打通缓考流程：借助 JAMN Bot 支持学生——一款基于检索增强生成的移动端政策助手

Mark Farinas*、Alesandra Garcia*、Maria Reyes*、Jethro Nueva*、Nawaf Qureshi*、Mahmoud Elsaadany 和 Shokry Shamseldin*MacEwan University* 计算机科学系加拿大埃德蒙顿{farinasm, garciaa37, reyesm26,nuevaj,qureshin4}@mymacewan.ca{elsaadanym, shamseldins}@macewan.ca

arXiv:2611.32266v1 [cs.CL] 24 Nov 2026

摘要——**在一些教育机构中，考试延期等流程可能缓慢、不便且缺乏一致性，因此有必要实施新型数字系统。本文介绍 Schedulater，这是一款用于管理大学考试延期的移动应用。JAMN Bot 是集成于 Schedulater 中、由大语言模型（LLM）驱动的聊天机器人，可提供结合上下文的支持，帮助用户了解缓考流程。该系统采用检索增强生成（RAG）流程，并使用 Ollama（Llama 3.2 3B）进行自托管推理，从而在不牺牲隐私的情况下提供基于政策的响应。我们使用一个包含 43 个条目的基准测试评估该系统，问题涵盖考试延期、隐私、日程安排、跨政策以及对抗性与范围外问题；同时报告人工评估者之间的一致性，以及检索命中率指标。随后，我们采用基于 GPT-5.4 的大语言模型裁判方法，发现其与评分者的吻合度从轻微到中等不等，且会随着回答质量的提高而改善；并将排名与包含 Gemini 3.5 Flash 和 Qwen3.5 397B 的评审组进行对照。研究结果展现了开放权重模型与前沿模型之间的权衡，也表明大语言模型与人工评估的一致性会随回答质量而变化。Schedulater 展示了通过结构化检索将轻量级大语言模型集成到大学行政工作流程中的可行性。**

索引词—**教育聊天机器人、移动应用程序、大语言模型（大语言模型）、检索增强生成（检索增强生成）、机构政策导航**

I. 引言

A. 研究动机

大学高度依赖数字系统来协调学术事务，但学生往往难以理解和遵循学校的政策。延期考试和排期流程等事项通常难以在没有指导的情况下理解。传统的支持渠道，包括电子邮件往来和当面咨询，可能响应缓慢，且得到的响应不一致，从而导致学生感到不确定 [1]。这些理解大学行政流程的障碍表明，用户显然需要一个便捷、可靠的系统，为其提供及时帮助 [1]。

我们开发了 Schedulater，这是一款考试延期移动应用，旨在简化大学延期流程。该应用支持三类用户：学生、教职员工和管理员。学生可以向教授发送消息申请延期，并获得提交延期申请的权限。教职员工可以批准或拒绝学生的申请，也可以自愿担任延期考试的监考员。管理员负责预订考试延期所需的考场，并指派监考员监督考试场次。

Schedulater 的一项核心功能是 JAMN Bot，这是集成于移动应用中的大语言模型（LLM）驱动的对话助手。该聊天机器人使用通过 Ollama 自行托管的开源大语言模型 [2], [3] 作为 JAMN Bot 的核心服务，从而为用户生成响应。JAMN Bot 作为政策助手，回答学生的查询，并指导用户完成提交延期申请和考试当天流程等事项。JAMN Bot 当前的实现侧重于学生体验，并帮助他们应对缓考流程中遇到的各种困惑。

B. 现有不足

许多大学都开发了个性化聊天机器人；然而，现有系统更侧重于辅导、互动或广泛提供大语言模型的使用权限，因此在政策导航支持方面仍有不足。密歇根大学使用两种不同的人工智能工具：Go Blue 和 Maizey。Go Blue 是一款服务于校园生活的人工智能助手，可实时提供校园服务信息，包括餐饮和公共交通 [4], ，并允许用户上传自己的数据和文档，以获得个性化体验。教职员工使用

*这些作者对本研究的贡献相同。本文是发表于论文集中的论文的扩展版本

该论文发表于 2026 IEEE 智慧世界大会（SWC），第 810–817 页，属于 IEEE SmartEdu 2026（DOI：10.1109/SWC69865.2026.11763986）的一部分。为满足页数限制，会议版本删去了一些内容；本版本补回了这些内容，包括温度敏感性分析、界面设计依据以及完整的局限性讨论。请引用已发表的会议版本。

第二种工具 Maizey 为课程提供辅导支持。Maizey 使用检索增强生成（RAG），确保响应直接来自上传的材料。加利福尼亚大学欧文分校推出了 ZotGPT ClassChat，教师可上传课程材料，让聊天机器人为学生的问题提供深入解答 [5]。亚利桑那州立大学推出了 CreateAI 平台，为学生、教职员工提供 50 多种生成式人工智能模型的使用权限，用户还可以添加指令和系统提示词进行定制 [6]。该校还推出了聊天机器人 Sunny，可协助学生处理课程导航和日程安排任务 [7], [8]。

JAMN Bot 与上述聊天机器人不同之处在于，它预先配置了相关政策并集成到工作流程中，能够立即向学生提供以机构政策文件为依据的响应。它将学术工作流程与人工智能相结合，使学生无需等待即可获得指导。

C. 贡献

本研究的贡献包括：在大学环境中设计并实现一个集成于考试延期安排系统中的基于大语言模型的聊天机器人，并通过案例研究评估其作为政策助手的有效性。通过该案例研究，我们报告了三项发现：我们的评分者在依据性方面具有较高一致性，但在正确性方面的一致性较低；在严格依据性评分标准下，模型规模无法预测回答质量；服务技术栈会影响大语言模型与人工在正确性方面的一致性，但不会影响二者在依据性方面的一致性。

II. 相关工作

A. 人工智能与移动端学习系统

得益于人工智能驱动的教育工具近期取得的进展，用于学术和行政任务的移动端助手生态系统持续发展壮大 [9], [10]。Duolingo 和 Quizlet 是这类学习应用的两个典型例子，这两个平台都展示了人工智能如何融入教育领域众多人群的日常工作流程 [10], [11]。这些系统提供自适应解释、练习支持和个性化反馈，凸显了人工智能工具对用户的有效性，因为它们提升了移动端学习环境的可及性和参与度 [9]。

B. 教育领域中由大语言模型驱动的助手与聊天机器人

从历史上看，教育机构早期采用人工智能聊天机器人，主要是为了回答学生的常见问题 [12]。然而，直到近些年，这类聊天机器人才开始采用大语言模型，以更好地处理学生的开放式问题 [9]。除了能力上的提升，这些基于大语言模型的系统还能在用户提示其解释概念、总结材料、根据课程内容生成练习题及相应答案时，带来更好的用户体验，并支持更自然的互动 [13]–[16]。尽管这类系统尚未广泛融入机构的工作流程，但对话式教育支持系统的需求日益增长，相关发展趋势也正转向这一方向。

C. 教育系统中的检索增强生成

检索增强生成（RAG）是一项关键技术，它将响应建立在外部知识来源的基础上，从而提高语言模型输出的事实可靠性 [17]。在这类系统中，用户查询会从结构化文档集合中检索相关上下文，并在生成过程中将其提供给模型 [17]。与独立大语言模型相比，这种方法能够减少幻觉并提高事实一致性 [18]。

在教育环境中，检索增强生成已应用于问题求解系统、辅导支持及其他学术辅助工具。将检索增强生成用于大语言模型驱动的工具，可以让模型基于讲义、教科书或机构文档生成解释 [9]。这些特性使检索增强生成特别适用于对准确性和可追溯性要求至关重要的机构应用。

