评分，证实结构化提示和参考依据至关重要。

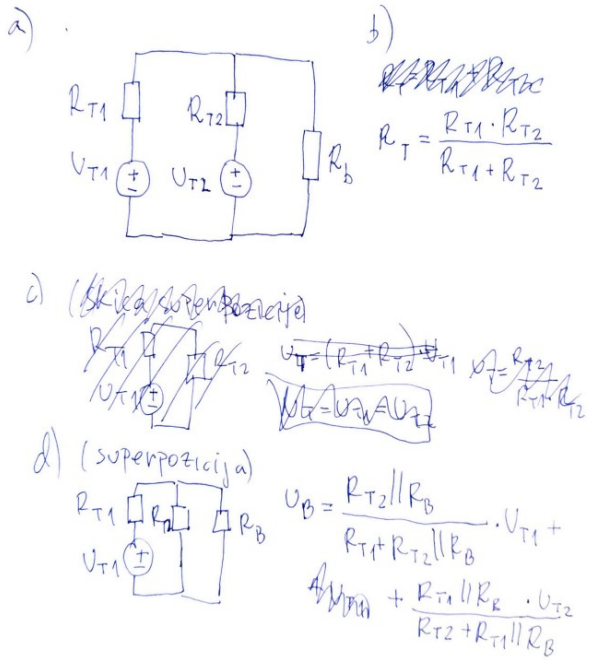

图1. 我们的系统旨在处理的手写答案示例，其中包含图表（如电路）。示例由作者提供（并非学生作业）。完整试卷及其评分结果的链接作为补充材料，列于论文最后一节。

1. 引言

手写纸笔考试在科学、技术、工程和数学教育中仍很常见，因为这类考试能自然地引出开放式推理、中间步骤和草图（例如电路图），而这些内容难以通过纯数字化评估来捕捉。图1展示了其中的挑战。然而，这类考试的人工评分既耗时，又难以扩大规模。Learnosity 报告称，在一项针对 258 名美国教师的在线调查中，受访者平均每周花 9.9 小时评分和批改 [14]。在大学层面，教师情绪研究也表明，评分带来的体验可能相对令人不悦：在一项大规模美国教师样本中，评分与相比研究和教学时较少的积极情绪和较多的消极情绪相关联 [24]。除了直接的时间成本，面向从业者的综述还指出，大量评分会耗费时间和精力，而这些时间和精力原本可以用于课程规划和改进教学 [28]。

这一问题在STEM课程中更为突出，因为笔试和技术作业需要在时间压力下评估多步骤解答和图形成果。一份聚焦STEM的综述指出，随着招生人数和工作负载的增加，教师可能难以及时为劳动密集型评估提供反馈；尽管这些评估具有教学益处，他们仍可能减少练习机会 [27]。更广泛地说，自动评分系统通常旨在减轻教育工作者的工作负载

并缩短反馈周转时间 [27]。Gradescope等工作流系统表明，将扫描后的手写提交作答数字化并加以组织，可以减少评分开销，并支持更一致地应用评分标准 [25], 但这些平台基本上保留了一个根本性瓶颈：人类仍需阅读并为每份回答评分。

大语言模型（LLM）的最新进展为实现更自动化的评分开辟了道路。在仅针对文本的简答题评分方面，针对K–12简答题数据开展的实验报告称，GPT-4通过相对简单的提示即可达到接近人工评分者一致性的水平 [10]。受控的自动简答题评分评估进一步表明，评分表现对提示上下文十分敏感——包括是否提供参考答案——且在不同数据集和设置下可能出现显著差异 [11]。对于纸质STEM考试，可以直接利用手写解答图像评估多模态大语言模型；近期结果显示，在提示中加入官方解答和评分标准可提升一致性，同时也指出其准确性可能仍不足以支持无需监督的真实场景部署[5]。混合工作流程仍存在实际限制：在评分前将手写内容转录为文本或LaTEX时，转录错误可能影响后续评分；近期研究将大语言模型生成的成绩视为需要进一步人工核验的结果，并探索通过重复采样进行置信度估计 [15]。此外，针对较长的多部分作答的实际应用指出，手写内容转文本反复成为瓶颈，并报告了在较长解答和图表密集型作答中应用细粒度评分细则所面临的挑战[12]。综合来看，这些发现表明，要实现可靠的考试评分，不仅需要强大的多模态模型，还需要系统层面的设计：采用结构化提示、验证与聚合，以及确定性后处理，以确保输出一致。

在本研究中，我们关注对包含自由文本和手绘图的扫描版手写 STEM 考试进行端到端评分。我们提出了一种多阶段、多提示评分工作流程，并通过确定性后处理确保输出可靠且可审计；同时，在要求学生手写文本答案并绘制电气原理图的本科开放题工程考试上对其进行评估。

2. 相关工作

自动考试评分的研究历史悠久，早于现代大语言模型。在大语言模型出现之前，许多实用系统要么（i）将答案限制为计算机可读的格式（从而可直接评分），要么（ii）依靠手写识别/光学字符识别，从扫描的答卷中提取文本，再应用基于文本的评分方法。这一领域反复面临的挑战是数据可用性：真实的考试答卷通常属于私人资料，难以共享；而可公开分发的数据集往往只能部分反映真实考试条件。

2.1. 大语言模型出现之前的考试评分

在多模态大语言模型出现之前，自动评估大体分为两类：（i）基于文本的自动简答题评分（ASAG），以及（ii）先转录手写文字的文档处理流程。自动简答题评分的综述介绍了早期系统，这些系统基于人工设计的词汇/句法特征以及与参考答案的语义相似度构建，通常使用已评分作答训练监督模型，并采用标准的一致性/相关性指标进行评估 [4]。具有代表性的特征整合方法将多种语言特征类别结合起来，构建判别式评分模型 [23], ；而基于向量的方法则利用分布式表示和相似度评分进行评分 [16]。近期综述着重指出，基于文本的自动简答题评分已转向深度学习和预训练语言模型（包括基于 Transformer 的方法） [9]。

