HalluMix：一项任务无关的、多领域基准测试，用于现实世界中的幻觉检测

Deanna Emery,
Michael Goitia, Freddie Vargus, Iulia Neagu
Quotient AI
{deanna, mike, freddie, julia}@quotientai.co

摘要

随着大型语言模型（LLMs）越来越多地部署在高风险领域，检测幻想内容——没有支持证据的文字——已成为一个关键挑战。现有的幻觉检测基准通常是合成生成的，专注于提取式问答，并且无法捕捉涉及多文档上下文和完整句子输出的真实世界场景的复杂性。我们引入了幻觉混合基准测试，这是一个多样化的、任务无关的数据集，包含来自不同领域和格式的例子。使用该基准测试，我们评估了七个幻觉检测系统——包括开源和闭源系统——突出了在任务、文档长度和输入表示方面性能的差异。我们的分析显示，在短上下文和长上下文中存在显著的性能差距，这对实际的检索增强生成（RAG）实现具有重要意义。商检测达到了最佳的整体性能，准确率为 0.82，F1 得分为 0.84。

1 介绍

随着大型语言模型（LLMs）在各个领域中的影响力持续增长，确保其输出的事实准确性已成为一个核心问题。在这种情况下，一个关键问题是幻觉，即模型生成的内容与给定的来源不符或矛盾。在法律、医学和金融等高风险领域，这种幻觉现象可能会破坏信任并导致有害后果 (Huang et al., 2025)。

尽管检测幻觉仍然是一个活跃的研究领域，但代表性基准的缺乏阻碍了进展。大多数现有的评估数据集都是任务特定的——通常专注于开放式问题回答——并且严重依赖合成示例或狭窄的上下文格式 (Huang et al., 2025; Ravi et al., 2024; Li et al., 2023; Niu et al., 2024; Yang et al., 2018)。这限制了它们在现实世界设置中的通用性，在这些设置中，大语言模型输出通常是基于多文档上下文的多句子或多段落响应。

我们引入了**幻觉混合**基准测试，这是一个大规模、领域多样的数据集，特别设计用于评估在现实生成场景中的幻觉检测。我们的数据集包括来自多个任务的示例——包括总结、问题回答和自然语言推理——涵盖了医疗保健、法律、科学和新闻等多个领域。每个实例包含一个多文档上下文和一个响应，并带有二元幻觉标签，以指示响应是否忠实于提供的文档。

我们进一步使用该基准系统地评估了七个最先进的幻觉检测系统，包括开源工具和商业工具。

我们的贡献有三个方面：

- 我们提出了一种统一的幻觉检测基准，该基准由跨越多个任务和领域的高质量人工整理数据集构建而成。
- 我们引入了一个一致的评估框架，该框架将幻觉检测与特定任务假设（例如问题的存在）解耦，反映了更多样化的 LLM 用例。
- 我们对现有的幻觉检测方法进行了比较评估，提供了对其优缺点以及在不同实际应用场景中适用性的见解。

以下各节详细介绍了我们的基准构建方

法，展示了我们对领先幻觉检测系统的比较评估，并讨论了我们的研究发现对于学术和工业应用的影响。

2 HalluMix 基准测试

现有的幻觉基准数据集通常由大语言模型生成，并且主要关注问答任务。在许多情况下，这些数据集中的参考答案仅限于直接从单个上下文中提取的单词跨度，这限制了它们在更复杂形式生成中的适用性。(Ravi et al., 2024; Li et al., 2023; Niu et al., 2024; Yang et al., 2018).

然而，幻觉不仅仅局限于问答任务；它们经常出现在其他任务中，如摘要、对话和开放生成。此外，现实世界中的大语言模型部署通常使用列表份文档（通常是通过检索增强生成 (RAG) 提取的）来生成全句输出，而不是简短的抽取片段。为了应对这些限制，我们构建了一个新的基准测试，将幻觉检测与问答格式分离。每个示例包括一个上下文（拆分为文本段落列表）和一个响应，使得评估仅基于两者之间的事实一致性。

为了评估幻觉检测器的性能，我们构建了幻觉混合基准数据集，该数据集综合了多个由人类整理的来源样本。这些任务包括摘要生成、自然语言推理（NLI）和问答（QA），涵盖了新闻、科学、法律、医疗保健、对话和长篇叙述等广泛领域。每个示例都被标记为忠实的或幻觉产生的。

