

GaRAGe: 一个用于 RAG 评价的带有基础标注的基准

Ionut-Teodor Sorodoc
Amazon AGI
csorionu@amazon.com

Leonardo F. R. Ribeiro
Amazon AGI
leonribe@amazon.com

Rexhina Blloshmi
Amazon AGI
blloshmi@amazon.com

Christopher Davis
Amazon AGI
davisjnh@amazon.co.uk

Adrià de Gispert
Amazon AGI
agispert@amazon.com

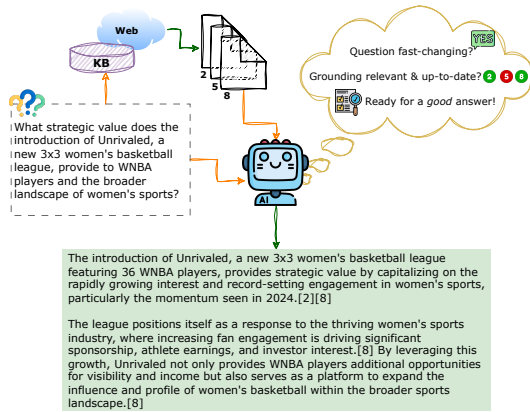


Figure 1: GaRAGe Q & A 示例一个复杂问题和人工整理的长答案的数据点。

Abstract

我们介绍了 GaRAGe，一个大型 RAG 基准，其中包含人工整理的详细答案和每个基础段落的注释，允许对 LLM 生成 RAG 答案时是否能识别相关基础进行细粒度评估。我们的基准包含 2366 个具有多样复杂性、动态性和主题的问题，并且包括从私人文档集和网络检索到的超过 35K 注释段落，以反映真实的 RAG 用例情况。这使其成为评估 LLM 识别必要相关信息以构成响应或在信息不足时提供转移性响应的理想测试平台。对 GaRAGe 上的多种先进 LLM 进行评估表明，模型倾向于过度总结而不是 (a) 严格基于注释的相关段落回答 (达到最多相关性识别事实性得分 60%)，或者 (b) 在没有相关基础时转移 (达到最多 31% 的转移真阳率)。对相关来源的归属度 F_1 最多为 58.9%，我们表明在回答时间敏感问题以及需要从稀疏的私人基础来源中汲取知识时，性能特别下降。¹

1 介绍

大型语言模型 (LLMs) 在许多需要自然语言理解、编码、数学或逻辑推理的任务中表现出一致的改进，并已在用户应用程序中广泛应用，尤其是作为检索增强生成 (RAG) 系统的一部分 (Lewis et al., 2020)。RAG 在实际应用中至关重要，因为用户 (a) 有自己的私人文档需要搜索，且 (b) 需要基于相关的、及时的和可归属的来源来回答。为了让 RAG 应用成功满足这些需求，LLMs 必须同时在两个核心能力上表现出色：查询生成和答案生成。在查询生成方面，LLMs 必须能够判断何时调用可用的外部资源，并生成可能返回良好搜索结果（基础信息）的正确查询，或多个查询，这取决于可用的（对话、情境、视觉）上下文。在答案生成方面，LLMs 必须能够基于可用信息，为用户问题提供流利、有帮助且可能篇幅较长的回答。重要的是，这个任务不能仅通过简单的总结步骤来完成，因为实际上无论是来自用户的私人文档还是互联网，得到的基础信息都会包含相关和无关的段落 (Oh and Thorne, 2023)。因此，LLM 在生成答案之前，需要分析每个基础片段的相关性，通常包含对所选源文档的引文标记以便于透明性和问责制。图 1 展示了一个反映该过程的例子。

无论使用何种特定查询生成方法，生成准确的最终答案的过程都是至关重要的。无论是使用单个查询还是多个查询，进行一遍或多遍，结合推理链，还是从各种信息源中提取，基本挑战仍然存在：LLM 必须仔细从噪声段落中提取相关信息，以提供准确且事实的响应，满足用户需求，同时避免幻觉。

尽管文献中存在大量的 RAG 基准，但可靠地评估这些能力中的每一个仍然是一个挑战，因为查询生成会影响可用的依据，而这个过程缺乏人工注释。以前的工作要么是直接评估最终回答，导致查询生成和答案生成混淆在一起；要么是使用一个或多个文档中固定的依据，

¹数据集可以在 <https://github.com/amazon-science/GaRAGe> 找到

	Human Intervention		Question			Answer			Grounding		
	Human Validation	Human Annotation	Temporal Dynamism	Complexity Variation	Detailed Annotation	Comprehensive	Contain Citations	Deflection	Public and Private	Annotated Grounding	Contain Metadata
MultiHop RAG (Tang and Yang, 2024)	×	×	✓	·	✓	×	×	✓	×	·	✓
CRAG (Yang et al., 2024b)	✓	·	✓	✓	✓	×	×	✓	·	·	✓
ConcurrentQA (Arora et al., 2023)	✓	✓	×	✓	×	×	×	✓	✓	·	×
Summary of a Haystack (Laban et al., 2024)	·	·	×	·	×	✓	✓	×	×	×	✓
RAG-QA Arena (Han et al., 2024)	·	✓	·	·	·	✓	✓	×	×	×	✓
Facts Grounding (Jacovi et al., 2024)	✓	✓	×	·	✓	✓	×	×	×	×	×
GaRAGe (ours)	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: GaRAGe 基准相对于现有 RAG 基准的特征。

这些文档的相关性要么是有保证的，要么是未知的，要么是合成混合的，从而错过了来自用户数据库和网络的现实检索细微差别。除此之外，它们至少具有以下一个缺点：(1) 基于无监督 LLM 的数据集创建，引发关于问题自然性和答案可靠性的担忧，(2) 为了更简单的评估而使用短的或多选题考试类型的答案，但代价是偏离了真实用户的期望，(3) 自动评估与依据的忠实性，却忽视了依据是否正确，以及(4) 不考虑从用户特定文档集和网络获取的混合依据。

为克服这些限制，我们引入了 GaRAGe，这是一项用于 RAG 评估的支持注释基准，包含 2.4K 个问题和超过 35K 个从私人文档和网络检索的手动注释支持段。由专业注释员全面注释和验证，GaRAGe 能够对大型语言模型的答案生成能力进行细粒度评估，适用于不同类型的问题（不同动态性、复杂性等）并有控制的支持噪声度。除了每个片段的支持相关性注释外，该基准还将问题根据多个维度进行分类：时间敏感性、主题受欢迎程度、复杂性或领域，并提供手工制作的长篇、全面的回答，这些回答包含对相关支持片段的归因，并反映了如今 RAG 用户的期望。表 1 展示了 GaRAGe 与其他 RAG 基准的比较。

得益于基础标注，GaRAGe 允许定义关联感知的真实性分数 (RAF)，该分数衡量大语言模型 (LLM) 响应是否依据严格相关的段落，而不是对未知价值的总结——这是对 Jacovi et al. (2024) 的真实性指标的扩展。此外，数据集的一个子集涉及 LLM 的偏转能力，其中提供的基础被标注为不足以导致良好的回答，而 LLM 被期望避免幻觉。此外，我们报告归因指标，利用我们的细粒度标注。使用我们的基准和提议的指标，我们评估了各种最先进的专有和开放重量 LLM，并发现所有模型在正确检测相关基础和过度总结方面都存在困难，RAF 仅达到 60%，而在基础不相关时的偏转率，最大达到 31% 真实正率。这些问题在从稀疏的私人知识库 (KB) 中检索文档，或在处理需要相关性和时间推理的时间敏感问题时变得更加明显。我们发布了用于评估的数据集和提示，以鼓励社区在基准中展示的维度上进行评估和改

