超越事实：评估大语言模型中的意图幻觉

Yijie Hao
Emory University
yhao49@emory.edu

Haofei Yu
UIUC
haofeiy2@illinois.edu

Jiaxuan You
UIUC
jiaxuan@illinois.edu

Abstract

当面对包含多个条件的复杂查询时，现今的大型语言模型（LLMs）往往生成的响应只能部分满足查询，同时忽略某些条件。因此，我们引入了意图幻觉的概念。在这种现象中，LLMs 要么省略（忽视某些部分的处理），要么误解（回应想象的查询部分）给定查询的元素，导致意图幻觉生成。为了系统地评估意图幻觉，我们引入了 FAITHQA，这是一个用于意图幻觉的新基准，共包含 20,068 个问题，涵盖了仅查询生成和检索增强生成（RAG）设置的多种主题和难度。FAITHQA 是第一个超越事实验证的幻觉基准，专为识别意图幻觉的根本原因而设计。通过对各种 LLMs 在 FAITHQA 上的评估，我们发现（1）即使是最先进的模型，意图幻觉也是一个普遍的问题，（2）这种现象源于 LLMs 的省略或误解。为了促进未来的研究，我们引入了一种自动 LLM 生成评估指标 CONSTRAINT SCORE，用于检测意图幻觉。人工评估结果表明，CONSTRAINT SCORE 在意图幻觉方面比基线更接近人类表现。

大型语言模型（LLMs）已在各类应用中展示了其效用。然而，幻觉仍然是一个显著的挑战。特别是，对于包含多个条件的复杂查询（图），LLM 的输出常常偏离查询，导致不理想的结果。我们将这一现象称为“意图幻觉”，这个问题在当前研究中鲜有关注。

与基于事实的幻觉不同，(Li et al., 2023; Cao et al., 2021)，研究人员可以通过基于搜索的事实核查直接检测出这种幻觉，(Sellam et al., 2020; Min et al., 2023)，而检测和评估意图幻觉则更具挑战。这是因为复杂的查询通常包含重复的意图，而大型语言模型（LLMs）通常只满足其中的一部分，导致不满难以被察觉或量化。此外，随着 LLMs 的不断进步，用户往往向这些更强大的 LLMs 提供越来越复杂的查询，即使是人类也难以理解。这一趋势凸显出 LLMs 不仅需要在事实层面上正确，还需要在意图上与人类保持一致。我们的论文提出了两个尚未被深入探讨的问题：（1）为什么 LLMs 倾向于表现出意图幻觉？以及（2）我们如何检测意图幻觉？回答这些问题对于依赖事实准确性和忠实意图对齐的 LLM 应用至关重要。

针对第一个问题，我们提出长语言模型在单词层面上的意义上省略（例如，忽略查询组件）或误解（例如，对虚构的查询组件作出回应）构成了意图幻觉的根本原因。为了进一步研究，我们引入了 FAITHQA，这是第一个专门设计用于解决意图幻觉的两个关键情景：省略和误解的基准。FAITHQA 由 20,068 个查询组成，这些查询通过广泛的人类评估进行了验证以确保质量。FAITHQA 涵盖了各种难度水平的广泛主题，即使是当前最先进的模型也证明其具有挑战性。我们的基准揭示了随着查询复杂性的增加，意图幻觉的发生率更高。

为了解决第二个问题，我们引入了 CONSTRAINT SCORE，这是一种新的评估指标，专注于检测意图幻觉。我们的方法涉及两个主要步骤：(1) 通过概念和动作将查询分解，然后将其转换为一系列短语句，每个短语句代表生成必须满足的特定要求；(2) 为每个约束分配一个重要性加权的二进制标签，从而实现细粒度评估。我们的人工评估显示，CONSTRAINT SCORE 显著优于以 LLM 为裁判的 (Manakul et al., 2023; Mishra et al., 2024; Sriramanan et al., 2024)，因为 LLM 裁判相比于人工判断往往提供偏向的评估。

综上所述，我们的主要贡献包括：(1) 我们提出了意图幻觉的概念，这一概念超越了现有的事实幻觉类别；(2) 我们开发了 FAITHQA，这是第一个专注于评估意图幻觉的基准。我们的结果显示，即使是最先进的 LLM，意图幻觉也是一种普遍现象；(3) 我们引入了 CONSTRAINT SCORE，这是一种新颖的评估指标，通过将查询分解为意图约束并计算加权分数来自动评估 LLM 生成的结果。我们的分析显示，CONSTRAINT SCORE 显著优于纯粹的 LLM 评分基准，因为后者往往存在偏见。

1 相关工作

在 LLMs 中，“幻觉”指的是那些不符合事实、无关紧要或虚构的输出。这个问题出现在诸如回答问题 (Sellam et al., 2020) 、翻译 (Lee et al., 2018) 、摘要 (Durmus et al., 2020) 和对话 (Balakrishnan et al., 2019) 的任务中，如多项研究所指出的那样 (Ji et al., 2023; Azaria and Mitchell, 2023; Huang et al., 2023; Cao et al., 2021) 。为了应对这个问题，许多努力旨在检测和减轻幻觉。Min et al. (2023) 通过将每个句子的核心事实（原子事实）与可靠来源（如维基百科）进行核对来评估事实的准确性。Hou et al. (2024) 提出了一种隐藏马尔可夫树模型，将陈述分解为前提，并基于所有上级前提的概率分配事实性分数。Manakul et al. (2023) 通过采样多个响应并利用自我一致性来检测幻觉，以识别不一致之处。

尽管这些努力很重要，但仍然存在局限性。大多数现有工作要么仅专注于事实精度，要么专注于上下文召回，忽略了生成中查询的作用 (Li et al., 2023; Yang et al., 2023; Niu et al., 2024) （例如，在图 1 中对两个输出进行相等评分），要么将查询视为一个整体 (Zhang et al., 2024a) ，这导致了粗粒度的评估。

Hallucination benchmarks. 最近关于 LLM 幻觉检测的工作包括 HaluEval (Li et al., 2023) （合成和自然响应）、FELM (Chen et al., 2023) （跨领域的自然响应）、RAGTruth (Niu et al., 2024) （RAG 幻觉）和 InfoBench (Qin et al., 2024) （通过查询分解进行指令跟随）。这些基准测试主要关注事实性幻觉或需要人工标注。相比之下，据我们所知，FAITHQA 是首个从查询中心的视角评估非事实性幻觉的工作。

尽管 Zhang et al. (2024b) 也讨论了一个相关主题，但他们的工作主要从训练语料库的角度探索意图幻觉的原因。相反，我们的论文提供了一个全面的评估指标以及一个用于系统测试的广泛基准。FaithEval (Ming et al., 2025) 通过评估模型输出在处理不可回答或矛盾证据的条件下，是否对外部检索的上下文保持忠实，来研究 RAG 设置中的幻觉。FAITHQA 采用了类似的 RAG 设置，但将焦点从上下文对齐转移到查询对齐上。它引入了一种创新的、以查询为中心的视角，评估模型响应是否准确满足用户的查询。我们的实验结果与 FaithEval 的发现一致，揭示出当 LLMs 被提供相关但不完整或噪声的检索结果时，它们常常表现出遗漏式意图幻觉，未能解决原始查询的所有方面。

对于包含多个条件的复杂查询，研究报告称模型产生的响应仅部分满足这些条件。为了进一步研究这一点，我们在本文中概述了关于意图幻觉的关键见解。

1.1 意图约束：一个基本单位

一个查询由多个概念和操作组成，每个代表不同的意图，并在给定的上下文中具有特定的意义。如图 **??** 所示，大型语言模型（LLMs）常常无法处理查询中提供的约束，这导致偏离查询意图的幻觉生成。为实现细粒度的、查询中心的评估，我们引入了意图约束的概念——简短的陈述，每个陈述表达生成必须解决的单一要求（参见图 1 中的示例）。一个查询，由上下文中的概念和操作定义，分解为这些意图约束，每个约束代表一个独特的概念或操作。解决每个这些约束有助于减少与查询意图不一致的幻觉响应的风险。

Definition 1.1 (Intent Constraint Mapping Function). 设 $\mathcal{Q}$ 为所有查询的集合，$\mathcal{I}$ 为所有可能的意图约束的集合。对于每个 $q \in \mathcal{Q}$ ，定义映射函数 $C : \mathcal{Q} \longrightarrow \mathcal{P}(\mathcal{I})$ 为

$$C(q) \;=\; C_m(q) \;\cup\; C_i(q) \;\cup\; C_o(q), \qquad (1)$$

，其中 $C_m(q) \subseteq \mathcal{I}$ 是强制约束的集合（这些约束必须首先解决），$C_i(q) \subseteq \mathcal{I}$ 是重要约束的集合（在强制约束之后解决），而 $C_o(q) \subseteq \mathcal{I}$ 是可选约束的集合（理想的但不是必须的）。这个映射确保了 $C(q)$ 捕获了所有需要的意图约束，以保持 q 的原意。