D. 部署中的小型模型与前沿模型之比较

JAMN Bot 由 Llama 3.2 3B 驱动，这是一种相对较小的语言模型 [19]。与大型前沿模型相比，小型模型更易于部署，也更适合大学等隐私敏感环境，同时效率更高 [20], [21]。此外，借助检索增强生成，行为可预测性得以维持，开销也有所降低 [17]。值得注意的是，小型本地模型与大型前沿模型之间始终存在取舍 [19]。大型模型具备更强的推理能力和更广博的世界知识；然而，它们的成本也更高，依赖外部云端 API 和服务提供商，可能引发隐私担忧，且需要更多计算资源 [22]。机构还需考虑到，通过应用程序接口使用的前沿大语言模型仍受版本管理和弃用的影响，这会影响其长期稳定性 [21]。

归根结底，尽管能力有限，本地部署的开放权重模型仍能提供足够的性能，并且由于模型文件固定，能够进行行为验证 [20]。虽然随着 vLLM [26], 等技术的发展，Kimi K 3 [23], 、GLM 5 [24], 和 Deepseek V4 [25] 等大型模型已变得更易于部署，但我们认为大多数机构仍无力使用这些模型。对于端侧部署而言，硬件限制则构成了更大的障碍。

E. 大语言模型评审方法与评估实践

近期关于大语言模型评估的研究探索了使用大语言模型作为其他模型输出的自动评估器，这通常被称为大语言模型评审框架 [27]。这种方法根据正确性、相关性和有用性等标准对输出进行评分。然而，既有研究表明，大语言模型评审在评估过程中会遇到校准问题，尤其是在有用性和推理质量等主观标准上，大语言模型的判断可能与人工标注系统性地出现偏差 [27]。使用单个大语言模型评审器评估多个模型时，一个值得注意的问题是，该评审器可能会偏向于自己的答案。遵循既有研究[28], , 我们采用由独立评审器组成的评审组进行评分，并在第 V-G 节报告结果。

III. 系统设计

A. 部署环境：*Schedulater* 平台

Schedulater 是一个旨在支持大学环境中考试延期的移动平台。它将学生、教师和管理人员整合到一个工作流程中，用于提交、审核和安排延期考试，并在整个流程中提供状态更新。该应用采用 Jetpack Compose Android 客户端和使用 Ktor 实现、部署在 Railway 上的 Kotlin 后端；Firebase 用于身份验证、推送通知（Firebase 云消息传递）和文档存储，而 SQLite 则作为检索操作的嵌入式向量存储（图 1）。后端负责包括延期申请、消息传递、教室安排以及系统组件之间的协调在内的核心功能，其中也包括 JAMN Bot 政策助手。JAMN Bot 内置于Schedulater中，使学生能够提出与缓考流程相关的问题。提交查询后，查询会发送至后端，由后端管理人工智能处理流程。该流程包括检索相关知识、构建提示词，以及将提示词转发给通过 Ollama 提供服务的自托管语言模型。随后，响应会返回至客户端并显示在界面上。将后端与大语言模型分开部署，可以让模型在本地运行或作为专用服务运行，从而更便于独立更新模型，并改善成本、隐私和性能。

B. 用户界面设计依据

JAMN Bot 直接嵌入学生界面中，并显示在消息选项卡顶部，以便访问。这一设计选择基于以下理念：情境内辅助能够减少任务切换，并提高学生在感到不确定时即时使用该工具的可能性[29]。JAMN Bot 界面提供对话式体验，包含消息历史记录和引文；引文会显示源文档及检索到的上下文所在章节，从而提高透明度，如图2所示。这使用户能够验证每个回答的依据，并减少对模型生成内容的过度依赖。该系统实施了明确的拒绝行为。当查询被护栏机制拦截，或检索未获得相关信息时，助手会返回标准化拒绝消息，而不会提供虚构来源。JAMN Bot 旨在通过优先确保正确性和机构可信度，而非追求完整性，来避免给出推测性或缺乏依据的回答。

IV. 检索增强生成流水线

A. 知识库

为了有效回答有关政策、隐私和日程安排的问题，JAMN Bot 采用了检索增强生成流水线。预训练的大语言模型具有令人期待的能力；然而，它们也容易产生幻觉 [17]。我们由大语言模型驱动的聊天机器人通过将响应建立在描述 MacEwan University 政策[1] 以及 Schedu-later 程序流程的 Markdown 文档之上，确保输出保持一致。

B. 检索与索引

在数据导入过程中，Markdown 文档会被分割为约 1500 个字符的文本块，块与块之间重叠 100 个字符。此过程采用标题感知分块策略。每个文本块都通过 Ollama 使用 nomic-embed-text 模型 [30] 生成嵌入，并存储在本地 SQLite 数据库中。在查询时，会将用户的问题转换为嵌入，并在 Kotlin 中计算余弦相似度，与已存储的文本块向量进行比较。检索器使用 0.3 的相似度阈值筛选结果，并选择最相关的前 6 个文本块来构建上下文。在实际应用中，对于当前语料库规模，该阈值从未起到筛选作用，因为我们观察到的最低相似度为 0.505，因此每个问题都会检索到六个文本块。

C. 提示构建与生成

提示词由三部分组成：系统指令、用户查询和检索到的上下文。系统指令要求模型仅使用 Schedulater 知识库中的信息，并避免给出缺乏依据或推测性的响应。响应由我们部署在生产环境中的大语言模型生成，温度为 $= 0.3$，可通过环境变量进行配置，从而控制不同评估轮次之间的变化。

D. 护栏与可靠性

如前所述，应用中实施了分层护栏系统以确保安全。在搜索答案之前，JAMN Bot 会检查是否存在恶意使用行为，例如通过提示词诱导系统泄露敏感数据。如果检测到此类行为，查询会立即被阻止。当检索未返回足够相关的上下文时，系统会生成受控

[1]大致参考了 MacEwan University 的延期政策：https://www.macewan.ca/data/policies/policies/academics/student_deferrals_procedure/student_deferrals_procedure.pdf

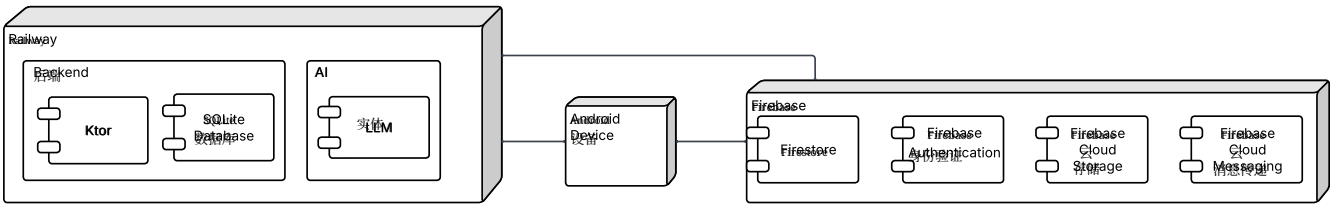


图 1. Schedulater 的部署架构：Android 客户端、Railway（托管 Ktor 后端、SQLite 存储和由 Ollama 提供服务的大语言模型）与 Firebase（用于身份验证、文档存储、云存储和 FCM 通知）之间的通信

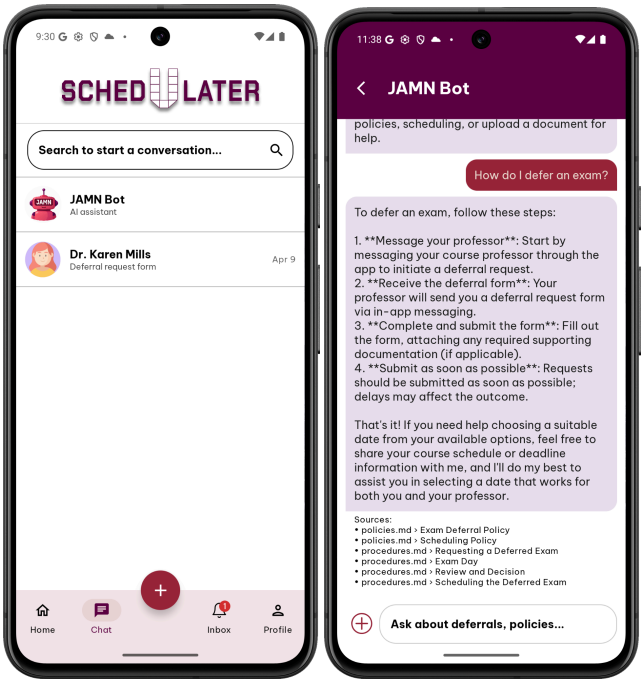


图2. Schedulater 的 JAMN Bot 用户界面设计集成在消息传递页面中（左），以及与人工智能聊天的示例（右）

拒答响应，而不是生成答案。这样可以确保在缺少上下文时，模型不会凭空编造缺乏依据的信息。正则表达式仅用于表层的初步检查，可能无法识别改写后的输入或提示词注入攻击。因此，稳健性来自将正则表达式预检与检索验证和受控拒答机制分层结合，而不是依赖任何单一检查 [31],[32]。这种分层方法还可以部分借助大语言模型 [33]，以减少反复迭代和改进系统安全与保障所需的人力成本。