进。

我们的创建方法 GaRAGe 的重点是生成能够以高复杂度挑战模型并反映真实世界 RAG 场景的问题。该数据集具有多样化的基础来源和全面的人类精心编辑的长格式答案。图 2 展示了这个过程，其主要元素包括如下：

动态和复杂问题。 一组多样化的问题，具有不同程度的时间动态性、复杂性，涵盖了趋势话题和尾部话题。对于时间动态性，我们构造了快速变化、慢速变化和静态类型的问题。这些不同类别需要与基础信息进行不同的交互，并对参数知识采取不同的宽容度。此外，当简单地从检索到的文档中提取信息不足以回答问题时，问题对于 RAG 来说会变得具有挑战性 (Gabburo et al., 2024)。这种复杂性可能源于多个因素，例如，当答案需要多个逻辑步骤，或需要结合来自不同检索片段的信息时。根据文献 (Pradeep et al., 2024)，我们在比较、多跳和后处理等维度上生成问题（见 §1.1）。

多样地基段落 与每个问题相关的显著数量的段落。基础需要反映问题的主题（例如，如果问题涉及该数据库中的特定实体，则应包含来自 SEC 文件的段落），同时还需包含足够的信息，以用于撰写全面的答案（参见 §1.2）。

每个组件都由专业标注员进行了标注和验证，最终目标是创建一个全面、可靠的答案，以解决问题并包含归因标识，以标示所使用的信息来源（见 §??）。

1.1 问题构建

为了生成复杂问题，我们使用大型语言模型通过多次推理开发以下四步流程：(1) 生成信息检索的网络搜索计划，(2) 搜索网络资源，(3) 通过信息融合生成问题，最后 (4) 筛选和优化。

受到以往工作 (Luo et al., 2025) 的启发，我们提示 LLM 生成一个针对网页探索的战略计划，重点在于识别相关的信息源。这个计划阶段包括将整体的信息搜寻任务分解为具体且可操作的目标，然后为每个目标生成针对性的搜索查询，以最大化获取信息的相关性和多样性。我们使用种子主题来引导生成过程。我们在附录 B.5 中添加了用于此步骤的提示。

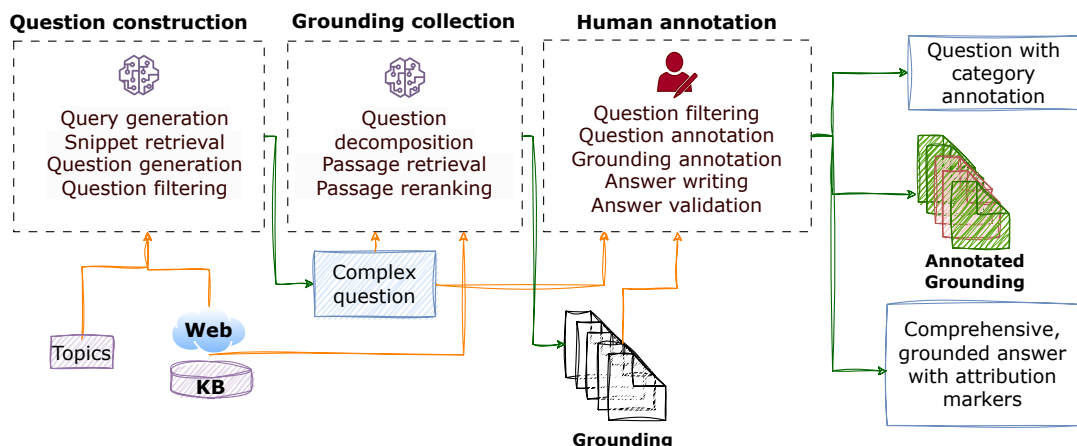


Figure 2: 构建 GaRAge 的流程主要包含 3 个步骤：(i) 通过多阶段框架生成复杂问题，(ii) 收集与查询相关的多样化基础片段，以及 (iii) 人工整理的长篇答案创作。

步骤二：搜索网页。 使用在前一步生成的查询，我们对每一个查询进行网页搜索。这个检索过程系统地收集了多样的信息片段，创建了一个全面的知识库，作为生成问题的基础。

步骤三：生成问题。 通过将不同类型的示例问题 (Yang et al., 2024b) 纳入上下文学习提示和从前一步收集的信息，系统学会了无缝整合来自多个来源的信息，同时保持连贯性和相关性，以生成多样化的问题。这一步聚焦于创建需要跨不同片段综合信息的问题，以促进更具挑战性和现实的信息检索场景。生成问题的提示显示在附录 B.6 中。

步骤 4：过滤。 第四阶段实施了一个严格的过滤框架，以确保生成问题的质量和实用性。这包括一个用于识别格式良好的查询的分类器²，过滤掉没有命名实体的问题³，以及使用 SentenceTransformer 进行的去重⁴。

1.2 基准集合

对于每个生成的问题，我们从多个来源收集基础信息。特别是，我们使用独立的网络搜索引擎⁵或结合网络结果和从私有知识库检索的信息的混合方法。我们在这项工作中包含的私有知识库包含恩荣员工电子邮件 (Klimt and Yang, 2004)、Arxiv 摘要、AWS DevOps 故障排除指南和证券交易委员会文件 (Guinet et al., 2024)。检索流程主要分为两个步骤。首先，查询分解将复杂的查询分解为重点子查询，从而能够有针对性地检索相关信息。这种分解是通

过使用一个精心设计的提示来鼓励详细解释的 LLM 来进行的，以在生成分解之前进行详细解释。为了保持质量和相关性，我们采用基于 SentenceTransformer (Reimers and Gurevych, 2019) 的语义文本相似性 (STS) 分类器来过滤掉偏离原始问题意图的子查询。

第二个步骤涉及文档检索和重新排序。使用分解后的查询，我们从选定的来源中检索初始文档集。然后使用一个交叉编码器 (Chen et al., 2024a)，对原始问题和每个检索到的片段之间的相关性进行评分以对这些文档进行重新排序。最终的基础由通过交叉编码器评分排名前 K 的文档组成，确保在基准评估中优先考虑最相关的信息。最后，我们加入一个不进行重新排序的子集，在其中随机选择检索到的片段，以包括具有较高噪声水平的困难示例。

我们对数据集中的每个样本进行人工注释以确保质量。专业注释员对四个维度中的 2,366 个问题进行了注释，根据其与问题的相关性对每个基础段落进行了分类，并撰写了详细而全面且有依据的答案。

问题注释。 我们首先过滤掉那些无效、有错误前提或不是信息求取的问题。一个问题被认为是有效的，如果它是可理解的、可以回答的，并且不具伤害性。然后，我们会从四个维度对每个问题进行注释：时间敏感性、复杂性、受欢迎程度和类别。图 3 显示了问题注释的分布。

关联的基础 每个依据包含一段文字、来源 (网络或私人知识库)，以及 (如果可用) 其时间。⁶ 总共从私人知识库中检索了 4,752 段落，从网络中检索了 30,599 段落，并进