对于手写作业，一种常见策略是先进行手写识别/光学字符识别，再基于文本评分。早期研究将手写识别与自动作文评分相结合，并指出端到端评分质量取决于转录质量 [26]。针对简短的手写答案，Rowtula等人提出了一种基于词定位的方法，避免完整转录，转而依靠检索式信号进行后续评估 [22]。与此同时，Gradescope 等系统级工具通过动态评分标准和答案分组，扩大了扫描版提交作答的评分工作流程 规模，但仍依赖人工给分 [25]。针对结构化 STEM 问题，研究者提出了基于聚类的方法，将解题结构相似的作答分组，以便教师为各组标注并据此发放（部分）分数 [13]。总体而言，LLM之前的评分系统要么假设文本清晰易读，要么依赖容易出错的转录前端；大多数系统都无法端到端稳健地处理不受约束的手写文字、图表和多页考试上下文。

2.2. 大语言模型时代的考试评分

自2023年以来，大语言模型一直被评估为通用评分器，最初用于评阅纯文本作答，后来逐渐扩展到手写作业的图像。在 K–12 简答题数据上，Henkel等人报告称，经过相对简单的提示，GPT-4 的评分结果与人工评分者一致性接近 [10]。其他自动短答案评分研究表明，结果取决于提示上下文和实验设置，包括是否提供参考答案 [11]。在高等数学场景中，Gandolfi 报告称，GPT-4 能够给出有用的解答和评分理由，同时也记录了可靠性问题，例如偶尔出现连贯性丧失和幻觉，因此需要明确的验证和护栏机制 [6]。

一个重要转变是直接对扫描手写解答进行多模态评分。针对大学数学考试，Caraeni等人评估了 GPT-4o 直接根据手写图像进行评分的效果，并报告称，在提示词中加入官方解答和评分标准后，评分结果的一致性有所提高，

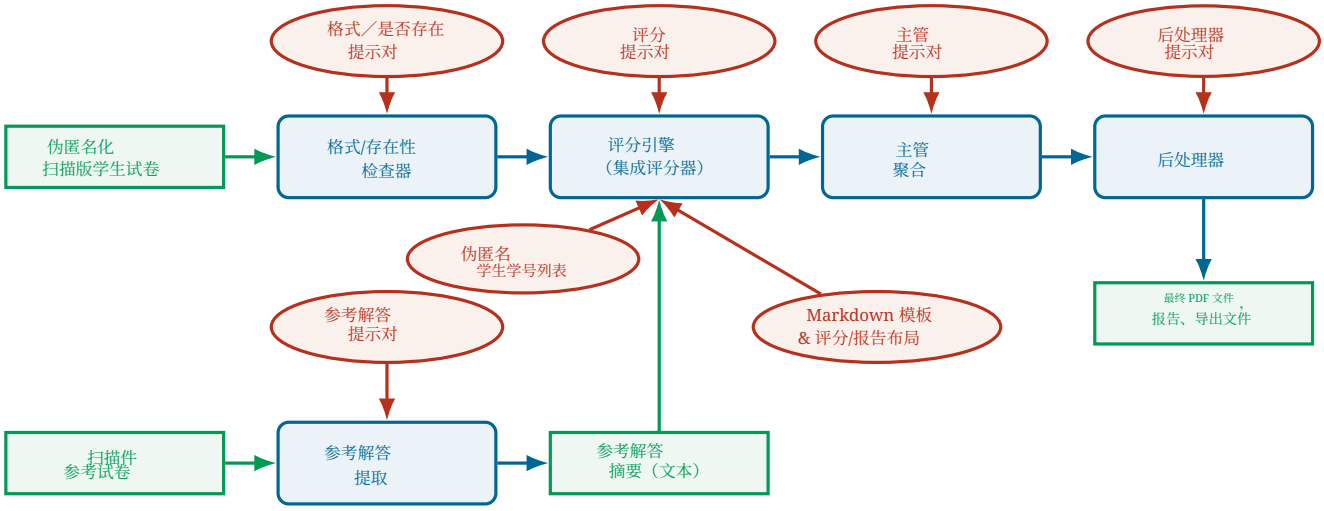

图 2. 系统概览。绿色方框表示数据产物，红色椭圆表示提示对和模板，蓝色方框表示处理阶段。所有蓝色处理阶段均调用多模态大语言模型后端（例如 GPT-4o、GPT-5.x、Gemini 或 Mistral），而参考解答扫描件会转换为纯文本摘要，并嵌入评分提示中。

同时指出，总体准确率仍是其实际应用的制约因素 [5]。混合处理流程仍然具有现实意义：Liu 等人研究了利用光学字符识别/LaTeX 转写阶段对大学数学手写考试进行人工智能辅助评分，并强调转写敏感性和人工核验的作用，同时还通过多次评分运行进行采样，探索置信度估计。[15]在更大规模的场景中（例如，数百份物理学试卷），工作流程研究仍然指出，将手写文字转换为文本是一项实际瓶颈；细粒度评分细则难以应用于较长的解答；而图示/图表的处理也持续存在困难 [12]。针对手绘图表等非文本输出，近期的比较研究评估了图表评分任务中元学习方法与视觉语言模型的表现 [21]。其他多模态评分范式则采用 CLIP 风格嵌入（可选择结合光学字符识别），将视觉信息纳入评分 [2], ；新的基准（例如 DrawEduMath）则系统地探查视觉语言模型对学生手绘数学图像的解读能力，并记录其仍然存在的不足 [3]。

