KScope：用于表征语言模型知识状态的框架

Yuxin Xiao[1], **Shan Chen**[2], **Jack Gallifant**[2],
Danielle Bitterman[2], **Thomas Hartvigsen**[3], **Marzyeh Ghassemi**[1]
[1]Massachusetts Institute of Technology, [2]Harvard University, [3]University of Virginia
yuxin102@mit.edu

Abstract

描述大型语言模型（LLM）对给定问题的知识是一个具有挑战性的任务。因此，之前的工作主要研究了在知识冲突情况下的 LLM 行为，其中模型的内部参数化记忆与外部环境中的信息相矛盾。然而，这并不能完全反映模型对问题答案的了解程度。在本文中，我们首先介绍了基于 LLM 知识模式的一致性和正确性划分的五种知识状态的分类法。接着，我们提出了 KScope，一个逐步细化知识模式假设的统计测试分层框架，并将 LLM 知识归入这五种状态之一。我们在四个数据集上的九个 LLM 中应用了 KScope，并系统性地确定了：（1）支持性背景缩小了模型之间的知识差距。（2）与难度、相关性和熟悉度相关的背景特征推动知识更新的成功。（3）LLM 在部分正确或存在冲突时表现出类似的特征偏好，但在持续错误时则显著分歧。（4）受限于我们的特征分析的背景摘要，以及增强的可信度，进一步提高了更新效能并在 LLM 间进行了推广。[1]

1 介绍

LLMs [15, 24, 55, 72] 从它们的训练语料中记忆信息作为参数化知识 [3, 4, 5, 58] 。它们可能进一步结合用户提示中的有根且最新的上下文知识，以完成知识密集型任务 [14, 34, 60] 。当 LLM 的参数化知识与上下文输入相矛盾时，可能会出现知识冲突 [71] 。现有工作试图在这些矛盾的条件下测量和理解 LLM 的行为 [11, 30, 39, 68, 69, 76] 。

然而，先前对知识冲突的研究并未完全描述大型语言模型对给定问题的潜在知识。最近的工作通常通过最可能的响应 [30, 62, 69] 来表示大型语言模型的知识，这可能会忽略答案分布中多个竞争模式的共存。此外，基于熵的不确定性度量 [11, 39] 捕捉的是整体不确定性而不是模式结构，并且会为分布 $[0.45, 0.45, 0.1]$ 和 $[0.6, 0.2, 0.2]$ 分配类似的熵值（约为 1.37 ）。然而，第一种分布反映了冲突的知识，而第二种则显示了一致的偏好。

为了解决这一差距，我们沿着两个关键维度定义了五种知识状态的分类法，并提出了 KScope，一个用于表征知识状态的层次测试框架。如图 1 所示，我们通过检查大型语言模型的模式集大小来评估知识的一致性，并相对于真实情况评估知识的正确性。基于这两个维度，我们识别出五种不同的知识状态：(1) 一致正确，(2) 冲突正确，(3) 缺失，(4) 冲突错误，和 (5) 一致错误。我们通过反复采样目标大型语言模型来构建实证响应分布，并利用 KScope 逐步完善关于其潜在知识模式的假设。

我们在四个数据集（章节 5 ）上评估了九个大型语言模型（LLM）的知识状态，发现支持上下文增加了所有数据集和模型中一致正确知识的比例。我们注意到两个与医疗相关的数据集（Hemonc [64] 和 PubMedQA [26] ）在提供上下文之前和之后展现的一致正确知识水平低于普通领域的数据集（NQ [32] 和 HotpotQA [74] ）。在每个模型系列中，较大的 LLM 展示了更高比例的一致正确参数化知识。Llama-3 [15] 在其中达到了最高比例，其次是 Qwen-2.5 [72] 和 Gemma-2 [55] ，尽管在引入上下文后这些差异有所缩小。

接下来，我们检查输入背景的哪些特征能够成功地将 LLM 知识更新为一致的正确状态（第 6 节）。我们在难度、相关性和熟悉度三个类别中选择了十一种背景特征，并展示了所有三个特征类别都有所贡献，其中背景长度和熵在不同状态中始终具有价值。我们还发现，LLM

[1]我们的代码可在 https://github.com/xiaoyuxin1002/KScope 获取。

Preprint. Under review.

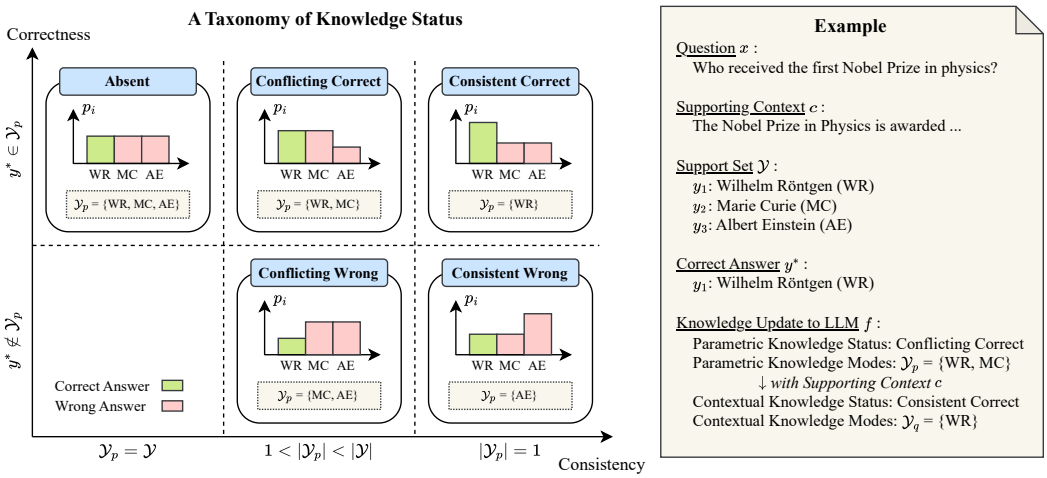

Figure 1: 我们提出了一个基于 LLM 知识模式的一致性和正确性的五个知识状态的分类法。我们在一个三选分类任务中使用 LLM 的参数知识模式 $\mathcal{Y}_p$ 来说明这个分类法。这个表述也适用于情境知识模式 $\mathcal{Y}_q$ ，并推广到具有更多选项的开放性问题或分类任务。

在部分正确或表现出矛盾知识时会类似地优先考虑背景特征，而不论正确与否。相比之下，一致错误状态与其他状态的相关性相对较低，这表明克服根深蒂固的错误信念可能需要不同的背景特征。

最后，我们探索了上下文增强策略以提高知识更新的成功率（第 7 节）。我们发现，在我们的特征重要性分析所引导的约束下进行的上下文摘要优于天真的摘要。当与增强的上下文可信度相结合时，这种方法平均提高了所有状态下成功知识更新的比例为 4.3% —即使在我们的特征分析中未包含的 GPT-4o [24] 中也是如此。

我们在本文中的贡献总结如下：

- 我们基于一致性和正确性定义了五种知识状态的分类法，并提出了 KScope，一个用于表征 LLM 知识状态的层次测试框架。
- 我们将 KScope 应用于四个数据集上的九个大型语言模型，并确定支持上下文实质上缩小了不同模型尺寸和家族之间的知识差距。
- 我们确定了与难度、相关性和熟悉度相关的关键背景特征，这些特征推动了成功的知识更新。
- 我们揭示了基于参数知识状态的 LLM 特征重要性差异，表明在冲突下的相似性，但在持续错误时的分歧。
- 我们验证了约束上下文摘要结合改进的可信度，显著提升了各状态的知识更新成功率，并且具有良好的泛化能力。

2 相关工作

知识特征化。已经显示，大型语言模型（LLMs）在其训练语料库中记忆了世界知识 [3, 4, 5, 10, 44, 47, 58] 。现有工作通过探测激活模式 [7, 8, 23, 33] 来解释这种记忆机制，并通过与知识图谱对比基准测试召回信息的事实准确性 [36, 78, 82] 。研究人员采用多轮提示来提高从 LLMs 中引出的知识的一致性 [43, 61, 77] 和校准性 [19, 31, 73] 。在这项工作中，我们确定了一个 LLM 可能在给定问题下展示的五种知识状态，并通过一个分层的统计检验框架明确地描述它们。

知识冲突。知识冲突可以在大型语言模型的参数记忆中发生，也可以在提供的上下文中发生，或在二者之间发生 [71] 。先前的研究已经考察了影响模型顺从程度的各种上下文因素 [53, 54, 57, 70] ，并引入了量化上下文说服力的指标 [11, 39] 。这些研究发现，当大型语言模型不确定 [45, 68, 76] 或当模型记忆与上下文之间存在确认偏差时，它们倾向于遵循上下文 [30, 69] 。还开发了大型评估基准和流程以促进该领域的研究 [20, 51, 62] 。然而，先前的研究通常使用单一的采样响应来代表模型记忆 [30, 62, 69] ，并应用基于熵的测量来评估冲突，两者均忽略了响应分布的模式结构。相比之下，我们定义了五种知识状态，以测量知识的一致性和正确性，严格测试不同的模式结构，并根据这些状态对模型行为的分析进行分层。

KScope: Knowledge Status Characterization via Hierarchical Testing

Step	Statistical Test	Null Hypothesis	Alternative Hypothesis	If Significant p-value	If Insignificant p-value
(1) Test for the Significance of Invalid Answers	One-Sided Exact Binomial Test	$\mathbb{P}(f(x) \in \mathcal{Y}) = \mathbb{P}(f(x) \notin \mathcal{Y}) = \frac{1}{2}$	$\mathbb{P}(f(x) \notin \mathcal{Y}) > \frac{1}{2}$	Absent Knowledge	Proceed ↓
(2) Test for Uniform Guessing	Two-Sided Exact Multinomial Test	$p_i = \frac{1}{\lvert\mathcal{Y}\rvert}, \forall y_i \in \mathcal{Y}$	$p_i \neq \frac{1}{\lvert\mathcal{Y}\rvert}, \exists y_i \in \mathcal{Y}$	Proceed ↓	Absent Knowledge
(3) Test for Conflicting Knowledge	Likelihood Ratio Test	$p_i = \frac{1}{\lvert\mathcal{Y}\rvert}, \forall y_i \in \mathcal{Y}$	(a) $p_1 = p_2 = \frac{\hat{p}_1+\hat{p}_2}{2} > p_3 = \hat{p}_3$ (b) $p_1 = p_3 = \frac{\hat{p}_1+\hat{p}_3}{2} > p_2 = \hat{p}_2$ (c) $p_2 = p_3 = \frac{\hat{p}_2+\hat{p}_3}{2} > p_1 = \hat{p}_1$	Proceed Accordingly ↓	Absent Knowledge
(4) Test for Consistent Knowledge	One-Sided Exact Binomial Test	$p'_1 = p'_2 = \frac{1}{2}$	(a) $p'_1 = \frac{\hat{p}_1}{\hat{p}_1+\hat{p}_2} > p'_2 = \frac{\hat{p}_2}{\hat{p}_1+\hat{p}_2}$ (b) $p'_1 = \frac{\hat{p}_1}{\hat{p}_1+\hat{p}_2} < p'_2 = \frac{\hat{p}_2}{\hat{p}_1+\hat{p}_2}$	Consistent Correct / Wrong Knolwedge (depending on correctness)	Conflicting Correct / Wrong Knowledge (depending on correctness)