²https://huggingface.co/Ashishkr/query_wellformedness_score

³我们使用 SPACY 库。

⁴all-MiniLM-L6-v2

⁵为了保持匿名性，我们不公开专有搜索引擎的名称。

⁶接地的年龄是问题日期和接地日期之间的差异。

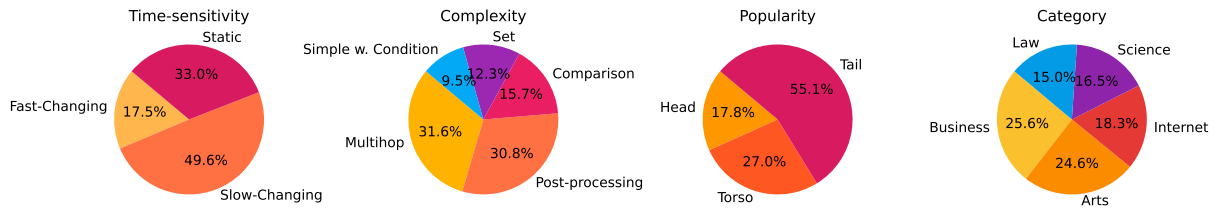


Figure 3: 问题统计。我们报告了关于时间敏感性、复杂性、受欢迎程度和类别维度的各种问题类型的百分比。我们的数据集总共包含 2366 个例子。

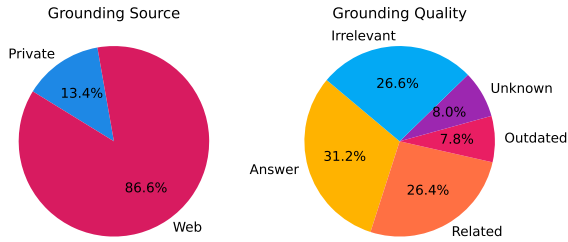


Figure 4: 落地统计。我们报告落地来源和落地质量的百分比。

行了标注。对于每个段落，标注者首先确定其是否与问题主题相关。如果相关，则将其标记为：ANSWER-THE-QUESTION、RELATED-INFORMATION、OUTDATED 或 UNKNOWN。图 4 显示了依据来源和相关性的分布。

答案注释。 标注者仅使用标记为相关的基础段落中的信息撰写长篇回答。⁷ 我们的指导方针为标注者提供了一定的自由，可以根据问题的时效性包括 OUTDATED 或 UNKNOWN 信息，只要它们被标注为相关。标注者从问题中提取论点，并添加引文标记以在支持性证据中定位每个回答论点。⁸ 对于缺乏足够信息以生成回答的问题，标注者撰写规避回答，例如“没有足够的基础来回答这个问题”。由于基础的不相关，427 个问题被标记为需要规避回答。这种模式在快速变化的问题上更为常见，因为此类问题的基础信息更可能已过时，此外也包括需要从私人来源获取信息的问题——由于主题的特性增加，导致检索设置更为困难。最后，为了确保基准中回答的高质量，我们对样本 (N=300) 进行了独立的人工标注流程，发现每个回答至少在一个基础段落中找到了依据。结果表明，基准确实以所提供信息为基础（在总计 2340 个论点中，97 % 至少基于 1 个信息块）。

⁷ 平均答案长度为 192 个词 ($\sigma = 73.5$)，不包括偏差。

⁸ 平均每个答案有 5 个独特的引用来源。

2 评估框架

得益于 GaRAGe 在问题、基础和答案层次上的精细标注，我们有机会评估模型性能的多个维度。为此，我们遵循 Jacovi et al. (2024) 并报告响应的适用性和真实性。我们在真实性指标的基础上，通过定义反映与相关段落的真实性的指标进行构建。然后，当没有提供相关基础时，我们评估转移行为与幻觉行为的对比，并评估生成最终响应时对每个提供的相关段落的出处能力。除非另有说明，我们使用 GPT-4o⁹ 作为评判标准来计算这些指标。

资格分数 衡量模型是否充分解决了用户请求，以人工编写的答案作为基准响应。LLM 评审提供的标签包括：无问题、次要问题或重大问题，合格分数计算为没有重大问题的数据点的百分比。用于此指标的提示在附录 B.2 中报告。

未经调整的事实性分数 通过将模型的回答分成句子来衡量答案中完全由提供的依据支持的百分比，然后判断每个句子是支持的、不支持的、矛盾的，还是不需要事实依据的。最终，只有当每个句子都有支持或不需要归因时，模型的答案才是事实性的。用于此指标的提示在附录 B.3 中。由于 GaRAGe 提供了每个依据片段的相关性注释，我们通过计算模型仅在作为上下文提供的相关段落中定位其答案的能力来优化该指标，以下称为未经调整的相关性认知事实性分数 (uRAF)。请注意，这是一种更受控的指标，不仅测量回答的事实性，还测量用以生成最终回答的信息的相关性。

事实性评分 结合了前两个指标，衡量答案中既合格又得到提供的依据支持的百分比。我们同样通过结合 uRAF 与 Eligibility 开发了相关性感知事实性评分 (RAF)，以计算既合格又得到相关段落支持的答案百分比。

偏转得分 衡量基准测试中期待偏转的子集中的偏转答案的百分比，即真正的正例，由于缺乏充足的基础信息。此外，我们报告在数据集

⁹ gpt-4o-2024-11-20 的温度为 0.2。

Model	Eligibility	Unadjusted Factuality	Factuality	uRAF	RAF
API-based models					
Claude Haiku	79.37	51.83	48.37	40.00	36.90
Claude Sonnet	86.07	68.48	64.67	51.75	48.91
Gemini 1.5	84.88	81.07	70.50	67.78	59.43
Gpt-4o	92.47	62.91	59.30	56.18	52.88
Nova Micro	90.97	48.83	45.02	39.94	37.16
Nova Lite	80.15	61.06	49.25	55.67	45.97
Nova Pro	87.77	74.47	66.63	68.29	60.67
Open-weights models					
Mistral	85.30	48.01	43.32	38.14	34.32
Mixtral	82.72	47.08	42.96	37.31	34.12
Qwen 14b	85.80	68.20	59.80	59.20	52.70
Qwen 32b	90.50	66.10	61.00	57.70	52.90

Table 2: 使用 LLM-as-a-judge 指标测量不同开放和封闭模型在 GaRAGe 上的性能。未调整的关联感知事实性分数 (uRAF) 衡量相关信息中响应的事实性，而关联感知事实性分数 (RAF) 则衡量既符合条件又被相关基础支持的答案的百分比。

中剩余部分中存在的错误的偏转答案部分，这些部分有足够的相关信息来提供非偏转的响应。用于此指标的提示在附录 B.4 中报告。

归因分数 衡量模型产生正确引用信息来源段落形式的能力。在此指标中，我们利用人类撰写的答案中的引注，并计算模型回答中引用的出现与人类撰写的回答中的比较的精确率、召回率和 F_1 -评分。

3 结果

我们在 GaRAGe 基准测试中对不同大小的强大专有模型和领先的开放权重模型进行评估。特别地，我们评估了 GPT-4o¹⁰ (Hurst et al., 2024)、Gemini 1.5 Flash (Team et al., 2023)、Claude 3.5 Haiku¹¹ 和 Sonnet¹² (Anthropic, 2024)、亚马逊 Nova Micro、Lite 和 Pro (Intelligence, 2024)，以及开放权重的 Mistral、Mixtral (Jiang et al., 2024)、Qwen2.5 14b 和 32b 模型 (Yang et al., 2024a)。我们使用贪婪解码和附录中报告的相同输入提示进行比较 B.1。

3.1 主要结果

在表 2 中，我们报告了答案合格性和事实性的主要结果。首先，GPT-4o 获得了最高的合格性得分。令人惊讶的是，模型的大小对这一指标没有明显影响：Nova Micro 和 Mistral 都胜过了同一系列中的更大模型，而 Claude 3.5 Haiku 落后于 Sonnet。此外，Gemini 1.5 Flash 在事实性得分中取得了最高分，其次是 Nova Pro，然后是 Claude 3.5 Sonnet 和 GPT-4o。模型的大小是事实性上的一种能力指标，表明更大的模型持续更好地使用提供的信息作为背景。GPT-4o

可能正在从参数知识中生成信息，虽然合格性得分报告中显示这种信息是有用的，但它超出了地基中提供的相关信息，这可能表明有出现幻觉的风险。通过结合合格性和未调整的事实性得分，事实性得分倾向于选择充分利用提供地基的合格答案。在这个指标中，Gemini 1.5 Flash 领先，紧随其后的是 Nova Pro 和 Claude 3.5 Sonnet。