2.3. 数据集

支持对手写考试评分进行端到端 评估（输入图像，输出数字分数）的公开数据集十分罕见，主要原因在于真实考试答卷涉及隐私，通常是在机构同意/IRB审批下收集的，因此限制了公开发布。因此，近期的多模态评分研究是在包含扫描页面、评分标准和人工评分的私有课程考试数据集中进行评估的 [5, 12], ，而以图表为重点的手写语料库也同样来自机构 [21]。少数包含手写文字的开放资源要么针对简答题 [7]，要么作为视觉语言模型理解基准，而非基于分数的评分数据集 [3]；包含图像作答内容的大型教育语料库通常无法再分发 [1]。因此，为了在真实的科学、技术、工程和数学场景下评估端到端处理流程

考试条件（多部分解答和图表），我们收集并评分了自己的考试数据。

3. 方法

该处理流程旨在尽量减少对考试过程的干扰。除使用标准 A4 纸外，不对学生或讲师组织考试的方式施加任何限制。唯一的额外要求是，讲师需提供一份代表满分（100%）答案的手写参考解答。这些要求直接影响了图 2 所示的架构。

3.1. 系统架构

图 2 所示的处理流程将确定性文档处理与少数几个调用大语言模型的阶段结合起来。尽管该系统由大量编排代码实现，但其主要方法学贡献在于处理流程的结构，以及至关重要的提示对和固定模板，它们使整体行为稳定且可由机器解析。受篇幅限制，我们不再复现完整的提示词和模板。

在不暴露参考答案扫描件的情况下进行参考答案条件化。讲师提供一份手写参考答案，代表满分（100%）答案。专门的参考答案提取阶段会将扫描的参考答案转换为纯文本摘要，并将其注入评分提示词；评分过程中不会使用参考图像本身。该阶段使用能力强大的多模态后端运行（我们在实验中使用 GPT-5.2-pro），以可靠地识别非约束手写文字和草图。

答案存在性保护机制。 评分前，格式／存在性检查器会预测哪些题目包含实际的学生答案。这项保护措施是在发现少数情况下，模型会在整块答案区域留白时幻觉内容后引入的——

此后，尽管这类情况不会反映在汇总指标中，但在实际运行中绝不可接受；存在性列表可防止评分器给未作答的题目打分。

集成评分与监督器汇总。对于每项任务，评分由 $K = 3$ 个相互独立的无状态模型调用组成的集成执行，这些调用根据固定模板生成结构化草稿。随后，监督模型将这些草稿合并为单一的考试级输出，强制确保模板合规性，应用存在性判定（未作答的任务得分为 0%），并标记不一致之处，以便进行可选的人工裁决。这种集成加监督器的设计降低了方差，并提高了对偶发模型故障的鲁棒性。

后处理与导出文件。最后，后处理器生成可直接用于展示的产物（例如报告组装和可选翻译），同时保留所有数值字段和模板结构，从而支持确定性解析、审计和导出。

隐私。要求学生不要在试卷任何位置写下姓名。相反，他们必须填写注册号，因此在调用大语言模型之前，数据已进行伪匿名化处理。大语言模型会获得经过清理的学生名册，其中包含参加考试的所有学生的注册号；它只需从这个有限的注册号集合中找出写在试卷第一页顶部的号码。所有去匿名化操作均在评分完成后于本地进行。强制使用的 Markdown 模板确保模型在输出中的特定位置生成学生的伪身份，以便进行解析。如果解析失败，则需要由人工从扫描件中读取该号码。

3.2. 提示词与模板

每个调用大语言模型的阶段都通过提示对（系统提示词 + 用户提示词）实现，并由严格的 Markdown 模板支撑。系统提示词规定角色、约束和禁止行为，用户提示词则注入实例特定上下文（任务标签、学生答卷扫描件、参考摘要以及可选的教师规则）。模板严格规定章节层级和数值字段，将原本具有概率性的输出转换为可由确定性后处理验证和解析的产物；相比之下，直接要求大语言模型“给考试评分”并不可靠，因为这会导致格式漂移以及评分标准应用不一致。

面向教师的配置。课程特定的输入只有两项：(i) 手写参考答案扫描件，以及 (ii) 一份简短的评分规则清单。处理流程中的其他所有阶段、提示词和模板都旨在跨课程保持不变。规则编号为 (`[R1]`, `[R2]`, ...)，并逐字附加到与评分相关的提示词中，同时要求评分者在解释中引用适用的规则编号，以提高可审计性并便于人工审核。在我们的实验中，大多数规则都适用于不同的 STEM 测验，只需进行少量课程特定调整（例如，根据拓扑结构而非绘图方向评估电路草图）。

模板约束。我们使用两种模板：单题评分器模板和主管模板。两者都禁止添加或删除章节，并强制采用固定的评分行及确定性数值模式（得分、权重、贡献分）。对于每道题，模板会分别列出题目文本、学生答案的纯文本摘要、附有必需短元标签（包括规则引用）的正确性说明，以及唯一一行评分。期末考试总分必须等于各题贡献分之和（不进行归一化），从而支持自动一致性检查。

语言。除非另有说明，我们实验中使用的提示词、规则和模板均以斯洛文尼亚语编写。处理流程本身与语言无关：适配其他语言只需翻译这些文本材料，无需更改处理阶段。

4. 实验

自动评分质量包含一些仅靠汇总指标无法充分体现的重要定性方面（例如生成的反馈的连贯性和教学实用性）。因此，除下文报告的定量评估外，我们还提供补充材料，其中包括“B班”考试的扫描版参考解答，以及一份相应的模拟解答；该模拟解答实际上由使用 GPT-5.2 且思考设置为 “ 高 ” 的处理流程进行了评分。我们提供这些材料，是为了展示所生成反馈的结构和深度。补充材料的链接位于手稿末尾。

