法律 R AG Bench：法律 RA 的端到端基准 G

Abdur-Rahman Butler、Umar Butler
Isaacus

2026年3月2日

arXiv:2603.01710v1 [cs.CL] 2 Mar 2026

摘要

我们推出 Legal RAG Bench，这是一套用于评估法律RAG系统端到端性能的基准与评估方法。作为基准，Legal RAG Bench 包含来自维多利亚州刑事指控手册的 4,876 个文本片段，以及 100 道复杂的人工编写问题，这些问题要求具备刑法与程序方面的专业知识。同时提供长篇答案和支撑段落。作为一种评估方法，Legal RAG Bench 采用全因子设计和新颖的分层错误分解框架，从而能够对检索增强生成中检索模型与推理模型的贡献进行同口径比较。我们评估了三种最先进的嵌入模型（Isaacus 的 Kanon 2 嵌入器、谷歌的 Gemini Embedding 001，以及 OpenAI 的 Text Embedding 3 Large）和两种前沿大语言模型（Gemini 3.1 Pro 与 GPT-5.2），发现信息检索是法律 RAG 性能的主要驱动因素，而大语言模型对正确性和依据性的效应则较为温和。尤其是 Kanon 2 Embedder 对性能的正向影响最大，使平均正确性提高了 17.5 分，依据性提高了 4.5 分，检索准确率提高了 34 分。我们观察到，法律RAG系统中许多被归因于幻觉的错误实际上是由检索失败引发的，并由此得出结论：检索为许多现代法律RAG系统的性能设定了上限。我们阐述了构建 Legal RAG Bench 的原因与方法，并一并呈现评估结果。我们还公开了代码与数据，以协助复现我们的发现。

1 引言

检索增强生成（RAG）是指让大语言模型以从语料库中检索到的当前上下文为条件。近年来，RAG已成为向人工智能系统注入新信息的最流行方法之一，尤其包括法律人工智能系统，这类系统往往要求更高水平的依据性和可验证性。然而，尽管法律科技行业对RAG的采用十分踊跃，面向法律RAG应用的高质量端到端基准测试仍然相对稀缺。大多数公开基准测试要么面向一般应用，要么只评估法律RAG的某一个组成部分，通常是检索或文本生成中的一项。更广泛地说，许多法律人工智能基准测试存在标签质量不佳、方法与设计缺陷，以及其所声称评估的内容与实际评估内容普遍不符等问题。

这种高质量端到端法律RAG评估的缺失，意味着应用构建者往往不得不依赖与真实世界表现仅有微弱相关、甚至毫无相关的评估，进而降低其应用对律师和终端消费者的价值。相应地，低质量的法律基准测试可能导致模型构建者为了一个关于“良好法律人工智能”应有形态的不可行模型，而牺牲真实世界表现。

因此，为了解决 t 大多数法律 RAG 基准测试的不足，从而提升基于法律基准测试构建的应用的整体 l

质量。为此，我们推出 Legal RAG Bench，这是一个新的数据集和评估方法，用于评估法律RAG系统的端到端性能。作为一个数据集，Legal RAG Bench 包含维多利亚州刑事指控手册中的4,876 个文本片段 [7]，以及 100 个由专家精心设计、具有实质挑战性的问题，要求具备维多利亚刑法与刑事诉讼程序方面的专家级知识。每个问题都配有一份长篇答案和一个支撑段落，从而能够同时评估检索与答案生成的性能。作为一种方法，Legal RAG Bench 针对多种前沿检索模型和生成模型开展全因子实验，并利用一种新颖的错误分类体系，将错误分解为幻觉、检索失败和推理失败，再将这些错误归因于特定模型。我们的分析明确表明，检索质量是法律 RAG 端到端性能的首要决定因素，而且生产环境中的法律RAG系统所产生的大多数幻觉都是由检索失败引起的。我们公开发布 Legal RAG Bench 的数据、代码和结果，以支持可复现的评估。

2 相关工作

与其他领域相比，法律人工智能评估仍然十分稀缺。现有的评估中，许多存在标签质量低、方法论缺陷，以及对其现实价值的假设不当等问题。这些问题背后，往往还存在一个更严重的缺陷：在基准测试的设计和标注过程中，没有充分引入并依靠法律领域专业知识。

2.1 检索基准测试

近来，法律检索评估领域变化不大，例外是2025年10月发布的大规模法律嵌入基准（MLEB） [2] 。迄今为止，MLEB 仍是规模最大、最多样且最全面的开源法律检索模型基准测试。在介绍 MLEB 时，我们将其构建方式与其他流行的检索基准测试进行了对比，并指出这些基准普遍存在严重缺陷，其中包括大规模文本嵌入基准（MTEB） [8]中的法律子集。