有趣的是，当限制基础评分仅限于相关基础时，我们注意到大多数模型的表现整体下降。这表明大多数模型在生成响应时利用了不相关或过时的基础，即，扮演了汇总者的角色，而不是仅识别新鲜且相关的信息以纳入最终响应。在 RAF 评分方面领先的模型是 Nova Pro 和 Gemini 1.5 Flash，它们在合格且相关信息为基础响应中均达到 $\sim 60\%$ 。

在表 ?? 中，我们报告了在偏转子集上作为真正阳性的平均偏转回答，以及在信息充足以提供非偏转回答的子集上的假阳性。首先，所有基线模型的假阳性率都很低（小于 3.5% ），这意味着当在基础段中提供了足够的相关信息时，模型能够正确地提供非偏转回答。另一方面，模型在识别生成良好答案所需的基础信息不足方面存在困难。虽然较大的模型整体表现出更好的偏转行为，但通过真正阳性率测量的偏转覆盖率相当低，表现最好的模型是具有 31.1% 的 GPT-4o。这是所有模型需要改进的重要差距，以避免幻觉和误导 RAG 用户。

在表 3 中，我们报告了归因分数的精确率、召回率和 F_1 分数。结果显示，较大的模型在归因上表现出更为保守的方法，召回率较低，但通过较高的精确率加以补偿。这表明较小的模型可能会过度引用提供的来源，而不是仅专注于相关的依据。总体而言，我们观察到大多数模型的表现相似， F_1 分数在 $50\text{--}60\%$ 之间，

¹⁰gpt-4o-2024-11-20

¹¹anthropic.claude-3-5-haiku-20241022-v1:0

¹²anthropic.claude-3-5-sonnet-20241022-v2:0

Model	Precision	Recall	F ₁
Claude Haiku	49.9	71.9	58.9
Claude Sonnet	51.8	67.5	58.6
Gemini 1.5	54.7	56.3	55.5
GPT-4o	57.9	59.0	58.4
Nova Micro	48.2	75.8	58.9
Nova Lite	50.0	46.4	48.1
Nova Pro	56.9	49.6	53.0
Mistral	44.4	60.0	51.0
Mixtral	46.3	56.3	51.0
Qwen 14b	55.2	43.6	48.7
Qwen 32b	55.2	47.0	50.8

Table 3: 模型的性能通过 GaRAGe 中的归因精度、召回率和 F₁ 进行测量。

表明模型在将生成的响应归因到相应的依据信息上存在差距。

3.2 分析

时间敏感性。 我们分析了基准中快变问题的性能，并假设 (1) 基础段落可能含有更多噪声并且过时，(2) 大型语言模型 (LLMs) 在推理关于问题的相对近期性和基础信息时存在困难。在图 5 的第一个图中，我们报告了大模型的 RAF 分数。结果显示了与主要结果类似的趋势 (§3.1)，其中 Nova Pro 取得了 51.7 % 的分数，其次是 Gemini (49.8 %)，Sonnet (46.7 %)，GPT-4o (45.7 %) 和 Qwen 32b (43.8 %)。有趣的是，对于这一组问题的总体表现较低 (比缓变或静态问题低 ~ 10 %)，这表明 LLMs 在推理时间相关事实时更加困难。

私人问题表现。 我们进一步分析了构建于特定领域的问题的基础质量，这些领域模拟了私人知识库，比较了反映一般网络搜索的问题。与网络问题相比，领域问题的相关基础较少，这可能使得它们对大型语言模型提取仅相关信息的挑战更大。例如，涵盖安然公司话题的问题平均有 47.8 % 的相关基础，而反映网络使用的更广泛话题的问题则平均有 85.6 % 的相关段落。

图 5 中的结果显示，即便是在评估的最强劲的 LLM 之间，其性能也显著下降，对于涉及更具体主题的问题，性能始终下降超过 10 %。这证实了我们的直觉，并进一步表明 LLM 难以仅根据相关信息创建有根据的答案。

基础质量。 我们研究了基础相关性如何影响答案质量，以了解模型对无关信息的稳健性。我们根据提供的相关段落的百分比 ($pass_r$) 将基准测试分成三个不同的组：低 ($pass_r < 33\%$)、中 ($33\% \leq pass_r < 66\%$) 和高 ($pass_r >$

66%)。

图 6 展示了我们在基准测试中报告的较大模型的 RAF 得分。我们观察到一个明显的趋势：随着基础质量的下降，模型的表现显著下降，从高到中下降了 10 %，从中到低下降了 20 %。这个模式突显了在 RAG 设置中，模型在构建答案时无法区分相关和无关信息的能力不足。

问题受欢迎程度。 当按照问题的受欢迎程度分析模型性能时，我们发现对尾部问题的 RAF 分数在模型平均时为 44.2 %。这显著低于躯干和头部问题，其平均 RAF 得分分别为 49.7 % 和 50.3 %。这些结果与之前分析中呈现的模式相似，其中尾部问题通常由于相关信息的稀缺性和潜在的噪声基础，使得其更具挑战性。

在过去的几年里，由于问题回答任务的影响以及解决此问题所需的复杂多样的推理，发布了大量令人印象深刻的基准。在阅读理解方面，从用于考察提取能力的基准如：Natural Questions (Kwiatkowski et al., 2019)、WikiQA (Yang et al., 2015) 或 TriviaQA (Joshi et al., 2017)，到易于自定义为新数据和模型的框架如：Ragas (Es et al., 2024)、Ragnarok (Pradeep et al., 2024) 或 Ares (Saad-Falcon et al., 2024)。

最近，为了挑战 LLM 在 RAG 上的能力，一个演变的维度是增加所提问题的复杂性。这是通过要求在多个文档上进行更复杂的推理来实现的，或者通过理解额外的语言外上下文信息，例如时间信息，以构建正确的答案。Gabburo et al. (2024) 提出了一种基于检索证据测量问题复杂性的度量，该度量捕捉了从多个基础来源找到准确答案的难度。

另一个对大型语言模型来说可能难以处理的组件是提供给模型的基础的多样性。这包括具有不同噪声水平的基础 (Chen et al., 2024b; Liu et al., 2023)，特定领域 (Guinet et al., 2024; Han et al., 2023)，或结合来自私人 and 公共来源的基础 (Arora et al., 2023)。

回答的全面性。 大多数 RAG 基准测试集中在非常简短的答案上。这种模式的形成既是由于创建长篇、全面且有依据的答案的复杂性，也是由于评估它们的难度。同时，LLMs 通常默认的行为是回答时较为冗长，因此在模型行为和 RAG 应用场景的评估框架之间存在不匹配。这种情况最近通过评估长篇问答 (Q & A) 的努力得到缓解，例如使用以下基准测试：Rag-QA Arena (Han et al., 2024)、Facts (Jacovi et al., 2024)、LongFact (Wei et al., 2024) 或 ELI5 (Fan et al., 2019)。