Figure 2: 我们提出了 KScope，这是一种分层测试框架，用于将大型语言模型的知识划分为五种识别出的状态之一。我们注意到，通过重复步骤 3 以迭代地细化关于知识模式集的假设，我们的框架可以推广到具有更大支持集的问题。

知识更新。为了纳入准确且最新的信息，检索增强生成系统 [6, 14, 17, 34, 46, 49] 从外部语料库中检索相关上下文，以支持大型语言模型（LLMs）完成知识密集型任务。可以进一步增强检索到的上下文，以提高知识更新 [53, 57, 70] 的有效性。其他方法包括直接编辑模型的参数化知识 [12, 18, 59, 60] 或在推理时应用机制干预 [27, 35, 80] 。在这项工作中，我们采用黄金检索设置，并系统地评估各种针对每种已识别知识状态量身定制的上下文增强策略。

3 LLM 知识状态及其表征方法

3.1 LLM 知识状态：一致性和正确性

在分析 LLM 的知识时，我们关注两个关键维度：

1. 一致性：模型的知识模式有多一致？也就是说，它是表现出一个单一的连贯信念还是多个矛盾的信念？
2. 正确性：模型的知识模式集合是否包含正确答案？

为了形式化这个问题，考虑一个问答对 (x, y^*) 。我们定义 LLM 的参数化知识 f ，针对问题 x ，作为条件多项分布 $\boldsymbol{p} = f(\cdot \mid x)$ 在支持集 $\mathcal{Y} = \{y_1, \ldots, y_d\}$ 上的分布，其中 $p_i = f(y_i \mid x)$ 。我们进一步定义参数化知识模式集合 $\mathcal{Y}_p = \text{modes}(\boldsymbol{p}) \subseteq \mathcal{Y}$ 为满足以下条件的子集：对于任何 $y_i, y_j \in \mathcal{Y}_p$ ，满足 $p_i = p_j$ ；对于任何 $y_k \notin \mathcal{Y}_p$ ，满足 $p_i > p_k$ 。本质上，知识模式构成支持集内高概率元素的一个高原，这些元素与其他元素可区分。

为了评估一致性维度，我们考察了 $\mathcal{Y}_p$ 的三种可能结构：(1) $\mathcal{Y}_p = \mathcal{Y}$ ，(2) $1 < |\mathcal{Y}_p| < |\mathcal{Y}|$ 和 (3) $|\mathcal{Y}_p| = 1$ ，其中 $|\cdot|$ 表示一个集合的基数。我们通过检查标准答案是否为 $y^* \in \mathcal{Y}_p$ 来评估正确性。基于这两个维度，我们制定了一个包含五种参数化知识状态的分类法 $P = \text{status}(\mathcal{Y}_p)$ ，如图 1 所示，它对关于问题 x 的 f 的知识进行了特征化：当问题答案对的支持上下文 c 可用时，我们将 LLM 的上下文知识定义为 $\boldsymbol{q} = f(\cdot \mid x, c)$ 。同样地，我们根据相应的知识模式 $\mathcal{Y}_q = \text{modes}(\boldsymbol{q}) \subseteq \mathcal{Y}$ 分配其上下文知识状态 $Q = \text{status}(\mathcal{Y}_q)$ 。

3.2 知识状态操作化的挑战

虽然在第 3.1 节中引入的分类法为 LLM 知识状态提供了一个原则视角，但在实践中应用它仍然会面临几个挑战。首先，LLM 知识的真实基础分布是不可观察的。其次，即使在相同的知识状态下，模型的行为可能也会有所不同。例如，当一个 LLM 对某个问题缺乏足够知识时，它可能会随机回应或完全拒绝回答 [22, 66] 。

为了解决第一个挑战，我们使用经验样本频率来近似潜在知识分布。具体来说，我们首先生成 M 个给定问题的释义，以减少提示敏感性 [48] ，然后使用这些释义从目标 LLM 收集 N 链式思维响应 [65] 。对于开放式生成，我们通过语义聚类 N 样本 [31] 来定义支持集 $\mathcal{Y}$ 。对于多项选择任务，$\mathcal{Y}$ 只是给定选项的集合。

然后，我们考虑潜在无效模型响应的可能性来解决第二个挑战。这些包括在支持集外产生幻觉式回答 [22] 或在高风险应用中拒绝回答 [66] 。设 N' 表示无效响应的数量。我们通过以下方式估计经验分布 $\hat{\boldsymbol{p}}$ （并类似地，$\hat{\boldsymbol{q}}$ ）：$\hat{p}_i = \frac{1}{N-N'} \sum_{n=1}^{N} \mathbb{1}[f_n(x) = y_i], \forall y_i \in \mathcal{Y}$ 。

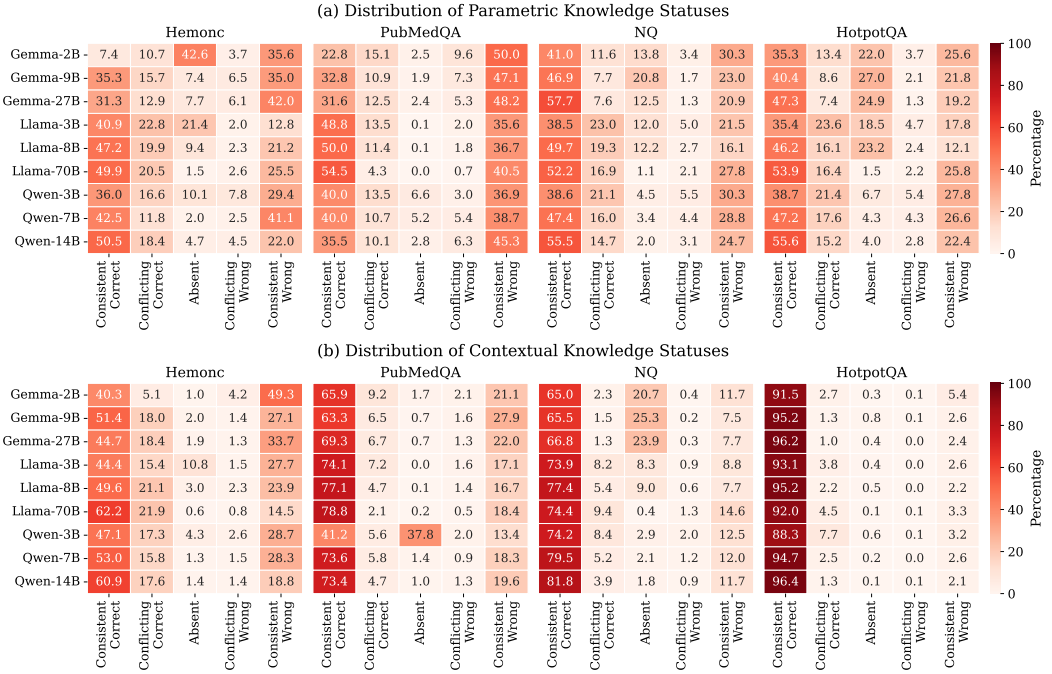

Figure 3: 通过将 KScope 框架应用于四个数据集中的九个大模型，得到了表征结果。总体而言，大多数大模型表现出最高比例的一致正确参数化知识状态，在提供支持性上下文时，这一比例进一步增加。

3.3 KScope: 基于分层测试的知识状态表征

为了弥合第 3.1 节中讨论的真实潜在知识分布与第 3.2 节中估计的经验分布之间的差距，我们引入了 KScope，一个用于知识状态表征的层级测试框架。我们在图 2 中展示了测试的细节。

步骤 1：检验无效答案的显著性。我们首先通过单边精确二项检验来评估模型是否表现出更高的倾向产生无效回答。

步骤 2：测试均匀猜测。接下来，我们使用双侧精确多项式测试来检验 LLM 的经验反应分布是否显著偏离均匀分布。

步骤 3：检测冲突知识。接下来，我们进行一组似然比检验，以优化模型的知识模式集。那些估计概率违背自身不等式约束的选项会被立即拒绝。如果在经过 Bonferroni 校正后仍有多个选项显著，我们选择贝叶斯信息准则（BIC）最低的那个。对于较大的支持集，我们重复此步骤以从模式集中移除低概率元素。

步骤 4：一致性知识测试。上一步将模型的知识模式集缩减为两个元素。在步骤 3 中选择的备选项的条件下，我们接着使用两个单边精确二项测试来测试模型是否为这两个剩余元素分配了显著不同的概率。如前所述，我们会舍弃无效的备选项，并在多个备选项被接受时选择 BIC 最低的那个。

4 实验设置

数据集。我们专注于四个任务，其中两个来自医疗领域，两个来自通用领域：

- Hemonc [64] 是一个从定期维护的肿瘤学参考数据库中提取的医疗数据集。它由 6,212 个临床研究实例组成，每个实例比较某种治疗方案与某一对照方案在特定医疗条件下的疗效，并标记为优于、劣于或无差异。为了减少位置偏差，我们对每个治疗方案-对照方案对的顺序进行排列。这些问题的背景信息由相关 PubMed 文章的摘要构成。
- PubMedQA [26] 由 1,000 个医学研究问题组成，每个问题都标记为是、否或可能的答案。上下文来自相应的 PubMed 摘要。
- NQ [32] 包含 3,596 个谷歌搜索查询，检索维基百科页面作为上下文。