以 MTEB 中包含的 AILA 案例文档和 AILA 法规 [1] 数据集为例。我们发现，这些数据集是通过一种“自动化方法”创建的：将“某些 [印度] 最高法院案件中陈述的事实”与“审理这些案件的律师所援引的”案件和法规配对 [1]。作者称，“实际上让法律专家参与进来（例如，寻找相关的先前案件／法规）需要投入大量资金和时间” [1, 第 4]页。只要对判决的撰写方式和法律引证的运作方式有基本了解，就会明白，仅凭援引某一案件的判决文本来检索该案件的事实，在大多数情况下是不可能的；即使能够做到，实际价值也微乎其微。举例来说，尽管*Donoghue* 诉 *Stevenson* 案 [5]在事实上涉及一位名叫 May Donoghue 的女士，她因饮用了含有腐烂蜗牛的姜汁啤酒而病倒，但世界各地的判决都曾援引、且至今仍在援引此案，以支持一些法律观点；客观而言，这些观点与蜗牛、姜汁啤酒或 Donoghue 女士毫无关系。

2.2 大语言模型和检索增强生成基准测试

即使在最受欢迎的文本生成和检索增强生成基准测试中，也能发现类似缺陷，例如人类最后的考试（HLE） [9], ——据称这是“同类中最后一个覆盖广泛学科的封闭式学术基准测试”，由“全球各地的领域专家开发”，耗资至少 50 万美元的一个评估集。我们对 HLE 法律

子集的审查发现，大多数示例要么不恰当，要么表述不清，要么标注错误。下面这组问题与答案就是一个典型例子 [3]。

人类最后的考试示例 #6736f7bc980211368f0f94eb

问题：
Tommy 退休后购置了一块土地。2016 年，Tommy 将这块土地与他的兄弟 James 分割。

律师提交并登记了 2 份契据。土地上有两辆拖车。土地和拖车分别有不同的税务地图编号。

现在，土地已合法登记并记录为 A地块归 Tommy 所有，B地块归 James 所有。
Tommy 的编号是 1234567890，拖车编号是 1234567890.1。A地块

James 的编号是 0987654321 和 0987654321.1。B地块

Tommy 于 2017 年去世。由于 A地块 的备案有误，Tommy 的遗产必须更正 Tommy 的契据。Tommy 的契据于 2017 年备案并重新登记至 B地块。

现在 Tommy 的编号是 1234567890，拖车编号是 1234567890.1，位于 B地块。

James 于 2016 年以其契据为 B地块设立了留置权。

按照律师于 2017 年的建议，James 从未将其契据更正为 A地块。James 需要其 B地块的原始契据以进行更正。

Tommy 和 James 的拖车都从未移动过。

A地块和 B地块分别归谁所有，编号是什么?

答案：
Tommy与1234567890共同拥有A地块和B地块。

不仅这个问题本身存在缺陷，所谓的答案也不正确。首先，该问题没有提供足够的上下文，因而无法被完整解答。不同司法管辖区对财产所有权如何转让和确认有着不同的规则。然而，该问题既未说明适用的司法管辖区，也未提供据以解释的规则。其次，无论如何，所谓的答案几乎可以肯定是不正确的。由于Tommy的遗产已更正其契据，将B地块而非A地块登记为与其土地相对应，Tommy不可能同时拥有这两块地。

在开源法律评估数据集中，未能使答案所依赖的假设能够从提供给模型的全部上下文中合理推断出来的情况屡见不鲜。但即便不存在明显的标注或方法论错误，基准测试声称评估的内容与其实际上评估的内容之间仍可能存在致命错位。LegalBench [6]和 LegalBench-RAG [10]都存在后一类问题。尽管它们被宣传为检验大语言模型推理能力和检索能力的有效压力测试，但其绝大多数数据实际上

由价值较低、相对简单的文本分类和情感分析任务构成，这些任务只需回答“是”或“否”，例如：“该条款是否描述了向被许可人（包括再许可人）及该被许可人／再许可人的关联方授予许可？”以及“请考虑 Exact Sciences Corporation 与辉瑞公司之间的 Cologuard推广协议；本合同是否包含任何优先拒绝权、优先要约权或优先谈判权？”

同样，斯坦福RegLab的HousingQA和BarExamQA [12] 评估虽然比LegalBench更具实质性挑战 [6] ，也比LegalBench-RAG更具挑战性 [10], ，但仍未能真正评估其声称要评估的能力。具体而言，由于只使用封闭式问题评估模型，这些评估无法模拟现实世界法律RAG系统中混乱得多的情形；在这种情形下，大语言模型不仅可能无法生成连贯的答案，有时还会产生看似正确、实则根本缺乏依据的幻觉答案。作者本人也承认了这一不足，并指出：“我们的数据集在主题领域方面受到限制，且仅限于多项选择答案形式，以便自动评估下游任务。这些数据集可能无法真实近似法律问题完整的自然分布。”[12]。