4.1. 实验方案

我们的评分流水线不可训练，因此完全运行于零样本设置下。为限制迭代式提示词设计和评分标准设计带来的潜在偏差，我们采用洁净室方案，使用两门并行课程：**A班**（开发）和**B班**（留出评估）。

学期期间，系统每周都会用于 A班测验，其输出由讲师和学生审阅，从而为系统行为提供形成性反馈。第9周之后，处理流程和所有提示词均已冻结。随后，将冻结后的系统应用于一份 B班扫描测验，处理流程、提示词或模板均未作任何修改。与 A班相比，唯一的变化是面向教师的评分规则调整，以反映以下方面的标准做法：

电路草图：电路答案依据电气拓扑结构而非绘图方向进行评估。本文报告的所有定量和定性结果均仅基于这项留出的 B班评估，并与 B班讲师给出的成绩进行比较。

为确保隔离，B班测验在学期期间仅由讲师评分，并扫描归档。除第 2 周进行的有限试点运行（未用于分析）外，在留出评估之前，系统开发人员未接触或检查过 B班的材料。

4.2. 数据集

A班是一门本科通信技术课程（约有 30 名注册学生），B班则是一门工业电子学课程（约有 15 名注册学生）。两门课程均每周安排一次简短的笔试测验，在每节课开始时进行，时长 20 分钟，内容涵盖前一周的课程材料。测验包含开放式问题，要求学生手写作答，并在适当情况下绘制图表或电路示意图。

每周测验并非严格要求必须参加，但学生必须通过其中若干次测验才有资格参加期末考试；如果每周表现出色，也可以选择以此替代期末考试成绩。这既提供了有意义的激励，又使整体压力保持在适中水平。出于隐私原因，我们不会公开学生提交材料、扫描件或成绩；我们仅报告汇总的性能指标以及留出评估问题的完整文本。

语言。 测验和学生答案均使用非英语语言（斯洛文尼亚语）。因此，主要评估中使用的所有提示对、评分规则和报告模板均以斯洛文尼亚语编写。该处理流程本身与语言无关：特定语言的内容仅限于提示词、规则和模板，这些内容可以翻译成其他语言（也可以借助现代基于大语言模型的工具），而无需更改处理流程的各个阶段。

留出考试内容。 表1 列出了用于评估的留出B班测验中的三道题（含权重）。

权重	题目文本
25%	使用分压器时，我们会遇到一种权衡—— 断开：出于某些原因，我们希望用电阻器构成分压器， 并且使用尽可能小的电阻值 ；但另一方面，我们又希望 电阻值尽可能大。请解释这一矛盾 及其背后的原因。
25%	连接时必须满足什么条件? 由电阻器 R_1组成的分压器的负载 和 R_2，以使负载成为电流驱动?
50%	两个戴维南电压分别为 U_{t1} 和 U_{t2}的电池 以及戴维南内阻分别为 R_{t1} 和 R_{t2} 的电池 并联连接后再连接到一个负载 R_b. • 画出相应的电路。 • 写出戴维南电阻的表达式 组合电源的。 • 写出戴维南电压的表达式 组合电源。 • 写出负载两端电压的表达式 负载。

表1. 留出评估中使用的考试题目。仅公开题目文本和分值；不共享学生数据。题目由斯洛文尼亚语翻译而来；提供译文以便阅读。

4.3. 评估指标

我们使用考试层面的指标评估评分系统，这些指标用于量化其与人工评分的一致性。我们以B班讲师给出的考试成绩作为真实值，这是唯一可用的参考。

除非另有说明，所有指标均在整份考试（三项任务）层面计算，从而在真正的端到端设置下评估系统。重要的是，我们区分了两种变异性的来源：(i) 评分流水线将内部集成作为其固定设计的一部分；(ii) 对完整评估进行独立重复，这些重复仅用于估计实验变异性。

流水线参数：集成规模。 在单次流水线执行中，每项任务都由一个包含 $K = 3$ 个相互独立的无状态模型调用的集成进行评分，随后由监督模型将生成的草稿合并为单一的考试级评分。我们将学生 i的这一最终监督模型汇总输出记为 g_i^{LLM}。集成规模 K 是固定的流水线参数（不会在实验中改变），早期选定该规模时，旨在平衡推理成本与对偶发模型故障的鲁棒性。

设 N 表示学生人数，g_i^{LLM}表示系统为学生 i评定的考试成绩，g_i^{H} 表示人工教师评定的相应成绩。我们将有符号评分差异定义为 $\Delta_i = g_i^{\mathrm{LLM}} - g_i^{\mathrm{H}}$。

平均绝对差（MAD）。 总体评分准确率采用自动评分与人工评分之间的平均绝对差来衡量：

$$\mathrm{MAD} = \frac{1}{N}\sum_{i=1}^{N} |\Delta_i| \,. \tag{1}$$

绝对差的标准差。 为衡量学生评分误差的变异性，我们计算绝对差的标准差：

$$\sigma_{|\Delta|} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} \left(|\Delta_i| - \mathrm{MAD}\right)^2}. \tag{2}$$

评分偏差。我们通过有符号差值的均值来量化系统性的评分过高或过低：

$$\text{Bias} = \frac{1}{N}\sum_{i=1}^{N}\Delta_i. \tag{3}$$

人工复核触发率。除了与人工评分的一致性之外，我们还量化了当多个自动评分器意见不一致时所需的预期人工整合量。该指标不依赖人工参考评分，仅根据集成输出计算。

令 $s_{i,k}$ 表示集成中评分器 k 赋予学生 i 的考试层级分数，其中 $k = 1,\ldots,K$。对于给定的分歧阈值 $D_{\max}$，若评分器之间的最大成对分歧超过该阈值，则对学生 i 触发人工复核，即

$$\max_k s_{i,k} - \min_k s_{i,k} \geq D_{\max}. \tag{4}$$

我们将相应的触发指标定义为

$$T_i(D_{\max}) = \begin{cases} 1, & \text{if a review is triggered for student } i, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