建立了一个精细的、以查询为中心的视角之后，我们正式定义意图幻觉为大型语言模型未能处理词汇层面的概念或动作，并表现为忽略或误解意图约束。当大型语言模型忽略查询的某些部分（例如，未能处理特定概念或动作）或者误解查询（例如，回应未提及的概念或动作）时，生成的结果未能与原始查询对齐，无论其是否事实准确。在处理复杂、多条件的查询时，将意图约束视为意图幻觉的基本评估单元尤为重要。在这种情况下，大型语言模型往往只部分地回应查询而忽视其余部分。根据意图约束评估生成输出提供了一种有效的方法来识别和区分这些细微的不一致性。

Definition 1.2 (Intent Hallucination). 设 q 为用户查询，设 P_θ 表示我们的 LLM。用 $C(q) = \{c_1, \ldots, c_k\}$ 表示从 q 中提取的意图约束集。理想情况下，模型的分布仅依赖于这些约束，即

$$P_\theta\big(\cdot \mid q\big) \;=\; P_\theta\big(\cdot \mid C(q)\big). \qquad (2)$$

。然而，在实际操作中，模型通常隐含地以幻觉约束集

$$\widehat{C}(q) \;=\; \{\hat{c}_1, \ldots, \hat{c}_{k'}\}, \qquad (3)$$

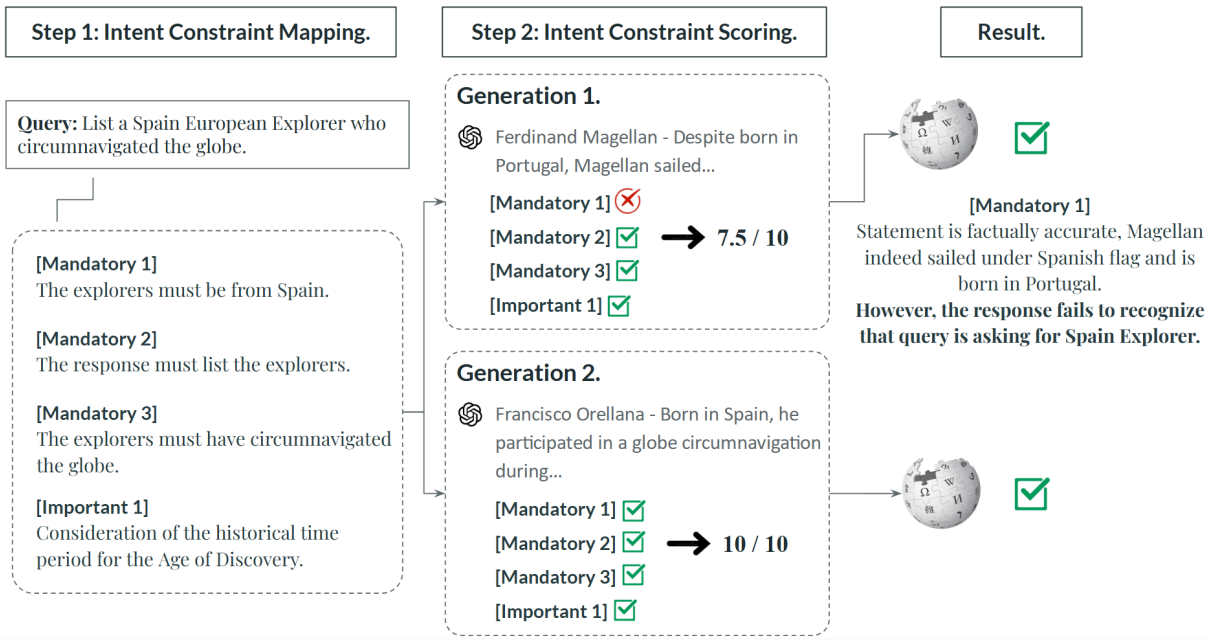

Figure 1: CONSTRAINT SCORE 计算过程。尽管两代都在事实的准确性上是正确的，但与第二代相比，第一代并不理想，因为第一代省略了“探险者必须来自西班牙”。

为条件，它与 $C(q)$ 不同（例如，通过用 $\hat{c}_i$ 替换约束 c_i，或者省略某个约束）。在这种情况下，实际响应遵循

$$y_h \sim P_\theta\big(\cdot \mid \widehat{C}(q)\big), \tag{4}$$

，而 y_h 与理想响应 $y \sim P_\theta(\cdot \mid C(q))$ 之间的偏差被定义为意图幻觉。

基于意图约束和意图幻觉的定义，我们引入了 CONSTRAINT SCORE，这是一种基于意图约束检测意图幻觉的新评价指标。为了实现前面定义的约束映射函数 $C(\cdot)$，我们开发了一个多步骤过程，系统地从查询中提取和分类意图约束集合。我们的方法具有很高的灵活性，能够适应涉及 RAG 的不同查询。提示模板出现在附录 §5 中。

1.2 意图约束映射

Step 1: Preliminary assessment. 大型语言模型首先分析查询 q，以验证是否存在足够的信息用于提取约束。这一步对于 RAG 查询至关重要，因为它减少了外部内容的影响 (Liu et al., 2023; Wu et al., 2024)。如果检测到信息不足，约束映射过程将停止并请求附加输入，以确保 $C(q)$ 定义明确。

基于语义角色标注 (SRL)，我们提取出 q 的基本组成部分：主语、动作和上下文。这种结构化分解使得在不同类型的实际查询中能够识别出稳健的约束。

Step 3: Constraint set extraction. 我们指导语言模型分析步骤二生成的给定提示的上下文，从七个类别——位置、时间、主题、动作、限定词和数量——进行分析，然后将其重新表述为三个约束集：$C_m(q)$，包括位置、时间、主题和动作的约束；$C_i(q)$，包括限定词和数量的约束；以及 $C_o(q)$，包括语言模型提供的任何其他约束，例如排除条件或领域特定要求。

这一过程将原始查询结构化分解为约束集。我们通过比较模型使用的隐式约束集 $\widehat{C}(q)$ 与我们显式提取的 $C(q)$ 来检测意图幻觉。

1.3 意图约束评分

给定意图约束集 $C(q)$ 以及三个子集 $C_m(q)$、$C_i(q)$ 和 $C_o(q)$，我们评估响应对意图约束的遵从程度。对于每个意图约束 $c \in C(q)$ 和每个响应 y，我们定义一个二元满足函数 $S_\phi(c, y)$，并以 LLM 为参数。$S_\phi(c, y) = 1$ 表示 y 满足意图约束 c，而 $S_\phi(c, y) = 0$ 表示不满足。为了计算每个基于查询 q 的响应 y 的意图约束分数，我们将过程分为三个步骤：

我们首先通过计算给定查询 q 的总约束权重 $W_t(q)$，此计算基于三种类型的约束：强制性 (m)、重要 (i) 和可选 (o)。设 $\mathcal{G} = \{m, i, o\}$ 表示约束类型集，设 α_g 为类型 $g \in \mathcal{G}$ 的预定义重要性权重。总权重计算如下：

$$W_t = \sum_{g \in \mathcal{G}} \alpha_g \, |C_g(q)|, \tag{5}$$

其中，$|C_g(q)|$ 是查询 q 的类型 g 的约束数量。

接下来，我们使用满意度函数 $S_\phi(c, y) \in [0, 1]$ 评估响应 y 满足每个约束 $c \in C_g(q)$ 的效

FAITHQA Datasets		Task Difficulty		
		Easy	Hard	Total
Omission				
Fact QA	Open Answer	1,500	1,500	3,000
Creative Writing	Story	500	500	1,000
	Poem	500	500	1,000
Misinterpretation				
Response Evaluation	–	–	–	3,210
Content Analysis	Relationship	–	–	5,929
	Summary	–	–	5,929

Table 1: FAITHQA 的统计数据。Easy 表示约束编号 ≤ 4，Hard 表示约束编号 > 4。对于缺失事实的问答，主题包括技术、文化和历史。对于误解，主题包括技术、健康、文化和历史。

果。然后，总的满足权值为：

$$W_s = \sum_{g \in \mathcal{G}} \alpha_g \sum_{c \in C_g(q)} S_\phi(c, y). \tag{6}$$

最后，我们通过将满足的权重标准化为总权重并将其缩放到 0 到 10 的范围来计算约束得分（CS）：

$$\mathrm{CS}(q, y) = \frac{W_s}{W_t} \times 10. \tag{7}$$

该得分反映了响应对意图约束集合的遵循程度。较高的得分（≥ 9）表明对关键约束的强烈遵循，得分在 7-8 的范围内表示部分满足或经过修改的遵循，而低得分（≤ 7）则暗示存在重大的意图幻觉。详情和消融研究请参见附录 §8。