我们选择了主要由人工标注或整理的数据集。如图 1 中的数据处理流程所示，我们应用了多种特定于数据集的转换来构建幻觉示例，同时尽可能使用原始注释来保留忠实样本。

2.1 数据变换

2.1.1 自然语言推理数据集

NLI 数据集通过重新解释其标签模式被重新用于幻觉检测。每个示例包含一个前提和一个假设，以及一个指示它们关系的标签。我们使用以下映射将 NLI 标签转换为幻觉标签：

- 忠实的：标记为蕴含关系的假设。

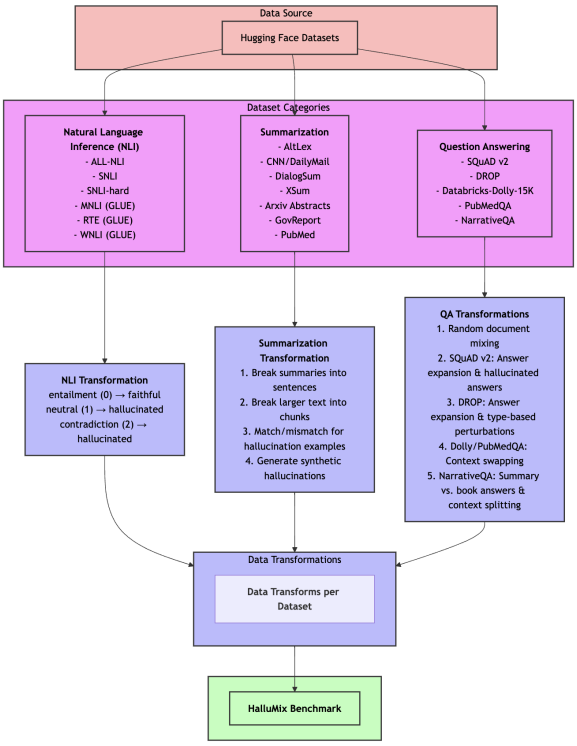


图 1: HalluMix 构建管道概述，显示数据集和转换策略。

- 幻觉生成的：标记为中性或矛盾的假设。

在具有二元 NLI 标签的数据集（蕴含关系对比非蕴涵关系）中，我们应用了类似的映射，将非蕴含关系视为虚构的。以下是对所使用的 NLI 数据集的说明：

- `sentence-transformers/all-nli` (Williams et al., 2018; Bowman et al., 2015)
- `stanfordnlp/snli` (Bowman et al., 2015)
- `snli-hard` (Gururangan et al., 2018)
- 粘合剂：`mnli`、`rte` 和 `wnli` (Wang et al., 2018)

2.1.2 摘要数据集

摘要数据集由长文档与人工编写的摘要配对组成。由于摘要旨在忠实于原始文档，我们默认将其标记为忠实的。为了创建虚假示例，我们应用基于置换的转换：将摘要随机错配到无关文档。

我们包括以下摘要数据集：

- 句子变换器/替代选项 (Hidey and McKeown, 2016)
- CNN/每日邮报 (See et al., 2017)
- 对话摘要 (Chen et al., 2021)
- X 总和 (Narayan et al., 2018)
- arXiv 摘要 (Cohan et al., 2018a)
- 政府报告摘要 (Huang et al., 2021)
- PubMed 概要 (Cohan et al., 2018b)

2.2 问题回答数据集

QA 数据集包含一个问题、一个上下文段落以及相应的答案。按设计，这些数据集中的答案是忠实的。幻觉变体通过多种特定于数据集的策略生成。在某些数据集中，答案由单个单词组成；我们使用大型语言模型将它们扩展为完整的陈述句（例如，问：汽车是什么颜色？答：红色 → 汽车是红色的。），以确保与现实世界用例的一致性，并分离对问题的依赖以确定幻觉。

我们使用了以下问答数据集：

- SQuAD-v2 (Rajpurkar et al., 2016, 2018) 无法回答的问题（空白答案）与仅基于问题（无上下文）由大语言模型生成的答案配对，并标记为虚构的。
- 删除 (Dua et al., 2019) 每个上下文包含多个带有类型化答案的问题（数值型、日期型、字符串型）。虚假示例是通过用同一上下文中相同类型的另一个可能的答案替换正确答案创建的。
- Databricks-Dolly-15K (Conover et al., 2023) 并且 PubMedQA (Jin et al., 2019)：通过将答案和上下文错配生成了幻觉示例。
- 叙事 QA (Kočiský et al., 2018) 该数据集包含书籍长度的文本、摘要、问题和答案。为了便于处理，我们主要使用文档摘要作为上下文。为了保持长上下文评估，我们保留了一小部分较短的全文样本。虚构的例子是通过将答案与无关的摘要或段落不匹配生成的。