阈值 $D_{\max}$ 处的人工复核触发率则由下式给出

$$\text{TR}(D_{\max}) = \frac{1}{N}\sum_{i=1}^{N} T_i(D_{\max}). \tag{6}$$

在本研究中，TR($D_{\max}$) 在考试层级进行估计，使用各评分器给出的期末考试成绩。在实际部署中，同一准则可应用于更细的粒度，例如单个题目或答案的层级，以进一步定位并减少所需的人工干预量。

实验参数：评估重复次数。与流水线集成分开，我们将整个 评估重复 $R = 3$ 次，以衡量由随机模型输出引起的各次运行之间的变异性。每次重复对应于在系统配置不变的情况下对评分流水线进行完整重跑（包括所有模型调用和监督器汇总）。对于每个指标，我们报告三次逐次运行的数值，以及它们在 R 次重复中的均值和标准差。选择$R = 3$ 反映了稳健性与重复多模态推理的计算成本之间的实际折中。

5. 结果

所有结果均基于保留的 B班考试（第 4 节）进行报告。我们使用第 4.3 节中定义的考试级指标（平均绝对偏差、$\sigma_{|\Delta|}$、偏差）；在适用情况下，还报告基于评分者分歧的人工复核触发率。除非

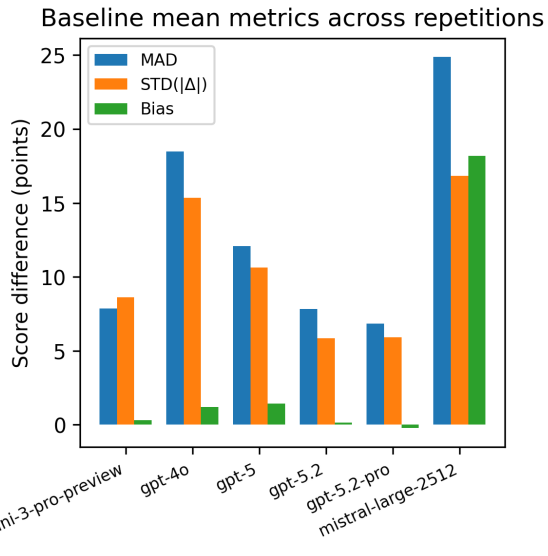

图3. 各后端在考试层面的基线表现（在可用情况下，取 $R = 3$ 次重复运行的均值；GPT-5.2-pro：单次运行）。

另有说明，每个后端均使用完整流程配置进行评估，并重复运行 $R = 3$ 次，以估计实验变异性；由于成本原因，GPT-5.2-pro 的结果基于单次运行。OpenAI 模型通过官方 OpenAI API [17]，访问，而其他模型则通过 OpenRouter [20] 访问。

模型可行性筛选。图 3比较了完整流程下各后端的表现。表现最强的后端（GPT-5.2、GPT-5.2-pro 和 Gemini-3 Pro）平均绝对偏差为个位数且偏差较小，表明其评分与讲师给出的考试成绩高度一致。相比之下，GPT-4o 和 Mistral 3 的偏离明显更大；Mistral 3 还表现出显著的正偏差，这与其在本次考试中系统性地给出过高评分相符。

流程指导的消融实验。为单独考察提示工程和参考答案条件化的影响，我们在简单提示设置下评估两个强大的后端（图 4）。在这种设置中，提示词仅要求采用 Markdown 输出格式，但不包含评分规则和参考解答摘要，也不启用主管聚合。由于未启用聚合，我们将三位评分者各自的输出视为相互独立的评分尝试，并报告其均值。对于 GPT-5.2 和 Gemini-3 Pro，简单提示都会增大平均绝对偏差并引入显著的正偏差，这证实了结构化提示和参考依据对于获得可靠评分是必不可少的。