3 数据集

为解决当代法律 RAG 评估所面临的全方位问题，我们推出了LegalRAGBench，这是一个基准及评估方法，用于稳健地评估法律RAG系统在真实世界端到端场景中的表现。

作为数据集，Legal RAG Bench 由从维多利亚司法学院《刑事指控手册》 [7] 抽样的 4,876 个文本片段组成，并配有 100 个复杂、具有实质挑战性且手工精心设计的问题；要正确回答这些问题，需要具备维多利亚刑法与程序方面的专家级知识。就此而言，Legal RAG Bench 是首个评估集，用于评估更大规模检索增强生成系统中检索与生成模型的表现，这些系统旨在提供切实可行的现实世界法律咨询，尤其是在资源不足但至关重要的领域，即刑法领域。

独特之处在于，法律与人工智能领域的专业知识贯穿并指导了 Legal RAG Bench 设计与开发的每一个阶段，从精心构建一套多样化的现实假设情景，一直到选择维多利亚《刑事指控手册》作为该基准的基础，因其与刑事律师的日常工作具有核心相关性。

在构建 Legal RAG Bench 时，我们下载了《刑事指控手册》的每一节 [7]，将其保存为微软Word文档，并转换为 Markdown。我们运用一套复杂的启发式方法，将各节拆分为完整的层级结构，例如章与分章；随后在必要时，使用 semchunk [11] 语义分块算法对各节进一步分块，使每个分块的长度不超过 512 个词元（由 Kanon法律分词器测定）。

在构建由 4,876 个段落组成的语料库后，我们随机抽取文本片段，手工编写了 100 个复杂且具有实质挑战性的问题；这些问题在最大程度上要求仅凭各自对应的段落即可正确作答。编写问题时，我们尽可能使其与相关段落在词汇上有所区别，以严格检验受评估模型的语义理解能力。

领域专家还为每个问题手工编写了长篇答案，为 Legal RAG Bench 中的每个示例构建了问题-答案-证据三元组。这种独特的结构使法律检索增强生成系统的检索组件和生成组件能够分别接受评估，既

分别考察并结合考察。

4 评估

作为一种方法，Legal RAG Bench 是首个全因子实验，用于评估法律检索增强生成系统的检索准确率、正确性和依据性，从而能够以实证方式公平比较检索模型和生成模型对性能的相对影响。具体而言，我们评估了三种最先进的嵌入模型——Isaacus 的 Kanon 2 嵌入器、谷歌的 Gemini Embedding 001 和 OpenAI 的 Text Embedding 3 Large——与两种前沿大语言模型 Gemini 3.1 Pro 和 GPT-5.2 的所有组合。选择这些模型是基于它们的普及度，以及它们在法律检索和推理方面据称的表现。为尽量减少混杂变量，我们在所有模型中使用相同的精简版、基于 Langchain 的 [4] 检索增强生成流水线，且不更改任何超参数的默认值。我们使用高推理模式下的 GPT-5.2 作为“评判者”，根据答案和检索上下文评估大语言模型的表现。我们对内部的大语言模型评判评估流水线进行审查后发现，使用高推理模式下的 GPT-5.2 作为评判者，准确率达到了 99%，我们认为这一结果可以接受。我们之所以能达到如此高的准确率，是因为为评判者提供了明确的评估量规，以及一组简单、尽可能明确无歧义的二元结果供其预测。

具体而言，对于每个问题 i、嵌入模型 e 和大语言模型 l，我们评估了以下维度：

1. **正确性** (c_{eli})：如果模型的回答蕴含参考答案，则记为 1；否则记为 0。2. **依据性**(g_{eli})：如果模型的回答得到提供给模型的检索段落支持，则记为 1（无论这些段落实际上是否相关）；否则记为 0。3. **检索准确率** (r_{ei})：如果嵌入模型检索到了经过标注的支撑段落，则记为 1；否则记为 0。

这三个信号共同展现了检索增强生成系统各组成部分的有效性，使我们能够发现仅考察整体 RAG准确率时会被掩盖的失效模式，例如幻觉。事实上，任何对 RAG性能的严肃评估都必须考虑幻觉，包括那些尽管产生了正确答案、但仍属于幻觉的情况。法律研究与分析在很大程度上以证据为导向——法律结论的可验证性往往与其真实性同等重要，甚至更为重要。不真实但可验证的结论总能被证明是错误的，而真实但不可验证的结论则永远无法被证明为真实。

因此，为了涵盖幻觉等可能碰巧得出正确答案的失效模式，我们进一步提出一种新的分层错误分解分类法，其中包括以下错误类型：

1. **幻觉**：生成模型编造了其所提供的上下文中不存在的事实（$g_{eli} = 0$）。2. **检索错误**：嵌入模型未能检索到相关段落，导致生成模型给出有依据但不正确的答案（$g_{eli} = 1 \wedge c_{eli} = 0 \wedge r_{ei} = 0$）。3. **推理错误**：嵌入模型检索到了相关段落，但生成模型仍然生成了不正确的答案（$g_{eli} = 1 \wedge c_{eli} = 0 \wedge r_{ei} = 1$）。