V. 评估

A. 设置

评估基于一个包含 43 项的基准数据集，涵盖与 Schedulater 系统相关的五个操作类别：考试延期（10 项）、隐私（5 项）、日程安排（5 项）、跨政策（15 项）以及对抗性/超出范围（8 项）。每个问题都配有一份人工编写的标准答案。该数据集旨在评估大学政策情境中的事实检索和工作流推理。基准测试包含一个结构化的对抗性划分，其中包括强对抗性提示（第 5、34、36、39 项）以及软对抗性/超出范围（OOS）提示（第 37、40–43 项）。

本研究在两个不同环节进行了评分：首先，将 JAMN Bot 的生产环境部署版本（通过 Ollama 在 Railway 上运行的 Llama 3.2 3B Q4-K-M）与人工评审者及大语言模型评审器进行比较；其次，通过 OpenRouter [34]评估不同规模模型之间的性能差距。在这两个评分环节中，人工智能生成的针对基准测试问题的答案均依据相同的评分标准，以 0–2 分制从依据性（G）和正确性（C）两个维度进行评分。依据性衡量响应是否得到知识库的支持，而正确性衡量响应是否准确且符合事实。

评分者间共识取三位评分者（R1、R2、R3）评分的平均值。对于 Cohen 线性加权卡帕系数 [35],（以有序类别为输入），共识取三位评分者评分的中位数，并四舍五入至最接近的整数。通过 OpenRouter 使用的 GPT-5.4 大语言模型评审器收到了一份系统提示，其中包含三个校准示例，分别为高（G=2,C=2）、中（G=1, C=1）和低（G=0, C=0）。这些示例用于帮助其评估结果与人工评分标准保持一致。所有评审均在温度设为 0 的条件下进行，以使评分尽可能接近托管 API 所允许的确定性，并确保我们观察到的差异来自生成的答案，而非评估过程 [36]。

对于多模型比较，使用相同的 43 项基准测试评估了七个模型，并为每个问题预先计算检索片段，以确保公平性。在这些运行中，检索和提示阶段均使用 Python 重新实现，采用与已部署的 Kotlin后端相同的分块参数、嵌入模型、相似度阈值和预检模式，从而确保每个模型接收到完全相同的上下文。评估的模型包括 GPT-5.4 [37] 和 Opus 4.5 [38]，等前沿模型，以及 Llama 3.2 和 3.3 [19]，Command R7B [39],、DeepSeek V4 Pro [25] 等开源权重模型。选择这些模型，主要是为了体现模型规模与性能之间可能存在的部署权衡。在两项测试中，无论是架构内服务比较还是模型间能力比较，每个模型的每次运行均进行三次重复，且生成温度始终保持为 0.3。每个查询都会记录检索片段、相似度分数、延迟以及最终响应，以便在基准测试流程中进行分析。

B. 基准测试的人工评分

表 I 分节人工评分表现：通过 Railway 使用 Ollama 上的 Llama 3.2 3B Q4 K M

章节	依据性	正确性	备注
第 1 节：考试延期（10）	2.000	1.367	表现最佳的类别
第 2 节：隐私（5）	1.800	1.333	一致性一般
第 3 节：日程安排（5）	0.867	0.733	表现最弱的类别
第 4 节：跨政策推理（15）	1.714	1.311	综合能力较弱
第 5 节：对抗与分布外（8）	1.389	1.208	偏向软对抗
总体（43）	1.632	1.240	合计（$G + C$）/2 = 1.436

表 I 中的数据阐明了生产级模型在 43 项基准测试中的表现。如表 I 所示，Llama 3.2 3B Q4 K M 在有据可依性（1.632）方面的表现优于正确性（1.240），表明该模型在知识库相关内容上的表现比回答复杂问题更可靠。此外，综合得分（1.436）体现了 JAMN Bot 扎实的基础及其改进能力。

根据表 I 中的数据，日程安排政策在人工评估中的表现最弱（G = 0.867，C = 0.733），而考试延期政策的结果最佳（G = 2.000，C = 1.367）。此外，对抗性／超出范围部分整体上难度较高（G =1.389，C = 1.208），而预检正则表达式成功拦截了全部四个高难度注入提示（第 5、34、36、39 项）。

C. 评审者间一致性

在将人工共识与大语言模型评审器进行比较之前，我们将先确定三位评分者彼此之间的一致程度。我们报告 Krippendorff 序数 α[40] 以及 Fleiss κ [41]，因为与无加权测度相比，加权序数测度更适用于多评审者 0–2 量表 [42]。依据性基于 39 个项目进行报告，因为四个困难对抗性项目（5、34、36、39）在检索前被预检正则表达式拦截，因此未定义。正确性则基于全部 43 个项目进行报告。

表 II 三位人工评审者之间的评审者间一致性

维度	α (序数)	Fleiss κ	解释
依据性（n=39）	0.894	0.848	几乎完美
正确性（n=43）	0.687	0.519	显著／中等

[a]α = 克里彭多夫序数 α；两两线性加权 Cohen's κ 的平均值分别为 0.87（依据性）和 0.60（正确性）。κ 其平均值按两两计算，因为加权 κ不适用于两名以上评分者 [42]。

如表 II 所示，评分者在依据性方面的一致性很高（α = 0.894），89.7% 的项目得到了全体评分者的一致评分。正确性方面的一致性较低（α = 0.687），55.8% 的项目得到了全体评分者的一致评分。Fleiss' κ 的读数在这两个维度上都低于 α，因为它将评分视为名义类别，且评分者给出相邻分值时不提供部分得分；这种情况在正确性评分中经常出现，评分者的意见往往分歧于 1 分和 2 分之间。各评分者在依据性方面的平均分相近（1.641、1.615、1.641），而其中一位评分者对正确性的评分略严于另外两位（1.140、1.279、1.302）。值得注意的是，评分者之间高度一致，这表明人工共识是一个稳定的基准，可用于与大语言模型评审器进行比较。

D. 大语言模型（*LLM*）评审评估

为补充人工评估，研究采用 GPT-5.4 作为大语言模型评审器，实施了自动评估流程，并使用相同的 0–2 分评分标准，同时在评审提示词中提供了三个校准示例。评审器以检索到的文本块作为上下文，从而依据与模型所看到的相同材料评估依据性。

如表 III 所示，大语言模型评审器与人工评审者在依据性方面高度一致，二者之间的差距仅为 0.035。正确性方面的差距较大，为 0.202，其中大语言模型评审器的评审分数高于人工共识。与我们的评分者相比，该评审器更倾向于为正确性给予认可，这与既有文献关于大语言模型评审器在准确性类标准上较为宽松的结论一致 [27]。

表 III 大语言模型评审与人工评估及评审者间一致性对比

指标	人工 Avg.	大语言模型评审器 Avg.	κ	解释
依据性	1.632	1.667	0.400	中等一致性
正确性	1.240	1.442	0.144	轻微一致性

[a]在 Railway 上通过 Ollama 运行 Llama 3.2 3B Q4 K M 的运行 1。[b]κ = Cohen 线性加权卡帕系数：人工共识与 GPT-5.4 评判结果的比较（温度为 0，已提供分块文本）。