估计的人工审核工作量。图5 报告了提交作答中需要人工整合的比例，该比例是阈值分歧度的函数——

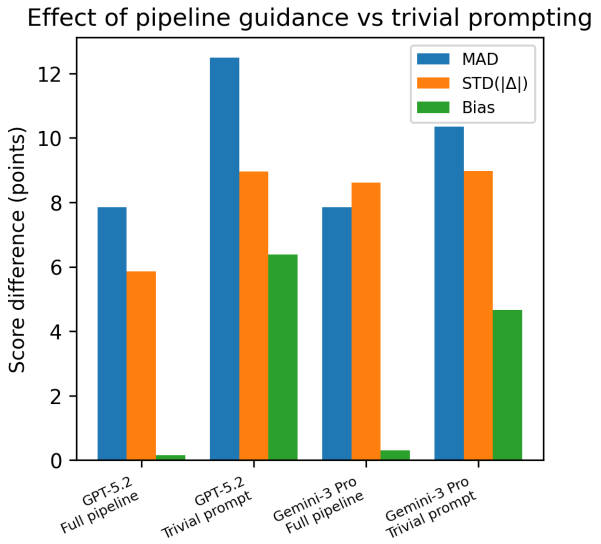

图4. 两种强力后端的完整流程与简单提示对比（简单提示：无规则、无参考、无主管）。

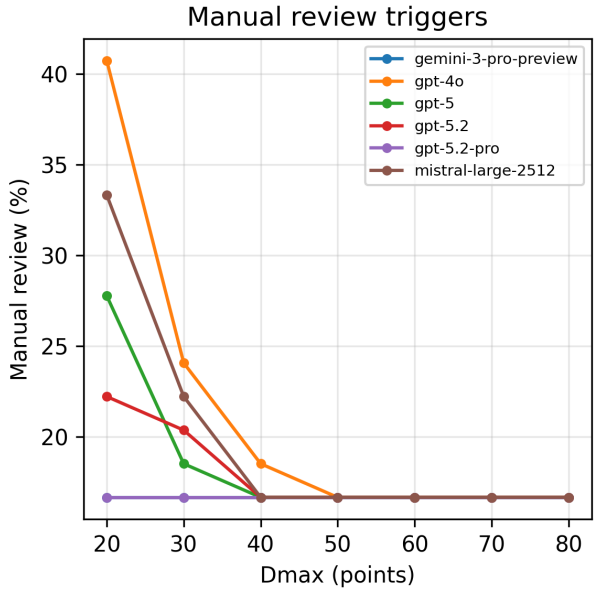

图 5. 估算的人工复核触发率与分歧阈值 $D_{\max}$（对可用重复实验取均值）。

old $D_{\max}$，根据集成评分器输出在考试层级上计算得出。在严格阈值下（例如， $D_{\max} = 20$–30 分），较弱的后端会产生显著较高的审核率，反映出评分稳定性较差。阈值较大时，所有模型的触发率都会下降，表明只有少数提交作答存在严重的评分者分歧。这项分析通过量化安全部署该系统所需的人工工作量，对准确率指标进行了补充。

参考条件消融。 表2 对两个表现最佳的后端（ GPT-5.2、Gemini-3 Pro）在三种制度下的参考资料使用情况进行消融实验：完整流程（将提取出的参考资料转换为文本并注入提示词）、无参考和图像参考（仅使用参考图像，不进行文本提取）。每个单元格报告 run1/run2/run3 以及均值±标准差；每个模型和指标的最佳结果以粗体标示。我们根据各评分者得分报告平均绝对偏差（MAD）、$\sigma_{|\Delta|}$、偏差以及考试级人工复核触发率 TR（$D_{\max} = 40$）。对于 GPT-5.2，移除参考资料会增大平均绝对偏差并引入明显的正偏差；仅使用图像参考可在一定程度上降低平均绝对偏差，但仍存在偏差，这凸显了文本提取对于实现校准评分的作用。对于 Gemini-3 Pro，两种消融主要都会增大正偏差，而平均绝对偏差/$\sigma_{|\Delta|}$的改善有限；这与流程开发主要基于 GPT-5 模型进行的情况相符，因此 GPT-5.2 从中获益更多，Gemini 则获益较少。

学生反馈（A班）。在系统实际用于 8 次每周人工智能评分测验后，我们收集了 A班学生的初步反馈。共有 14 名学生填写了匿名问卷；调查时，学生已收到详细的 PDF反馈，并且可以使用申诉流程。学生对人工智能优先评分的态度总体积极：64% 表示持积极态度或更倾向于采用该方式，21% 持消极态度，14% 持中立态度。问及是否从该系统中受益时，71 % 的学生回答是（29% 回答否）。学生报告的益处（可多选）主要包括详细的解释（41%）、感知公平性／公正性（34%）以及快速反馈（21%）。主要担忧包括漏掉答案（35%）、错误多于教授（26%）以及考试难度发生变化（22% 认为要求降低，17% 认为要求提高）。总体而言，43% 认为优点大于缺点，29% 认为缺点更多，21% 表示没有差别。

6. 讨论

据我们所知，现有文献尚未描述过类似的端到端框架：该框架使用多模态大语言模型对扫描的、多页的、带有图表的手写*STEM*试卷进行评分，同时生成可确定性解析的输出，并具备明确的护栏机制、汇总功能和可审计报告。我们的结果表明，通过这样的工作流设计，现代多模态后端能够对简短的工程测验进行评分，且一致性已足够高，可以在有限人工干预下实现实际应用。在与教师的非正式讨论中，我们反复遇到这样的质疑：对于非约束手写文字和草图，能否达到本文所报告的准确度；这些实验表明，只要围绕模型失效模式而非理想化提示词来设计系统，这种能力如今已经成为现实。一个关键的背景因素是时间：根据我们的经验，这类端到端方法直到2025年末一代具备多模态推理能力的模型发布后，才在实践中真正变得可行

模型	制度	MAD	$\sigma_{\lvert\Delta\rvert}$	Bias	TR($D_{\max}=40$) [%]
GPT-5.2	完整流程	**7.7/8.4/7.4** (7.8±0.4)	**4.2/6.5/6.9** (5.9±1.2)	**0.6/-0.4/0.3** (0.2±0.4)	**16.7/16.7/16.7%** (16.7±0.0)%
	无参考	9.8/11.2/9.5 (10.2±0.8)	8.4/6.4/6.0 (6.9±1.1)	6.5/6.4/6.1 (6.4±0.2)	22.2/16.7/16.7% (18.5±2.6)%
	图像参考	7.7/8.2/8.2 (8.1±0.2)	5.9/4.9/7.1 (6.0±0.9)	3.9/3.4/4.8 (4.0±0.5)	**16.7/16.7/16.7%** (16.7±0.0)%
Gemini-3 Pro	完整流程	**7.1/9.8/6.7** (7.9±1.4)	5.0/15.1/5.7 (8.6±4.6)	**3.0/-2.7/0.7** (0.3±2.3)	**16.7/16.7/16.7%** (16.7±0.0)%
	无参考	7.4/8.9/9.6 (8.6±0.9)	6.8/8.2/8.6 (7.9±0.8)	6.6/8.4/9.0 (8.0±1.0)	**16.7/16.7/16.7%** (16.7±0.0)%
	图像参考	7.9/8.5/8.4 (8.3±0.3)	**6.9/6.2/7.4** (6.9±0.5)	7.1/7.8/7.7 (7.5±0.3)	**16.7/16.7/16.7%** (16.7±0.0)%