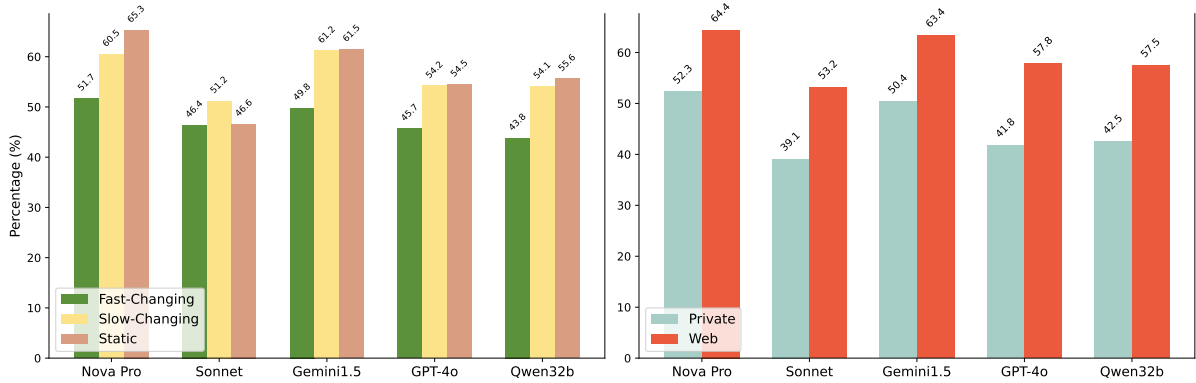


Figure 5: 不同模型的性能通过 RAF 得分来衡量：1) 在左侧是不同时间性问题的动态变化；2) 在右侧是私有问题与 Web 问题的比较。

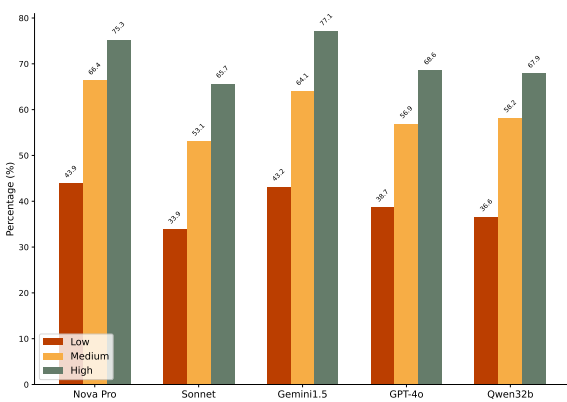


Figure 6: 不同模型的性能通过 RAF 评分相对于基础质量进行衡量。

4 结论

我们引入了 GaRAGe，这是一个大型 RAG 基准测试，具有人工编制的长格式答案和单个基础段落的全面注释。由于数据集的详细注释，我们能够开发出一套全面的实验和一项新的指标：参考感知事实性 (RAF) 得分。该框架帮助揭示了开放和封闭的大型语言模型在处理嘈杂基础时的多个差距。此外，我们对不同数据集维度上模型的性能进行了深入分析，加深了我们对模型局限性的理解。

我们承认该基准仅包含英文数据点。我们计划在未来探索多语言环境。

该数据集可能会被包含在未来的大型语言模型训练中，这在一定程度上会使基准测试的评估失去效力。即使未来可能会包含这些数据 (或者部分数据已经包含在内)，但其设计旨在衡量模型遵循提供的依据的能力，这确保了该基准测试的有效性。

注释的一些组件有些主观，例如主题的受欢迎程度可能在不同的标注者之间有所不同。虽然这可能会产生一些偏离对齐的标签，但我们认为数据集的显著规模可以缓解这些异常值。

由于基准测试的很大一部分反映了现实中的互联网问题，它们可能涉及个人。为了确保不存在冒犯性内容，问题标注的首要过滤条件是，问题不应要求敏感信息，也不应包含冒犯性、侮辱性或有害的语言。它不应基于性别或种族进行歧视。

当我们多次运行大型语言模型 (LLM) 时，它们可能会表现出不同的行为。由于模型推理的运行成本，我们报告一个语句的性能，同时我们也承认多次运行可能会出现一些波动。

我们仅使用 GPT-4o 作为评判，因此结果可能在评估 (Liu et al., 2024) 时对 GPT 模型有偏向。这可以通过运行多个 LLM 评判并汇总他们的评估以获得更稳健的报告，或者通过人工评估来缓解。

5

伦理声明 我们使用公开可用的新闻数据，这些数据可能包含来自不同视角的观点以及之前在出版物中使用过的私有知识库。论文中的输出结果不代表作者的观点。

References

- AI Anthropic. 2024. Introducing the next generation of claude.
- Simran Arora, Patrick Lewis, Angela Fan, Jacob Kahn, and Christopher Ré. 2023. Reasoning over public and private data in retrieval-based systems. *Transactions of the Association for Computational Linguistics*, 11:902–921.
- Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. [Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). *Preprint*, arXiv:2402.03216.
- Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024b. Benchmarking large language models in retrieval-augmented generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17754–17762.
- Shahul Es, Jithin James, Luis Espinosa Anke, and Steven Schockaert. 2024. Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 150–158.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. Eli5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567.
- Matteo Gabburo, Nicolaas Paul Jedema, Siddhant Garg, Leonardo F. R. Ribeiro, and Alessandro Moschitti. 2024. [Measuring retrieval complexity in question answering systems](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14636–14650, Bangkok, Thailand. Association for Computational Linguistics.
- Gauthier Guinet, Behrooz Omidvar-Tehrani, Anoop Deoras, and Laurent Callot. 2024. Automated evaluation of retrieval-augmented language models with task-specific exam generation. *arXiv preprint arXiv:2405.13622*.
- Rujun Han, Peng Qi, Yuhao Zhang, Lan Liu, Juliette Burger, William Yang Wang, Zhiheng Huang, Bing Xiang, and Dan Roth. 2023. Robustqa: Benchmarking the robustness of domain adaptation for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4294–4311.
- Rujun Han, Yuhao Zhang, Peng Qi, Yumo Xu, Jinyuan Wang, Lan Liu, William Yang Wang, Bonan Min, and Vittorio Castelli. 2024. Rag-qa arena: Evaluating domain robustness for long-form retrieval augmented question answering. *arXiv preprint arXiv:2407.13998*.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Amazon Artificial General Intelligence. 2024. The amazon nova family of models: Technical report and model card.
- Alon Jacovi, Andrew Wang, Chris Alberti, Connie Tao, Jon Lipovetz, Kate Olszewska, Lukas Haas, Michelle Liu, Nate Keating, Adam Bloniarz, Carl Saroufim, Corey Fry, Dror Marcus, Doron Kukliansky, Gaurav Singh Tomar, James Swirhun, Jinwei Xing, Lily Wang, Madhu Gurumurthy, Michael Aaron, Moran Ambar, Rachana Fellingner, Rui Wang, Zizhao Zhang, Sasha Goldshtein, and Dipanjan Das. 2024. FACTS grounding: A new benchmark for evaluating the factuality of large language models.
- Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.
- Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.
- Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. In *CEAS*, volume 4, page 1.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.
- Philippe Laban, Alexander Fabbri, Caiming Xiong, and Chien-Sheng Wu. 2024. [Summary of a haystack: A challenge to long-context LLMs and RAG systems](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9885–9903, Miami, Florida, USA. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Yi Liu, Lianzhe Huang, Shicheng Li, Sishuo Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. 2023. Recall: A benchmark for llms robustness against external counterfactual knowledge. *arXiv preprint arXiv:2311.08147*.