Cohen 线性加权卡帕系数报告于表 III（κ_G = 0.400， κ_C = 0.144）。根据 κ值，Landis 与 Koch 解释标准 [43] 对依据性和正确性的判定分别为“中等”和“轻微”，这为生产级模型的评审与人工评分的一致性提供了下限。

E. 多模型比较

我们使用七种不同的模型，基于同一组预先计算的检索片段，并将生成温度设为 0.3，针对 43 项数据集进行了三次基准测试。随后，GPT-5.4 在温度为 0 的条件下，依据相同的评分标准对每个 AI 生成的输出进行了评判。

表 IV 多模型排行榜（按综合 = (G+C)/2 排名，3 次运行的均值 ±标准差）

Rank	模型	依据性	正确性	综合
1	GPT-5.4	1.829	1.837	1.833 ± 0.018
2	Llama 3.3 70B Instruct	1.812	1.721	1.766 ± 0.029
3	Opus 4.5	1.726	1.775	1.751 ± 0.039
4	DeepSeek V4 Pro	1.675	1.597	1.636 ± 0.115
5	Llama 3.2 3B Instruct	1.718	1.481	1.600 ± 0.041
6	Command R7B (12-2024)	1.615	1.543	1.579 ± 0.044
7	Llama 3.2 1B Instruct	1.385	0.961	1.173 ± 0.051

根据表 IV 的结果，全精度 Llama 3.2 3B Instruct 按综合结果排名，在 7 个模型中位列第 5 $^{G+C}_{2}$ 。表中的所有模型均通过 OpenRouter 以全精度提供服务，因此此排名针对的是我们所部署模型的全精度版本，而非实际部署的配置；我们会在第 V-F 节中将两者进行比较。GPT-5.4 在七个模型中表现最佳，但其综合结果与 Llama 3.2 3B Instruct 相差 0.233。相比之下，Llama 3.2 1B Instruct 排名垫底，与前者的差距相当显著。这是因为 1B 是规模较小的 Llama 模型，在遵循指令和基于检索到的上下文进行推理方面的能力较弱。值得注意的是，第 V-I 节的温度消融显示，在此基准测试中，不同运行结果的波动幅度约为 0.08。表 IV 中相邻排名之间的六个差距，只有两个大于这一波动幅度，因此，排名前三的模型以及排行榜中段的模型，与其说是严格的名次顺序，不如说更适合视为若干组。

我们按章节计算每个模型的依据性和正确性，以观察各模型在每个章节中的表现。表 V 显示，对抗与 OOS 部分是大多数模型表现最弱的类别，说明这是检索增强生成系统面临的普遍挑战。这种表现并不意外，因为许多对抗与 OOS 提示无法匹配知识库中的任何相关文本块，

表 V 各节综合得分（按总体得分排序）

模型	第 1 节	第 2 节	第 3 节	第 4 节	第 5 节	总体
GPT-5.4	2.000	1.833	1.800	1.865	1.542	1.833
Llama 3.3 70B Instruct	2.000	1.567	1.734	1.773	1.576	1.766
Opus 4.5	1.965	1.634	1.633	1.795	1.500	1.751
DeepSeek V4 Pro	1.791	1.700	1.567	1.727	1.222	1.636
Llama 3.2 3B Instruct	1.967	1.534	1.500	1.615	1.167	1.600
Command R7B (12-2024)	1.724	1.400	1.367	1.681	1.424	1.579
Llama 3.2 1B Instruct	1.537	1.000	1.067	1.184	0.841	1.173

第 1 = 节考试延期政策，第 2 = 节隐私政策，第3 = 节日程安排政策，第 4 = 节跨政策推理，第5 = 节对抗与分布外

导致模型缺乏有依据的上下文，无法生成正确响应，无论模型规模或能力如何。当被拒绝的查询存在相应政策时，我们的检索结果（第 V-H 节）表明，检索器并不总能找到该政策；这是与缺少相关文本块不同的另一种失败情况。所有模型在考试延期政策上的表现始终最强，这表明该政策的结构最清晰、规则性最强。尽管跨政策推理要求模型同时从 policy.md 和 procedure.md 知识文件中提取信息，但表现也不错。此外，我们在人工评估（表 I）中发现的日程安排政策方面的弱点，并未出现在其他模型上，这说明问题出在我们的量化基线上，而非该章节本身存在难点。聚焦考试延期政策时，DeepSeek V4 Pro 的表现不如 Opus 4.5（排名第 3）和 Llama 3.2 3B Instruct（排名第 5），但总体排名仍为第 4。在隐私政策方面，它的表现优于这两个模型，这或许解释了它为何总体排名介于两者之间。

F. 全精度 *Llama 3.2 3B Instruct* 的评分

我们此前指出，根据 Landis 与 Koch 的解释，对经过基准测试并接受评审的 Llama3.2 3B Q4 K M（表 III），Cohen κ 系数的 κ值分别为“中等”（κ_G = 0.400）和“轻微”（κ_C = 0.144）。尽管 κ_C 显示大语言模型评审器与人工评审者之间存在不一致，但这并不意味着评审器不可靠。相反，我们需要使用生成回答质量更高的模型开展更多基准测试，以观察这是否能减少大语言模型评审器与人工评分之间的分歧。

我们对 Llama 运行 1 的基准测试进行了人工评分通过 OpenRouter 使用 3.2 3B Instruct，采用相同的温度和三位评审者协议，并依据 0–2 分制的评分标准进行评估。我们的评分者对 Instruct 配置的平均 G = 1.701 和 C = 1.434 进行评分，而 Q4 K M 配置的平均 G = 1.632 和 C = 1.240。逐项比较两种配置后发现，43 个项目中有 20 个项目的正确性提高、8 个项目降低；依据性在 9 个项目中提高、8 个项目中降低。

由于我们在两次运行中保持生成温度和评审配置不变，因此能够将服务技术栈对评审与人工评分的一致性的影响单独分离出来。不过，这两种配置在不止一个方面存在差异。两者运行的是同一个指令微调模型，但在数值精度、推理服务提供商方面有所不同，

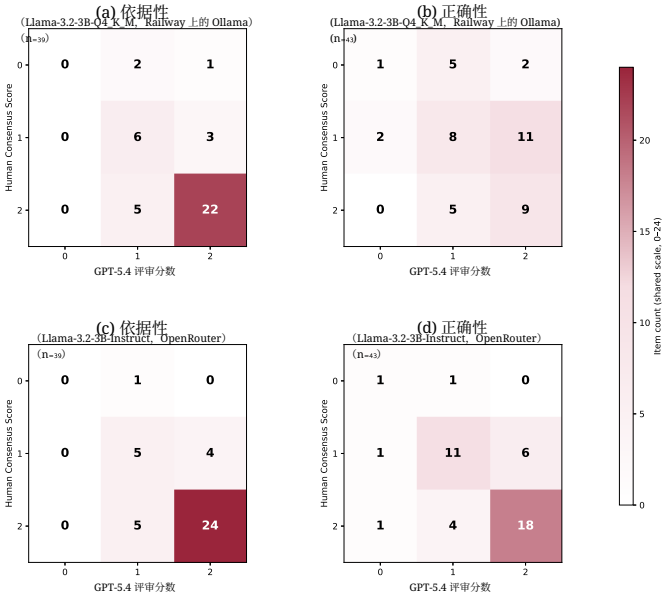


图 3。人工共识与 GPT-5.4 评判结果：在 Railway 上通过 Ollama 使用 Llama 3.2 3B Q4 K M（a、b）以及通过 OpenRouter 使用 Llama 3.2 3B Instruct（c、d）时的混淆矩阵。所有面板使用统一的颜色刻度。