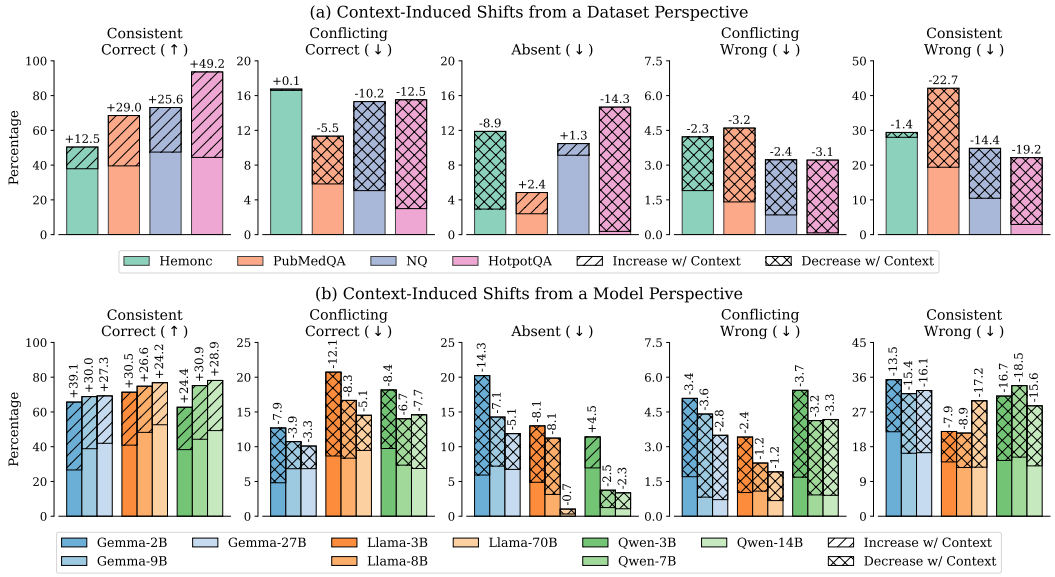

Figure 4: 上下文引发的知识状态分布变动。支持性上下文增加了所有数据集和模型中一致正确知识的比例。Llama 系列和各系列中的较大模型实现了更高比例的一致正确知识，尽管随着上下文的增加差距有所缩小。

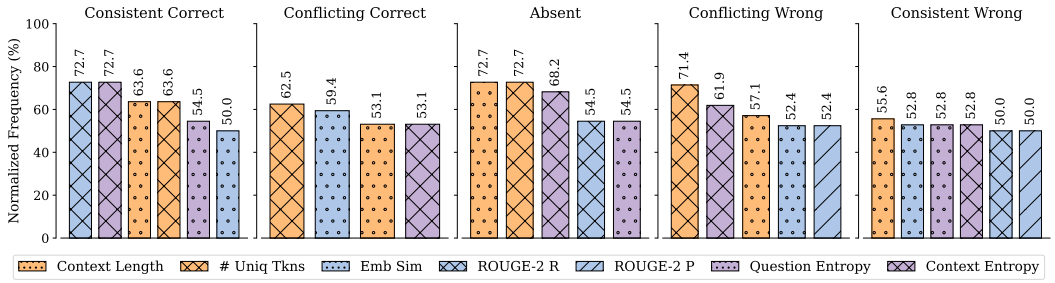

Figure 5: 在每个参数化知识状态的至少 50% 个案例中，情境特征出现在最重要的前五名中。每种颜色表示一个特征类别：橙色表示难度，蓝色表示相关性，紫色表示熟悉程度。在所有状态中，情境的长度和熵始终排名在最重要的特征之中。

- HotpotQA [74] 包含 6,119 道常规领域的多跳推理问题，使用从相关维基百科文章中提取的句子级支持事实作为上下文。

此处所有支持上下文都包含对真实答案的证据。为了直接比较跨数据集的 LLMs 的知识状态，我们将 NQ 和 HotpotQA 转换为三个选项的分类任务。根据 [70] 的方法，我们提示 GPT-4o [24] 为每个问题生成两个额外的错误选项（详情见附录 **??**）。我们注意到，尽管这里没有评估，KScope 仍然适用于开放式问题和更多选项的分类，如第 3 节所讨论。

实现细节。我们评估了九种经过指令微调的大型语言模型（LLM），涵盖了三个模型系列：Gemma-2 (2B, 9B, 27B)，Llama-3 (3B, 8B, 70B)，和 Qwen-2.5 (3B, 7B, 14B)。通过将温度设置为 1，我们保持 LLM 的采样分布不变，并将 KScope 中使用的显著性水平固定在 $\alpha = 0.05$。在 Hemonc 上的超参数搜索使用 Llama-8B（详情见附录）表明，知识状态表征在使用每个问题的 $M = 20$ 个释义后，经验上在 $N = 100$ 个样本后稳定下来。

5 Q1: 上下文如何更新大语言模型的知识状态？

使用第 4 节中的设置，我们首先描述 LLMs 参数和上下文知识状态的分布（第 5.1 节），然后分析上下文如何引起它们之间的转换（第 5.2 节）。

5.1 参数化与情境化知识状态的分布

我们展示了在图 3 中将 KScope 应用于四个数据集上的九个 LLM 所得的表征结果。当仅依赖参数知识时（图 3 (a)），LLM 表现出一致的知识——无论是正确的还是错误的——的频率要高于冲突知识。当冲突知识出现时，正确答案通常包含在知识模式中。一些异常值表现出较高比例的知识缺失，例如 Hemonc 上的 Gemma-2B。

当大型语言模型（LLM）被提供支持性上下文（图 3 (b)）时，一致的正确知识状态的比例在所有数据集和模型中显著增加，其中在 HotpotQA 中增幅最大，在 Hemonc 中最小。这突显了检索增强生成 [6, 14, 17, 34] 的有效性，该方法旨在通过相关的外部信息增强 LLM 的知识。然而，在某些情况下，上下文可能会使 LLM 困惑，导致它们随机猜测，并增加缺失知识的比例。例如，在 PubMedQA 中，Qwen-3B 出现这种情况，在 NQ 中，Gemma 家族也出现这种情况，这可能是由于这些数据集中的上下文长度较长——我们将在第 6 节对此问题进行进一步调查。

5.2 背景诱发的知识状态变化

我们进一步研究了上下文如何改变每个知识状态的分布。图 4 (a) 显示了数据集层面的变化，其中知识状态的分布在九个大型语言模型之间取平均。我们观察到，Hemonc 和 PubMedQA 展示了一致正确知识的比例低于 NQ 和 HotpotQA，无论是在提供上下文之前还是之后。此外，支持上下文略微增加了 Hemonc 的冲突正确知识和 PubMedQA 的缺失知识的比例，突显了医学领域上下文引入的挑战 [9, 50, 56, 63] 。

在每个 LLM 家族中，我们发现较大的模型在提供上下文之前和之后一致呈现出更高比例的一致正确知识（图 4 (b)） [5, 37] 。Llama 实现了一致正确参数化知识的最高比例，其次是 Qwen 和 Gemma，不过一旦引入上下文，这一差距缩小。附录 9 详细描述了上下文如何在每个数据集上使每个 LLM 的知识在不同状态间转换。

6 问题 2：哪些上下文特征推动所需的知识更新？

基于在第 5 节的评估结果，我们试图理解哪些上下文特征推动了一致正确知识的增加。我们在第 6.1 节介绍了三类上下文特征，并在第 6.2 节检查每种知识状态的特征重要性。

6.1 上下文特征

我们考虑了三个类别中的十一种上下文特征：

- 困难度：(1) 上下文长度；(2) 可读性，通过 Flesch-Kincaid 可读性测试 [29] 测量，该测试基于每句的平均单词数和每个单词的音节数；(3) 词形还原后的唯一标记数（# Uniq Tkns）。
- 相关性：(1) 嵌入相似度（Emb Sim），计算为每个问题及其相应上下文的嵌入之间的余弦相似度，使用 text-embedding-3-large [40] ；以及 (2–4) ROUGE-2 召回率、准确率和 F1（ROUGE-2 R/P/F），基于每个问题及其相应上下文之间的二元组重叠。
- 熟悉度：（1-2）问题和上下文困惑度，以及（3-4）问题和上下文熵，由每个大型语言模型测量。

我们定义了一个二元标签，指示上下文是否成功地将 LLM 的参数知识更新到一致的正确状态。如在第 4 节所述，对于每个数据集、LLM 和初始参数知识状态的组合，我们构建了一个分层的二元分类任务。我们使用带有 L2 正则化的逻辑回归对提取的特征进行分析，并舍弃在 Macro-F1 中表现超过虚拟基线的情况，因为存在极端的类别不平衡。实施细节和完整回归结果见附录 **??** 。

我们通过对每个分层组合中的每个特征的绝对 SHAP 值 [38] 进行平均来计算特征重要性。然后，我们计算每个特征在不同数据集和 LLMs 中出现在前五个最重要特征中的归一化频率。这产生了一个基于频率的排名，用于每个初始参数知识状态。

6.2 特征重要性对上下文驱动的知识更新

我们在图 5 中绘制了在我们实验中至少 50% 的案例中出现的前五个最重要的上下文特征。结果包括所有三个类别的特征：难度、相关性和熟悉度。这与 [11, 57] 的发现相符，即上下文的相关性会影响上下文的说服力。在五种参数化知识状态中，持续错误的知识状态展示了相比其他状态更加平缓的特征频率分布，这表明处于这种状态的 LLMs 对任何特定特征的强调较少。值得注意的是，上下文长度和熵在所有状态中始终排名靠前。

为了评估在不同参数化知识状态下的大模型是否同样地优先考虑上下文特征，我们计算了特征重要性排名的 Spearman 等级相关性，并进行 Bonferroni 校正。如图 **??** 所示，在一致正确和冲突正确、缺失知识之间的相关性在统计上显著。这表明了一种确认偏误 [30, 69]：当上下文至少部分与模型的知识模式对齐时，模型会以类似方式关注上下文特征。我们还观察到冲突正确和冲突错误之间存在显著的相关性，表明无论正确与否，知识冲突期间具有相似的特征偏好 [45, 68, 76] 。相比之下，一致错误状态与其他状态的相关性相对较低，这意味着要克服一个坚定持有的错误信念可能需要不同的上下文特征。

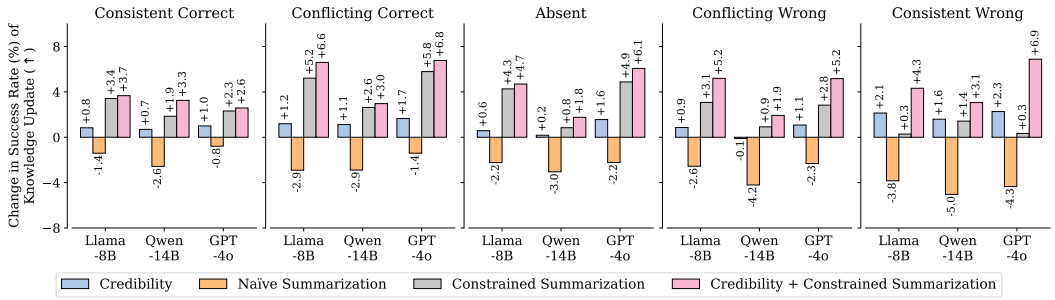

Figure 6: 相对于原始上下文，每种上下文增强策略在知识更新成功率上的绝对变化（%），并在数据集间取平均。将可信度元数据整合到受限摘要中，平均提高了在大语言模型（LLMs）和状态间的成功率 4.3%，并且对 GPT-4o 有良好的泛化效果。