表 2. 三种参考模式下表现最佳的两个后端的消融研究：完整流程 （将提取出的参考内容转换为文本并注入提示词）、无参考和图像参考 （仅提供参考图像，不进行文本提取）。每个单元格依次报告 run1/run2/run3，以及各次运行的均值±标准差。指标包括考试级平均绝对偏差、$\sigma_{|\Delta|}$、偏差和人工复核触发率 $TR(D_{\max}=40)$，均根据各评分者的考试分数计算。每个模型和指标的最佳结果均以粗体标出（均值最低；对于偏差，则为绝对均值最小）。

主要提供商推出的模型（例如 GPT-5/GPT-5.2 和 Gemini 3）[8, 18, 19]。我们筛选的早期后端表现出明显更大的偏离和更强的偏差（图3），进一步凸显了既需要高能力模型，也需要系统级防护措施。本文所介绍的处理流程最初是为实际教学用途而构建的，而不是面向基准优化的研究原型。因此，我们的评估仍属初步阶段：结果来自一次留出测验，并以一名人工评分者作为参照；由于隐私限制，底层考试数据无法以原始形式发布。因此，我们计划将验证扩展到规模更大、更加多样化的数据集，并在完成隐私审查和获得机构批准后，发布代码、提示词以及一个适用于标准化评估的配套数据集。对于真实手写考试评分而言，缺乏广泛可用的端到端数据集仍是该领域的一项实际障碍；我们将这项工作视为迈向实现此类评估可行且可重复开展的第一步。

7. 结论

我们提出了一套端到端工作流，用于为扫描版手写工程考试评分。这项工作的核心贡献是一种稳健的系统设计，将提示对和固定模板与确定性验证、集成评分及主管汇总相结合，从而把概率模型行为转化为可审计的评分材料。在一项留出的真实课程测验中，最先进的多模态后端所得成绩与讲师评分高度一致，预计人工复核率也在可控范围内，这表明该方法有望应用于简短的形成性评估。我们将此作为初步证据发布，表明在严格遵守工作流限制的情况下，如今已能对真实的手写科学、技术、工程和数学作业进行自动评分；待进一步验证并完成隐私审查后，我们计划开展更广泛的评估并开放源代码（同时提供配套数据集）。

伦理考量

本研究的伦理考量旨在保护研究参与者和拟议技术的未来用户。主要伦理问题包括敏感数据（如成绩）的处理、参与实验可能影响学生表现的风险，以及学生与教师之间的权力失衡。

我们的方法以《贝尔蒙报告》 [29]的三项原则为指导：

1. **尊重个人**：通过尊严、能动性和知情同意来保障自主权；2.**行善**：在最大限度发挥潜在益处的同时，尽量降低风险和伤害；以及 3. **公正**：确保公平待遇，并公平分配益处和负担。

在此背景下，一个核心考量是，预期益处是否足以抵偿施加于受影响群体的负担。学生可能获得的益处包括：(i) 更客观的评分，以及 (ii) 教师提供的高质量反馈。个性化反馈尤其有价值，因为教师通常难以大规模提供此类反馈。由此产生的反馈也可能有助于改进课程设计和教学实践。重要的是，自动评分系统并未影响学生的正式课程成绩或考试成绩；所有评估均像未开展本研究时一样由人工评分。

要全面部署，必须进行正式的伦理审查，包括对实验设计和数据保护措施的评估。尽管如此，我们认为，预期给学生带来的益处大于相关负担。

人工智能使用声明

整个编排流水线和实验代码（约 13.000 行 Python 代码）是在 OpenAI 的 GPT5-codex 工具和 GPT5-pro 模型的帮助下编写的。相关研究的检索使用了 OpenAI 的人工智能代理（DeepResearch），手稿则是通过向 GPT5-pro 口述内容撰写的。作者对手稿进行了全面检查，并在必要时作了修订。

补充材料

我们提供了“B班”考试的扫描版参考解答，以及相应的模拟解答。该模拟解答实际上是由使用 GPT-5.2、思考模式设为“高”的处理流程评分的。可通过以下链接访问：
h
t
tps://lmi.fe.uni-lj.si/en/janez-pers-2/supplementary-material/

致谢

我们感谢 EC/EuroHPC JU 以及斯洛文尼亚高等教育、科学与创新部通过 SLAIF 项目（资助编号 101254461）提供的支持。

参考文献

[1] Sami Baral、Anthony F. Botelho、John A. Erickson、Priyanka Benachamardi 和 Neil T. Heffernan。《提升数学学科学生开放式回答的自动评分》。载于第*14*届国际教育数据挖掘会议（*EDM 2021*）论文集，第130–138页，法国巴黎，2021年。国际教育数据挖掘学会。会议日期：2021年6月29日至7月2日。3

[2] Sami Baral、Anthony Botelho、Abhishek Santhanam、Ashish Gurung、Li Cheng 和 Neil Heffernan。利用图像自动评估学生的数学作答。载于第*16*届教育数据挖掘国际会议论文集，第362–369页，印度班加罗尔，2023年。国际教育数据挖掘学会。3[3] Sami Baral、Lucy Li、Robert Knight、Alice Ng、Luca Soldaini、Neil Heffernan 和 Kyle Lo。DrawEduMath：使用专家标注的学生手绘数学图像评估视觉语言模型。载于*2025*年计算语言学协会北美分会会议论文集：人类语言技术，第7110–7132页。计算语言学协会，2025年。arXiv 版本：https://arxiv.org/abs/2501.14877。3[4] Steven Burrows、Iryna Gurevych 和 Benno Stein。自动简答题评分的时代与趋势。《国际人工智能教育杂志》，25(1):60–117，2015年。2[5] Adriana Caraeni、Alexander Scarlatos 和 Andrew S. Lan。评估 GPT-4 在数学考试手写解答评分中的表现，2024年。2，3[6] Alberto Gandolfi。教育领域中的 GPT-4：评估其在求解微积分问题和作业评分方面的适用性、可靠性及连贯性丧失。《国际人工智能教育杂志》，35(1):367–397，2024年。2