- Yiqi Liu, Nafise Moosavi, and Chenghua Lin. 2024. Lms as narcissistic evaluators: When ego inflates evaluation scores. In *The 62nd Annual Meeting of the Association for Computational Linguistics*.
- Haoran Luo, Haihong E, Yikai Guo, Qika Lin, Xiaobao Wu, Xinyu Mu, Wenhao Liu, Meina Song, Yifan Zhu, and Luu Anh Tuan. 2025. [Kbqa-o1: Agentic knowledge base question answering with monte carlo tree search](#). *Preprint*, arXiv:2501.18922.
- Philhoon Oh and James Thorne. 2023. [Detrimental contexts in open-domain question answering](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11589–11605, Singapore. Association for Computational Linguistics.
- Ronak Pradeep, Nandan Thakur, Sahel Sharifymoghadam, Eric Zhang, Ryan Nguyen, Daniel Campos, Nick Craswell, and Jimmy Lin. 2024. Ragnar\": ok: A reusable rag framework and baselines for trec 2024 retrieval-augmented generation track. *arXiv preprint arXiv:2406.16828*.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2024. Ares: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 338–354.
- Yixuan Tang and Yi Yang. 2024. [Multihop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries](#). In *First Conference on Language Modeling*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.
- Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, RuiBo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.
- Xiao Yang, Kai Sun, Hao Xin, Yushi Sun, Nikita Bhalla, Xiangsen Chen, Sajal Choudhary, Rongze Daniel Gui, Ziran Will Jiang, Ziyu Jiang, et al. 2024b. Crag—comprehensive rag benchmark. *arXiv preprint arXiv:2406.04744*.
- Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

A 注释定义

我们在下面给出了所有步骤注释指南中使用的定义。

A.1 问题维度

A.1.1 问题过滤器

问题有效性。如果一个问题可以通过在网络上寻找信息，或者使用其他信息来源或常识来回答，则该问题是有效的。此外，它不应请求敏感信息，也不应包含攻击性、侮辱性或有害的语言。它不应基于性别或种族实行歧视。

错误前提。错误前提问题是指那些包含错误假设事实的问题，这些事实并未直接陈述，但很可能被提问者相信。

信息寻求。信息寻求问题是指为了获取关于某个特定主题的知识、清晰性或见解，寻求事实性信息或详细解释而提出的问题。

A.1.2 时间动态性

时间敏感问题。时间敏感问题是指其答案会随时间变化的问题，并且需要实时内容来回答。如果您不熟悉该主题，您应通过搜索互联网来决定它是否具有时间敏感性。

静态问题。静态问题是指答案不会随时间变化的问题，或不需要实时内容来回答的问题。

慢变问题。对于回答此类问题，信息的时效性超过一周是可以接受的。这包括超过一个月、超过一年等等。注意，如果有关于同一问题的热门新闻，则信息应视为不到一周的时效性。

快速变化的问题。所需信息需要在七天内更新才能回答该问题。

A.1.3 问题复杂性

简单。询问简单事实的问题

简单的条件问题。询问具有某些给定条件的简单事实的问题，例如某个日期的股票价格和某导演在某个类型中的近期电影。

集合。期望以一个实体或对象集合作为答案的问题。

比较。比较两个实体的问题。

聚合。需要聚合检索结果来回答的问题。

多跳。需要将多条信息连接起来以形成答案的问题。

后处理繁重。需要对检索到的信息进行推理或处理以获得答案的问题。

A.1.4 问题受欢迎程度

头条。涵盖众所周知、经常讨论的主题，这些主题受到大量媒体报道。它类似于高频搜索词或热门话题。

躯干。涵盖了适度受欢迎的话题，但不是主流。它类似于中等频率的搜索词。

尾部。涉及小众或专业课题，并涵盖高度技术性 or 专业化领域。

A.1.5 问题类别

我们涵盖以下问题类别：艺术 & 娱乐、计算机 & 电子、健康、工作 & 教育、家居 & 园艺、法律 & 政府、旅行、科学、商业 & 工业、爱好 & 休闲、书籍 & 文学、体育、新闻、美容 & 健身、金融、人物 & 社会、汽车 & 车辆、游戏、时间 & 天气、在线社区、互联网 & 电信、本地信息、宠物 & 动物、股票及其他。

A.2 基础标注

无关。这段文字不包含与问题相关的信息，因此无法用于生成答案。该段内容未能回答具体问题，缺乏能够帮助形成全面回应的相关细节或见解。

相关。该片段与当前查询或话题直接相关。它包含详细和重要的信息，可以对答案的创建作出显著贡献。这意味着片段的内容应该直接回答查询，提供相关的见解、数据或解释，从而增强模型 XMathXSQ 生成准确和全面响应的能力。

回答问题。片段中的信息是最新的，并且正确回答了问题。确保您检查问题提出的日期和文档的日期（证据日期）。当不同片段中的信息相互矛盾时，执行谷歌搜索以确定正确答案。

相关信息。该片段本身并不能直接回答问题。然而，它包含与查询相关的信息，可以用来扩展详细的回答。内容提供了细节和背景，可以整合到一个更全面和彻底的答案中。

过时。这段文字包含的信息已经过时，不再当前。这意味着所呈现的数据或细节来自以前的时间段，不反映与主题相关的最新进展或发现。因此，这些信息在当前情境中可能不准确或不相关，如果依赖可能导致误解或错误结论。

未知。源年龄未知，无法从片段中推断年龄。

A.3 答案质量

正确性。一个正确的答案应该是内容丰富的，并且能够为用户提供有效的信息，直接回答所提的问题。

自然性。答案应该是自然的，即根据标准英语语法/讲话规则是构造良好的、流利的、语法正确的。回应应该接近人类对问题的回答方式。

有用性。答案应具有实用性和可操作性，使客户在阅读回复后能够做出明智的决定。

客观性。一个答案应当是客观的，不带有偏见或个人意见，并且基于可以用证据支持的可验证事实。

有用性。答案应该有帮助，诚实，并且不会对读者/人/客户造成伤害。

B 提示

B.1 大型语言模型提示

Example Inference Prompt for Baseline Models

A chat between a curious User and an artificial intelligence Bot. The Bot gives helpful, detailed, and polite answers to the User's questions.
In this session, the model has access to search results and a user's question, your job is to answer the user's question using only information from the search results.
Model Instructions:

- 当答案直接包含在搜索结果中时，你应该对简单问题提供简明的回答，但对于是/否问题，则应提供一些细节。
- 如果问题需要多跳推理，你应从搜索结果中找到相关信息，并基于相关信息通过逻辑推理来总结答案。
- 要忠实于搜索段落中提供的信息，如果查询无法通过搜索结果中的信息来回答，请指出“模型找不到足够的信息来回答问题”。
- 将引用标记为内嵌标记（例如，% [1] %，% [2] % 等），直接放在响应文本的相关部分中。不要在响应后的单独部分中包含引用。
- 不要使用搜索结果中没有的信息！

Context: The current time is { query_time } .
User: { query } Bot:
Resource: Search Results: { search_results } Bot:

B.2 事实资格提示

Example Eligibility Prompt for the LLM judge

Your mission is to judge the response from an AI model, the **test** response, calibrating your judgement using a **baseline** response. Please use the following rubric criteria to judge the responses:

<START OF RUBRICS>

Your task is to analyze the test response based on the criterion of "Instruction Following". Start your analysis with "Analysis".

****Instruction Following****

Please first list the instructions in the user query.

In general, an instruction is VERY important if it is specifically asked for in the prompt and deviates from the norm. Please highlight such specific keywords.

You should also derive the task type from the user query and include the task-specific implied instructions.

Sometimes, no instruction is available in the user query.

It is your job to infer if the instruction is to autocomplete the user query or is asking the LLM for follow-ups.

After listing the instructions, you should rank them in order of importance.

After that, INDEPENDENTLY check if the test response and the baseline response meet each of the instructions.

You should itemize, for each instruction, whether the response meets, partially meets, or does not meet the requirement, using reasoning.

You should start reasoning first before reaching a conclusion about whether the response satisfies the requirement.

Citing examples while reasoning is preferred.

Reflect on your answer and consider the possibility that you are wrong.

If you are wrong, explain clearly what needs to be clarified, improved, or changed in the rubric criteria and guidelines.

In the end, express your final verdict as one of the following three json objects:

```
{ { "Instruction Following": "No Issues" } }
```

```
{ { "Instruction Following": "Minor Issue(s)" } }
```

```
{ { "Instruction Following": "Major Issue(s)" } }
```

<END OF RUBRICS>

Your task

User query

< |begin_of_query| >

{ query }

< |end_of_query| >

Test Response:

< |begin_of_test_response| >

{ answer }

< |end_of_test_response| >

Baseline Response:

< |begin_of_baseline_response| >

{ expected_answer }

< |end_of_baseline_response| >

Please write your analysis and final verdict for the test response.