7 Q3: 哪些上下文增强在不同知识状态下效果最好？

在本节中，我们利用 6 中关于特征重要性分析的见解来改进 LLMs 中的知识更新。

7.1 上下文增强策略

我们对第 4 节中数据集的原始上下文应用以下增强策略：

- 可信度 [70]：为了增强上下文的可信度，我们为 Hemonc 和 PubMedQA 在上下文中添加了相关的 PubMed 元数据，为 NQ 和 HotpotQA 添加了相应的 Wikipedia 文章标题。在使用上下文进行响应采样时，我们指示 LLM 优先考虑可信的上下文，而不是其内部的参数化知识（详情见附录 10）。
- 朴素摘要：我们利用 GPT-4o [24] 直接总结上下文。
- 限制性摘要：我们通过附加的约束来指导摘要，根据分析结果调整上下文特征。具体来说，我们提示 GPT-4o 在摘要过程中减少上下文长度和唯一标记的数量。此外，指示 GPT-4o 保留语义内容（保持与问题的嵌入相似性）、保留词汇级别的重叠（保持 ROUGE-2 召回率和精确度），并确保流利度（保留上下文困惑度和熵）。
- 综合：我们将可信度信息整合到从受限摘要生成的上下文中。

然后，我们调查每种增强策略如何影响知识更新的成功率，相较于原始上下文，针对每种知识状态。我们在三个不同大小和系列的 LLM 上评估这些策略：Llama-3.1-8B-Instruct [15] (Llama-8B)、Qwen2.5-14B-Instruct [72] (Qwen-14B) 和 GPT-4o。尽管 GPT-4o 未包含在第 6 节的功能分析中，我们在此评估它是为了检验我们的发现是否可以推广到其他 LLM。我们在图 6 中展示了四个数据集的平均结果，完整的细节在附录 10 中。

7.2 上下文增强的有效性

如图 6 所示，约束性摘要提高了除一致错误状态之外的所有知识状态的成功率。这与我们在图 **??** 中的发现一致，即修正一致错误的信念可能需要与其他情况下不同的上下文特征。此外，这种约束性的摘要策略在 GPT-4o 上具有良好的普适性。

相比之下，简单的总结总是会降低性能。我们在图 7 中绘制了对 Hemonc 有影响的九个特征的归一化特征空间（问题困惑度和熵不受上下文总结影响）。两种总结方法都减少了上下文长度和唯一词元的数量，同时增加了上下文困惑度和熵。然而，简单总结未能保存流畅性和关键语义内容，导致可读性更差且 ROUGE-2 召回率更低。相比之下，约束总结更有效地提高了嵌入相似性、ROUGE-2 精度和 F1。这些在增强特征空间中的差异解释了成功率的差距，并强调了我们在第 6 节中的特征分析的重要性。其他数据集的详细信息在附录 10 中。

另一方面，正如图 6 所示，可信度对于一致的错误状态更有效。这表明，当一个大语言模型（LLM）持续持有错误的信念时，向上下文添加可信度元数据会使其更有说服力 [70]。这种组合策略保留了对前四种状态的约束总结的好处，并进一步提高了一致错误状态的成功率，超出了单靠可信度所能达到的效果。总体而言，与使用原始上下文相比，它平均提高了所有状态和 LLM 的成功率 4.3%。

8 讨论

结论。在本文中，我们首先基于一致性和正确性提出了五种知识状态的分类法。然后，我们介绍了 KScope，这是一种分层测试框架，通过逐步细化关于 LLM 知识模式的假设来表征

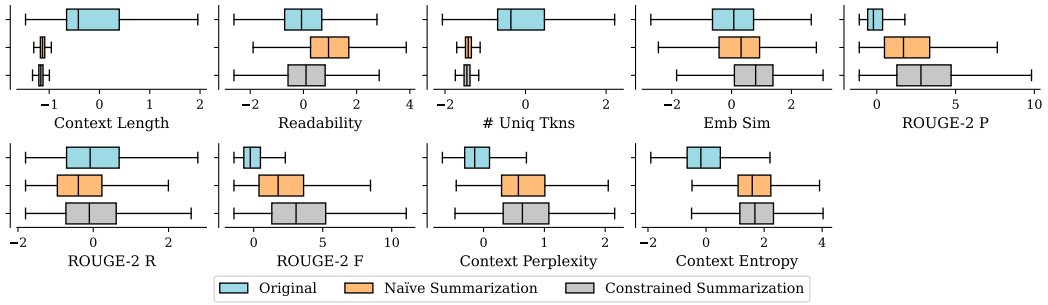

Figure 7: Hemonc 的原始和总结上下文的归一化特征空间，使用 Llama-3.1-8B-Instruct 进行测量。简单的摘要会降低可读性和 ROUGE-2 召回率，而受限的摘要在嵌入相似度、ROUGE-2 精确度和 F1 方面表现更高。

知识状态。通过将 KScope 应用于四个数据集上的九个 LLM，我们得出以下结论：(1) 支持性上下文显著缩小了不同 LLM 之间的知识差距。(2) 与上下文难度、相关性和熟悉度相关的特性推动了成功的知识更新。(3) 当 LLM 的知识模式部分与正确答案对齐或内部矛盾时，LLM 会以相似的方式关注特性，但在持续错误时则会明显不同。(4) 由特性分析引导的受限上下文总结，结合增强的可信度，进一步提升了更新效果并在不同模型间广泛适用。这些发现为 LLM 的知识机制 [58] 提供了有价值的见解，并强调了在未来工作中定制知识更新策略 [14, 60] 以适应不同知识状态的重要性。

限制。我们在第 6 节的特征重要性分析集中在三个类别的十一种特征上。然而，更加细致的风格化上下文特征 [11, 57, 70] 如何影响大型语言模型（LLM）的知识状态仍未被充分探讨。虽然 KScope 框架支持开放式问题 [28, 67] 和具有任意选项数的分类任务 [52, 81]，但由于计算限制，我们的实验限制在三选项分类上。我们还专注于输入上下文总是包含支持正确答案证据的设置。这使我们能够就这些上下文如何更新 LLM 知识到一致性正确状态生成实际的见解。在现实世界的检索增强生成系统中 [6, 14, 17, 34, 46, 49]，检索到的上下文可能包含噪声或冲突的信息 [20, 51]，这可能带来额外的挑战。我们将调查在这些更复杂条件下上下文如何影响 LLM 知识状态的问题留待未来研究。

更广泛的影响。LLM 被广泛应用于日常用户--聊天机器人应用 [1, 24] 和高风险领域，如医疗保健 [42, 56] 和法律服务 [13, 16]。然而，之前的工作 [11, 39, 53, 54, 57, 70, 71] 缺乏一个正式的框架来描述 LLM 的知识状态，特别是由于输入上下文的变化可能导致这些状态发生变化。当训练数据包含错误信息时 [2, 41]，这一挑战变得更加令人担忧，因为这可能导致模型产生一贯错误的信念。我们在第 6 节中的分析表明，纠正这些信念往往需要与其他知识状态不同的上下文特征。所提出的 KScope 框架也与 LLM 中的幻觉检测 [22, 25, 79] 和不确定性量化 [21, 31, 75] 相关。通过识别知识状态，它有助于区分由于知识缺失导致的幻觉与由于知识冲突导致的不确定性。这些联系突显了 KScope 的实用性及其在提高 LLM 可靠性方面的更广泛影响。

References

[1] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint*, 2023.

[2] D. A. Alber, Z. Yang, A. Alyakin, E. Yang, S. Rai, A. A. Valliani, J. Zhang, G. R. Rosenbaum, A. K. Amend-Thomas, D. B. Kurland, et al. Medical large language models are vulnerable to data-poisoning attacks. *Nature Medicine*, pages 1–9, 2025.

[3] Z. Allen-Zhu and Y. Li. Physics of language models: Part 3.1, knowledge storage and extraction. In *ICML*, 2024.

[4] Z. Allen-Zhu and Y. Li. Physics of language models: Part 3.2, knowledge manipulation. In *ICLR*, 2025.

[5] Z. Allen-Zhu and Y. Li. Physics of language models: Part 3.3, knowledge capacity scaling laws. In *ICLR*, 2025.

[6] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *ICLR*, 2024.

[7] T. Bricken, A. Templeton, J. Batson, B. Chen, A. Jermyn, T. Conerly, N. Turner, C. Anil, C. Denison, A. Askell, R. Lasenby, Y. Wu, S. Kravec, N. Schiefer, T. Maxwell, N. Joseph, Z. Hatfield-Dodds, A. Tamkin, K. Nguyen, B. McLean, J. E. Burke, T. Hume, S. Carter, T. Henighan, and C. Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023.

[8] C. Burns, H. Ye, D. Klein, and J. Steinhardt. Discovering latent knowledge in language models without supervision. In *ICLR*, 2023.

[9] F. Busch, L. Hoffmann, C. Rueger, E. H. van Dijk, R. Kader, E. Ortiz-Prado, M. R. Makowski, L. Saba, M. Hadamitzky, J. N. Kather, et al. Current applications and challenges in large language models for patient care: a systematic review. *Communications Medicine*, 2025.

[10] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang. Quantifying memorization across neural language models. In *ICLR*, 2023.

[11] K. Du, V. Snæbjarnarson, N. Stoehr, J. White, A. Schein, and R. Cotterell. Context versus prior knowledge in language models. In *ACL*, 2024.

[12] J. Fang, H. Jiang, K. Wang, Y. Ma, J. Shi, X. Wang, X. He, and T.-S. Chua. Alphaedit: Null-space constrained model editing for language models. In *ICLR*, 2025.

[13] Z. Fei, X. Shen, D. Zhu, F. Zhou, Z. Han, A. Huang, S. Zhang, K. Chen, Z. Yin, Z. Shen, et al. Lawbench: Benchmarking legal knowledge of large language models. In *EMNLP*, 2024.

[14] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint*, 2023.

[15] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint*, 2024.