并且检索和提示阶段究竟是通过已部署后端还是 Python 重新实现来运行的也不同。因此，我们无法将其中任何一个因素单独确定为原因。

我们记录了 Instruct 模型配置在比较 GPT-5.4 与人工评分时的 Cohen 线性加权卡帕系数，分别为 κ_G = 0.389 和 κ_C = 0.463。再次依据 Landis 与 Koch 量表，这些数值对于依据性和正确性分别被解读为“一般”和“中等”。在依据性方面，两种配置的结果基本相同（Q4 K M 配置为κ_G = 0.400，Instruct 配置为 0.389），因此服务技术栈似乎不会改变大语言模型评审器与我们的评分者在依据方面的一致程度。此外，Instruct 配置的 κ_C 提高了 0.319，使其明显高于 Q4 K M 配置，从“一致性轻微”提升至“一致性中等”。

图 3 所示的 Llama 3.2 3B Instruct 模型混淆矩阵呈现出明显的对角线分布：依据性项目中有 29/39 项、正确性项目中有 30/43 项达到了大语言模型评审与人工评审完全一致。另一方面，从图 3 可以直观看出，Q4 K M 模型配置在正确性方面有更多不一致的情况，其对角线显示仅有 18/43 项评分一致。Q4 K M 配置在依据性方面的对角线数值为 28/39，与 Instruct 配置接近，也与我们从 κ值中观察到的情况一致。需要注意的是，依据性的报告值为 n=39，因为四个困难对抗性项目（5、34、36、39）在检索之前触发了预检正则表达式，导致这四个项目未定义。

尽管两次运行中的温度均保持不变，但计算得出的κ值提升以及评审者与人工评分之间表面上改善的一致性，无法归因于单一因素。不过，我们认为数据表明，改善后的

评审者与人工在正确性量表上的评分一致性会随回答质量提高而增强，而在依据性方面的一致性则在两种配置下都保持稳定。

G. 独立多评委小组

在多模型比较中，我们既将 GPT-5.4 用作评委，也将其作为候选模型；它在表 IV 的排行榜上将自己排在第一位。为检验这一排名是否取决于评委，我们另外选取了候选池之外的两位评委——Gemini 3.5Flash 和 Qwen3.5 397B——对全部七个候选模型重新评分。鉴于此，先前研究建议采用评委小组 [28]。三位评委均使用相同的评分标准和三个校准示例，并将温度设为 0。此外，各评委默认生成的中间推理量有所不同，因此我们在每位评委的整个评测过程中保持其配置不变。值得注意的是，三者都是专有托管模型，因此我们不声称它们具有可复现性或开放权重。

表 VI 多评委小组：按评委列出的每个候选模型的综合$(G+C)/2$

候选模型	GPT-5.4（现任）	Gemini 3.5 Flash	Qwen3.5 397B
GPT-5.4[a]	1.833	1.967	1.857
Llama 3.3 70B Instruct	1.766	1.932	1.819
Opus 4.5	1.751	1.884	1.824
DeepSeek V4 Pro	1.636	1.735	1.655
Llama 3.2 3B Instruct	1.600	1.643	1.525
Command R7B (12-2024)	1.579	1.719	1.582
Llama 3.2 1B Instruct	1.173	0.961	0.828

[a]现任评审者为其自身输出评分。正文中报告了秩相关性和自偏好测试的结果。

三位评审器均将 GPT-5.4 排在首位，其他六个候选模型的排序也基本保持不变，如表 VI 所示。排除每位评审器自身对应的候选模型后，GPT-5.4 与 Gemini 之间的斯皮尔曼秩相关系数为 0.943，GPT-5.4 与 Qwen 之间为 0.886，两位独立评审器之间为 0.964。各评审器的评分严厉程度有所不同：Gemini 的评分整体偏高，而 Qwen 的评分整体偏低。有两对相邻模型互换了位置，即 Opus 4.5 与 Llama 3.3 70B，以及 Llama 3.2 3B 与 Command R7B；这两对模型之间的差异均小于第 V-I 节所测得的采样噪声。

为检验自我偏好，我们比较了 GPT-5.4 与每位独立评审者对各个候选模型评分之间的差异。GPT-5.4 对自身输出的评分差异，相较于对其他六个候选模型的评分差异并非离群值（$z = -0.51$与 Gemini 相比，$z = -0.44$ 与 Qwen 相比）。我们还在匹配项目集上计算了三位评审者之间的一致性，范围限于评分者也评过的那个候选模型，结果为：依据性 $\alpha = 0.608$，正确性 $\alpha = 0.711$，与三位人工评审者的评分（$\alpha = 0.661$ 和 $\alpha = 0.643$）接近。值得注意的是，GPT-5.4 相对于排名第二的模型的优势为 0.067，低于 0.08 的采样噪声下限。因此，我们无法在这一基准测试规模上确认或排除自我偏好；我们将评审小组的结果作为排名稳健性的证据，而非偏差测量结果。

H. 检索指标

到目前为止，我们的基准测试衡量的是生成答案的质量，但无法说明检索器本身的表现如何。为直接进行衡量，三名人工标注员分别独立标注了每个条目中包含答案的金标准文本块。每位标注员都以固定顺序查看全部十个文本块，且看不到分数、排名，也不知道检索器返回了哪些内容，因此标注结果不会受到被评分输出的影响。这些标注的一致性较高（$\alpha = 0.553$），并且经裁定，三个条目在知识库中没有任何支持性文本块。由于检索结果是在所有七个模型之间预先计算并共享的，命中率是检索器的属性，而不是任何单个模型的属性，因此我们针对 36 个可回答条目报告一组数值。

表 VII 检索指标与随机检索器基线对比（$n = 36$）

指标	检索器	随机[a]	阅读
Hit@1	**0.667**	0.167	判别性
Hit@3	0.889	0.438	判别性
Hit@6	0.889	0.732	近乎饱和[b]
MRR	**0.773**	0.343	判别性
Recall@6	0.875	0.600	近乎饱和[b]

[a]精确值，基于每个项目的标准文本块数量进行条件化。

[b]在包含 10 个文本块的语料库中固定取前 6 个；见正文。

如表 VII 所示，检索器的 Hit@1 = 0.667 和平均倒数排名 = 0.773，均优于随机基线 0.167 和 0.343。这表明，当存在相关文本块时，它通常会被排在第一位。Hit@3 和 Hit@6 均达到 0.889，Recall@6 达到 0.875。不过，我们的知识库仅包含十个文本块，而检索器固定返回前 6 个，因此随机检索器在 Hit@6 上也能达到 0.732。我们报告集合成员指标以求完整，但在这一语料库规模下，真正具有区分度的是 Hit@1 和平均倒数排名。

检索器表现不佳的唯一部分是对抗性与 OOS，其中 Hit@1 和 Hit@6 均为 0.500，而根据该部分标准文本块数量组合计算的随机 Hit@6 为 0.692。对于第 38 项和第 41 项，检索器未能检出处理该请求的隐私政策文本块，因此在上下文中缺少相关政策时，助手会拒绝请求。鉴于涉及的项目数量较少，我们将此作为一项观察结果报告。

I. 温度敏感性分析

所有基准测试均采用 0.3 的生成温度，以便观察七个模型针对查询以非确定性方式生成的答案之间的差异。通过对每项基准测试进行三次运行，我们有意呈现样本方差，从而能够在真实条件下直观地观察模型行为的变化。不过，我们还希望测试温度敏感性，因此仅使用 Llama 3.2 3B Instruct 模型进行另一项测试，以便将生成温度作为唯一变量。我们以 0.1 为步长，在温度从

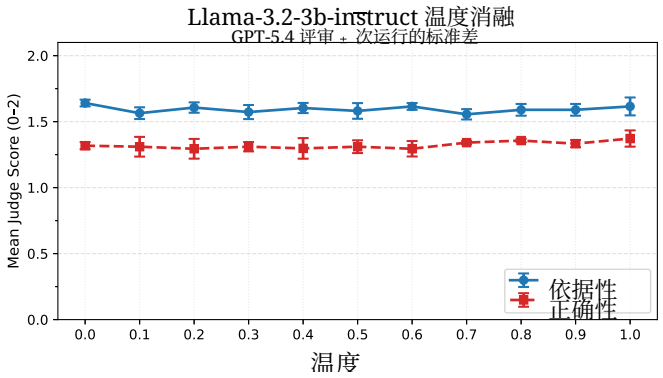


图 4. Llama3.23B Instruct 的温度消融：11 种采样温度（0.0 至 1.0）下依据性和正确性的均值。

0.0 到 1.0 的范围内进行了三次运行。结果显示，在 11 个温度设置下，平均依据性的总变化范围为 0.085（从 1.556 到 1.641），而平均正确性的总变化范围为 0.078（从 1.295 到 1.372）。如图 4 中的误差条所示，这一趋势在统计上并无显著意义。