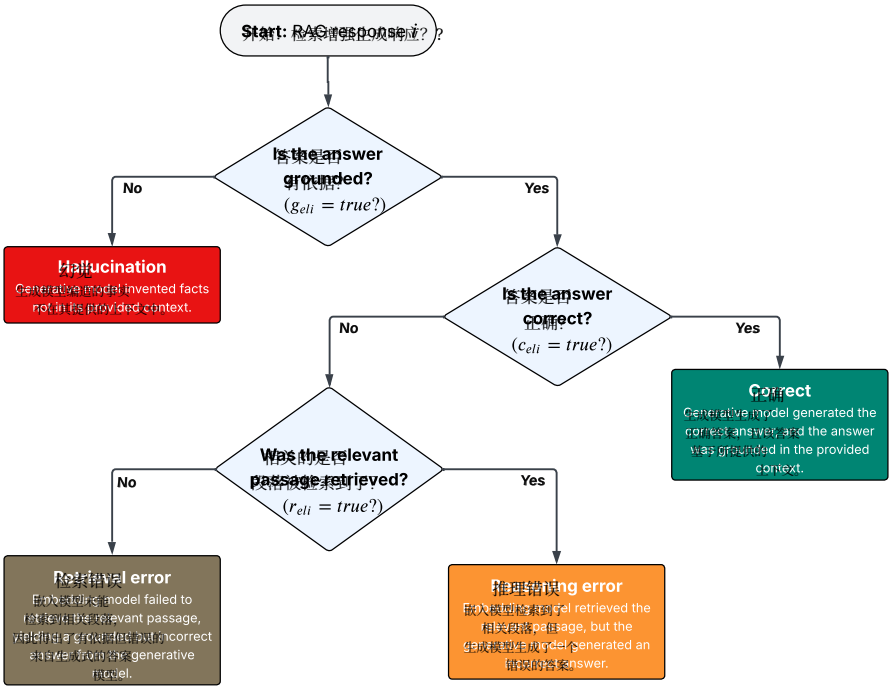

图1：表示我们的分层检索增强生成错误分解分类体系的流程图。

在我们的分类体系中，幻觉被有意列为检索增强生成流水线中首要的潜在失败模式，甚至排在检索失败之前。正如前文所述，在现实世界中，出现幻觉或缺乏依据的答案本身无法验证其正确性。我们也只将导致错误答案的情况认定为检索失败，因为从逻辑上讲，如果生成模型的回答完全基于其检索到的上下文，且回答正确，那么检索到的上下文很可能是相关的，即使其中恰好没有包含 Legal RAG Bench 标注为最相关的特定段落；毕竟，其他段落也可能相关。

5 结果

5.1 评估维度

表1 报告了每种评估过的嵌入模型与大语言模型组合的正确性、依据性和检索准确率。 表2 报告了相应的模型层面平均值：嵌入模型的分数是跨大语言模型取平均值，大语言模型的分数则是跨嵌入模型取平均值。

LLM	嵌入器	正确率（%）	依据充分度（%）	检索准确率（%）
Gemini 3.1 Pro	Kanon 2 Emb.	**95.0**	95.0	**86.0**
	Text Emb. 3 L.	77.0	**96.0**	52.0
	Gem. Emb. 001	75.0	92.0	53.0
GPT-5.2	Kanon 2 Emb.	**93.0**	**97.0**	**86.0**
	Text Emb. 3 L.	76.0	87.0	52.0
	Gem. Emb. 001	73.0	82.0	53.0

表1：按大语言模型和嵌入模型划分的各维度得分。

模型	正确性（%）	依据性（%）	检索准确率（%）
嵌入器			
Kanon 2 Embedder	**94.0**	**96.0**	**86.0**
Text Embedding 3 Large	76.5	91.5	52.0
Gemini Embedding 001	74.0	87.0	53.0
LLMs			
Gemini 3.1 Pro	**82.3**	**94.3**	**63.7**
GPT-5.2	80.7	88.7	**63.7**

表2：嵌入模型和大语言模型维度得分的平均值。

从这些结果可以清楚看出，对法律RAG系统性能起主导作用的是嵌入模型的选择，而非大语言模型。表现最佳的嵌入模型 Kanon 2 Embedder 尤其突出，其正确性比表现最佳的大语言模型 Gemini 3.1 Pro 高出 11.7 个百分点，依据性高出 1.7 个百分点，检索准确率高出 22.3 个百分点。改用 OpenAI 的 Text Embedding 3 Large 后，正确性相应下降 17.5 个百分点，依据性下降 4.5 个百分点，检索准确率下降 34 个百分点。