[16] N. Guha, J. Nyarko, D. E. Ho, C. Re, A. Chilton, A. Narayana, A. Chohlas-Wood, A. Peters, B. Waldon, D. Rockmore, D. Zambrano, D. Talisman, E. Hoque, F. Surani, F. Fagan, G. Sarfaty, G. M. Dickinson, H. Porat, J. Hegland, J. Wu, J. Nudell, J. Niklaus, J. J. Nay, J. H. Choi, K. Tobia, M. Hagan, M. Ma, M. Livermore, N. Rasumov-Rahe, N. Holzenberger, N. Kolt, P. Henderson, S. Rehaag, S. Goel, S. Gao, S. Williams, S. Gandhi, T. Zur, V. Iyer, and Z. Li. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *NeurIPS Datasets and Benchmarks Track*, 2023.

[17] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M.-W. Chang. Realm: retrieval-augmented language model pre-training. In *ICML*, 2020.

[18] T. Hartvigsen, S. Sankaranarayanan, H. Palangi, Y. Kim, and M. Ghassemi. Aging with grace: lifelong model editing with discrete key-value adaptors. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 47934–47959, 2023.

[19] B. Hou, Y. Liu, K. Qian, J. Andreas, S. Chang, and Y. Zhang. Decomposing uncertainty for large language models through input clarification ensembling. In *ICML*, 2024.

[20] Y. Hou, A. Pascale, J. Carnerero-Cano, T. T. Tchrakian, R. Marinescu, E. M. Daly, I. Padhi, and P. Sattigeri. Wikicontradict: A benchmark for evaluating LLMs on real-world knowledge conflicts from wikipedia. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[21] H.-Y. Huang, Y. Yang, Z. Zhang, S. Lee, and Y. Wu. A survey of uncertainty estimation in llms: Theory meets practice. *arXiv preprint arXiv:2410.15326*, 2024.

[22] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM TOIS*, 2025.

[23] R. Huben, H. Cunningham, L. R. Smith, A. Ewart, and L. Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2023.

[24] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint*, 2024.

[25] C. Jiang, B. Qi, X. Hong, D. Fu, Y. Cheng, F. Meng, M. Yu, B. Zhou, and J. Zhou. On large language models' hallucination with regard to known facts. In *NAACL)*, pages 1041–1053, 2024.

[26] Q. Jin, B. Dhingra, Z. Liu, W. Cohen, and X. Lu. Pubmedqa: A dataset for biomedical research question answering. In *EMNLP-IJCNLP*, 2019.

[27] Z. Jin, P. Cao, H. Yuan, Y. Chen, J. Xu, H. Li, X. Jiang, K. Liu, and J. Zhao. Cutting off the head ends the conflict: A mechanism for interpreting and mitigating knowledge conflicts in language models. In *ACL Findings*, 2024.

[28] E. Kamalloo, N. Dziri, C. Clarke, and D. Rafiei. Evaluating open-domain question answering in the era of large language models. In *ACL*, 2023.

[29] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Technical Report, Naval Technical Training Command Research Branch*, 1975.

[30] E. Kortukov, A. Rubinstein, E. Nguyen, and S. J. Oh. Studying large language model behaviors under context-memory conflicts with real documents. In *COLM*, 2024.

[31] L. Kuhn, Y. Gal, and S. Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*, 2023.

[32] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, et al. Natural questions: A benchmark for question answering research. *TACL*, 2019.

[33] W. Laurito, S. Maiya, G. Dhimoïla, O. Yeung, and K. Hänni. Cluster-norm for unsupervised probing of knowledge. In *EMNLP*, 2024.

[34] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *NeurIPS*, 2020.

[35] K. Li, O. Patel, F. Viégas, H. Pfister, and M. Wattenberg. Inference-time intervention: eliciting truthful answers from a language model. In *NeurIPS*, 2023.

[36] X. Liu, F. Wu, T. Xu, Z. Chen, Y. Zhang, X. Wang, and J. Gao. Evaluating the factuality of large language models using large-scale knowledge graphs. *arXiv preprint*, 2024.

[37] X. Lu, X. Li, Q. Cheng, K. Ding, X.-J. Huang, and X. Qiu. Scaling laws for fact memorization of large language models. In *EMNLP Findings*, 2024.

[38] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, 2017.

[39] S. Marjanovic, H. Yu, P. Atanasova, M. Maistro, C. Lioma, and I. Augenstein. Dynamicqa: Tracing internal knowledge conflicts in language models. In *EMNLP Findings*, 2024.

[40] OpenAI. Text-embedding-3-large - OpenAI Platform. `https://platform.openai.com/docs/models/text-embedding-3-large`. [Accessed 2025-03-12].

[41] Y. Pan, L. Pan, W. Chen, P. Nakov, M.-Y. Kan, and W. Wang. On the risk of misinformation pollution with large language models. In *EMNLP Findings*, pages 1389–1403, 2023.

[42] C. Peng, X. Yang, A. Chen, K. E. Smith, N. PourNejatian, A. B. Costa, C. Martin, M. G. Flores, Y. Zhang, T. Magoc, et al. A study of generative large language model for medical research and healthcare. *NPJ digital medicine*, 2023.

[43] S. Pitis, M. R. Zhang, A. Wang, and J. Ba. Boosted prompt ensembles for large language models. *arXiv preprint*, 2023.

[44] U. S. Prashanth, A. Deng, K. O'Brien, J. S. V, M. A. Khan, J. Borkar, C. A. Choquette-Choo, J. R. Fuehne, S. Biderman, T. Ke, K. Lee, and N. Saphra. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. In *ICLR*, 2025.

[45] C. Qian, X. Zhao, and T. Wu. "merge conflicts!"' exploring the impacts of external knowledge distractors to parametric knowledge graphs. In *COLM*, 2024.

[46] O. Ram, Y. Levine, I. Dalmedigos, D. Muhlgay, A. Shashua, K. Leyton-Brown, and Y. Shoham. In-context retrieval-augmented language models. *TACL*, 2023.

[47] A. Schwarzschild, Z. Feng, P. Maini, Z. C. Lipton, and J. Z. Kolter. Rethinking LLM memorization through the lens of adversarial compression. In *NeurIPS*, 2024.

[48] M. Sclar, Y. Choi, Y. Tsvetkov, and A. Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *ICLR*, 2024.

[49] W. Shi, S. Min, M. Yasunaga, M. Seo, R. James, M. Lewis, L. Zettlemoyer, and W.-t. Yih. Replug: Retrieval-augmented black-box language models. In *NAACL*, 2024.

[50] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 2023.

[51] Z. Su, J. Zhang, X. Qu, T. Zhu, Y. Li, J. Sun, J. Li, M. Zhang, and Y. Cheng. $\texttt{ConflictBank}$: A benchmark for evaluating the influence of knowledge conflicts in LLMs. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[52] S. A. Tabatabaei, S. Fancher, M. Parsons, and A. Askari. Can large language models serve as effective classifiers for hierarchical multi-label classification of scientific documents at industrial scale? In *COLING Industry Track*, pages 163–174, 2025.

[53] H. Tan, F. Sun, W. Yang, Y. Wang, Q. Cao, and X. Cheng. Blinded by generated contexts: How language models merge generated and retrieved contexts when knowledge conflicts? In *ACL*, 2024.

[54] Y. Tao, A. Hiatt, E. Haake, A. Jetter, and A. Agrawal. When context leads but parametric memory follows in large language models. In *EMNLP*, 2024.

[55] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint*, 2024.

[56] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting. Large language models in medicine. *Nature medicine*, 2023.

[57] A. Wan, E. Wallace, and D. Klein. What evidence do language models find convincing? In *ACL*, 2024.

[58] M. Wang, Y. Yao, Z. Xu, S. Qiao, S. Deng, P. Wang, X. Chen, J.-C. Gu, Y. Jiang, P. Xie, et al. Knowledge mechanisms in large language models: A survey and perspective. In *EMNLP Findings*, 2024.

[59] P. Wang, Z. Li, N. Zhang, Z. Xu, Y. Yao, Y. Jiang, P. Xie, F. Huang, and H. Chen. Wise: Rethinking the knowledge memory for lifelong model editing of large language models. In *NeurIPS*, 2024.

[60] S. Wang, Y. Zhu, H. Liu, Z. Zheng, C. Chen, and J. Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 2024.

[61] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. In *ICLR*, 2023.

[62] Y. Wang, S. Feng, H. Wang, W. Shi, V. Balachandran, T. He, and Y. Tsvetkov. Resolving knowledge conflicts in large language models. In *COLM*, 2024.

[63] Y. Wang, Y. Zhao, and L. Petzold. Are large language models ready for healthcare? a comparative study on clinical language understanding. In *ML4H*, 2023.

[64] J. L. Warner, D. Dymshyts, C. G. Reich, M. J. Gurley, H. Hochheiser, Z. H. Moldwin, R. Belenkaya, A. E. Williams, and P. C. Yang. Hemonc: A new standard vocabulary for chemotherapy regimen representation in the omop common data model. *Journal of biomedical informatics*, 2019.

[65] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 2022.

[66] B. Wen, J. Yao, S. Feng, C. Xu, Y. Tsvetkov, B. Howe, and L. L. Wang. Know your limits: A survey of abstention in large language models. *TACL*, 2025.

[67] Z. Wen, Z. Tian, Z. Jian, Z. Huang, P. Ke, Y. Gao, M. Huang, and D. Li. Perception of knowledge boundary for large language models through semi-open-ended question answering. In *NeurIPS*, 2024.

[68] K. Wu, E. Wu, and J. Zou. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[69] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *ICLR*, 2024.

[70] R. Xu, B. Lin, S. Yang, T. Zhang, W. Shi, T. Zhang, Z. Fang, W. Xu, and H. Qiu. The earth is flat because...: Investigating llms' belief towards misinformation via persuasive conversation. In *ACL*, 2024.

[71] R. Xu, Z. Qi, Z. Guo, C. Wang, H. Wang, Y. Zhang, and W. Xu. Knowledge conflicts for llms: A survey. In *EMNLP*, 2024.

[72] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2. 5 technical report. *arXiv preprint*, 2024.

[73] A. X. Yang, C. Chen, and K. Pitas. Just rephrase it! uncertainty estimation in closed-source language models via multiple rephrased queries. In *NeurIPS Workshop*, 2024.

[74] Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*, 2018.

[75] F. Ye, M. Yang, J. Pang, L. Wang, D. F. Wong, E. Yilmaz, S. Shi, and Z. Tu. Benchmarking LLMs via uncertainty quantification. In *NeurIPS Datasets and Benchmarks Track*, 2024.

[76] J. Ying, Y. Cao, K. Xiong, L. Cui, Y. He, and Y. Liu. Intuitive or dependent? investigating llms' behavior style to conflicting prompts. In *ACL*, 2024.

[77] O. Yoran, T. Wolfson, B. Bogin, U. Katz, D. Deutch, and J. Berant. Answering questions by meta-reasoning over multiple chains of thought. In *EMNLP*, 2023.