2.3 最终数据集结构

HalluMix 的结构支持强大和灵活的幻觉检测评估。

为了更好地反映现实世界的信息检索场景（即 RAG），每个上下文都被分割成由完整句子组成的等大小块，确保没有一个块包含部分或碎片化的句子。这种方法保持了语法的完整性，同时维持了可管理的块长度，并防止过度分割。此外，我们随机打乱文档块的顺序以消除任何排序优势，因为现实世界中的检索系统通常会返回不保留原始叙述顺序的文档。

为了模拟真实的检索噪声，我们在基准测试中将忠实示例与从无关文档中随机选择的十个不相关的文档片段进行了扩充。增加的内容提升了识别相关信息的难度，而没有改变可用于支撑假设的证据。通过仅对忠实示例应用这种扩充，我们避免了无意中向虚构案例引入支持性证据。这种方法创建了一个评估环境，该环境反映了现实世界条件，在这些条件下，幻觉检测系统必须在存在噪声文档检索的情况下取得成功。

最终数据集中的每个示例包括：

- A **文档**字段：作为文本块列表表示的上下文（例如，分词后的句子或段落块），
- 一个**回答**：待评估的假设，例如摘要句子、答案或声明，
- 一个二进制**幻觉标签**：其中 0 表示忠实的，而 1 表示幻觉产生的，
- A **源标识符**：表示用于来源追踪的原始数据集。

具有代表性的幻觉和忠实数据点的示例分别在附录的表 5 和表 6 中提供。

忠实示例（标签 = 0）直接来自人类标记或人类整理的数据集。虚构示例（标签 = 1）在某些情况下是通过控制变换构造的，例如摘要不匹配、QA 上下文排列或 NLI 重新标记。由于这些变换——包括分块、混洗、插入干扰项和标签重新分配——HalluMix 中的每个数据

点都已大幅修改，不应被视为等同于其原始来源，即使保留了源标识符用于跟踪目的。

最终数据集进行了去重处理，并收集了6500 个分层随机样本，以实现具有幻觉标签、数据类型和来源均衡表示的平衡数据集。每个来源在每种数据类型（NLI、QA、摘要）中大致有相等的代表性，且每种数据类型在整个基准数据集中也有大致相等的代表性。

所得到的数据集为跨多个领域、格式和任务设置的幻觉检测提供了一个统一且可扩展的基准。

3 基准测试方法论

使用幻觉混合基准测试，我们比较了不同幻觉检测器的性能，选择了基于实际部署考虑的方法，包括模型大小（≤8B 参数）、推理成本和延迟要求。我们的评估包括开源方法和闭源方法。

- *llama-3-守护神-lynx-8b-instruct-v1.1* (Ravi et al., 2024) - 经过幻觉检测数据集微调的 Llama-3.1 模型。输出是一个二进制分数。
- 拉加斯的忠诚 (Es et al., 2024) - 一种两步方法，使用大型语言模型识别模型响应中的不同主张，然后将大型语言模型用作裁判来判断每个主张是否忠实于源文档。输出是忠实于文档的主张的比例（即，值小于 1 表示存在幻觉）。
- *Azure* 基础性 (Azure AI Content Safety, 2024) - 一个基于封闭源代码 API 的幻觉检测器。输出是一个二进制分数。
- *Vectara HHEM-2.1-Open* (Forrest Bao and Mendelevitch, 2024) - HHEM-2.1 模型的开放权重版本。输出是一个忠实度的概率（介于 0-1 之间）。在本文中，我们设置了一个阈值，使得小于 0.5 的值被预测为幻觉产生的。
- 顶点 *AI* 定位 (Google Vertex AI, 2025) - 一个基于闭源 API 的幻觉检测器。输出是一个介于 0 到 1 之间的可信度概率。在本文中，我们设置了一个阈值，使得小于 0.5 的值被预测为幻觉产生的。
- 量身定制-迷你检查-*7B*(Tang et al., 2024; Bespoke Labs, 2024) - 一个经过微调的 70 亿参数模型，用于接受句子和文档对进行评估。多句回应会先被分解成单个句子再进行评估。该模型为每个句子返回一个二元评分。在本文中，如果有任何一句被预测为虚构内容，我们将整体预测设置为虚构的。
- 商检测 - 一个作为裁判的 LLM，它采用基于句子的方法来识别幻觉。输出是一个二进制分数，表示至少有一个句子包含幻觉。