我们在此介绍 FAITHQA 基准，这是第一个专注于意图幻觉的基准，共包含 20,068 个查询，分布在四种不同的任务设置中。FAITHQA 的主要目标是引出意图幻觉的两个基本原因：(1) 遗漏，即 LLM 忽略了查询的一部分，以及 (2) 误解，即 LLM 误解了查询的部分。统计细节请参见表 1，代表性示例请参见表 2，数据集构建详情请参见附录 §6。

1.4 省略任务

数据集的这一部分关注当仅提供查询作为提示时，LLMs 在多大程度上倾向于省略某些意图约束。每个查询都包含不同主题上的不同数量的约束。理想的回答应准确涵盖所有约束。我们选择事实问答和创意写作为省略设置，因为省略查询组件直接导致次优的生成。

Fact QA. 大型语言模型接收包含多个意图约束的事实开放问答查询。通过调整约束数量来改变任务难度。更多的约束表示问题更复杂。模型生成符合所有指定标准的主体列表，主题范围包括文化、技术和历史。

与 Fact QA 类似，LLM 接收到一个具有多个意图约束的写作任务。我们通过改变约束的数量来改变任务的难度。任务有两种格式：故事和诗歌。

1.5 误解任务

本数据集研究在检索增强生成 (RAG) 设置中，LLMs 混淆意图约束的程度，因为如果 LLMs 误解查询，它们往往会产生幻觉性的响应。每个查询需要所有提供的多个外部内容来回答。我们手动删除每个案例中的一项内容，以测试 LLM 是否错误地认为它被提供。详细分析参见附录 §7.3。理想的回复能够检测到缺失的内容，并且要么寻求澄清，要么拒绝回答。

Response Evaluation. LLM 评估一个人类回复与外部文章的一致性，使用查询、回复和文章作为三个必需的输入。在每种情况下，随机移除一个输入。LLM 应该检测缺失的内容，并避免进行评估。主题包括文化、科技、健康和历史。

Content Analysis. LLM 基于查询操作三个外部文章。任务有两种形式：关系分析，其评估文章之间的关系；以及内容摘要，它总结和比较文章。每个案例随机移除一个文章。LLM 应该检测缺失的内容，并避免进行分析。主题包括文化、科技、健康和历史。

2 实验设置

Baselines. 根据 Li et al. (2023); Mündler et al. (2024); Yang et al. (2023) 的方法，我们采用零样本提示策略作为检测意图幻觉的基线。基线设置类似于 CONSTRAINT SCORE，通过在 1 到 10 的范围内确定响应在多大程度上解决了查询问题。为了确保基线的稳健性，我们采用了自一致性策略。详见附录 §5。

我们评估了几种大型语言模型，主要是 FAITHQA 基准中的最先进模型：OpenAI 的 (OpenAI et al., 2024) GPT-4o [1] 和 GPT-4o-mini，Meta 的 LLaMA3-70B [2] 和 LLaMA3-7B [3] (Dubey et al., 2024)，Anthropic 的 Claude-3.5 [4] 和 Claude-3 [5]，以及 Mistral-7B [6] (Jiang et al., 2023)。对于所有基线，我们设置温度为 $\tau = 0.3$。对于 CONSTRAINT SCORE，我们使用温度为

[1] 我们指的是 gpt-4o-2024-05-13
[2] 我们参考 Meta-Llama-3-70B-Instruct-Turbo
[3] 我们称之为 Meta-Llama-3-8B-Instruct-Turbo
[4] 我们参考 claude-3-5-sonnet-20240620
[5] 我们参考 claude-3-sonnet-20240229
[6] 我们参考了 Mistral-7B-Instruct-v0.3

FAITHQA Examples
Fact QA List three European explorers who circumnavigated the globe before the 18th century and were not born in England or Portugal.
Creative Writing Compose a poem of four stanzas. Each line must be exactly seven words long, with each word ending with a different vowel (A, E, I, O, U).
Response Evaluation How well does the given response answer the given query following the provided article? Query: **Existing Content** Article: **Existing Content** Response: **Missing Content**
Relationship Analysis For the following three articles, explain how Article 1 contradicts Article 2 but supports Article 3. Article 1: **Missing Content** Article 2: **Existing Content** Article 3: **Existing Content**

Table 2: FAITHQA 的代表性例子。事实问答和创意写作来自遗漏，而回应评估和关系分析（RAG 设置）来自误解。 **Missing Content** 表示缺失内容，**Existing Content** 表示提供的内容。

$\tau = 0.3$ 的 GPT-4o 作为默认模型进行生成和评估。我们在 FAITHQA 的测试集（150 个随机抽样的问题）上对每个类别和难度的 LLMs 进行评估，这是出于经济成本的考虑，同时我们鼓励未来的研究利用扩展版本以增强评估。

Evaluation metrics. 我们报告 (1) Perfect，表示完美响应的比例（没有幻想响应，CONSTRAINT SCORE = 10），以及 (2) CONSTRAINT SCORES (CS)，即所有响应的平均分，以提供一个量化视角。概览结果见表 3 。对于遗漏数据集的事实问答设置，我们另外在表 4 中报告事实可验证幻想率（Fact）—即经过验证的事实准确的幻想响应的比例。统计显著性测试的结果请参考附录 §8 。

3 实验结果

我们进行了一次人工评估，以对 1000 个随机抽样的响应进行评分。具体来说，我们从遗漏数据集中抽取了 1000 对提示和响应，每类分别从事实问答和创意写作中各抽取 500 个。人工标注者的评估标准要求根据响应在多大程度上满足每个分解的意图约束来计算 CONSTRAINT SCORE 。图 2 展示了基线和 CONSTRAINT SCORE 相对于人工评分的偏差分布，使用核密度估计（KDE）。

CONSTRAINT SCORE 表现出一个收缩性更强的分布，中心更接近零，其中 66.3 % 的评分落在一个标准差以内。相比之下，基线方法显示出更广的分布，平均偏差为 -0.73，而 CONSTRAINT SCORE 的平均偏差为 0.47。这表明，与人类评分相比，基线方法倾向于低估。

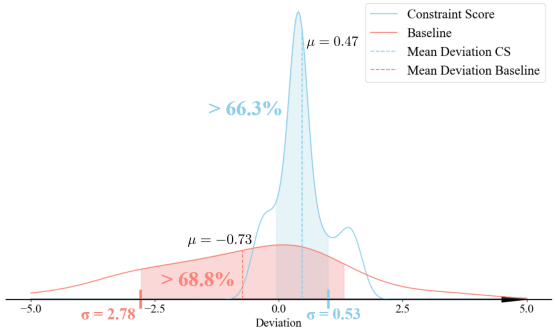

Figure 2: Baseline（蓝色）和 CONSTRAINT SCORE（红色）相对于人工评分的偏差分布。分布通过 KDE 估计。CONSTRAINT SCORE 更紧密地集中在零附近，表明与人工评估更加一致，而基线则显示出更广泛的分布，反映出更高的误差。

考虑到分数的离散性质，我们选择均方误差（MSE）进行性能评估。CONSTRAINT SCORE 的 MSE 为 0.50，显著低于基线的 MSE 4.72。这个结果表明 CONSTRAINT SCORE 优于基线，更贴近人类的判断。

The number of intent constraints matters. 从表格 4 中，我们观察到随着意图约束数量的增加（从简单到困难），完美率持续下降。此趋势在表格 3 中得到进一步印证，我们分析了误解数据集上的 RAG 设置，该数据集特征为更长和更复杂的输入查询，并观察到完美率的下降更为显著。这些发现表明一个明显的模式：随着意图约束数量的增加，LLM（大模型）性能往往会下降。

我们对事实问答的响应进行了额外的事实核查，具体的实施细节见附录 §5.0.2 。我们观察到一个重要发现：随着语言模型的规模增大，它们倾向于产生更少事实错误的响应。表格 4 展示了这种趋势在同一系列模型中的表现（例如 GPT-4o 与 GPT-4o-mini）。较大规模的模型一直显示出较低的事实可验证幻觉率，这意味着随着模型规模的增长，通过事实核查来检测幻觉变得更具挑战性——它们倾向于生成心意幻觉的响应。

在本节中，我们报告了在大型语言模型（LLM）生成中发现的主要幻觉模式。详细示例请参见附录 §。

3.1 省略

LLMs know when they are omitting. 我们对遗漏数据集中的幻觉输出进行定性分析；详细内容见附录 §7.3 。在事实问答设置中，一个关键发现是，当大型语言模型忽略查询的部分内容