我们的数据与 Renze 和 Guven 得出的无显著结果一致：他们在不同温度设置（0.0 到 1.0 [44]）下测试五个大语言模型处理选择题的性能时，数据缺乏统计显著性。我们推测，我们的数据遵循与 Renze 和 Guven 类似的机制，因为提示模板明确限制模型检索某些文本块，导致评分标准结果没有偏离，也使温度变化无法偏离主导轨迹。然而值得注意的是，Renze 和 Guven 提到，取平均值和统计噪声可能会掩盖真正的温度敏感性结果[44]。

这些数据使我们得出以下结论：首先，生产环境采用的 0.3 设置处于生成质量平稳的区间；其次，在测试范围内，该模型没有明确的性能崩溃点；最后，模型内部 0.08 的依据性和正确性离散幅度，可作为排行榜采样噪声的下限（表 IV）。对教育机构而言，这意味着，如果选择采用基于RAG的系统，温度对结果方差的影响很小；调整工作更应侧重于提示结构、模型选择和服务精度，因为表 IV 和第 V-F 节表明，这些因素对分数的影响远大于采样温度。

VI. 讨论

A. 部署配置的影响

面临成本或硬件限制的教育机构可能会优先考虑更小、更便宜的本地模型，并且通常会选择在本地运行的量化模型。从我们部署的配置切换到全精度托管配置后，评分者记录到正确性有所提升，而大语言模型评审器记录到的提升幅度小得多，依据性方面的一致性则完全没有变化（第 V-F 节）。这两种配置在精度、提供商和流水线实现方面均有差异，因此我们报告这一影响，但不将其归因于任何单一因素。值得注意的是，这意味着评审器与评分者的一致性最低的，恰恰是机构为节省成本而会部署的配置。实际启示是，量化模型的准确性应由人工评审者而非仅由大语言模型评审器进行验证，因为自动评判器低估了量化带来的代价。

B. 检索增强生成与指令遵循的原始能力

许多人可能会认为，大模型天生就比小模型表现更好。然而，我们的研究结果表明，这一假设并非总是成立。观察比较七个模型在依据性和正确性方面表现的多模型排行榜，除了少数例外，大多数排名都在意料之中。前沿模型（GPT-5.4 和 Opus 4.5）的表现良好，正如预期。不过，非前沿模型的排名为“大不一定更好”这一论点提供了佐证。按大语言模型的大致规模降序排列，顺序为：GPT-5.4、Opus 4.5、DeepSeek V4Pro、Llama 3.3 70B、Command R7B、Llama 3.2 3B 和 Llama 3.2 1B；需要指出的是，闭源模型的参数量尚未公布。值得注意的是，表 IV 显示，规模更大的 DeepSeek V4 Pro 表现不佳，仅排名第 4（G&C = 1.636），比规模小得多的 Llama 3.3 70B（G&C = 1.766）低 0.130。这个差距大于我们测得的约 0.08 的多次运行波动，因此并非抽样造成的假象。尽管 DeepSeek V4 Pro 的思维链内容冗长，其评分标准得分仍低于预期 [45]。由于评分标准会惩罚超出检索片段内容进行推断的情况，不忠实遵循指令的模型表现较差。

总体而言，RAG 基准测试更重视指令执行规范，而非原始能力。了解模型规模仍有助于推断其在不同程度上的实用性，这一点对于 Llama 3.2 1B Instruct 尤其如此。其 1.173 的综合 G&C 分数可作为有用的性能下限。由于 Instruct 配置意味着该模型已经过微调，以提升指令遵循能力，因此规模低于这一水平的模型可能更容易产生无效输出和幻觉。无论模型规模如何，都存在一种持续且系统性的模式：对抗与分布外类别中的项目始终会降低表 V 所示的 GPT-5.4 评审分数。尽管预检正则表达式层能够准确捕获全部四个高难度对抗性查询，但超出范围查询会采用释义改写或间接注入，从而对与 G&C 分数相关的表层模式匹配产生负面影响，凸显了分层防御系统的重要性 [32]。对于教育机构的部署决策而言，实际选择介于 Llama 3.2 3B Instruct 这类经过微调的开放权重模型与前沿模型之间。然而，隐私、成本和延迟等与规模相关的部署限制，将影响各机构决定采用哪种模型。

C. 教育中的人工智能的更广泛影响

我们相信，我们的工作能够为高等教育采用基于大语言模型的助手提供一些关键见解。在不同模型上收集的关于依据性和正确性的实证数据，使我们能够审视本地部署的小型开放权重模型与通过 API 调用访问的前沿模型之间所作的权衡。由于 1B 模型（G&C =1.173）很可能不足以支持任何具有实际意义的面向学生的产品，因此教育机构至少必须依赖与 3B 和 7B 开放权重模型相近的模型。尽管表 IV 中的分数差距表明，前沿模型的性能优于其他模型，但由此带来的权衡包括更高的成本和延迟，以及对外部 API 的依赖，而这些因素在大规模应用时可能构成限制。此外，我们的量化结果表明，模型的选择并非唯一重要的决策，因为模型的服务方式会影响其准确性被可靠测量的程度。我们还认为，面向学生的人工智能助手（即使其应用领域超出考试延期）自然会收到可能令人困惑或具有对抗性的查询，这与我们的测试类似。因此，我们提出，我们的分层防护栏模式（预检正则表达式、相似度阈值和检索结果为空时拒答）是一种合理的设计，可作为不同领域聊天机器人的第一道防线。

VII. 局限性与未来工作

JAMN Bot 的知识由我们的团队整理，用于呈现 Schedulater 工作流程，并不能完全代表任何一所大学针对延期考试制定的政策和程序。为了提高 JAMN Bot 的实用性，未来的工作应包括扩展聊天机器人的知识库，使其能够为考试延期之外的政策提供帮助。如果 JAMN Bot 能够就学术诚信、成绩申诉、主修与辅修专业申报、更换专业项目、学习计划、压力管理或休学提供指导，它就有可能成为一个集中式平台，帮助学生处理学业事务。此外，进一步加强与教职员工的合作，将有助于确保聊天机器人能够回答学生的常见查询，从而可能减轻教职员工的工作负担。另外，还应增强聊天机器人对教职员工和管理人员需求的支持，因为当前的实现主要面向学生。然而，扩展 JAMN Bot 的功能可能需要更大的大语言模型。与此同时，JAMN Bot 目前依赖简单的基于正则表达式的防护栏，在处理更复杂的查询方面能力有限。随着其应用范围扩大，需要更强的护栏机制来防止负面结果和不良用户体验 [32]。

为了让 JAMN Bot 尽可能便于使用，我们可以对用户进行 A/B 测试，以了解他们更喜欢将 JAMN Bot 放置在哪里。目前，JAMN Bot 是查看消息选项卡时首先出现的对话，更像是一项独立功能。另一种方案是将 JAMN Bot 嵌入导航栏。

在人工测试期间，我们也记录了延迟，每次查询平均用时略低于一分钟。若要将 Schedulater 部署到真实的大学环境中，我们会集成一个向量数据库，以存储文本块，从而提高检索效率 [46]，并缓存常见查询以降低延迟。改进检索效果还可以解决第 V-H 节中提到的问题：相关的依据文本块从未被检索出来。这不仅是效率问题，也是正确性问题。未来可以开展更严格的人工评估，纳入纵向研究并增加人工评审者。如果进行更严格的测试，我们会采用七点李克特量表，更精确地评估依据性和正确性。