我们进一步观察到，大语言模型的选择会对正确性和依据性产生中等程度的影响。Gemini 3.1 Pro 的平均正确性比 GPT-5.2 高 1.6 个百分点，依据性高 5.6 个百分点。值得注意的是，当 Kanon 2 Embedder 与 Gemini 3.1 Pro 和 GPT-5.2 搭配使用时，两者在依据性上的差距显著缩小——GPT-5.2 最终比 Gemini 3.1 Pro 高出 2 个百分点。这表明，在获得无关上下文时，GPT-5.2 产生幻觉的可能性可能高得多。

5.2 错误分解

借助我们新颖的错误分解框架，我们不仅报告正确性、依据性和检索准确率，还进一步将错误分解为幻觉、推理失败和检索失败，如图 2所示。这种分解使我们能够交叉定位错误的精确来源。

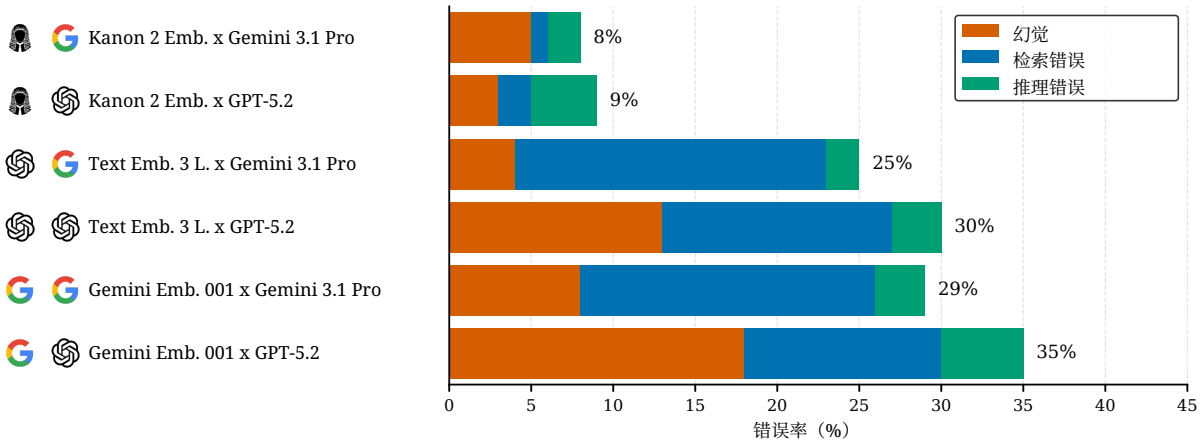

图 2：每种嵌入模型与大语言模型组合的分解错误率。

我们评估中最重要的发现之一是，较差的检索往往与幻觉的增加密切相关。相对于 Text Embedding 3 Large，Gemini Embedding 001 平均使幻觉增加 4.5 个百分点。而 Kanon 2 Embedder 与其通用替代方案相比，平均使幻觉降低 6.75 个百分点。我们由此推断，生成模型或许能够判断某个段落何时更可能正确或至少相关，并在这种情况下更不容易编造新事实来帮助自己给出答案。

切换生成模型也会对幻觉产生中等程度的效应，其中 Gemini 3.1 Pro 的平均幻觉率为 5.7%，而 GPT-5.2 的幻觉率为 11.3%。

使用 Kanon 2 Embedder 时，生成错误所占比例更高；不过，这很可能是因为上游检索错误大幅减少，使 RAG 流水线中的失败转移到了生成层。换言之，过去是表现不佳的检索模型拖累生成模型最终的 RAG 性能，如今则是表现欠佳的生成模型拖累高质量检索模型最终的 RAG 性能。

我们还测量了各模型的平均RAG准确率，如图3所报告，从而能够在嵌入模型与生成模型之间进行同类比较。我们将RAG准确率定义为所有错误之和的补数。据我们所知，该分析是迄今为止对嵌入模型与生成模型对最终法律 RAG 性能之效应所做的最为稳健的比较。