这些幻觉检测器所需的输入和输入格式各不相同。有些需要单个上下文，而有些接受文档列表；有些需要问题作为输入，而有些只需要上下文和响应。表 1 列出了每种幻觉检测方法的输入要求。

因为我们的基准数据集是与问题无关的，并非所有示例都有适用的问题；在这种情况下，当检测器需要时，我们将问题输入为无。对于检测器不接受文档列表的情况，我们改为使用两个换行符将每个文档之间的文档连接起来。Patronus Lynx 8B 和 Vectara HHEM-2.1-Open 都需要单一的上下文输入。

4 结果

我们在 HalluMix 基准上评估了七个幻觉检测系统，性能指标如表 2 所示。

商检测 达到了整体最佳性能，在准确率（0.82）和 F1 分数（0.84）方面领先，同时保持了精确度（0.76）和召回率（0.93）之间的良好平衡。虽然 Quotient Detections 没有在个体的精度或召回率上取得最高分，但在某一项指标中表现出色的方法在另一项指标上出现了显著下降。例如，*Azure* 基础性 [1] 展示了高精确度

[1] Azure Groundedness 的性能可能由于其对输入文档长度的限制而被高估。我们无法获得 304 个最长上下文示例的 Azure Groundedness 评估。通常，这些长上下文示例更具挑战性。

	Single Context	List of Documents	Question	Response
Quotient Detections	-	✓	-	✓
Patronus Lynx 8B	✓	-	✓	✓
Ragas Faithfulness	-	✓	✓	✓
Azure Groundedness	-	✓	可选	✓
Vectara HHEM-2.1-Open	✓	-	-	✓
Vertex AI Grounding	-	✓	-	✓
Bespoke-Minicheck-7B	-	✓	-	✓

表 1: 输入要求和每种幻觉检测方法的格式。勾号表示该字段是必填项。破折号表示该字段不被接受。对于 Azure Groundedness，问答任务和摘要任务有不同的 API 请求格式，可以在有无问题的情况下进行幻觉检测。

	Accuracy	F1	Precision	Recall
Quotient Detections	**0.821**	**0.840**	0.764	0.932
Bespoke Minicheck 7B	0.808	0.832	0.744	0.944
Patronus Lynx 8B	0.808	0.828	0.754	0.919
Ragas Faithfulness	0.787	0.818	0.719	**0.950**
Azure Groundedness*	0.784	0.788	**0.781**	0.795
Vectara HHEM-2.1-Open	0.749	0.771	0.715	0.836
Vertex AI Grounding	0.727	0.772	0.668	0.915

表 2: 幻觉检测性能在完整基准数据集上的评估结果。商检测获得了最高的总体准确率和 F1 分数，显示出平衡的精确度和召回率。Azure Groundedness[1] 达到了最高的精确度但召回率较低，而 Ragas Faithfulness 则以牺牲精确度为代价实现了最高的召回率。

(0.78)，但召回率较低（0.79)。相反，拉加斯的忠诚显示了高召回率（0.95)，但在精度方面有所牺牲（0.72)。

在考察不同数据源的性能（表 3）时，我们观察到方法之间和数据集之间的显著差异。

在摘要任务中，最显著的模式是性能出现剧烈分化。帕特罗努斯 赋形兽 *8B* 在长格式摘要任务中始终优于其他方法。例如，在 PubMed 摘要上，它达到了 0.91 的准确率，相比之下商检测为 0.63，量身定制-迷你检查-*7B* 为 0.58。

表 3 显示了在 HalluMix 的每个子数据源中幻觉检测方法的准确率得分。分数在不同方法和不同数据集之间的高变异表明，每种检测方法可能有不同的优缺点。值得注意的是，在 NLI 和问答子集上表现一般最好的方法在这项总结子集上的表现往往较差，反之亦然。

表 4 展示了不同数据类型在内容长度上的显著差异。NLI 示例简短（响应平均为 11 个标记，文档为 88)，而摘要生成示例涉及更长的文本（响应平均为 174 个标记，文档为 439)。这些长度差异与检测性能密切相关，表明质量幻觉检测方法依赖于内容长度。