我们的结果还存在两项限制。包含十个文本块的知识库与固定的前六检索器，使集合成员检索指标接近随机基线；而 NDCG 等指标则需要规模大得多的语料库 [46]。此外，尽管评委小组基本维持了模型排名，且没有出现异常的自评分数，但 GPT-5.4 相较第二名的优势小于我们测得的运行间差异，因此在这一基准规模下，既无法确认也无法排除自我偏好 [28]。

第八节：结论

在本研究中，我们介绍了 JAMN Bot，这是一款集成于 Schedulater 的大语言模型驱动的政策助手；Schedulater 是一款用于大学考试延期工作流程的应用程序。JAMN Bot 通过 RAG，利用由 Ollama 提供服务的自托管大语言模型，为学生提供具有上下文感知能力的考试延期指导。对 JAMN Bot 的评估表明，它在结构良好的政策指导方面表现最为有效，但在处理超出范围的提示时仍面临挑战。我们的三名评分者彼此高度一致（依据性方面的 Krippendorff's α = 0.894），为大语言模型评审比较提供了稳定的参照；由三名评审组成的独立评审小组也基本维持了模型排名，未检测到自我偏好。然而，在我们严格依据评分标准的评估中，模型规模并不能预测回答质量：DeepSeek V4 Pro 的综合得分比规模小得多的 Llama 3.3 70B 低 0.130，且这一差距大于我们测得的运行间差异，因为该评分标准更重视指令遵循，而非原始能力。我们还发现，服务技术栈会影响正确性方面的评审与人工评分的一致性，但不会影响依据性方面的一致性；对于正在权衡是否进行量化的机构而言，这一点十分重要。研究结果表明，具有明确护栏机制的大语言模型助手能够减少大学延期流程中的困惑。此外，尽管闭源模型与开源模型之间在性能上存在一定差异，但这些差异也增强了我们在教育场景中部署开源模型的信心，因为排行榜前三名模型之间的差距小于我们测得的运行间差异。总体而言，考虑采用类似系统的机构应注意，在将人工智能聊天机器人引入学生行政工作流程时，JAMN Bot 突显了平衡隐私、准确性、货币成本与人力成本、安全性，以及准确信息可用性的重要性。

致谢

我们谨此感谢 GaganpreetJhajj 提供专业知识和意见、应用程序接口额度以及校对支持。我们还

要感谢 Yumeko Nomura 和 Lindsey Searles 协助校对。

数据和代码的可用性

基准数据集（含 43 个带有真实答案的条目）、知识库及其衍生的分块语料库、附有校准示例的大语言模型评审提示词、匿名化的逐项人工评分、全部模型输出及三位评审器的评审分数、检索金标准标签，以及分析脚本均已公开[2]（存档于 [47]）。只需运行一个脚本（analysis/reproduce.py），即可仅使用 NumPy，根据发布的文件重新生成论文的核心结果。

[2]https://github.com/JAMNSoftware/jamn-rag-benchmark

参考文献

[1] Z. Nadir、H. M. Al Lawati、R. A. Mohammed、M. Al Subhi 和 A. Hossen，《Squbot：通过个性化聊天机器人系统增强学生支持》，Technologies，第 13 卷，第 9 期，第 416 页，2025年9月。 [ ] 10 3390 13090416 在线。可查阅：

[2] 2026 http://dx.doi.org/ . /technologies Ollama Inc.，《Ollama》，https://ollama.com/。[3] F. S. Marcondes、A. Gala、R. Magalh˜aes、F. Perez de Britto、D. Dur˜aes 和 P. Novais，《使用 Ollama》。Springer Nature Switzerland，2025年，第 23–35 页。 [在线]。可查阅：http://dx.doi.org/10.1007/978-3-031- 76631-2 3[4] 密歇根大学 ITS，《Go Blue AI：面向校园生活的生成式人工智能助手》，https://its.umich.edu/computing/ai/go-blue-ai，2024年。[5] UC Tech News，《通过 ZotGPT 的 ClassChat 试点项目拥抱课堂中的人工智能》，2024年。 [在线]。可查阅：https://uctechnews.ucop.edu/uc-irvine-embracing-ai-in-the-classroom-with-zotgpts-classchat- pilot/[6] 亚利桑那州立大学，《CreateAI Builder：校园生成式人工智能的技术基础》，https://ai.asu.edu/technical-foundation/creat eai-builder，2024年。[7] Y. Gao，《高等教育中聊天机器人的兴起：变革教学、学习与学生支持》，2025年7月，访问日期：2026- 05-01。 [在线]。可查阅：https://ascode.osu.edu/news/rise-chatbots-higher-education-transforming-teaching-learning-and-student-support [8] C. Hennick，《三所大学如何开发自己的聊天机器人》，5月 2025 [ ] 20 。在线。可查阅：25 05 https://edtechmagazine.com/higher/article/ / / how-three-universities-developed-their-chatbots [9] L. Chen、P. Chen 和 Z. Lin，《教育中的人工智能：综述》，IEEE Access，第 8 卷，第 75264–75278 页，2020年。 [在线]。可查阅：http://dx.doi.org/ 10 1109 2020 2988510 . /ACCESS. . S. Loewen、[10] D. Crowther、D. R. Isbell、K. M. Kim、J. Maloney、Z. F. Miller 等，《移动辅助语言学习：Duolingo 案例研究》，ReCALL，第 31 卷，第 3 期，第 293–311 页，2019年5月。 [在线]。可查阅：http://dx.doi.org/ 10 1017 0958344019000065 . /S O. Ozdemir 和 H. Sec¨kin，[11] 《量化 Quizlet 对学习成果的认知与情感影响：系统综述与全面荟萃分析》，Frontiers in Psychology，第 15 卷，2024年3月。 [在线]。可查阅：http://dx.doi.org/ 10 3389 2024 1349835 . /fpsyg. . L. C. Page 和 H. [12] Gehlbach，《人工智能虚拟助手如何帮助学生踏上大学求学之路》，AERA Open，第 3 卷，第 4 期，2017年10月。 [在线]。可查阅：http://dx.doi.org/ 10 1177 2332858417749220 . / M. R. Kabir 和 F. O. Lin，《基于大语言[13] 模型的自适应练习系统》，载于 LLM@ AIED，2023年，第 43–52 页。[14] G. Jhajj 和 F. Lin，《通过大语言模型与自动语音识别增强日语习得》，载于 2025年 IEEE 智慧世界大会（SWC）。IEEE，2025年8月，第 289–293 页。 [在线]。可查阅：http://dx.doi.org/10.1109 /SWC65939.2025.00066[15] N. F. Davar、M. A. A. Dewan 和 X. Zhang，《教育中的人工智能聊天机器人：挑战与机遇》，Information，第 16 卷，第 3 期，第 235 页，3月 2025 [ ] 10 3390 16030235 。[16] 在线。可查阅：http://dx.doi.org/ . /info Y. Hicke、A. Agarwal、Q. Ma 和 P. Denny，《AI-ta：基于开源大语言模型的智能问答教学助理》，2023 [ ] 2311 02775 。在线。可查阅：[17] https://arxiv.org/abs/ . P. Lewis、E. Perez、A. Piktus、F. Petroni、V. Karpukhin、N. Goyal 等，《面向知识密集型自然语言处理任务的检索增强生成》，载于《第 34 届神经信息处理系统国际会议论文集》，系列：NIPS '20。美国纽约州雷德胡克：Curran Associates Inc.，2020年。[18] K. Shuster、S. Poff、M. Chen、D. Kiela 和 J. Weston，《检索增强可减少对话中的幻觉》，载于《计算语言学协会 EMNLP 2021 论文集》，M.-F. Moens、X. Huang、L. Specia 和 S. W.-t. Yih 编。多米尼加共和国蓬塔卡纳：计算语言学协会，2021年11月，第 3784–3803 页。 [在线]。可查阅：https://aclanthology.org/2021.findings- emnlp.320/ [19] A. Grattafiori、A. Dubey、A. Jauhri、A. Pandey、A. Kadian、A. Al-Dahle 等，《Llama 3 模型家族》，2024年。 [在线]。可查阅：https://arxiv.org/abs/ 2407 21783 . M. P.-C. Lin、[20] D. Chang、S. Hall 和 G. Jhajj，《教育领域开源大语言模型的初步系统综述》。Springer Nature Switzerland，2024年，第 68–77 页。 [在线]。可查阅：http://dx.doi.org/10.1007/978-3-031-63028-6 6