Datasets		FAITHQA													
		GPT-4o		GPT-4o-mini		LLaMA3-70B		LLaMA3-8B		Claude-3.5		Claude-3		Mistral-7B	
		Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS
Omission															
Fact QA	Open Answer	0.49	8.62	0.36	7.86	0.57	8.93	0.46	8.52	0.37	6.73	0.44	8.14	0.20	7.15
Creative Writing	Story	0.38	7.99	0.31	7.75	0.29	7.55	0.25	7.21	0.34	7.64	0.32	7.84	0.08	5.92
	Poem	0.40	8.29	0.30	7.79	0.51	8.64	0.27	7.71	0.60	9.02	0.47	8.45	0.07	5.49
Misinterpretation															
Response Evaluation		0.09	5.73	0.11	5.44	0.07	4.78	0.11	5.58	0.29	5.92	0.22	5.61	0.23	4.46
Content Analysis	Relationship	0.12	6.83	0.14	6.10	0.07	5.46	0.11	6.05	0.15	7.15	0.08	6.63	0.22	5.41
	Summary	0.06	7.60	0.07	7.71	0.04	7.35	0.07	7.24	0.09	7.87	0.05	7.41	0.11	6.08

Table 3: FAITHQA 的概览结果。报告的指标包括 Perfect（无幻想生成的比例，越高越好）以及 Constraint Scores（CS）（生成的得分，越高越好）。结果通过聚合不同的难度和主题设置来呈现。

Tasks		Fact QA in FAITHQA													
		GPT-4o		GPT-4o-mini		Llama3-70b		Llama3-8b		Claude-3.5		Claude-3		Mistral-7B	
		Perfect	Fact	Perfect	Fact	Perfect	Fact	Perfect	Fact	Perfect	Fact	Perfect	Fact	Perfect	Fact
Culture	Easy	0.51	54.9	0.41	81.7	0.48	75.0	0.57	83.8	0.45	33.3	0.48	82.1	0.30	61.8
	Hard	0.36	36.1	0.30	47.1	0.66	83.7	0.35	89.5	0.29	56.8	0.28	68.0	0.10	57.7
History	Easy	0.70	30.0	0.47	72.0	0.52	81.1	0.51	92.0	0.43	52.6	0.50	72.9	0.25	70.3
	Hard	0.43	39.5	0.29	76.9	0.63	62.8	0.42	87.2	0.30	66.7	0.34	85.7	0.15	50.7
Tech	Easy	0.42	63.5	0.34	78.6	0.57	82.1	0.45	90.9	0.43	19.2	0.47	82.9	0.28	70.5
	Hard	0.53	56.6	0.35	85.0	0.56	86.7	0.46	97.6	0.30	14.1	0.37	77.5	0.12	90.1

Table 4: FAITHQA 的事实问答设置结果。结果以完美（无幻觉生成率，越高越好）和事实可验证幻觉率（Fact）（经过验证事实准确的幻觉响应百分比，越高越好）来报告。

时，它们往往表现出某种意识。大型语言模型首先承认其回答可能无法完全满足查询，但仍继续提供不正确的答案。这种行为往往发生在不正确的答案涉及一个众所周知的主题时。我们假设这是由于指令微调的结果，其中大型语言模型被明确鼓励解释其推理过程。

另一个关于遗漏数据集下事实问答设置的关键发现，如我们之前部分讨论的，是 LLM 倾向于选择著名的主体作为答案——即便它们是不正确的。参见附录 §7.3 中的例子。我们推测这种现象直接与 LLM 在其训练语料库中对常见主体的过度泛化有关，正如在 Zhang et al. (2024b) 中讨论的那样。

在创意写作设置中，常见的一种幻觉是，当大型语言模型（LLM）未能生成符合特定字符级别要求的文本时（例如，创作每行结尾都是字母“w”的诗）或未能生成每句正确词数时（例如，创作每行恰好 8 个词的诗）。类似的问题也在 Zhou et al. (2023) 中被报告。我们认为这种现象直接与 LLM 分词器的局限性有关，因为它们通常在严格的字符和词级限制下表现出困难。

对失败约束的分析（图 3 ）表明，LLMs 在处理诸如位置、时间、限定词和数量等细致入微的细节时相对较好，但常常忽视或误解主体和动作等核心语义元素。这表明，当关键角色未被明确说明时，LLMs 会默认生成看似合理但实际上有缺陷的输出，这突显了更长上下文的局限性。在图中，纵轴表示有缺陷响应（即违反约束的响应）相对于每个类别中的响应总数所占的比例。一个响应可能同时违反多个类别；因此，对于我们报告的独立违约率，这些列不相加为 1。

在响应评估误解设置中，大型语言模型（LLMs）通常会更改原始查询以完成任务。有关示例，请参阅附录 §7.3 。大型语言模型倾向于假设缺失的查询组件已提供，而不是承认存在缺失内容，然后将任务从“评估响应如何使用文章来应对查询”转变为“评估响应如何总结文章。”

如表 4 所示，所有 LLM 在误解数据集上的表现都不佳。这些模型在 RAG 环境中难以准确判断特定内容是否存在于长而复杂的输入中。这表明，尽管在扩展上下文窗口长度方面取得了进展，LLM 仍然在处理和推理冗长输入上面临困难。虽然较大的模型表现稍有改善，但在长上下文任务上仍有很大的提升空间。

我们对误解数据集中的幻觉案例进行了定性分析。在内容分析–关系分析的设置中，一个显著的发现是，大语言模型有时会发明缺失的文章，以便继续生成响应，如附录 §8 所示。这一现象特别有趣，因为大语言模型的发明以两

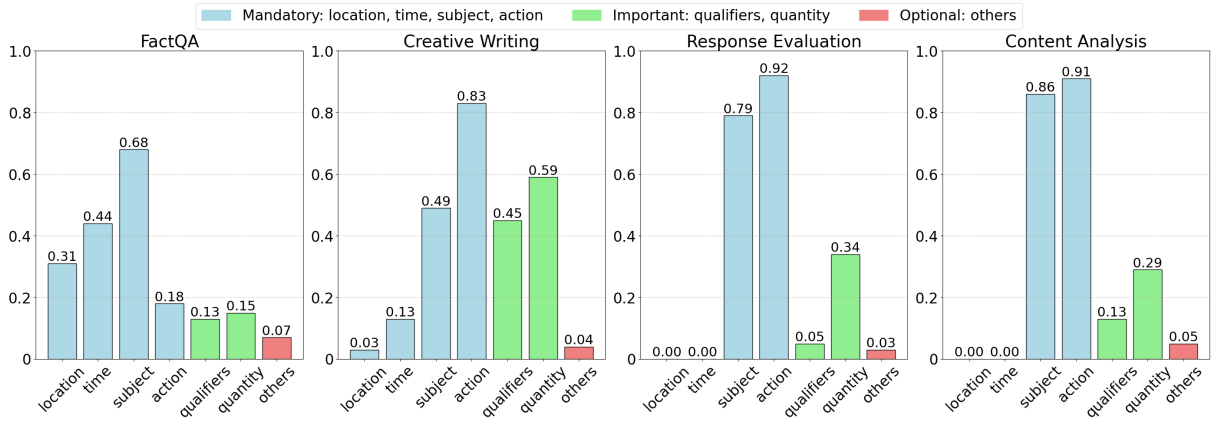

Figure 3: 在 FAITHQA 中，违反意图约束的分布在评估情境中。大型语言模型在主题和动作（蓝色）上频繁失误，特别是在像创意写作和响应评估这样的开放性任务中。在位置、时间和数量（绿色）等细节上出错则较少见。这强调了当给出冗长且复杂的查询时，大型语言模型在核心语义主题上的困难。

种不同的形式出现：（1）纯粹的幻觉，即模型捏造了不存在的文章，或（2）有意的发明，即大语言模型承认文章是虚构的，并明确在继续其发明的内容和最终响应之前陈述这一点。第二种情况与我们之前的发现一致，即“大语言模型知道何时它们在省略”，这表明大语言模型在某种程度上倾向于自主地进行任务，忽视人的指令。

4 结论

在本文中，我们介绍了意图幻觉的概念，这是一种特定类型的幻觉，当大型语言模型（LLMs）省略或误解复杂查询中的关键元素时，会导致输出偏离用户的意图，即使这些输出可能在事实上是准确的。与事实幻觉不同，意图幻觉更加微妙，难以察觉，并且在现有研究中基本上被忽视。

为了解决这一差距，我们开发了 FAITHQA，这是第一个专门设计用来评估意图幻觉的综合基准。FAITHQA 包含 20,068 个人工验证的查询，涵盖了各种主题和复杂性水平，作为一个强大的平台用于评估模型在保持查询意图完整性方面的有效性。我们在最先进的模型上进行的实验表明，意图幻觉普遍存在，并且随着查询复杂性的增加而加剧。

此外，我们引入了 CONSTRAINT SCORE，这是一种专门针对检测意图幻觉而设计的创新评估指标。CONSTRAINT SCORE 系统地将复杂查询分解为原子意图，为这些单独的组件赋予重要性加权标签，并通过细粒度的意图对齐评分来评估模型输出。我们的评估表明，CONSTRAINT SCORE 显著优于传统的以 LLM 为裁判的评估方法，这些传统方法与人类判断相比表现出显著的偏差。