[78] J. Yu, X. Wang, S. Tu, S. Cao, D. Zhang-Li, X. Lv, H. Peng, Z. Yao, X. Zhang, H. Li, et al. Kola: Carefully benchmarking world knowledge of large language models. In *ICLR*, 2024.

[79] Y. Zhang, Y. Li, L. Cui, D. Cai, L. Liu, T. Fu, X. Huang, E. Zhao, Y. Zhang, Y. Chen, et al. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*, 2023.

[80] Y. Zhao, A. Devoto, G. Hong, X. Du, A. P. Gema, H. Wang, X. He, K.-F. Wong, and P. Minervini. Steering knowledge selection behaviours in llms via sae-based representation engineering. In *NAACL*, 2025.

[81] C. Zhou, J. Dong, X. Huang, Z. Liu, K. Zhou, and Z. Xu. Quest: Efficient extreme multi-label text classification with large language models on commodity hardware. In *EMNLP Findings*, 2024.

[82] Y. Zhu, J. Xiao, Y. Wang, and J. Sang. Kg-fpq: Evaluating factuality hallucination in llms with knowledge graph-based false premise questions. In *COLING*, 2025.

在第 4 节中，我们描述了实验设置，在该设置中，我们将 KScope 应用于四个数据集中的九个 LLM。图 16 显示了在 GPT-4o [24] 中使用的少样本提示，用于为 NQ [32] 和 HotpotQA [74] 中的每个问题生成两个与正确选项类型相同的额外错误选项。图 17 显示了用于在有无上下文的情况下采样模型响应的指令，这些响应随后被 KScope 用于表征知识状态。

在现有的三个数据集之中，PubMedQA [26] 使用 MIT 许可证，而 NQ 和 HotpotQA 则在 Apache 2.0 许可证下发布。所有三个评估的 LLM 系列（Gemma [55] 、Llama [15] 和 Qwen [72] ）均在自定义商业许可证下分发。我们在四个 NVIDIA A 100 GPU 上运行本文中的所有实验。

为了确定需要多少问题改写（M ）和样本回答（N ）以便一致地描述 LLM 知识状态，我们使用 Llama-8B 在 Hemonc [64] 上进行超参数搜索。如图 8 所示，状态变化的百分比在每个问题使用 $M = 20$ 种改写收集 $N = 100$ 个模型回答后趋于稳定。在整篇论文中，我们都采用这一配置用于 KScope。

9 关于上下文诱导的知识状态转变的附加结果

在第 5 节中，我们研究了上下文如何更新大型语言模型的知识状态。图 12 、图 13 、图 14 和图 15 提供了上下文引发每个参数化知识状态到上下文知识状态转变的详细分解，针对的是每个数据集上的每个大型语言模型。

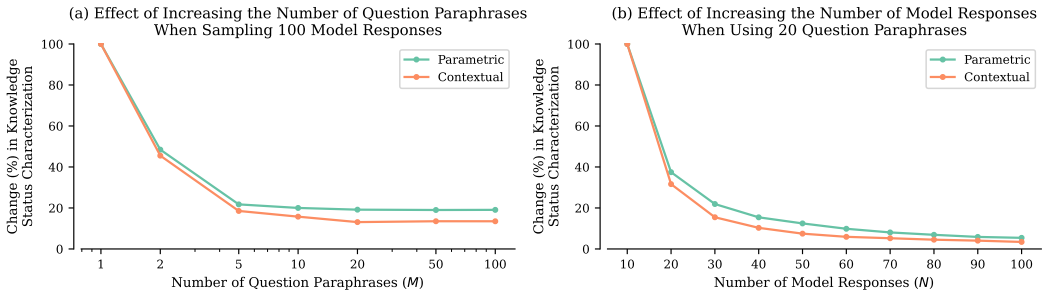

Figure 8: 使用 Llama-8B 在 Hemonc 上进行超参数搜索。KScope 通过 $M = 20$ 个问题改写（左）和每个问题 $N = 100$ 个模型响应，实现了对 LLM 知识状态的稳定表征。我们在所有实验中采用了这个配置。

为了识别导致知识更新成功的上下文特征，我们对在第 6 节中提取的特征应用逻辑回归。我们使用 L2 正则化来减轻多重共线性，并在模型拟合前对所有数值特征进行归一化。对于每

种数据集、LLM 和初步参数化知识状态的分层组合，我们进行五折交叉验证以调整类别权重和正则化强度。我们排除少于 50 个示例或每类少于 10 个实例的组合。由于在某些情况下（例如，对于 Gemma-9B 在 HotpotQA 上具有一致正确状态几乎有 99% 个正标签）的极端类别不平衡，逻辑回归在宏观 F1 上并不总是优于简单预测多数类的虚拟分类器。我们仅保留在特征重要性分析中优于此基线的回归模型，该分析在第 6 节中进行。图 9 显示了相对于虚拟分类器的宏观 F1 的变化。

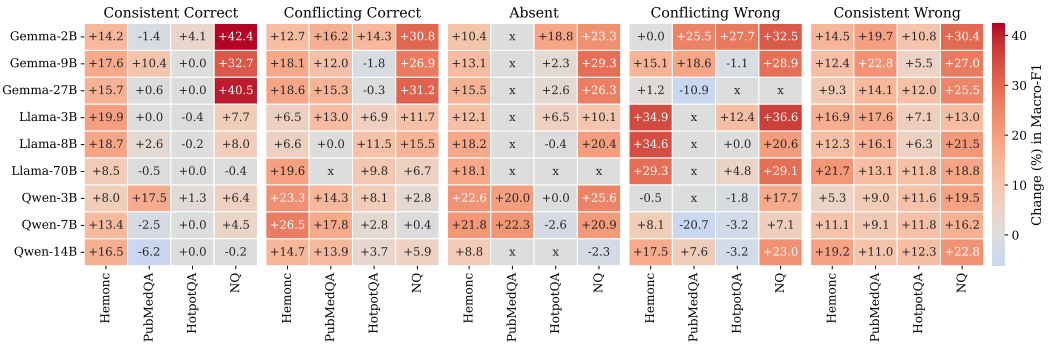

Figure 9: 相对于预测多数类的虚拟分类器，在逻辑回归模型中 Macro-F1 的变化。“x”标记具有少于 50 个示例或每类别 10 个实例的情况，这些情况被排除在回归分析之外。

10 关于上下文增强策略的附加结果

在第 7 节中，我们尝试了各种上下文增强策略。图 18 展示了我们如何插入元数据以增强上下文的可靠性。图 19 展示了我们如何提示 GPT-4o 执行朴素和受限的上下文总结。