[21] ´E. Ollion、R. Shen、A. Macanovi´c 和 A. Chatelain，“将专有大语言模型用于研究的危险”，Nature Machine Intelligence，第 6 卷，第 1 期，第 4–5 页，2024年1月。 [在线]。可访问：http://dx.doi.org/ 10 1038 42256 023 00783 6 。 /s - - - S. Bubeck、V. [22] Chandrasekaran、R. Eldan、J. Gehrke、E. Horvitz、E. Kamar 等，“通用人工智能的火花：GPT-4 的早期实验”，2023 年。 [在线]。可访问：https://arxiv.org/abs/2303.12712[23] K. Team、T. Bai、Y. Bai、Y. Bao、M. C.、J. Cai 等，“Kimi K3：开放前沿智能”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2607 24653 。GLM- -Team、：、A. Zeng、X. Lv、Z. Hou、[24] 5 Z. Du、Q. Zheng 等，“GLM-5：从氛围编程到智能体工程”，2026年。 [在线]。可访问：https://arxiv.org/abs/ 2602 15763 。[25] DeepSeek-AI、A. Xu、B. Lin、B. Xue、B. Wang、B. Xu 等，“DeepSeek-v4：迈向高效的百万词元上下文智能”，2026年。 [ ] 2606 19348 在线。可访问：[26] https://arxiv.org/abs/。W. Kwon、Z. Li、S. Zhuang、Y. Sheng、L. Zheng、C. H. Yu 等，“采用分页注意力的大语言模型服务高效内存管理”，载于《第 29 届操作系统原理研讨会论文集》，系列：SOSP '23。美国纽约州纽约：Association for Computing Machinery，2023 年，第 611–626 页。 [在线]。可访问：https://doi.org/ 10 1145 3600006 3613165 。/ 。L. Zheng、W.-L. [27] Chiang、Y. Sheng、S. Zhuang、Z. Wu、Y. Zhuang、Z. Lin、Z. Li、D. Li、E. P. Xing、H. Zhang、J. E. Gonzalez 和 I. Stoica，“使用 MT-Bench 和 Chatbot Arena 评审大语言模型评审”，载于《第 37 届神经信息处理系统国际会议论文集》，系列：NIPS '23。美国纽约州雷德胡克：Curran Associates Inc.，2023年。[28] P. Verga、S. Hofstatter、S. Althammer、Y. Su、A. Piktus、A. Arkhangorodsky、M. Xu、N. White 和 P. Lewis，“以陪审团取代评审者：使用多样化模型组成的评审团评估大语言模型生成结果”，2024 [ ] 2404 18796 。在线。可访问：[29] https://arxiv.org/abs/。R. D. Rogers 和 S. Monsell，“简单认知任务之间可预测切换的成本”，《实验心理学杂志：综合》，第 124 卷，第 2 期，第 207–231 页，1995年。[30] Z. Nussbaum、J. X. Morris、A. Mulyar 和 B. Duderstadt，“Nomic Embed：训练可复现的长上下文文本嵌入器”，《机器学习研究汇刊》，2025年，可复现性认证。 [在线]。可访问：https://openreview.net/forum?id=IP mzyQSiQE [31] Y. Liu、Y. Jia、R. Geng、J. Jia 和 N. Z. Gong，“提示注入攻击与防御的形式化及基准测试”，载于《第 33 届 USENIX 安全研讨会（USENIX Security 24）》，2024年，第 1831–1847 页。[在线]。可访问：https://aclanthology.org/2022.emnlp-main.225/ [32] H. Bastani、O. Bastani、A. Sungu、H. Ge、 ¨Ozge Kabakcı 和 R. Mariman，“缺乏护栏机制的生成式人工智能可能损害学习：来自高中数学的证据”，《美国国家科学院院刊》，第 122 卷，第 26 期，第 e2422633122 页，2025年。 [在线]。可访问：https://www.pnas.org/doi/abs/ 10 1073 2422633122 。/pnas。E. [33] Perez、S. Huang、F. Song、T. Cai、R. Ring、J. Aslanides、A. Glaese、N. McAleese 和 G. Irving，“使用语言模型对语言模型进行红队测试”，载于《2022 年自然语言处理实证方法会议论文集》。Association for Computational Linguistics，2022年，第 3419–3448 页。 [在线]。可访问：https://aclanthology.org/2022.emnlp-main.225/ [34] OpenRouter, Inc.，“OpenRouter：大语言模型的统一接口”，https: //openrouter.ai/，2026年，访问日期：2026-05-01。[35] J. Cohen，“加权卡帕系数：具有等级化分歧或部分得分处理的名义尺度一致性”，《心理学通报》，第 70 卷，第 4 期，第 213–220 页，1968年。[36] J. Gu、X. Jiang、Z. Shi、H. Tan、X. Zhai、C. Xu 等，“大语言模型评审综述”，The Innovation，第 101253 页，2026年。 [在线]。可访问：https://www.sciencedirect.com/science/article/pii/S2666675825004564 [37] A. Singh、A. Fry、A. Perelman、A. Tart、A. Ganesh、A. El-Kishky 等，“OpenAI GPT-5 系统卡”，2025年。 [在线]。可访问：https://arxiv.org/abs/ 2601 03267 。Anthropic，“介绍 Claude Opus 4.5”，[38] https://www.anthropic.com/ news/claude-opus-4-5，2025年11月，访问日期：2026-05-01。[39] T. Cohere、：、Aakanksha、A. Ahmadian、M. Ahmed、J. Alammar、M. Alizadeh 等，“Command A：企业级大语言模型”，2025 [ ] 2504 00698 。在线。可访问：[40] https://arxiv.org/abs/。K. Krippendorff，《内容分析：方法导论》，第 2 版。美国加利福尼亚州千橡市：Sage，2004年。

[41] J. L. Fleiss，“衡量多名评分者之间的名义尺度一致性”，《心理学通报》，第 76 卷，第 5 期，第 378–382 页，1971年。[42] J.-Y. Antoine、J. Villaneau 和 A. Lefeuvre，“加权 Krippendorff's alpha 是一种更可靠的多编码者序数标注指标：关于情绪、观点和共指标注的实验研究”，载于《第 *14* 届欧洲计算语言学协会分会会议论文集（*EACL*）》，瑞典哥德堡，2014年，第 550–559 页。 [在线]。可访问：https://aclanthology.o rg/E 14 1058 - J. R. Landis 和 G. G. Koch，“分类数据的观察者一致性测量”，《生物统计学》，第 33 卷，第 1 期，第 159–174 页，1977年。[44] M. Renze，“采样温度对大语言模型问题解决能力的影响”，载于《计算语言学协会论文集：*EMNLP 2024*》。计算语言学协会，2024年，第 7346–7356 页。 [在线]。可访问：http://dx.doi.org/ 10 18653 1 2024 432 。/v / .findings-emnlp。[45] Y. Wu、Y. Wang、Z. Ye、T. Du、S. Jegelka 和 Y. Wang，“越多反而越少：理解大语言模型中的思维链长度”，2025年。 [在线]。可访问：https://arxiv.org/abs/ 2502 07266 。Y. Gao、Y. Xiong、X. Gao、K. Jia、[46] J. Pan、Y. Bi 等，“面向大语言模型的检索增强生成：综述”，2024年。 [ ] 2312 10997 在线。可访问：[47] https://arxiv.org/abs/。M. Farinas、A. Garcia、M. Reyes、J. Nueva、N. Qureshi、M. Elsaadany 和 S. Shamseldin，“Jamn bot RAG 基准测试”，2026年。 [在线]。可访问：https://zenodo.org/doi/10.5281/zenodo.21569635