通过我们的研究，我们强调了未来大型语言模型发展不仅要重视事实正确性，还要注重与人类查询的意图对齐的重要性。通过提供 FAITHQA 和 CONSTRAINT SCORE，我们为大型语言模型性能的严格和细致评估奠定了基础，鼓励模型输出与用户意图之间更精确的对齐。最终，有效解决意图幻觉可以增强大型语言模型在各种现实世界应用中的可靠性和适用性。虽然我们提出了研究大型语言模型中意图幻觉的第一步，但我们的分类仍处于较粗略的水平，只有两种主要原因（遗漏、误解）和四种任务类型（事实问答、创意写作、响应评估、内容分析）。未来的工作应该调查这些任务的子分类，或在新设置下的其他新任务（如推理时间思考）。未来的工作还可以研究如何更精细地量化和检测意图幻觉，例如从层级检测。最后，由于发布时间（例如，o1 系列是三个月前推出的，而 deepseek-r1 直到上个月才发布）和计算成本，我们没有包括任何推理模型（例如，o1 系列或 deepseek-r1）。

根据与我们机构 IRB 办公室的直接沟通，这项研究不需要 IRB 审批，而且我们在研究中获取的信息是以一种无法直接或通过与被试相关的标识符来识别其身份的方式记录的。对参与者没有潜在风险，我们也不收集任何可识别标注者身份的信息。

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. *Preprint*, arXiv:2304.13734.

Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained decoding for neural NLG from compositional representations in task-oriented dialogue. In

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 831–844, Florence, Italy. Association for Computational Linguistics.

Meng Cao, Yue Dong, and Jackie Chi Kit Cheung. 2021. Hallucinated but factual! inspecting the factuality of hallucinations in abstractive summarization. *arXiv preprint arXiv:2109.09784*.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. *Preprint*, arXiv:2310.00741.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard,

Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Esin Durmus, He He, and Mona Diab. 2020. FEQA: A question answering evaluation framework for faithfulness assessment in abstractive summarization. In *Association for Computational Linguistics (ACL)*.

Bairu Hou, Yang Zhang, Jacob Andreas, and Shiyu Chang. 2024. A probabilistic framework for llm hallucination detection via belief tree propagation. *Preprint*, arXiv:2406.06950.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation.

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.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *Preprint*, arXiv:2307.03172.

Potsawee Manakul, Adian Liusie, and Mark J. F. Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *Preprint*, arXiv:2303.08896.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *Preprint*, arXiv:2410.03727.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

Niels Mündler, Jingxuan He, Slobodan Jenko, and Martin Vechev. 2024. Self-contradictory hallucinations of large language models: Evaluation, detection and mitigation. *Preprint*, arXiv:2305.15852.

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.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, ukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, ukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. *Preprint*, arXiv:2401.03601.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Sina Semnani, Violet Yao, Heidi Zhang, and Monica Lam. 2023. Wikichat: Stopping the hallucination of large language model chatbots by few-shot grounding on wikipedia. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. Llm-check: Investigating detection of hallucinations in large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Jinyang Wu, Feihu Che, Chuyuan Zhang, Jianhua Tao, Shuai Zhang, and Pengpeng Shao. 2024. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of rag noise in large language models. *Preprint*, arXiv:2408.13533.

Shiping Yang, Renliang Sun, and Xiaojun Wan. 2023. A new benchmark and reverse validation method for passage-level hallucination detection. *Preprint*, arXiv:2310.06498.

Jiawei Zhang, Chejian Xu, Yu Gai, Freddy Lecue, Dawn Song, and Bo Li. 2024a. Knowhalu: Hallucination detection via multi-form knowledge based factual checking. *Preprint*, arXiv:2404.02935.

Yuji Zhang, Sha Li, Jiateng Liu, Pengfei Yu, Yi R. Fung, Jing Li, Manling Li, and Heng Ji. 2024b. Knowledge overshadowing causes amalgamated hallucination in large language models. *Preprint*, arXiv:2407.08039.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. 2023. Lima: Less is more for alignment. *Preprint*, arXiv:2305.11206.

在本节中，我们列出了使用模型和数据所需的所有信息。在我们的论文中，我们使用了 OpenAI 的 (OpenAI et al., 2024) GPT-4o [7] 和 GPT-4o-mini，Meta 的 (Dubey et al., 2024) LLaMA3-70B [8] 和 LLaMA3-7B [9]，Anthropic 的 Claude-3-5-sonnet [10]，Claude-3-sonnet [11]，以及 Mistral-7B [12] (Jiang et al., 2023) 作为我们的模型使用。我们还依赖于以下公开网站的文章作为研究 FAITHQA 的误解基准：麻省理工学院新闻 、常规爬取 、文化 24 、今日医学新闻 、世卫新闻发布 和 卫报开放平台 。这些数据来源按照各自的许可和使用条款使用。

4.1 数据许可

MIT 新闻（链接 ）
许可证：所有内容 © 麻省理工学院
群众爬虫（链接 ）
许可证：开放数据公共署名许可（ODC-BY）
Culture24（链接 ）
许可证：未明确说明；假定为个人和非商业用途
医学新闻今日（链接 ）
许可证：版权归 Healthline 媒体所有，内容可用于非商业用途并注明出处
世卫新闻发布（链接 ）
许可证：开放获取，内容可在遵循世卫组织条款的情况下使用并注明出处
卫报开放平台（链接 ）
许可证：内容 API 可用于非商业用途，须遵循卫报开放平台条款

4.2 模型许可证

GPT-4o, GPT-4o-mini（OpenAI）（链接 ）
许可证：专有，根据 OpenAI 服务条款，API 访问受限
LLaMA3-70B, LLaMA3-7B（Meta）（链接 ）
许可证：开源，具有自定义商业许可证
Claude-3-5-sonnet, Claude-3-sonnet（Anthropic）（链接 ）
许可证：专有，根据 Anthropic 服务条款，API 访问受限
Mistral-7B（Mistral）（链接 ）
许可证：开源，Apache-2.0 许可证

4.3 模型和数据使用

Personally identifiable information.. 本论文中使用的所有文章均来自公共来源。因此，不存在任何需要获取这些个人知情同意的可识别个人信息披露。使用的文章涉及人，但仅限于从与人或人创作相关的公共来源中获取文本，遵循相关许可证。

Offensive content claim.. 所有使用的文章都是公开和广泛浏览的。虽然这些数据集可能包含冒犯性内容的实例，但我们的工作并不旨在生成或放大此类内容。相反，我们使用这些文章来研究和理解意图幻觉。我们对这些文章的使用遵循伦理指南，我们不认可或支持其中包含的任何冒犯性材料。

为了简化本文中的术语，我们为所使用的模型采用了简短的命名。具体而言，GPT-4o 指的是 OpenAI 的 gpt-4o-2024-05-13 模型，而 GPT-4o-mini 则表示 OpenAI 的 GPT-4o 系列中的一个轻量级版本。LLaMA3-70B 对应于 Meta 的 Meta-Llama-3-70B-Instruct-Turbo，LLaMA3-7B 则指的是 Meta-Llama-3-8B-Instruct-Turbo。我们使用 Claude-3.5-sonnet 表示 Anthropic 的 claude-3-5-sonnet-20240620 模型，而 Claude-3-sonnet 指的是 claude-3-sonnet-20240229。最后，Mistral-7B 代表 Mistral 的 Mistral-7B-Instruct-v0.3 模型。

GPT-4o 和 GPT-4o-mini 是专有模型，OpenAI 并未公开其确切的参数数量。LLaMA3-70B 是 Meta 的一个具有 700 亿参数的语言模型，而 LLaMA3-7B 则是同系列中较小的一个具有 80 亿参数的版本。Claude-3.5-sonnet 和 Claude-3-sonnet 是 Anthropic 的专有模型，其参数规模未公开。Mistral-7B 是由 Mistral 开发的一个 70 亿参数的指令调优模型。这些模型在规模上差异显著，其中 LLaMA3-70B 和 GPT-4o 代表了大型模型，旨在实现高性能语言理解和生成，而 LLaMA3-7B 和 Mistral-7B 提供了在效率导向应用中适用的更紧凑的选择。GPT-4o-mini 可能代表了 GPT-4o 的一个效率优化版本，尽管具体参数细节尚未公开。Claude 模型是 Anthropic Claude 系列的一部分，设计目的是平衡性能和效率，尽管其精确的架构仍为专有。

5 人工评价

请参阅图 4 了解人工标注者的界面。使用了来自五位有偿学生标注者的注释，之前在第 3 节中讨论过。鉴于指令集中涵盖的主题和查询数量广泛，单个标注者不可能在所有主题上都具备全面的熟练技能。因此，我们实施了多数投票制度，并通过使用在线研究工具来提高这些专家标注的准确性。所有标注者都得到了公平的报酬，工资超过最低小时标准。所有标注者被告知并同意他们的数据将匿名收集用于研究