B.3 事实真实度提示

Example Factuality Prompt for the LLM judge

You are a helpful and harmless AI assistant. You will be provided with a textual context and a model-generated response. Your task is to analyze the response sentence by sentence and classify each sentence according to its relationship with the provided context.

Instructions:

- Decompose the response into individual sentences.
- For each sentence, assign one of the following labels:
 - 'supported'**: The sentence is entailed by the given context. Provide a supporting excerpt from the context. The supporting excerpt must *fully* entail the sentence. If you need to cite multiple supporting excerpts, simply concatenate them.
 - 'unsupported'**: The sentence is not entailed by the given context. No excerpt is needed for this label.
 - 'contradictory'**: The sentence is falsified by the given context. Provide a contradicting excerpt from the context.
 - 'no_rad'**: The sentence does not require factual attribution (e.g., opinions, greetings, questions, disclaimers). No excerpt is needed for this label.
- For each label, provide a short rationale explaining your decision. The rationale should be separate from the excerpt.
- Be very strict with your 'supported' and 'contradictory' decisions. Unless you can find straightforward, indisputable evidence excerpts in the context that a sentence is 'supported' or 'contradictory', consider it 'unsupported'. You should not employ world knowledge unless it is truly trivial.

Input Format:

The input will consist of two parts, clearly separated:

Context: The textual context used to generate the response.

Response: The model-generated response to be analyzed.

Output Format:

Your output should be in json format as follows:

The key should be "grounding_quality" and the value should be a list of json objects with an object for each sentence in the response, containing the following fields:

- "sentence"**: The sentence being analyzed.
- "label"**: One of 'supported', 'unsupported', 'contradictory', or 'no_rad'.
- "rationale"**: A brief explanation for the assigned label.
- "excerpt"**: A relevant excerpt from the context. Only required for 'supported' and 'contradictory' labels.

Example:

Input:

Context: Apples are red fruits. Bananas are yellow fruits.

Response: Apples are red. Bananas are green. Bananas are cheaper than apples. Enjoy your fruit!

Output:

```
{ "grounding_quality":
  [ { "sentence": "Apples are red.", "label": "supported", "rationale": "The context explicitly states that apples are red.", "excerpt": "Apples are red fruits." },
    { "sentence": "Bananas are green.", "label": "contradictory", "rationale": "The context states that bananas are yellow, not green.", "excerpt": "Bananas are yellow fruits." },
    { "sentence": "Bananas are cheaper than apples.", "label": "unsupported", "rationale": "The context does not mention the price of bananas or apples.", "excerpt": null },
    { "sentence": "Enjoy your fruit!", "label": "no_rad", "rationale": "This is a general expression and does not require factual attribution.", "excerpt": null } ] }
```

Now, please analyze the following context and response:

User Query: { query }

Context: { context_document }

Response: { answer }

B.4 偏转判断提示

Example Deflection Prompt for the LLM judge

You are a human expert in analyzing responses to questions given by a model.

Your task is to assign the response into exactly one of the following categories:

- missing**: the response indicates that it is not sure, does not know the answer, or states that it cannot provide an answer.
- attempted**: the response attempts to answer the question.

QUESTION: { query }

MODEL RESPONSE: { answer }

CATEGORY:

Your response should be in json format as follows:

```
{
  "justification": (Explain why the MODEL RESPONSE is missing or attempted. Use one or two sentences maximum. Keep the answer as concise as possible.)
  "grade": (missing or attempted),
}
```

先提示

B.5 搜索网络

生成研究（提示研究任务）任务，然后扩展为搜索查询（提示创建搜索查询）。

Research Tasks

You are an AI assistant tasked with breaking down a given question into a list of research tasks. Your goal is to help organize the research approach and ensure all necessary information is gathered to answer the question effectively.

Here's the question you need to analyze:

<question>

Create complex questions that require reasoning about { seed_topic }. Search the Web for recent news and documents about seed_topic to be used as input for this creation. Those queries you will generate must require an answer that needs reasoning over evidence from multiple documents.

</question>

Carefully analyze the question to identify the key components that require research. Consider the following:

1. The main topic or subject of the question
2. Any specific details, dates, or names mentioned
3. The type of information needed (e.g., factual, comparative, historical)
4. Any implicit sub-questions within the main question

Based on your analysis, create a list of research tasks. Each task should:

1. Be clear, specific, and concise
2. Focus on one aspect of the research
3. Be actionable and help in gathering relevant information
4. Be ordered logically, if there's a natural sequence to the research
5. Not use pronouns to refer to something in the question. Use its actual name.

The number of tasks can range from 1 to 3, depending on the complexity of the question. Simpler questions may only require one or two tasks, while more complex ones might need up to three.

Present your list of research tasks in the following format:

<research_tasks>

1. [First research task]
2. [Second research task]
3. [Third research task]

... </research_tasks>

If a subsequent task specifically refers to an entity or other information that a preceding task can provide, mark that entity or item with a note in parentheses that indicates the task number that provides it. Format: (from task N)

Make sure to place the parenthesis inline in the text after the entity. Example:

1. Identify the name of ...
2. Find the ... of that person (from task 1).

It is important to use the correct format. If entity X was identified in task 1 and used in task 2, say X (from task 1).

Creating Search Queries

You are an AI research assistant tasked with understanding and researching a given task. Your goal is to analyze the task and generate appropriate Web search queries for further research.

Here is the task you need to research:

<task>

{ instruction }

</task>

Follow these steps to complete your assignment:

1. Carefully read and analyze the given task and the results of related tasks.
2. Think about the key aspects of the task that require further research. Consider what information you need to gather to fully address the task.
3. Based on your analysis, generate three Web search queries that will help you research the task effectively. Each query should focus on a different aspect of the task or seek different types of information.
4. Today's date is date . You may use the current year in your Web search queries where appropriate.
5. Format your output as follows: a. First, provide a brief explanation of your understanding of the task and how you've incorporated information from related tasks. Write this explanation inside <task_analysis> tags. b. Then, list your three Web search queries, each wrapped in <QUERY> tags.

Here's an example of how your output should be structured:

<task_analysis> [Your analysis of the task] </task_analysis>

<QUERY>[Your first Web search query]</QUERY>

<QUERY>[Your second Web search query]</QUERY>

<QUERY>[Your third Web search query]</QUERY>

Remember, your queries should be specific and targeted to gather the most relevant information for the task at hand. Avoid overly broad or vague queries.

B.6 生成问题

Generating Complex Questions

You are tasked generating questions, based on the results to several sub-questions. Your goal is to create detailed and focused queries that will require reasoning to be answered, addressing the main instruction while incorporating key information from each sub-question.

Here is the main instruction you need to address: <main_instruction>

Create complex questions that require reasoning about { seed_topic }. Search the Web for recent news and documents about seed_topic to be used as input for this creation. Those queries you will generate must require an answer that needs reasoning over evidence from multiple documents.

</main_instruction>

Below are the sub-questions and their corresponding answers:

<sub_questions_answers>

{ % for task, result in prev_results.items() % }

<task>

- Task { task } : { result }

</task>

{ % endfor % }

</sub_questions_answers>

To synthesize your response, follow these steps:

1. Look for common themes or connections between the sub-answers.

2. Based on the high level topics, generate complex questions which needs reasoning to obtain answers (e.g., "how many days did Thurgood Marshall serve as a Supreme Court justice?", "how many Oscar awards did Meryl Streep win?").

3. Avoid to create simple questions that can be answered by just composing and summarizing the answers.

4. The questions must require chaining multiple pieces of information to compose the answer (e.g., "who acted in Ang Lee's latest movie?")