图 2 进一步展示了内容类型如何影响相对性能。在仅评估短上下文示例（NLI 和 QA 子集，面板 b）时，*Patronus* 莱尼克斯 *8B* 的准确率从第三名下降到第五名，而商检测保持领先。这种变化凸显了基准组成对性能排名的影响。

Data Type	Data Source	Quotient Detections	Bespoke Minicheck 7B	Patronus Lynx 8B	Ragas Faithfulness	Azure Groundedness[1]	Vectara HHEM-2.1-Open	Vertex AI Grounding
自然语言推理	au123/snli-hard	0.856	**0.921**	0.686	0.813	0.878	0.789	0.694
	nyu-mll/glue/mnli	0.847	0.847	0.662	0.759	**0.912**	0.802	0.751
	nyu-mll/glue/rte	0.900	**0.966**	0.746	0.870	0.852	0.678	0.818
	nyu-mll/glue/wnli	0.850	**0.902**	0.758	0.766	0.568	0.488	0.502
	sentence-transformers/all-nli	0.901	0.908	0.673	0.821	**0.936**	0.821	0.755
	stanfordnlp/snli	**0.885**	0.880	0.699	0.858	0.869	0.833	0.749
问题回答	PubMedQA	0.624	0.572	**0.928**	0.586	0.596	0.670	0.542
	databricks-dolly-15k	**0.880**	0.766	0.826	0.848	0.842	0.864	0.855
	DROP	0.806	0.766	**0.878**	0.736	0.708	0.478	0.586
	narrativeqa	0.886	0.858	**0.916**	0.864	0.748	0.832	0.758
	squad_v2	**0.920**	0.912	0.890	0.892	0.912	0.818	0.890
总结	sentence-transformers/altlex	0.883	0.838	0.730	0.820	**0.932**	0.865	0.869
	arxiv_summarization	0.614	0.591	**0.926**	0.702	0.568	**0.926**	0.633
	cnn_dailymail	0.753	0.813	0.822	0.808	0.817	0.881	**0.936**
	dialogsum	0.876	0.814	0.814	0.850	**0.920**	0.690	0.606
	govreport_summarization	0.597	0.509	**0.943**	0.703	0.500	0.915	0.882
	pubmed_summarization	0.629	0.582	**0.911**	0.695	0.615	0.892	0.803
	xsum	0.715	0.720	0.725	0.617	**0.798**	0.606	0.601

表 3: 幻觉检测方法在基准数据源中的准确率得分。**粗体**值表示每个数据源的最高准确性，而红色 值表示最低准确性。数据集之间的显著差异表明了特定方法的优势，某些检测器在特定数据类型上表现出色而在其他类型上表现不佳——突显了当前幻觉检测方法中存在的潜在专业化或过拟合问题。

Data Type	Avg Response Token Count	Avg Document Token Count
NLI	11	88
QA	32	167
Summ.	174	439

表 4: 每种数据类型（NLI、问答和摘要）在 HalluMix 中的平均标记数量。

5 讨论

我们的综合评估揭示了关于当前幻觉检测系统状态的几个关键见解。尽管表现最佳的模型总体上达到了令人尊敬的准确性，但它们的效果因任务类型、内容长度和输入格式而异。这些变化既反映了我们基准数据集的多样性，也体现了每种检测方法中嵌入的设计决策。

5.1 子源过拟合的证据

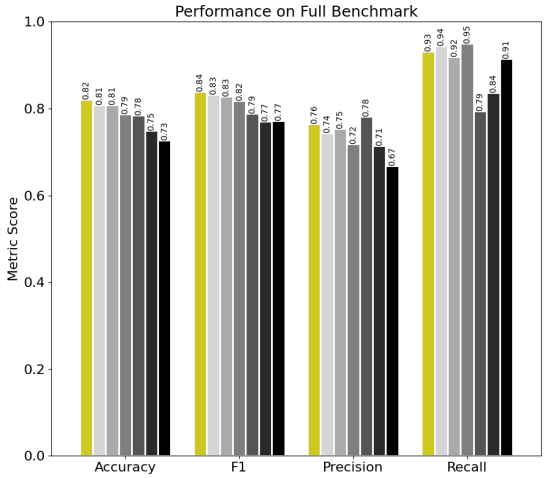
表 3 表明某些检测系统在特定数据集上表现非常出色，而在其他数据集上则表现不佳。这一模式表明，某些幻觉检测方法可能是在某个子数据集上进行过训练或受到了特别影响，特别是在 NLI 和 QA 类别中。例如，在 SNLI 或 SQuAD 等知名数据集上的高准确率可能表明在预训练或微调过程中有过接触。虽然这不一定否定其性能，但它确实引发关于通用性的问题——尤其是在不太常规或特定领域的生成场景中。