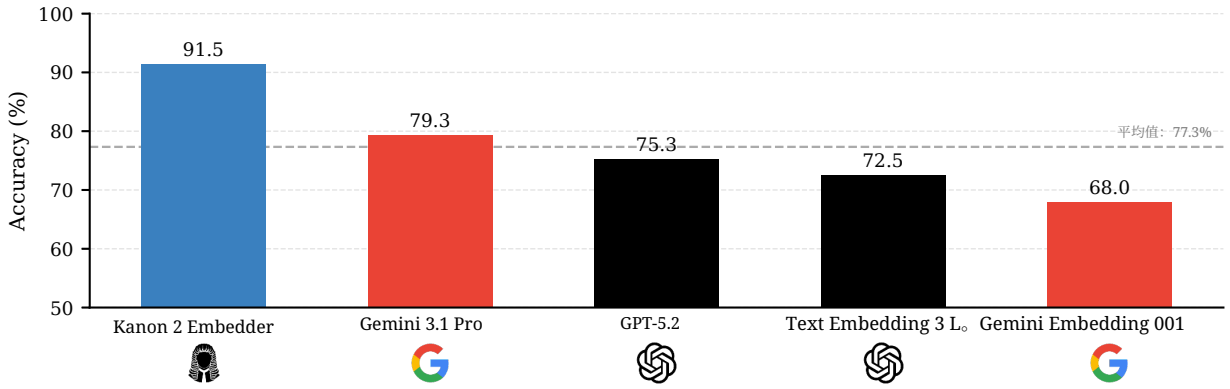

图3：模型对检索增强生成准确率的影响（相对于样本平均值）。

我们的结果显示，Kanon 2 Embedder相对于样本平均值，整体RAG准确率高出18%。相比之下，GPT-5.2和Gemini 3.1 Pro对准确率的影响分别为-3%和 $+3$%。这些结果进一步表明，嵌入模型的选择是RAG性能的主要驱动因素，并且在计入假阳性幻觉后，这一效应更为显著。

5.3 统计显著性与交互效应

实验可能揭示正向关系，但并不能据此确定该关系具有因果性或统计显著性。在像我们这样的析因实验中，还存在另一个问题：分数差异可能是由交互效应而非各模型的独立表现所驱动。

此类交互作用的一个可能来源，在于我们样本中所纳入的嵌入模型和生成模型的选择。例如，Gemini Embedding 001 可能与 Gemini 3.1 Pro 特别匹配，因为二者都是同一家模型开发商谷歌的产品。如果确实如此，我们不能仅凭 Gemini 3.1 Pro 在本次实验中的表现更好，就声称它在 Legal RAG Bench 上总体优于 GPT-5.2。在另一项实验中，例如将 Gemini Embedding 001 换成 Cohere Embed v4，GPT-5.2 反而可能胜过 Gemini 3.1 Pro。

尽管交互效应会威胁法律评估的外部效度，但大多数评估者既不检验也不报告交互效应。而我们与众不同之处，恰恰在于确实做了这两件事。

为量化主效应、检验交互作用并考虑估计不确定性，我们针对 $y \in \{c_{eli}, g_{eli}, b_{eli}\}$拟合线性概率模型，其中 $b_{eli} = 1$ if $c_{eli} \wedge g_{eli}$:

$$y_{eli} = \alpha_i + \beta_e + \gamma_l + \delta_{el} + \varepsilon_{eli}.$$

这里，α_i 表示问题固定效应，β_e 和 γ_l 分别表示嵌入模型和大语言模型的指示变量，而δ_{el} 则捕捉嵌入器与大语言模型之间的交互作用。我们使用普通最小二乘法估计模型，并报告按问题聚类的聚类稳健标准误。

我们还报告与析因设计相对应的 ANOVA式沃尔德检验：

- **交互效应：**对所有嵌入器–大语言模型交互项进行检验 $H_0 : \delta_{el} = 0$ 。在我们的基线设定下，这是一个联合检验，用于检验非基线嵌入模型（Gemini Embedding 001 和 Kanon 2 Embedder）对应的 GPT-5.2 交互系数是否为零。

- **大语言模型主效应**。检验在所有嵌入模型上，平均而言将 Gemini 3.1 Pro 切换为 GPT-5.2 的简单效应。
- **嵌入模型主效应**。检验各嵌入模型相对于基准嵌入模型（Text Embedding 3 Large）的对比，并对大语言模型取平均。具体而言，对于每个非基准嵌入模型 e，我们检验在这种平均方案下，β_e 是否与零存在差异。

表3 报告了线性概率模型的 ANOVA式沃尔德检验结果，该模型包含问题固定效应，并采用按问题聚类的聚类稳健标准误。对于全部三项结果，嵌入模型主效应均具有统计显著性，表明嵌入模型的选择解释了相当大比例的性能差异。相比之下，对于正确性，或同时考察正确性和依据性时，大语言模型主效应在统计上均无法与零区分。

	正确	有依据	Both
交互作用	$\chi^2(2) = 0.08$ ($p = 0.962$)	$\mathbf{\chi^2(2) = 8.12}$ **($p = 0.017^{*}$)**	$\chi^2(2) = 1.47$ ($p = 0.479$)
大语言模型主效应	$\chi^2(1) = 0.46$ ($p = 0.499$)	$\mathbf{\chi^2(1) = 5.36}$ **($p = 0.021^{*}$)**	$\chi^2(1) = 2.43$ ($p = 0.119$)
嵌入器主效应	$\mathbf{\chi^2(2) = 28.63}$ **($p < 0.001^{***}$)**	$\mathbf{\chi^2(2) = 8.95}$ **($p = 0.011^{*}$)**	$\mathbf{\chi^2(2) = 32.00}$ **($p < 0.001^{***}$)**