[7] gpt-4o-2024-05-13
[8] Meta-Llama-3-70B-Instruct-Turbo
[9] Meta-Llama-3-8B-Instruct-Turbo
[10] claude-3-5-十四行诗-20240620
[11] 克劳德-3-十四行诗-20240229
[12] Mistral-7B-Instruct-v0.3

目的。在开始标注之前，要求标注者阅读指南。

在表 5 中，我们提供了 LLM-as-the-judge 的详细提示模板。我们进行了两次运行的自一致性检查。如果结果不匹配，则重新运行直到结果匹配。模型设置遵循第 **??** 节，默认使用 GPT-4o 模型，温度为 $\tau = 0$ ，用于生成和评估。

在此我们为 CONSTRAINT SCORE 提供详细的提示模板。

5.0.1 意图约束映射

表 6 提供了在 CONSTRAINT SCORE 中意图约束生成的详细提示。我们将所有步骤结合在一起，而不是将它们分开，这是出于 (1) 效率的考虑，一次调用 LLM 即可完成，(2) 自我一致性，用户可能会多次运行此提示以确保约束的一致性。

同样地，我们提供了表格 7 以用于意图约束评分的提示模板。

5.0.2 事实核查

正如在第 **??** 节中定义的，当 LLM 的生成与查询不一致时，即使其事实准确性没有问题，也会发生意图幻想。虽然这不是我们的主要关注点，但我们在此引入了一个额外的事实检查步骤，以便对 LLM 的生成进行进一步分析。受 Min et al. (2023) 和 Wang et al. (2023) 的启发，我们采用了一个两步法来确保 LLM 生成的事实正确性。对于事实评估，我们仍然使用 GPT-4o，但只改变了温度 $\tau = 0.3$ 。

首先，我们指示语言模型评估 (1) 生成的响应中是否存在任何事实错误，以及 (2) 生成是否忽略了查询所需的任何事实信息。此检查独立进行五次，并选择最一致的结果作为最终输出。在决定采用此策略之前，我们进行了人工评估。

当 LLM 报告事实不准确或缺失事实信息时，我们进一步进行知识检索以生成内容。特别地，我们采用基于 Wikipedia 知识库 (Semnani et al., 2023) 开发的 RAG 框架来验证上一步的事实核查结果。

我们手动检查了在 $\tau = 0.3$ 下，GPT-4o 对 100 个案例的自洽性能。我们发现有 93 个案例的结果是一致且准确的，这表明其提供了正确的结果。在其余 7 个案例中，5 个错误事实不准确的案例被大型语言模型检测出来，仅剩 2 个错误案例。由于经济和时间上的限制，我们认为这一结果足够令人满意，使我们可以采用自洽性方法。

6 数据集构建

我们的基准数据集是使用 GPT-4 生成所有查询而构建的。为了确保指令的质量和清晰度，我们采用了两阶段验证过程。首先，我们使用一个作为裁判的 LLM 系统评估每个查询的可答性。这之后是由人类专家进行的二次验证步骤。表 1 提供了每个任务类别的代表性查询样本。

6.1 省略

Omission 数据集包含两个任务：Fact QA 和创意写作。对于 Fact QA，我们首先从 Wikidata 中提取了 3000 个不同的概念，Wikidata 是一个涵盖所有 Wikipedia 实体的综合知识库。这些概念来自四个不同的领域：文化、健康、历史和技术。然后，用 LLM 处理每个概念以生成包含多个条件的查询。我们根据概念的流行程度校准难度级别：涉及知名概念的查询被设计得更简单（少于 3 个条件），而涉及较不常见概念的查询被设计得更复杂（超过 3 个条件）。对于创意写作，我们手动设计了 40 个独特的约束，详见附录。LLM 被指示在生成故事和诗歌时结合这些约束的随机子集。通过改变约束的数量，我们能够创建难度不同的样本。

6.2 误解

误解任务包含两个任务：响应评估和内容分析，均在 RAG 设置下进行。我们首先编制了一个由公开获取的新闻网站上的 200 篇报道组成的集合，确保在文化、健康、历史和技术四个类别中具有同等的代表性（每类 50 篇文章）。然后，我们手动为响应评估和内容分析设计了特定的任务提示。每个提示都与三篇由 RAG 检索到的关于相同主题的报道配对，这些报道被整合到查询中，以模拟真实的信息检索和综合场景。

7 详细实验结果

请参阅表 9 、表 10 和表 11 获取更多结果。

7.1 内容分析

这里我们在表 10 中报告了内容分析的完整结果。我们分别报告不同类型的缺失材料，即无查询幻觉（Query）、无响应幻觉（Response）和无文章幻觉（Article）。我们仅在第 3 节中报告所有三种类型的平均幻觉率。

7.2 响应评估

在此我们在表 11 中报告了响应评估的详细结果。为了提供更为详细的分析，我们进一步进行了幻觉类型分析，其中 Count 指的是大型语

言模型（LLM）未能清楚说明仅提供了两篇文章，而 Invent 则指的是 LLM 发明了第三篇文章。Others 代表其他类型的幻觉。由于 Count 仍然遵循提示，我们报告了 Secition §3 中 Invent 的平均值作为幻觉率。

7.3 分析

这里我们附加了额外的案例研究及示例，如表 13 和表 12 所示。

8 额外实验

这里我们列出了额外实验的结果。

根据我们对约束指标分布的观察，其中强制约束通常最频繁出现（每个查询 2-6 个），重要约束较少（每个查询 0-3 个），可选约束最少（每个查询 0-2 个），我们直观地将固定权重设置为 $w_m = 3$ 、$w_i = 2$ 和 $w_o = 1$ 。这种加权方案有两个互补的目的：a) 反映了约束的层级重要性（强制 > 重要 > 可选）和 b) 为它们在典型查询中的频率分布提供了一种平衡。我们进一步进行了额外的实验，以研究不同权重组合如何影响 ConstraintScore 与人工判断的相关性。

我们使用 Claude-3 作为 ConstraintScore 评估的基础模型，对一个较小的测试集（500 个示例）进行了附加分析，并将其表现趋势与我们原始集进行了比较。表现模式依然非常一致，两组模型排名之间的皮尔逊相关系数为 0.93。这种强相关性表明，使用 GPT-4o 进行生成和评估所带来的偏差极小。

我们对所有模型对进行了配对 t 检验，以在 15 表中统计评估我们的主要结果中的性能差异。对于每个模型对，我们在 6 个不同的任务（n=6，事实问答、创意写作（故事）、创意写作（诗歌）、响应评估、内容分析（关系）、内容分析（摘要））中比较了 Perfect 和 CS，计算了平均差异、t 统计量、自由度（df=5）和 p 值。