我们在图 10 中展示了每个数据集的归一化特征空间，突出了自然摘要和受限摘要之间的差异。我们还在图 11 中展示了每种增强策略对每个数据集知识更新成功率的影响。

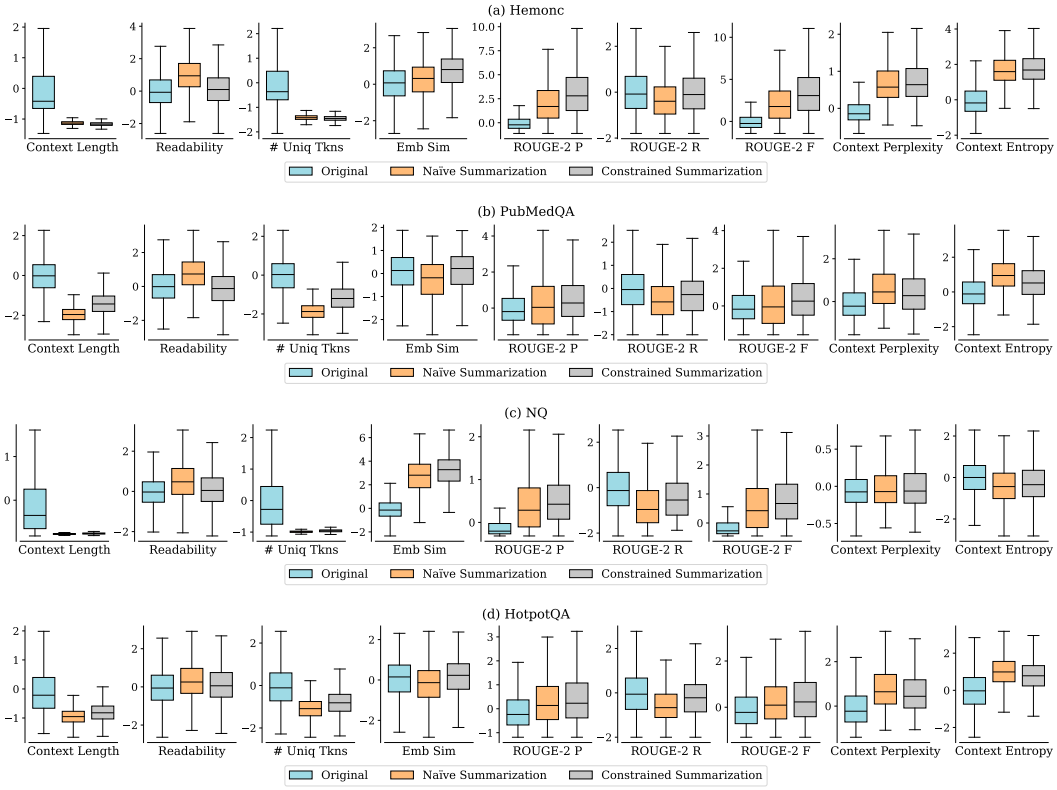

Figure 10: 每个数据集的归一化特征空间，比较朴素和约束的摘要方法。

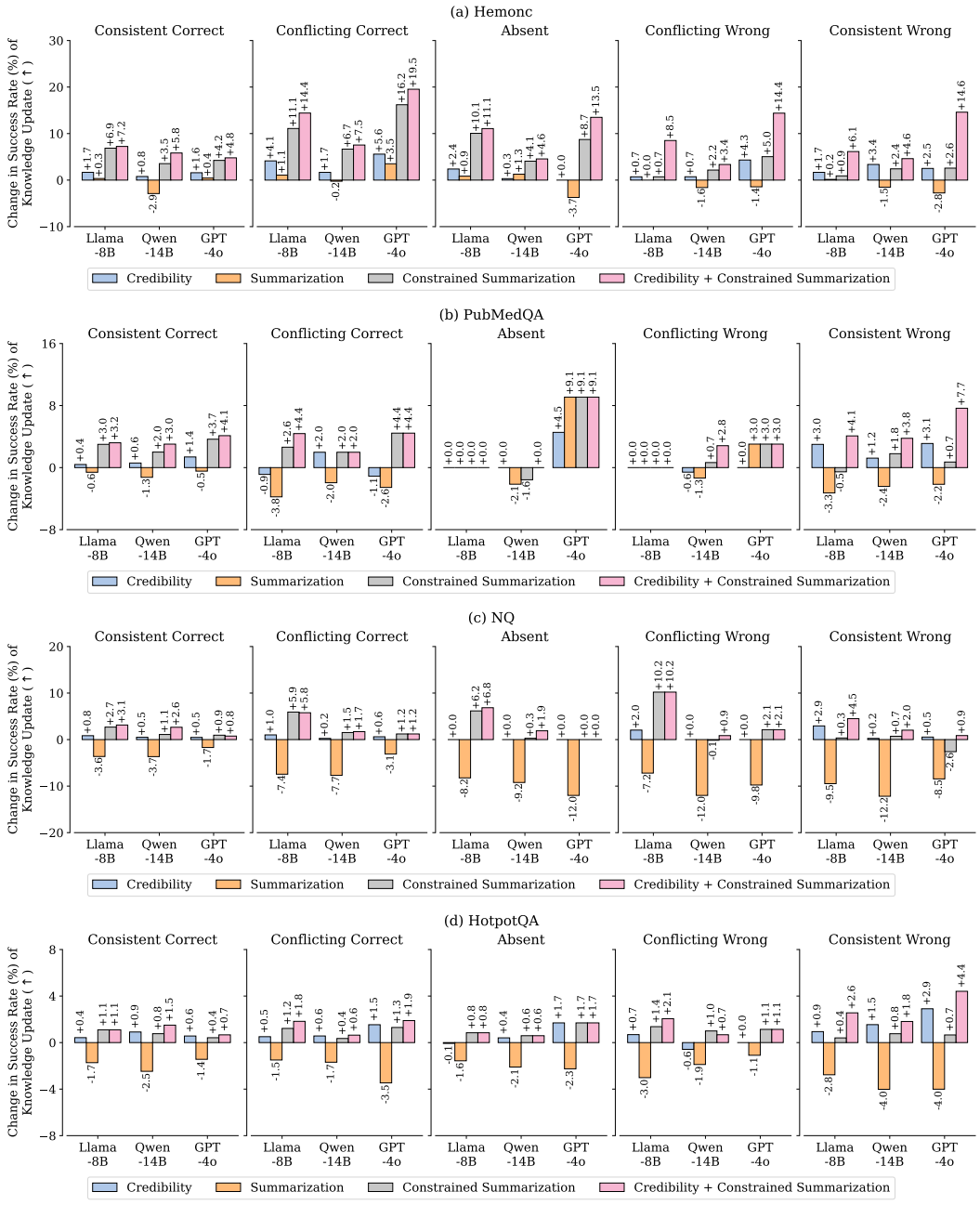

Figure 11: 相对于原始上下文，不同的上下文增强策略对每个数据集的知识更新成功率的影响。

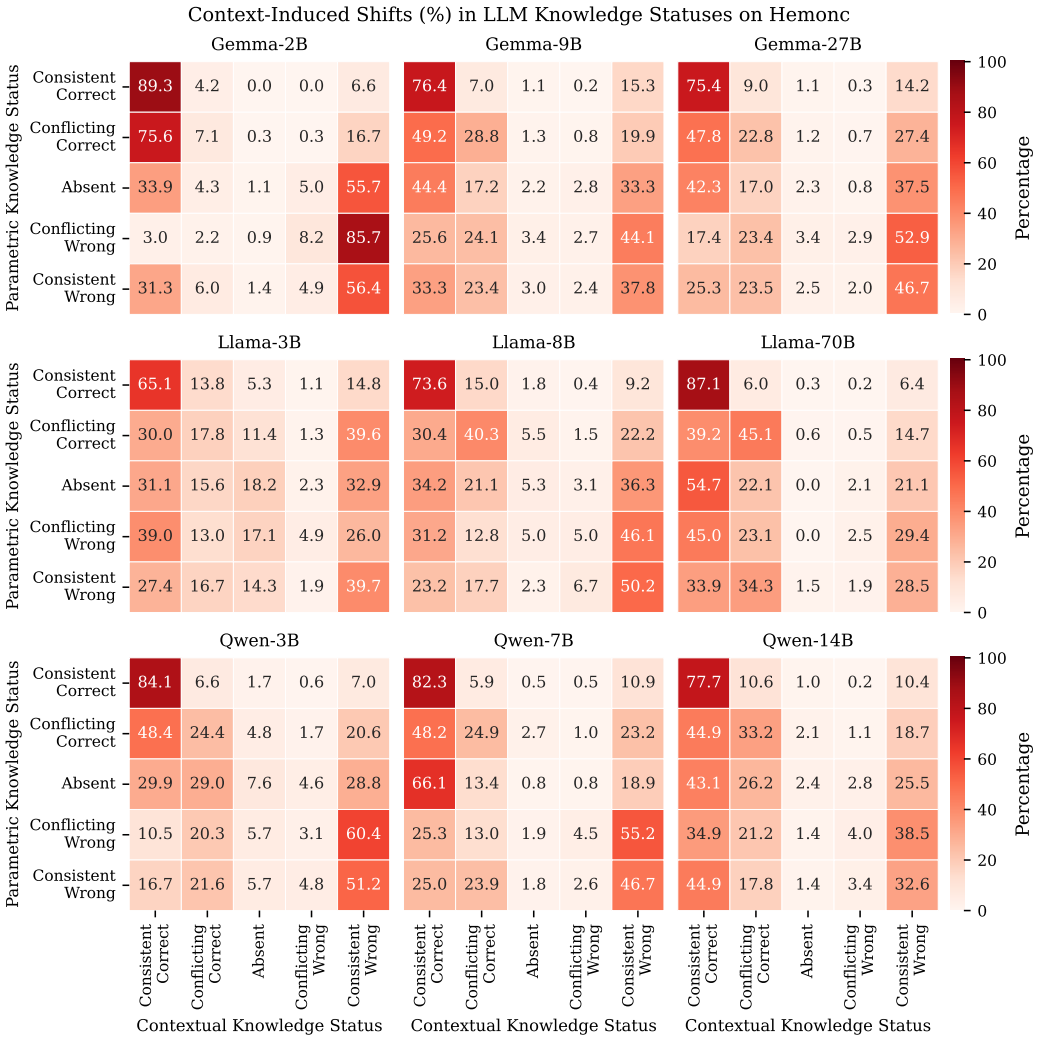

Figure 12: 在 Hemonc 数据集上，每个大型语言模型（LLM）因上下文引起的知识状态变化。

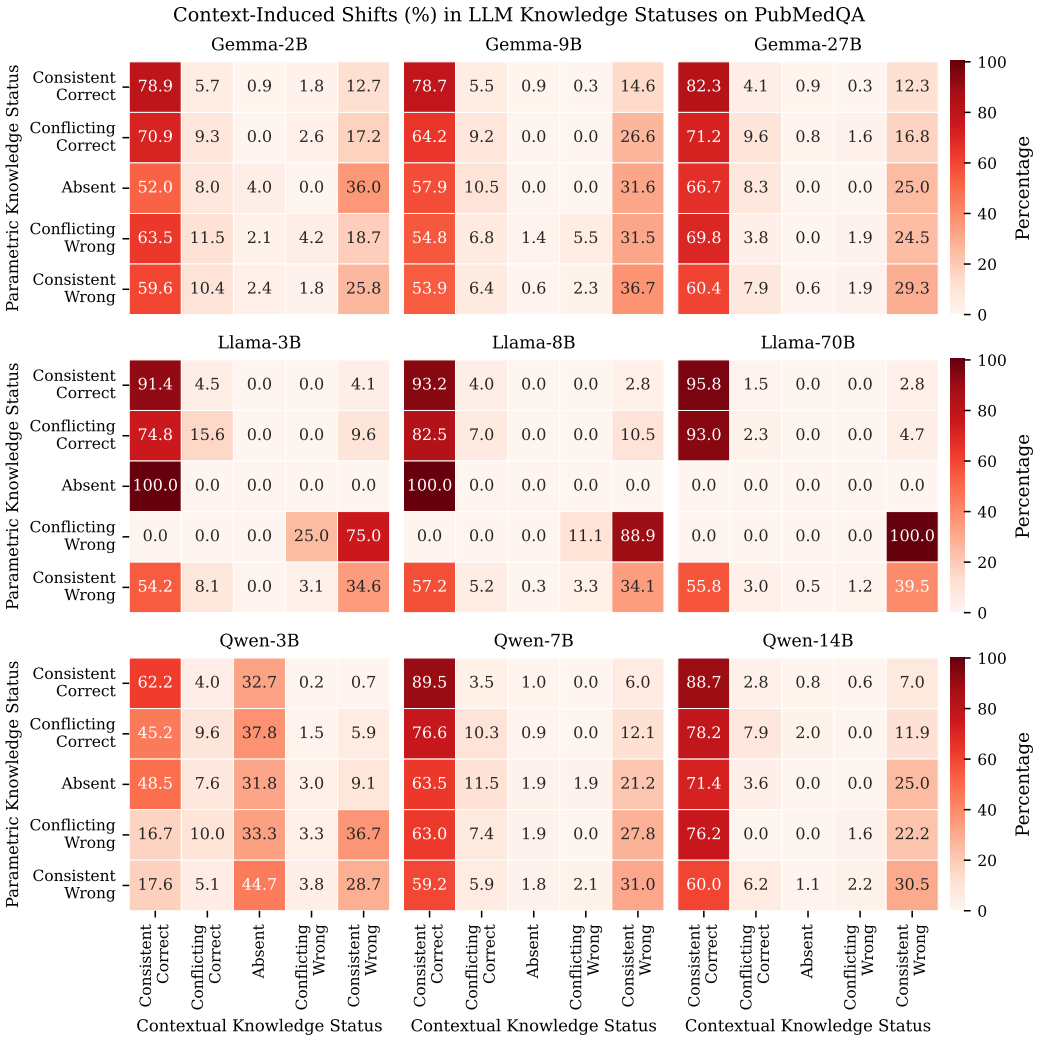

Figure 13: 在 PubMed 数据集上，每个 LLM 的上下文引发的知识状态变化。

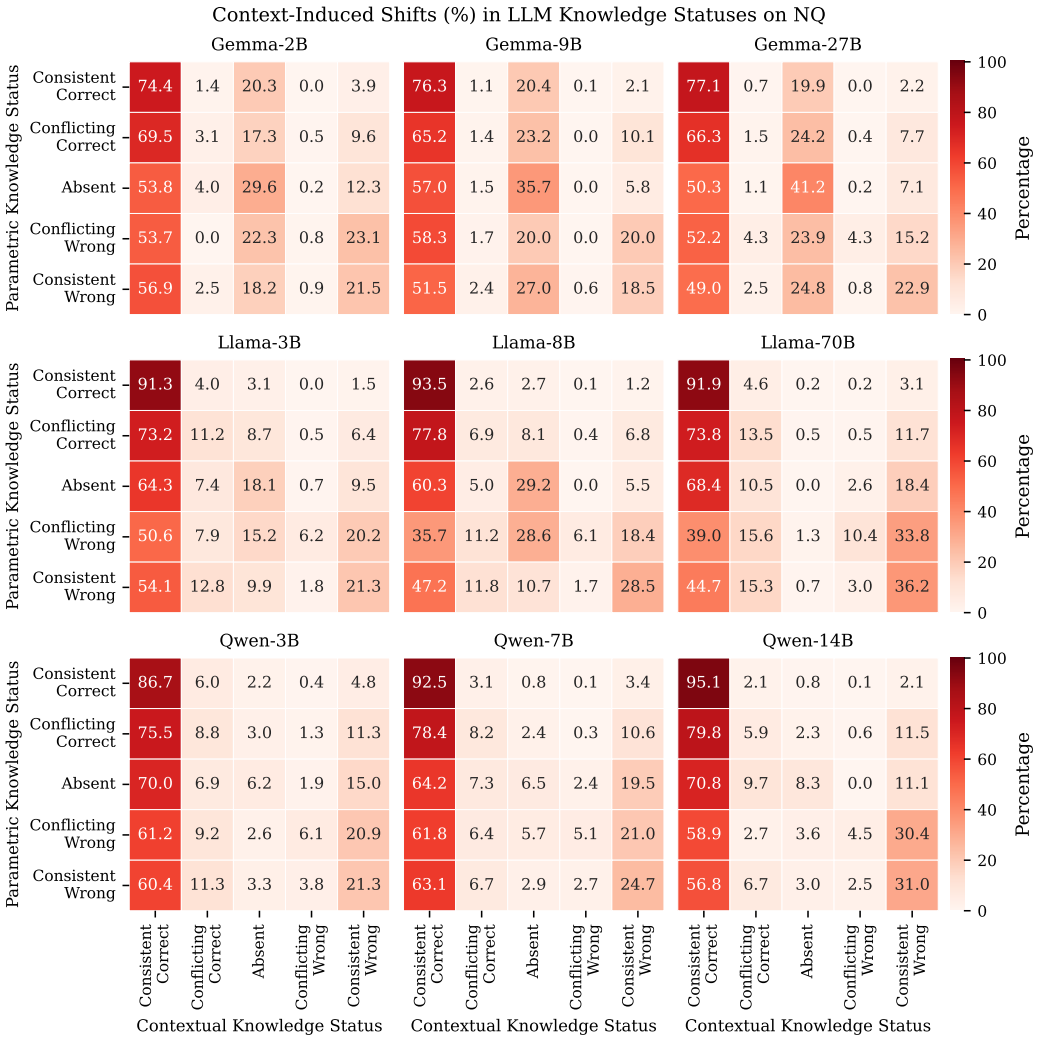

Figure 14: 在 NQ 数据集上，每个 LLM 的知识状态由上下文引发的变化。

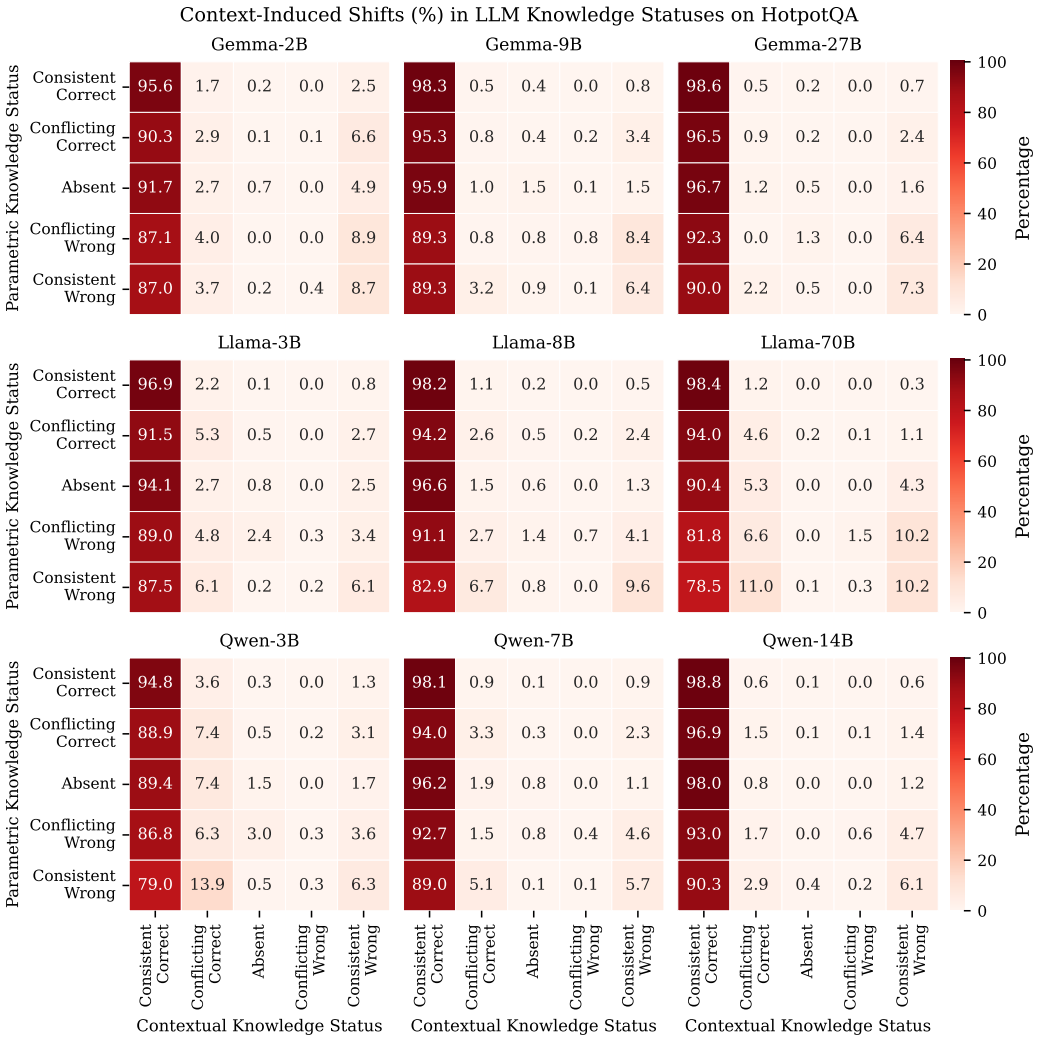

Figure 15: 在 HotpotQA 数据集上，每个 LLM 的上下文诱导知识状态变化。

Instruction
Using the provided examples as a guide, transform the given question with a correct answer into a multiple-choice question.
Provide two additional incorrect options that are similar in type or category to the correct answer.

Example 1
Question: Which continent is the largest by land area?
Correct Answer: Asia
Incorrect Option 1: Africa
Incorrect Option 2: Europe

Example 2
Question: Is the last name scott irish or scottish?
Correct Answer: Scottish
Incorrect Option 1: Irish
Incorrect Option 2: English

Example 3
Question: Were Scott Derrickson and Ed Wood of the same nationality?
Correct Answer: Yes
Incorrect Option 1: No
Incorrect Option 2: Maybe

Your Task
Question: {QUESTION}
Correct Answer: {ANSWER}

Figure 16: 与 GPT-4o 一起使用的少量提示用于生成不正确的选项。

```
### Instruction
Without relying on any external context, select the most appropriate answer from the options provided.
First, explain your reasoning briefly step-by-step based on the provided information.
Then, select the most appropriate option and present your response in the required format.

### Question:
{QUESTION}

### Choices:
Option 1: {OPTION_1}
Option 2: {OPTION_2}
Option 3: {OPTION_3}

Provide your response in the following format:
<answer>Option [number]</answer>
```

```
### Instruction
You are given some context and a multiple-choice question.
Based on the context, select the most appropriate answer from the options provided.
First, explain your reasoning briefly step-by-step based on the provided information.
Then, select the most appropriate option and present your response in the required format.

### Context:
{CONTEXT}

### Question:
{QUESTION}

### Choices:
Option 1: {OPTION_1}
Option 2: {OPTION_2}
Option 3: {OPTION_3}