注：星号：$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$。粗体表示 $p < 0.05$。

表3：线性概率模型 $y_{eli} = \alpha_i + \beta_e + \gamma_l + \delta_{el} + \varepsilon_{eli}$的 ANOVA 风格 Wald χ^2 检验，采用问题固定效应，并按问题聚类的聚类稳健标准误，分别报告 c_{eli}（正确）、 g_{eli} （有依据）和 b_{eli} （两者皆是；仅当 $c_{eli} = g_{eli} = 1$时 $b_{eli} = 1$ ）。

唯一在平均意义上受大语言模型选择影响的结果是依据性，其中大语言模型主效应显著。有趣的是，表3 还表明，依据性是唯一可检测到嵌入模型与大语言模型的交互效应的指标。这意味着依据性差异在不同嵌入模型之间并不稳定，不应由单一的全局大语言模型排名来概括。更一般地说，这些结果促使我们在对RAG 流水线进行基准测试时，将交互效应检验与主效应一并报告。

效应	对比	估计值（百分点）（标准误）	p
ΔLLM	与 Text Emb. 3 L.	**-9.0 (4.2)***	**0.031**
ΔLLM	与 Gem. Emb. 001	**-10.0 (4.8)***	**0.038**
ΔLLM	与 Kanon 2 Emb.	2.0 (2.7)	0.459
ΔΔ	Kanon 2 Emb. − Text Emb. 3 L.	**11.0 (5.2)***	**0.033**
ΔΔ	Kanon 2 Emb. − Gem. Emb. 001	**12.0 (5.0)***	**0.017**
ΔΔ	Gem. Emb. 001 −Text Emb. 3 L.	-1.0 (6.2)	0.871

注： Δ大语言模型表示在同一嵌入器内，将生成器从 Gemini 3.1 Pro 切换到 GPT-5.2 的简单效应。ΔΔ 表示相对于基线嵌入器（Text Embedding 3 L.）的相应双重差分。星号：
$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$, $^{\dagger}p < 0.1$。粗体表示 $p < 0.05$。

表4：依据性的简单大语言模型效应及嵌入器–大语言模型交互作用对比（g_{eli}）。

表4 通过报告在每种嵌入模型下切换生成模型的简单效应，以及相对于基线的交互对比，进一步拆解了依据性交互作用。从Gemini 3.1 Pro切换到GPT-5.2时，与Text Embedding 3 Large搭配会使依据性降低9.0分，与Gemini Embedding 001搭配会降低10.0分，但与Kanon 2 Embedder搭配时未产生统计上可检测到的变化。交互对比表明，这种异质性由Kanon 2 Embedder驱动。与Kanon 2 Embedder搭配时，大语言模型切换效应比与Text Embedding 3 Large搭配时正向高出11分，比与Gemini Embedding 001搭配时正向高出12分，而Text Embedding 3 Large与Gemini Embedding 001之间并无有意义的差异。

综上所述，表3 和表4 进一步印证了我们的主要结论。嵌入模型的选择是影响端到端法律 RAG 性能的首要因素，也是统计上最稳健的因素。检索会显著影响幻觉，不过嵌入模型与大语言模型之间的交互作用可能会掩盖这种关系。最后，通过在检验主效应的同时检验交互作用，我们得到了组件级估计，并表明，对于正确性，嵌入模型和大语言模型带来的性能提升呈线性相加关系。在实践中，这意味着，在 Legal RAG Bench 上选择表现最佳的嵌入模型和大语言模型，并将二者共同部署于分布相似的语料库中，便能获得可预测的性能提升。

6 结论

法律RAG系统的应用场景往往十分重视依据性，其重要性不亚于正确性，甚至有过之而无不及。可验证的错误结论总能被证伪，而无法验证的正确结论则永远无法被证明为真。随着检索增强生成在现代法律人工智能应用中日益不可或缺，从业者依赖的法律基准测试不仅应当质量上乘、与现实表现相关，还应评估最重要的失效模式，包括幻觉和推理失败。迄今为止，基准测试往往在质量、方法严谨性以及与真实法律工作的相关性方面有所欠缺，令从业者失望。为弥补可靠的法律 RAG 评估方面的不足，我们推出了 Legal RAG Bench。

作为一个数据集，Legal RAG Bench 将《维多利亚州刑事指控手册》中的 4,876 个文本片段 [7]与 100 个具有实质挑战性的专家编写的问题及长篇答案配对。作为一种方法论，Legal RAG Bench 采用全因子设计和新颖的分层错误分解分类法，对 RAG 流水线进行端到端评估。实际上，Legal RAG Bench 构成了对法律RAG系统现实世界性能最有用、最稳健的评估之一。