我们分析的一个有趣发现是，尽管专门优化用于幻觉检测的模型如 *Patronus 胡狼 8B*、*Vectara HHEM-2.1-Open* 和定制迷你检查-*7B* 在设计上针对特定任务进行了微调，但它们通常并不比利用通用语言模型并采取适当提示策略的方法（例如拉加斯的忠诚和商检测）表现得更好。

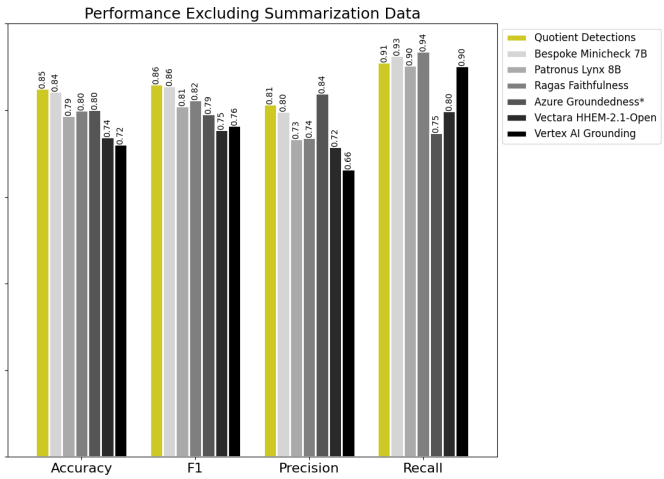
5.2 内容长度和上下文表示挑战

如表 4 所示，摘要子集涉及的上下文和响应比 NLI 或 QA 长得多。大多数模型在摘要示例上的性能下降表明，长篇幅生成为幻觉检测引入了额外的挑战，例如跟踪指代对象、保持话语连贯性以及在整个大文本跨度中对声明进行定位。

图 3 揭示了不同架构方法在处理内容长度



(a) 在完整的 HalluMix 基准测试（包括摘要数据）上的表现。



(b) 在排除摘要数据的情况下，在 HalluMix 基准测试中的表现。

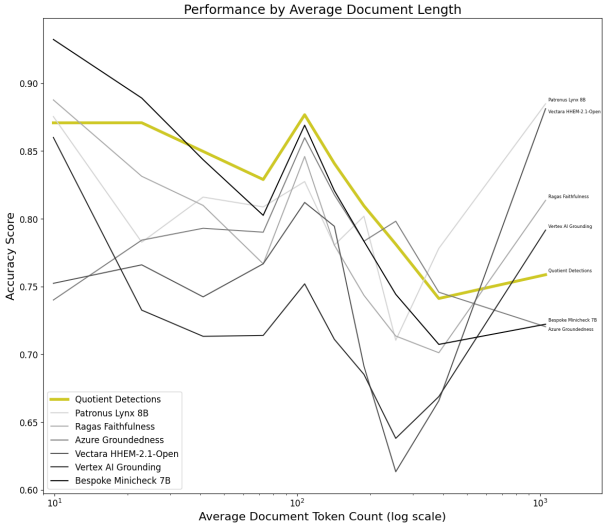
图 2: 幻觉检测性能指标（准确性、F1、精度、召回率）的比较，涵盖了所有评估的方法。面板 (a) 展示了在完整基准数据集上的表现，而面板 (b) 则显示了排除总结示例后的表现。商检测在两种情境下均实现了最高的准确性和 F1 值。

时的重要模式。*Vectara HHEM-2.1-Open* 和守护神猫科 *lynx 8B*——这两种经过微调的模型，处理连续而不是分块上下文的能力——始终表现出在较长内容上的优越性能，但在较短的例子上却表现不佳。相比之下，基于句子的方法如商检测和定制迷你检查-*7B* 在典型 NLI 和 QA 任务中的较短内容（~200 个标记）中表现出色，但在长篇摘要示例上的性能下降。