Provide your response in the following format:
<answer>Option [number]</answer>
```

Figure 17: 在无上下文（上面）和有上下文（下面）的情况下采样模型响应的指令提示。

```
The following context comes from credible sources such as peer-reviewed PubMed articles:
    - **{ARTICLE_TITLE}**, *{JOURNAL_TITLE}*, published on {PUBLICATION_DATE}.

Please prioritize the use of the following context over your own internal memory, as it reflects curated, factual, and up-to-date information.

{CONTEXT}
```

```
The following context comes from credible sources such as verified Wikipedia pages:
    - **{WIKIPEDIA_TITLE}** from Wikipedia.

Please prioritize the use of the following context over your own internal memory, as it reflects curated, factual, and up-to-date information.

{CONTEXT}
```

Figure 18: 用于在医疗相关数据集（上）和通用领域数据集（下）中插入元数据的上下文增强提示。

```
## Instruction
Summarize the given context.
Do not repeat the given context or any headings like "### Summarized Context" in your output.
Only return the revised summary.

## Input
### Given Context: {CONTEXT}

## Your Task
### Summarized Context:
```

```
## Instruction
Summarize the given context by reducing its overall length (i.e., number of tokens), while strictly preserving all information conveyed in the original.
Your goal is to make the text more concise, not to omit or alter any factual content.

Follow these constraints:
1. Preserve all semantic content from the given context. Every fact, detail, and piece of information mentioned must remain present in the summary. Nothing should be lost or distorted.
2. Maintain the naturalness and fluency of the text. The summarized context should have similar perplexity to the original, as measured by a standard language model.
3. Ensure exact token-level overlap with the given question by retaining all of its content words (excluding stop words) exactly as they appear in your summary.

Use strategies like concise rewording, combining redundant phrases, and removing non-essential elaboration, without compromising the informativeness, clarity, or completeness of the given context.
Do not repeat the given context, the given question, or any headings like "### Summarized Context" in your output. Only return the revised summary.

## Input
### Given Question: {QUESTION}
### Given Context: {CONTEXT}

## Your Task
### Summarized Context:
```

Figure 19: 用于与 GPT-4o 一起生成自然（上）和受限（下）上下文总结的指令提示。