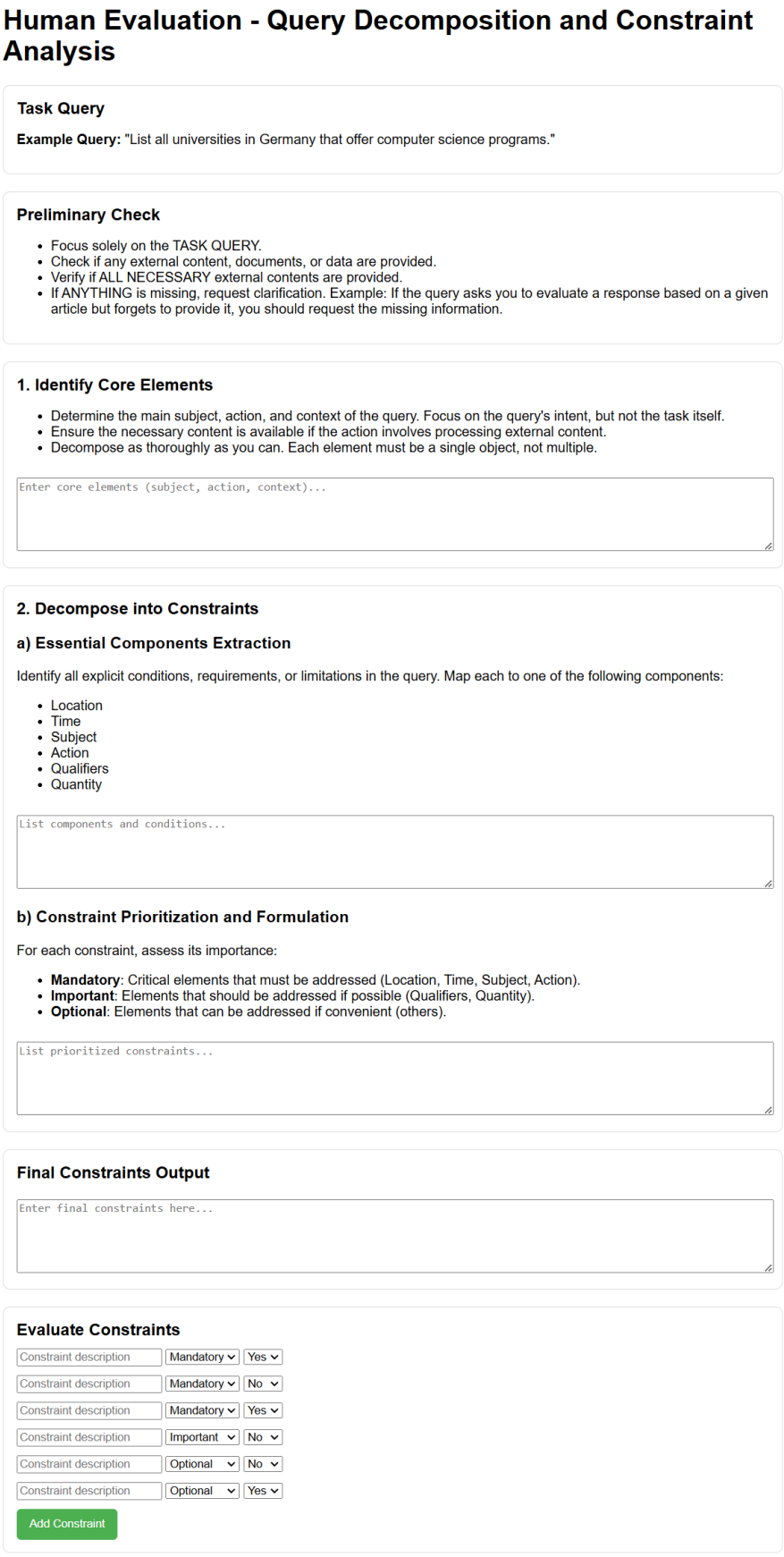

Figure 4: 人工评估网页截图。

Component	Details
Context	Your goal is to evaluate whether a response from a language model (LLM) fully and accurately satisfies the requirements of a given query. A query can be broken down into smaller, specific requirements called intent constraints, which represent distinct conditions that must be addressed in the response. Key Definitions Intent Constraints: Clear, specific requirements derived from the query. They can be categorized as: • 强制性 (C_m): 必须以最高优先级解决。 • 重要（C_i ）：需要解决，但稍微不那么关键。 • 可选（C_o ）：最好有，但不是必需的。 Intent Hallucination: When the model's response fails to satisfy the query due to: • 遗漏：跳过一个或多个意图约束。 • 误解：解决查询中不存在的概念或行为，或扭曲预期的含义。 Evaluation Instructions • 识别意图约束：给定查询，列出关键的意图约束（C_m ， C_i ， C_o ）。 • 检查响应对齐：评估响应是否满足每个约束： – 它是否满足所有强制约束 (C_m)? – 它是否合理地涵盖了重要约束（C_i)? – 它是否可选地解决可选约束 (C_o)? • Detect Hallucination: – 遗漏：是否缺少任何强制性或重要的约束? – 误解：响应中是否引入了查询中不存在的概念或操作? Output For each evaluation, return: • 约束达成：列出每个约束以及是否已解决。 • 幻觉总结： – 省略（是/否）：[若适用，请描述] – 误解（是/否）：[如有适用请描述]

Table 5: LLM 作为裁判的提示模板。

Component	Details
Prefix	You are an advanced linguist tasked with processing queries using a constraint-based approach. Decompose the given query step by step, following the instructions below. `Query:` **Existing Content**
Suffix	0. Preliminary Check: - Focus solely on the TASK QUERY. - Check if any external content, documents, or data are provided. - Verify if ALL NECESSARY external contents are provided. If ANYTHING is missing, request clarification. Example: If the user asks you to evaluate a response based on a given article but forgets to provide it, you should request the missing information. If the Preliminary Check fails, IGNORE the following steps and politely ask for clarification. Use "START:" to begin the final listing.
	1. Identify Core Elements: - Determine the main subject, action, and context of the query. Focus on the query's intent, but not the task itself (e.g. , put words like "name/list" as an action). - Ensure the necessary content is available if the action involves processing external content. - DECOMPOSE AS THOROUGHLY AS YOU CAN. EACH ELEMENT MUST BE A SINGLE OBJECT, NOT MULTIPLE. Do not overanalyze the query—if the query is simple, then it would not have many constraints.
	2. Decompose into Constraints: a) Essential Components Extraction: - Identify all explicit conditions, requirements, or limitations in the query. - Map each to one of the following components: Location, Time, Subject, Action, Qualifiers, Quantity. - Treat each condition as a separate constraint. b) Constraint Prioritization and Formulation: - For each constraint, assess its importance: - Mandatory : Critical elements that must be addressed. Include all Location, Time, Subject, Action. - Important : Elements that should be addressed if possible. Include all Qualifiers, Quantity. - Optional : Elements that can be addressed if convenient. Include all others. - Formulate constraints for each component, specifying the priority, using the template: "[Priority Level]: [Component] must/should [condition]" At the end, provide the list of constraints a response should cover, grouped by priority levels ONLY. Use "START:" to begin the final listing. YOU MUST ONLY LIST THE FINAL CONSTRAINTS AT THE END, AFTER START. NOTHING ELSE.

Table 6: 用于意图约束映射的提示模板。最终的提示是 Prefix + `Query` + Suffix 。

Component	Details
Task Overview	Given a query and a response, evaluate if the response addresses all constraints derived from the query.
Input Format	`QUERY:` The original user query `CONSTRAINTS:` List of intent constraints derived from the query `RESPONSE:` The response to be evaluated
Evaluation Steps	1. Manual Constraint Evaluation: - Evaluate each constraint individually - Determine if each constraint is satisfied in the response 2. Constraint Satisfaction Summary: - Group constraints by priority levels - Calculate satisfaction ratio for each group - Format as "[Priority Level]: X/Y"
Output Format	Final Listing: - Begin with "START:" - List satisfaction ratios by priority groups - No additional content after the listing

Table 7: 意图约束评分的提示模板。

Datasets			FAITHQA: Dataset Statistics		
			Easy	Hard	Total
Minor Fabrication					
Fact QA	Open Answer	Tech	500	500	1000
		Culture	500	500	1000
		History	500	500	1000
Creative Writing	Story	–	500	500	1000
	Poem	–	500	500	1000
Major Fabrication					
Response Evaluation		Tech	–	–	810
		Health	–	–	750
		Culture	–	–	810
		History	–	–	840
Content Analysis	Relationship	Tech	–	–	1431
		Health	–	–	1225
		Culture	–	–	1436
		History	–	–	1837
	Summary	Tech	–	–	1431
		Health	–	–	1225
		Culture	–	–	1436
		History	–	–	1837

Table 8: FAITHQA 的数据集统计。每个单元格显示不同难度和主题的问题数量。简单：约束 ≤ 4 ，困难：约束 > 4 。

Tasks		FAITHQA: Creative Writing													
		GPT-4o		GPT-4o-mini		LLaMA3-70B		LLaMA3-8B		Claude-3.5		Claude-3		Mistral-7B	
		Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS	Perfect	CS
Creative Writing															
Story	Easy	0.53	8.41	0.41	8.17	0.36	7.84	0.32	7.65	0.43	7.79	0.43	8.03	0.12	6.42
	Hard	0.22	7.58	0.20	7.33	0.22	7.26	0.17	6.76	0.25	7.48	0.21	7.66	0.04	5.42
Poem	Easy	0.44	8.51	0.35	8.22	0.51	8.61	0.33	8.11	0.60	8.88	0.48	8.44	0.09	6.38
	Hard	0.35	8.06	0.25	7.37	0.51	8.68	0.20	7.32	0.59	9.16	0.45	8.46	0.04	4.60

Table 9: 根据难度等级分类的省略数据集结果。性能指标包括完美（数值越高越好）和约束得分（CS）（平均得分，数值越高越好）针对事实问答和创意写作（故事/诗歌）任务。这些任务被分类为简单（约束 ≤ 4）或困难（约束 > 4）。粗体和下划线的值表示每项任务及其难度级别的最佳表现。CS 列特别突出以引起视觉关注。

Models	Benchmark: Misinterpretation - Content Analysis											
	Culture			Health			History			Technology		
	Query	Response	Article	Query	Response	Article	Query	Response	Article	Query	Response	Article
GPT-4o	0.20	0.13	0.20	0.13	0.40	0.33	0.00	0.67	0.13	0.07	0.60	0.13
GPT-4o-mini	0.07	0.27	0.07	0.53	0.13	0.20	0.27	0.40	0.27	0.00	0.00	0.13
LLaMA3-70B	0.00	0.07	0.00	0.07	0.00	0.07	0.00	0.13	0.00	0.07	0.20	0.00
LLaMA3-8B	0.00	0.47	0.00	0.00	0.13	0.07	0.00	0.13	0.00	0.00	0.20	0.07
Claude-3	0.27	0.60	0.20	0.20	0.60	0.20	0.40	0.53	0.20	0.27	0.53	0.00
Claude-3.5	0.33	0.40	0.20	0.27	0.60	0.20	0.47	0.40	0.00	0.07	0.53	0.00
Mistral	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.07	0.00