[7] Christian Gold 和 Torsten Zesch。手写 ASAP 简答题评分。Zenodo 数据集，2020年。版本 1.0；访问日期：2025-12-14。3[8] Google。Gemini 3 智能新时代。https://blog.google/products/gemini/gemini-3/，2025年。访问日期：2025-12-23。8[9] Stefan Haller、Adina Aldea、Christin Seifert 和 Nicola Strisciuglio。深度学习自动简答题评分综述：从词嵌入到 Transformer，2022年。2[10] Owen Henkel、Adam Boxer、Libby Hills、Bill Roberts 和 Zachary Levonian。大语言模型能胜任评分吗？一项实证研究，评估大语言模型在 K-12 教育中评阅简答题的能力。载于第*11*届 *ACM* 大规模学习会议（*L@S*）论文集，第300–304页，2024年。arXiv 版本：https://arxiv.org/abs/2405.02985。2

[11] Gerd Kortemeyer。预训练大型语言模型 GPT-4 在自动简答题评分中的表现。*Discover Artificial Intelligence*，4(1):47，2024年。也可通过 ETH Research Collection 获取开放版 PDF（链接请见 Springer 页面）。2

[12]Gerd Kortemeyer、Julian N¨ohl 和 Daria Onishchuk。使用人工智能为手写热力学考试提供评分辅助：一项探索性研究。《Physical Review Physics Education Research》，20(2)：020144，2024年。2，3[13] Andrew S. Lan、Divyanshu Vats、Andrew E. Waters 和 Richard G. Baraniuk。数学语言处理：对开放式数学题进行自动评分和反馈。载于第二届 ACM 大规模学习会议（L@S）论文集，第 167–176 页，2015年。2–。[14] Learnosity。三分之一的美国教师表示，过去 12 个月中曾考虑离开教育行业，原因是评分工作量过大。Learnosity EdTech 博客（新闻稿），2025年。发布于2025年3月26日。1[15] Tianyi Liu、Julia Chatain、Laura Kobel-Keller、Gerd Kortemeyer、Thomas Willwacher 和 Mrinmaya Sachan。人工智能辅助的大学数学考试手写简答题自动评分，2024年。HTML：https://arxiv.org/html/2408.11728v1。2，3[16] Ahmed E. Magooda、Mohamed A. Zahran、Mohsen A. Rashwan、Hazem M. Raafat 和 Magda B. Fayek。简答题评分的向量技术。载于第 29 届佛罗里达人工智能研究学会会议（FLAIRS）论文集，第 238–243 页，2016年。2–。[17] OpenAI。OpenAI API 参考文档。在线文档，2025年。访问日期：2025-12-23。6[18] OpenAI。推出 GPT-5。https://openai.com/index/introducing-gpt-5/，2025年。访问日期：2025-12-23。8[19] OpenAI。推出 GPT-5.2。https://openai.com/index/introducing-gpt-5-2/，2025年。访问日期：2025-12-23。8[20] OpenRouter。OpenRouter API 文档。在线文档，2025年。访问日期：2025-12-23。6[21] Behnam Parsaeifard、Martin Hlosta 和 Per Bergamin。学生手绘图表的自动评分：一种

元学习与视觉大语言模型的比较，2025年。 3[22] Vijay Rowtula、Subba Reddy Oota 和 C. V. Jawahar。迈向手写评估的自动化评估。载于*2019*年国际文档分析与识别会议研讨会（*ICDARW*），第 19–24 页，2019年。 2[23] Keisuke Sakaguchi、Michael Heilman 和 Nitin Madnani。将特征有效整合用于自动简答题评分。载于*2015*年计算语言学协会北美分会会议论文集：人类语言技术，第 1049–1054 页，科罗拉多州丹佛，2015年。计算语言学协会。 2[24] Carolin Schwab、Anne C. Frenzel、Jordan Jaeger、Allison BrckaLorenz 和 Robert H. Stupnisky。大学教师如何看待评分？控制—价值理论视角下的见解。《高等教育研究》，49(8)：1486–1503，2024年。在线发表于2023年10月15日。 1[25] Arjun Singh、Sergey Karayev、Kevin Gutowski 和 Pieter Abbeel。Gradescope：一种快速、灵活且公平的手写作业可扩展评估系统。载于 第四届 *ACM* 大规模学习会议（*L@S*）论文集，第 81–88 页，2017年。 2[26] Sargur N. Srihari、Rohini Srihari、Pavithra Babu 和 Harish Srinivasan。论手写作文的自动评分。载于第 *20* 届国际人工智能联合会议（*IJCAI-07*）论文集，第 2880–2884 页，2007年。 2[27] Le Ying Tan、Shiyu Hu、Darren J. Yeo 和 Kang Hao Cheong。基于人工智能技术的科学、技术、工程和数学领域自动评分系统综合综述。《数学》，13 (17)：2828，2025年。 1，2[28] Youki Terada 和 Stephen Merrill。教师为何应减少评分频率。Edutopia，2024年。发布于2024年11月8日。 1[29] 生物医学与行为研究人类受试者保护国家委员会。《贝尔蒙报告：保护研究人类受试者的伦理原则与指导方针》，1979年。发布于1979年4月18日。美国卫生、教育与福利部。由美国卫生与公众服务部人类研究保护办公室（OHRP）托管。 8