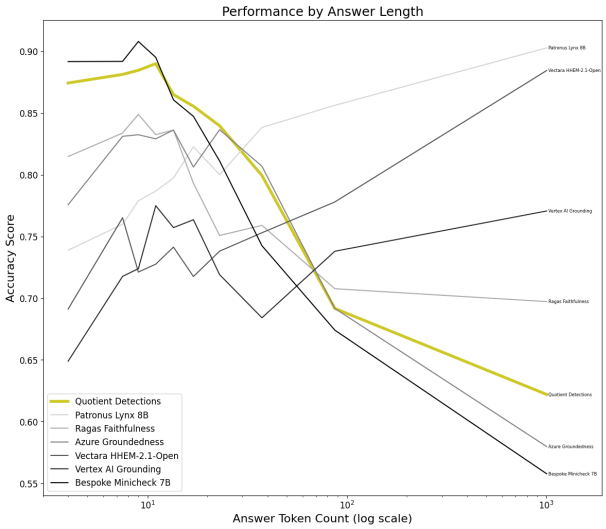
这种性能差异突显了在幻觉检测中上下文表示的基本权衡。连续上下文方法可能更好地保持文档连贯性，维持支持长文本准确忠实度评估的关键话语信号和跨句子依赖关系。然而，基于句子的方法对于短上下文中细粒度声明验证提供了更高的精度，但在处理较长文档时会遭受信息损失。

两种基于句子的方法在摘要示例上实现了接近 1.0 的召回值，表明它们在评估长篇内容时倾向于过度预测幻觉——这可能是因为句子隔离破坏了核心指代链和对准确评估至关重要的其他跨句上下文信号。

这些发现提出了几种可能改进句子基础检测器的方法。在评估过程中纳入包含相邻句子



(a) 每个幻觉检测方法按平均文档标记数的准确性。



(b) 按每个幻觉检测方法的答令牌数量计算的准确性。

图 3: 幻觉检测方法的性能作为内容长度的函数。图 (a) 展示了准确率与平均文档标记数量的关系，揭示了不同方法如何处理越来越复杂的情况。图 (b) 展示了准确率与答案标记数量的关系，展示了在较长回复上的表现。两个图表都显示了不同的性能模式：一些方法在整个长度范围内保持一致的准确性，而其他方法则随着内容变长表现出明显的下降。

的滑动窗口上下文可能会有助于保持局部连贯性。或者，分层验证方法可以首先评估单个句子，然后使用整个段落的上下文进行二次验证。这样的方法可以在保持句子级别检测的优势的同时解决上下文碎片化问题。

5.3 面向鲁棒的幻觉检测

未来的工作将专注于提升商检测在长上下文场景中的性能。这包括探索在保持句子级别颗粒度的同时利用全局文档上下文的混合方法。目标是开发一个检测系统，使其在整个生成长度范围内都有效，从简短的事实声明到多段落摘要。

总体而言，我们的研究结果强调了需要能够可靠地在不同输入类型、文档长度和生成格式下运行的幻觉检测系统。HalluMix 为此类研究提供了坚实的基础，而我们的评估揭示了在未来改进模型架构和输入处理策略方面的关键方向。

6 结论

在这项工作中，我们介绍了幻觉混合，一个大规模、任务多样且跨领域的基准数据集，用于评估现实语言生成环境中幻觉检测的表现。与之前的基准数据集不同，我们的数据集反映了多文档基础和长篇回复在现代大语言模型部署中常见的挑战。我们系统地评估了七种检测方法，并揭示了它们的效果根据输入格式、内容长度以及底层任务存在显著差异。总体而言，商检测在准确率和 F1 值上表现最佳。

我们的分析揭示了几个关键发现。一些检测模型似乎对已知数据集过度拟合，这引发了关于泛化能力的担忧。基于句子级别的方法在较短内容的检测中表现出色，但在较长上下文中的表现则较差，可能是由于失去了句间连贯性。同时，在完整连续文本上评估的模型可能会从保留的上下文中受益，突显了检索增强生成管道中典型分段输入的局限性。这一发现强调了细粒度声明验证与文档级连贯性评估之间的关键权衡。

这些发现对生产系统中 LLM 的部署具有重要意义。实施基于 RAG 解决方案的组织应仔细考虑现有幻觉检测器的局限性，特别是在处理特定领域内容或长格式输出时。

未来的研究方向包括提高在不同内容长度下的幻觉检测鲁棒性，更好地建模话语层面的依赖关系，并改进检测器以更有效地处理现实世界中的分块输入。我们公开了我们的基准测试，以便促进这一关键领域的持续研究，提升大语言模型的安全性和可靠性。

References

Azure AI Content Safety. 2024. Groundedness detection.