我们的结果确凿地表明，在所有评估维度上，嵌入模型的选择对 RAG性能起着主导作用。我们通过全面分析其统计显著性及潜在的交互效应，使这一发现得以具体呈现。我们还证实，信息检索效果不佳往往会引发法律领域幻觉增加，这表明生成模型可能意识到自己正在编造事实来回答那些缺乏相关事实依据的问题。因此，换用高质量的领域适配的法律嵌入模型（例如 Kanon 2 Embedder），能够提升性能上限，直至推理模型的质量成为流水线的瓶颈，而非嵌入模型。

为提升透明度，我们公开发布了 Legal RAG Bench 背后的代码、数据和结果。我们还将 Legal RAG Bench 纳入 MLEB[2], ，进一步推动整体法律信息检索评测的发展。我们鼓励他人复现我们的结果。

7 数据可用性

Legal RAG Bench 数据集可在 Hugging Face 上公开获取，地址为 https://huggingface.co/datasets/isaacus/legal-rag-bench。Legal RAG Bench 背后的代码可在 GitHub 上获取，地址为https://github.com/isaacus-dev/legal-rag-bench。最后，我们在宣布该项目的博客文章中，通过交互式浏览器展示了 Legal RAG Bench 的所有结果： https://isaacus.com/blog/legal-rag-bench。

8 披露

Isaacus 是一家专注于基础研究的法律人工智能研究公司，作者是该公司的创始人。该公司创建了 Kanon2 Embedder，并资助了 LegalRAGBench 和 MLEB 的开发 [2]。

参考文献

[1]Paheli Bhattacharya、Kripabandhu Ghosh、Saptarshi Ghosh、Arindam Pal、Parth Mehta、Arnab Bhattacharya 和 Prasenjit Majumder。《FIRE 2019 AILA 专题概述：用于法律援助的人工智能》。载于 Parth Mehta、Paolo Rosso、Prasenjit Majumder 和 Mandar Mitra 编，《FIRE 2019 工作论文集——信息检索评测论坛》，印度加尔各答，2019年12月12日至15日，CEUR研讨会论文集第 2517 卷，第 1–12 页。CEUR-WS.org，2019年。

[2] Umar Butler、Abdur-Rahman Butler 和 Adrian Lucas Malec。《大规模法律嵌入基准（MLEB）》，2025年10月。

[3] 人工智能安全中心。cais/hle：人类最后的考试（hle）数据集。Hugging Face Datasets。访问日期：2026年3月2日。

[4] Harrison Chase。Langchain（Python软件包）。PyPI，2026年2月。版本1.2.10。[5] Donoghue 诉 Stevenson 案。Donoghue 诉 Stevenson 案。判例，1932年。 [1932] AC 562（HL Sc）。

[6] Neel Guha、Julian Nyarko、Daniel E. Ho、Christopher Ré、Adam Chilton、Aditya Narayana、Alex Chohlas-Wood、Austin Peters、Brandon Waldon、Daniel N. Rockmore、Diego Zambrano、Dmitry Talisman、Enam Hoque、Faiz Surani、Frank Fagan、Galit Sarfaty、Gregory M. Dickinson、Haggai Porat、Jason Hegland、Jessica Wu、Joe Nudell、Joel Niklaus、John Nay、Jonathan H. Choi、Kevin Tobia、Margaret Hagan、Megan Ma、Michael Livermore、Nikon Rasumov-Rahe、Nils Holzenberger、Noam Kolt、Peter Henderson、Sean Rehaag、Sharad Goel、Shang Gao、Spencer Williams、Sunny Gandhi、Tom Zur、Varun Iyer 和 Zehua Li。LegalBench：协作构建的用于衡量大语言模型法律推理能力的基准，2023年。

[7] 维多利亚司法学院。维多利亚州刑事指控手册。https://resources.judicialcollege.vic.edu.au/article/1053858。《法官手册》。访问日期：2025年11月27日。

[8] Niklas Muennighoff、Nouamane Tazi、Loïc Magne 和 Nils Reimers。Mteb：大规模文本嵌入基准，2023年。

[9] Long Phan 等人。《用于评估人工智能能力的专家级学术问题基准》。《自然》，649(8099):1139–1146，2026年1月。

[10]Nicholas Pipitone 和 Ghita Houir Alami。《Legalbench-rag：面向法律领域的检索增强生成基准》，2024年。

[11] Umar Butler。semchunk：一个快速、轻量且易于使用的 Python 库，可将文本拆分为语义连贯的文本块。GitHub 仓库，2025年10月。版本 v3.2.5（提交 6b796e3）。[12] Lucia Zheng、Neel Guha、Javokhir Arifov、Sarah Zhang、Michal Skreta、Christopher D. Manning、Peter Henderson 和 Daniel E. Ho。《面向推理的法律检索基准》。载于计算机科学与法律研讨会论文集，CSLAW '25，第169–193页。ACM，2025年3月。