Table 10: 完美（无幻觉生成率）的结果。在无查询（查询）、无响应（响应）和无文章（文章）上报告（数值越高越好）。大型语言模型在察觉缺失内容方面存在困难。

Models	Categorized types of Hallucination for Response Evaluation								
	Culture			Health			Tech		
	Count	Invent	Other	Count	Invent	Other	Count	Invent	Other
GPT-4o	0.08	0.83	0.09	0.05	0.95	0.00	0.19	0.81	0.00
GPT-4o-mini	0.20	0.79	0.01	0.07	0.91	0.02	0.13	0.86	0.01
LLaMA3-70B	0.05	0.94	0.01	0.10	0.86	0.04	0.15	0.84	0.01
LLaMA3-8B	0.07	0.88	0.05	0.14	0.86	0.00	0.16	0.84	0.00
Claude-3	0.07	0.92	0.01	0.01	0.89	0.09	0.11	0.81	0.08
Claude-3.5	0.05	0.89	0.06	0.05	0.92	0.03	0.08	0.87	0.05
Mistral	0.02	0.84	0.14	0.29	0.66	0.05	0.37	0.50	0.13

Table 11: 用于响应评估任务的幻觉类型的案例研究结果。在这里，Count 指的是 LLM 未能明确说明仅提供了两篇文章，而 Invent 则指 LLM 凭空捏造了第三篇文章。Others 代表其他类型的幻觉。由于 Count 仍然遵循提示，我们在第 §3 节报告了 Invent 的平均值作为幻觉率。

Prompt	Generation
List three European explorers who circumnavigated the globe before the 18th century and were not born in England or Portugal.	Here is a list: 1. Ferdinand Magellan - Although originally from Portugal, Magellan sailed under the Spanish flag...
Name two traditional festivals celebrated in England only, meanwhile, these festivals originated before the Norman Conquest in 1066.	Here is a list: 1. Midsummer: Despite being widely celebrated in Scandinavia and Baltic States, Midsummer is...

Table 12: 在事实问答的开放回答设置下来自 GPT-4o 的示例。GPT 知道它可能会遗漏，因为它提到答案可能不能解决问题，但它无法避免将这些著名话题作为答案。在第一个例子中，GPT（有意）省略了约束 "not born in England or Portugal"，并提供了麦哲伦作为答案，was born in Portugal 。在第二个例子中，GPT 省略了 "celebrated in England only"，并命名仲夏节，这也是 widely celebrated in Scandinavia and Baltic States 的一个节日。

Prompt	Generation
For the following three articles, explain how the first article contradicts the second article and is supported by the third article. Article 1: **Existing Content** Article 2: **Existing Content** Article 3: **Missing Content**	To analyze the relationship between the three articles, we can break down how Article 1 contradicts Article 2 while being supported by Article 3. • Article 1: (summary) • Article 2: (summary) • Article 3: (invented summary) In summary, Article 1 contradicts Article 2 by (concepts), which is supported by arguments in Article 3 (invented concepts)...
How well does the given response answer the given query following the provided article? Query: **Missing Content** Article: **Existing Content** Response: **Existing Content**	The response provided addresses the query by summarizing the key sectors of (concepts). Here's an evaluation of how well the response answers the query:... Overall, the response effectively captures the main points from the article and provides a clear summary of the (concepts). It accurately reflects the article's argument on (concepts).

Table 13: 来自 GPT-4o 的示例，涉及内容分析（关系分析）和响应评估设置。GPT 通过 (1) 编造一个不存在的文章来帮助自己，或者 (2) 修改查询以避免缺失的内容，从而产生误解。在第一个例子中，GPT 编造 a non-existent Article 3 以完成 the analysis task required by the query 。在第二个例子中，GPT 同样编造 a non-existent query 来提供答案，但最终声称 the Response offers a clear summary of the Article ——从而修改了原本旨在 evaluate how well the Response addressed the Query with the provided Article 的查询。

Model Pairs	Complete Statistical Significance Tests for Model Comparisons							
	Perfect Score (PS)				Constraint Score (CS)			
	Mean Diff ś SD	t-value	p-value	Sig.	Mean Diff ś SD	t-value	p-value	Sig.
GPT-4o vs GPT-4o-mini	0.0417 ś 0.0668	1.5288	0.1263	n.s.	0.4017 ś 0.3305	2.9766	0.0029	**
GPT-4o vs LLaMA3-70B	-0.0017 ś 0.0773	-0.0528	0.9579	n.s.	0.3917 ś 0.6832	1.4043	0.1602	n.s.
GPT-4o vs LLaMA3-8B	0.0450 ś 0.0680	1.6199	0.1052	n.s.	0.4583 ś 0.3016	3.7221	0.0002	***
GPT-4o vs Claude-3.5	-0.0500 ś 0.1287	-0.9517	0.3412	n.s.	0.1217 ś 0.9327	0.3195	0.7493	n.s.
GPT-4o vs Claude-3	-0.0067 ś 0.0766	-0.2132	0.8312	n.s.	0.1633 ś 0.2044	1.9572	0.0503	n.s.
GPT-4o vs Mistral	0.1050 ś 0.2231	1.1526	0.2491	n.s.	1.7583 ś 0.5788	7.4412	0.0000	***
GPT-4o-mini vs LLaMA3-70B	-0.0433 ś 0.1302	-0.8154	0.4149	n.s.	-0.0100 ś 0.7592	-0.0323	0.9743	n.s.
GPT-4o-mini vs LLaMA3-8B	0.0033 ś 0.0554	0.1474	0.8828	n.s.	0.0567 ś 0.4376	0.3172	0.7511	n.s.
GPT-4o-mini vs Claude-3.5	-0.0917 ś 0.1212	-1.8522	0.0640	n.s.	-0.2800 ś 0.8591	-0.7984	0.4247	n.s.
GPT-4o-mini vs Claude-3	-0.0483 ś 0.0866	-1.3674	0.1715	n.s.	-0.2383 ś 0.3409	-1.7125	0.0868	n.s.
GPT-4o-mini vs Mistral	0.0633 ś 0.1611	0.9631	0.3355	n.s.	1.3567 ś 0.6623	5.0177	0.0000	***
LLaMA3-70B vs LLaMA3-8B	0.0467 ś 0.1117	1.0238	0.3059	n.s.	0.0667 ś 0.6515	0.2507	0.8021	n.s.
LLaMA3-70B vs Claude-3.5	-0.0483 ś 0.1370	-0.8640	0.3876	n.s.	-0.2700 ś 1.3402	-0.4935	0.6217	n.s.
LLaMA3-70B vs Claude-3	-0.0050 ś 0.0916	-0.1337	0.8936	n.s.	-0.2283 ś 0.7061	-0.7921	0.4283	n.s.
LLaMA3-70B vs Mistral	0.1067 ś 0.2681	0.9746	0.3297	n.s.	1.3667 ś 1.1188	2.9921	0.0028	**
LLaMA3-8B vs Claude-3.5	-0.0950 ś 0.1452	-1.6031	0.1089	n.s.	-0.3367 ś 1.1088	-0.7437	0.4570	n.s.
LLaMA3-8B vs Claude-3	-0.0517 ś 0.0924	-1.3698	0.1708	n.s.	-0.2950 ś 0.4320	-1.6728	0.0944	n.s.
LLaMA3-8B vs Mistral	0.0600 ś 0.1691	0.8690	0.3848	n.s.	1.3000 ś 0.5175	6.1534	0.0000	***
Claude-3.5 vs Claude-3	0.0433 ś 0.0668	1.5882	0.1122	n.s.	0.0417 ś 0.7643	0.1335	0.8938	n.s.
Claude-3.5 vs Mistral	0.1550 ś 0.2201	1.7252	0.0845	n.s.	1.6367 ś 1.2559	3.1920	0.0014	**
Claude-3 vs Mistral	0.1117 ś 0.2115	1.2932	0.1959	n.s.	1.5950 ś 0.7408	5.2741	0.0000	***

Table 15: 模型比较的统计显著性检验。我们在所有模型对之间进行了配对 t 检验，以在表 15 中对我们的主要结果的性能差异进行统计评估。

Model	Accuracy (%)
GPT-4o	98.23 ś 0.31
Claude-3.5	97.30 ś 0.12

Table 16: 不同基础模型在约束满足中作为 LLM-评判的性能表现。

Type	Task	Easy	Hard
Omission	Fact QA (Tech)	150	150
	Fact QA (Culture)	150	150
	Fact QA (History)	150	150
	Creative Writing	–	300
Misinterpretation	Response Eval.	150	–
	Content Analysis	150	–

Table 17: FAITHQA 基准测试的测试集分布。