Bespoke Labs. 2024. Bespoke-minicheck-7b.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5062–5074, Online. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018a. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018b. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2368–2378, Minneapolis, Minnesota. Association for Computational Linguistics.

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, St. Julians, Malta. Association for Computational Linguistics.

Rogger Luo Forrest Bao, Miaoran Li and Ofer Mendelevitch. 2024. HHEM-2.1-Open.

Google Vertex AI. 2025. Check grounding with rag.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Christopher Hidey and Kathy McKeown. 2016. Identifying causal relations using parallel Wikipedia articles. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1424–1433, Berlin, Germany. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *Preprint*, arXiv:2104.02112.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *Preprint*, arXiv:1909.06146.

Tomáš Kočiský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The NarrativeQA reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. *Preprint*, arXiv:2305.11747.

Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, Kashun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. Ragtruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. *Preprint*, arXiv:2401.00396.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Selvan Sunitha Ravi, Bartosz Mielczarek, Anand Kannappan, Douwe Kiela, and Rebecca Qian. 2024. Lynx: An open source hallucination evaluation model. *Preprint*, arXiv:2407.08488.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *Preprint*, arXiv:1809.09600.

A 附录

文档	• 由于前一天钢人队输给了乌鸦队，辛辛那提猛虎队以美联北区冠军的身份进入本场比赛。猛虎队在上半场凭借麦卡隆的触地传球和莫哈迈德 · 桑努的跑动取得了 14-0 的领先优势，但丹佛队通过布兰登 · 麦克马纳斯的一记 23 码短距离射门将分差缩小到 11 分，比赛结束前 18 秒。下半场，在第三局米奇 · 内格恩错失射门后，形势发生了巨大变化。埃曼努埃尔 · 桑德斯从布鲁克 · 奥斯维勒手中接住了一个 8 码的传球，将比分差距缩小至 14-10，丹佛队在第四节还剩 11 分 17 秒时凭借 C.J. 安德森的一个 39 码触地得分首次领先。猛虎队通过米奇 · 内格恩的一记赛季最长 52 码射门将比分扳平，使比赛结束后的常规时间比分为 17-17。在加时赛中，疲惫的猛虎队未能得分，让麦克马纳斯的 37 码射门帮助丹佛队以 20-17 领先。随后猛虎队的一次失误使得野马队获得了球权，结束了比赛和辛辛那提队季后赛首轮轮空的梦想。随着这场失利，猛虎队在本赛季的成绩变为 11 胜 4 负。这也是自 1975 年以来猛虎队在丹佛的连续第 10 场失利。
响应	The first field goal was by the Ravens.
标签	Hallucinated

表 5: An example of a hallucinated datapoint in the HalluMix Benchmark.

文档	• 最终幻想是由坂口博信创作并由史克威尔艾尼克斯（原名 Square）开发和拥有的日本科幻奇幻多媒体特许经营权。 • 彼得 · 赖特，DNR 的一名法律监督员，告诉 WLUC-TV，警官只是在履行职责。他说，警官认为那是一头野猪，因为它没有任何识别标记来证明它是一只宠物。“我想非常清楚地表明，我们部门绝不会希望射杀人们的宠物。”赖特说道。“如果他有任何一丝怀疑那是宠物的话，他就绝对不会开枪了。”令人不安的是：这家人现在正试图找回凯撒的遗体以便安葬，但被告知他们只能领回他的骨灰。布兰迪 · 塞维尔和托尼 · 杰瓦西正在努力把凯撒的遗体拿回来。然而，他们被告知只能领取骨灰。塞维尔女士要求从这一事件中得到某种补救措施。“如果这是一个如此大的错误，那么我们希望看到更好的培训。”她说。“让我们学会识别不仅是猪，还有所有宠物。” • 上帝恨我们所有的人是美国鞭击金属乐队斯莱尔的第八张录音室专辑。 • 那是对的，完全正确，但是越来越多的女性开始自己的生意，我注意到了这一点。 • 该特许经营权围绕一系列奇幻和科幻角色扮演游戏的视频游戏展开。该系列中的第一款游戏于 1987 年发布，迄今为止已发布了 15 款主要作品。 • 大约在公元前 3600 年左右，埃及社会开始迅速向精致的文明发展和进步。 • 男孩推着装有两个南瓜的手推车
响应	Final Fantasy was created by Hironobu Sakaguchi
标签	Faithful

表 6: An example of a faithful datapoint in the HalluMix Benchmark.