5. Generate a list of elaborate questions using an unordered list.

6. The questions should NOT have more than 20 words, they must be succinct. Don't make the generated questions too long!

Example of generated elaborate questions:

what age did ferdinand magelan discovered the philippines?

how many grammy awards were won by the song plan b until 62nd grammy?

how many feet is the place with the lowest elevation in vermont?

when did meta announce the release of the meta quest 4?

what christmas song is the most streamed, all i want for christmas is you or jingle bells?

who portrayed the younger character, gillian lynne in phantom of the opera: behind the mask or brian blessed in prisoner of honor?

was the lion king the highest-grossing film of all time when it was released?

how many american players ranked are in the top ten of the wta?

the 1984 maze game devil world has an amazing soundtrack. who composed it?

which movie has a higher budget, poor things or bob marley one love?

how many films are in the hunger games film series?

what was the pricing of the most recent ipo for a major social media company?

which feid album is the one that has "normal" on it?

what was the total value of all mergers and acquisitions in the healthcare sector?

what are the top 3 movies on hbo max?

how many regional confederations does fédération internationale de volleyball have?

how many american music awards has taylor swift won throughout her career?

what is the age difference between angelina jolie and billy bob thornton?

how much is the worst performing stock, amazon?

how many albums has the band radiohead released that have reached number one on the uk albums chart?

who was the first actress to play the role of wonder woman in a live-action movie?

the three countries with the highest oil production are...?

which artist has been performing for longer, shakira or machine gun kelly?

how many times has meryl streep been nominated for an academy award for best actress?

which female tennis players have won all four grand slam tournaments in a single calendar year?

which five dow jones companies have a debt-to-equity ratio of less than 0.1?

give at least 3 etf funds with over 1 billion total of assets

which movie did jennifer lawrence co-star in with bradley cooper where she played a character with a mental health condition?

what are the five highest-grossing movies at the box office directed by nancy meyers?

C 示例

在下文中，我们展示了来自不同数据集特征的问题（例如，数据点需要偏转）或我们提供了来自数据集的详细实例，以及相关的注释和其中一个评估模型的回应。

C.1 私人（领域特定）问题

不同私人知识库所涉及问题的示例，与更一般的网络问题形成对比：

- DevOps：“分数阶导数如何增强海洋工程中的 NLS 模型？”
- SEC：“欧盟的 CE 认证与美国 FDA 批准在医疗器械和营养产品方面的监管要求有什么区别？”
- 安然公司：“PG & E 对于 DWR 的电力采购收入需求提出了哪些主要关切和反对意见？”
- Arxiv：“像峰值信噪比 (PSNR) 和结构相似性指数 (SSIM) 这样的性能指标如何评估图像去噪算法的有效性，它们有哪些局限性？”
- 网络：“考虑到 2025 年对基于体验的事件的关注，像 Maha Kumbh Mela 这样世界上最大的宗教集会在规模和重要性方面如何比较？”

C.2 不同复杂程度的问题

- 简单来说：亚马逊 Bedrock 的新缓存功能提供了哪些改进？
- 简单的条件：唐纳德·特朗普目前多大年龄，他担任了多久的美国总统？
- 比较：“蓝色起源”的新格伦号成功发射可能会如何影响 SpaceX 在私人航天领域的主导地位？
- 后处理繁重：加沙停火谈判对地区稳定有什么影响，BBC 新闻如何报道这些谈判的复杂性？
- 多跳：由于马丽娜·萨蒂与 ERT 关系紧张而缺席 2025 年欧洲歌唱大赛国家决赛，这对希腊在即将到来的比赛中的参与有什么潜在影响？
- 问题：为了确保远北地区互联网服务的可靠性和可负担性，根据 CRTC 的概述，可以实施哪些策略？
- 聚合：乐队 Radiohead 发行了多少张专辑在英国专辑榜上排名第一？

C.3 涵盖不同流程序度的问题

- 头条：拜登决定将近 2500 名因非暴力毒品犯罪而服长刑的囚犯的刑期减免，这对刑事司法系统和公众舆论产生了什么影响？
- 躯干：谷歌最新的算法更新如何体现公司对提高搜索质量和提供有价值内容的承诺？
- 尾部：鉴于竞争环境和最近的趋势，Natixis 资产管理公司与忠利保险之间的新资产管理合资企业将如何影响欧洲金融的格局？

C.4 具有不同时间动态程度的问题

- 常青：雅培实验室是如何在 2009 年为收购 Visiogen 公司的成本进行辩护的？
- 慢速变化：COVID-19 大流行如何影响 Matson 的运营和财务表现？
- 快速变化：哪两名四分卫在 NFL 季后赛中传球超过 100 码？

C.5 GaRAGe 详细示例

GaRAGe example 1

Question: What trends are shaping RSU compensation in 2024?

- 时间动态性：缓慢变化
- 复杂性：集合
- 领域：股票
- 流行度：尾部

Grounding:

- 片段 1:
 - 文本: RSU 作为竞争性薪酬方案的组成部分 RSU 通常与基本工资、奖金和其他福利一起作为竞争性薪酬方案的一部分被包含在内。RSU 可以提升薪酬方案的整体吸引力...
 - 年龄: 11 个月零 8 天
 - 标签: 无关
- Snippet 2:
 - 文本: 股票奖励计划, 如限制性股票单位 (RSUs)、员工股票期权计划 (ESOPs) 和员工股票购买计划 (ESPPs), 正日益增加.....
 - 年龄: 1 周零 2 天
 - 标签: 无关
- Snippet 10:
 - 文本: 随着人工智能 (AI) 不断改变商业的各个方面, 其与 RSU 薪酬的融合正在成为一种新兴趋势。在 2024 财年, 公司正在寻求以下几种方式在 RSU 计划中利用基于 AI 的应用:
 - 年龄: 1 年 1 个月零 12 天
 - 标签: 相关

Human answer: In 2024, several trends are shaping RSU compensation. Companies are increasingly integrating Artificial Intelligence (AI) into RSU plans, using AI-based applications to improve the design, distribution, and management of RSU programs [10]

Nova Pro answer: The integration of AI-based applications in RSU plans is an emerging trend shaping RSU compensation in 2024 [10].

GaRAGe example 2

Question: How do the recent findings on malaria outbreaks in the DRC affect global health strategies?

- 时间动态：缓慢变化
- 复杂性：后期处理繁重
- 领域：健康
- 受欢迎程度：Torso

Grounding:

- 片段 1:
 - 文本：与该小组的发现一致，世界卫生组织更新了其对三种关键的疟疾预防策略的建议：季节性疟疾化学预防（SMC），……
 - 年龄：2 年，0 个月，13 天
 - 标签：无关
- Snippet 2:
 - 需要深入探索以指导 RTS,S 疫苗的实施，特别是在疫苗接种计划覆盖率较低或中等的国家...
 - 年龄：6 个月零 22 天
 - 标签：不相关

Human answer: There is not enough grounding for an answer.

Qwen 32b answer: The recent findings on malaria outbreaks in the Democratic Republic of the Congo (DRC) have influenced global health strategies by prompting the World Health Organization (WHO) to update its recommendations for malaria prevention strategies. These strategies include seasonal malaria chemoprevention (SMC), perennial malaria chemoprevention (PMC), and intermittent preventive treatment in pregnancy (IPTp). [1] The updates aim to broaden the use of chemoprevention among young children at high risk and to expand access to preventive chemotherapy during pregnancy. These recommendations reflect a more informed approach to when and where chemoprevention is most effective, encouraging tailored deployment in local contexts. [3] Additionally, the WHO is developing a Framework to guide decisions on the allocation of limited malaria vaccine supplies, ensuring that national malaria control programmes can make informed decisions about vaccine prioritization. [4] These adjustments in global health strategies are driven by the latest evidence and aim to enhance the effectiveness of malaria control measures, particularly in high-risk areas like the DRC. [1], [2], [3], [4]