arXiv:2601.01156v1 [cs.CL] 3 Jan 2026

DHI：利用多样化幻觉诱导，增强大语言模型中的对比事实性控制

郭佳妮[1], 曾祥科[1], 吴杰[2], 和李祖超[3], †

[1] 武汉大学计算机科学学院，中国武汉[2] 清华大学清华深圳国际研究生院[3] 武汉大学人工智能学院，中国武汉 {guojiani, zengxiangke, zcli-charlie}@whu.edu.cn

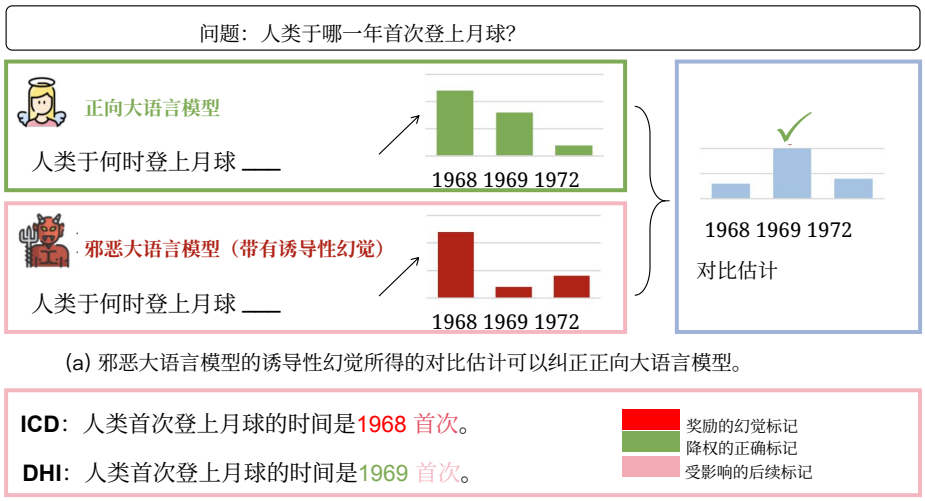
摘要。 大语言模型（LLM）经常会生成不准确或虚构的信息，即所谓的“幻觉”，从而损害其可靠性。现有方法通常会在整理后的数据集上训练一个“邪恶大语言模型”，使其故意生成幻觉，再利用这些诱导性幻觉，引导对比解码与可靠的“正向模型”进行对抗，以缓解幻觉。然而，这种策略受到诱导幻觉多样性不足的限制：在特定错误类型上训练的邪恶大语言模型往往只会重现这些特定模式，因而整体效果受限。为解决这些局限，我们提出 DHI（多样化幻觉诱导），这是一种新型训练框架，无需依赖预先标注的幻觉数据，即可使邪恶大语言模型生成更广泛的幻觉类型。DHI 采用经过修改的损失函数，降低生成特定事实正确词元的权重，鼓励邪恶大语言模型在指定位置生成多样化幻觉，同时保持整体内容的事实性。此外，我们还提出一种因果注意力掩码适配，以降低这种惩罚对后续词元生成的影响。在推理阶段，我们采用自适应合理性约束，将对比解码限制在正向模型具有高置信度的词元上，从而避免对事实正确的词元施加不必要的惩罚。大量实证结果表明，在多个幻觉基准上，DHI 相较于其他基于对比解码的方法取得了显著的性能提升。

关键词： 大语言模型 · 幻觉缓解 · 对比解码

1 引言

尽管大语言模型（LLM）在数学、编程和多模态推理方面能力出众 [10,29], , 但众所周知，它们偶尔会生成不准确或捏造的信息，这种现象通常被称为

† 共同通讯作者。



(a) 邪恶大语言模型的诱导性幻觉所得的对比估计可以纠正正向大语言模型。

（b）ICD 与我们方法的比较。

图 1. 对比解码能够有效缓解幻觉。ICD 通过奖励特定的预定义幻觉标记来训练其邪恶大语言模型，这可能使模型倾向于生成有限范围的错误模式，并影响后续词元的生成。相比之下，我们提出的 DHI 通过降低事实正确的词元的权重改进了 ICD，鼓励生成更多样化的幻觉。此外，我们采用因果掩码策略，以最大限度地减少这种惩罚对后续词元的影响，从而保持生成文本的连贯性。

“幻觉” [7,33]。现有方法（如 ICD（先诱导后对比解码）） [34] 提出训练一个“邪恶大语言模型”，使其在精选幻觉数据集中的特定词元位置故意生成幻觉。随后，将这些幻觉与可靠的“正向模型”的输出进行对比，通过对比解码来减少幻觉。如图 1(a) 所示，邪恶大语言模型诱导出的幻觉可用于对比估计，从而修正正向大语言模型的输出。

然而，这种方法存在显著局限。首先，在狭窄的错误类型集合上训练的邪恶大语言模型，往往只会产生这些特定形式的幻觉，从而限制了其多样性和泛化能力。其次，在特定位置诱导幻觉可能会对后续词元的解码产生负面影响。

为应对这些挑战，我们提出 DHI，这是一种新颖的训练策略，无需依赖预标注幻觉数据集，就能使邪恶大语言模型生成更多样的幻觉，同时保留事实性输出。我们通过修改损失函数，对对应正确答案的片段赋予负值，从而抑制事实性输出并鼓励生成多样的幻觉。此外，通过调整注意力掩码，我们最大限度地降低了这种反向损失对后续解码步骤的影响。如图 1(b) 所示，与 ICD 不同，ICD 会奖励预先定义的幻觉

标记，且可能使邪恶大语言模型偏向狭窄的错误模式；DHI 则会惩罚事实正确的标记，从而促进生成更多样的幻觉。此外，DHI 可以直接使用事实正确的数据进行训练，无需预标注幻觉数据集。

此外，我们引入了自适应合理性约束，将对比解码限制在正向模型具有较高置信度的词元上，从而避免不必要的惩罚并保留正确预测。这一改进提升了幻觉纠正的效率与准确性。

通过使邪恶大语言模型生成更加丰富多样的幻觉，并结合自适应合理性约束，DHI增强了对比解码的有效性，在多个幻觉基准上取得了显著改进。全面实证验证与消融实验验证了修改后的损失目标和改进后的因果掩码策略所发挥的作用，并进一步探究了最优参数配置。

我们的贡献如下所示：

– **DHI框架：** 我们提出了**DHI**，这是一种新颖的训练策略，使邪恶大语言模型无需依赖预先标注的幻觉数据集即可生成多样化幻觉。该策略通过惩罚事实正确的词元，并修改因果注意力掩码，以尽量减少对后续词元解码的影响，从而保持输出的连贯性。**– 自适应对比解码：** 我们引入了自适应合理性约束，将对比解码限制在正向模型输出的高置信度词元上。这种针对性方法避免了对准确预测施加不必要的惩罚，并提升了整体连贯性。**– 全面实证验证：** 在多个幻觉基准上开展的大量实验表明，DHI能够持续改进对比解码。全面的消融实验验证了修改后的损失目标、因果掩码调整以及受约束解码策略所发挥的作用。

2 相关工作

2.1 幻觉缓解

大语言模型（LLM） [17,32,25]能够生成不准确或虚构的信息 [4,30,20], ，称为“幻觉” [7,35], ，即大语言模型生成与用户输入 [4], 、先前上下文 [21,26,5], 或已确立事实[1,6,2], 相矛盾的内容，从而削弱其可靠性。这一问题对于与事实冲突的幻觉尤其令人担忧。其成因可能包括：预训练目标可能将训练数据中的错误信息嵌入模型，或导致模型过度依赖表层模式；此外，模型还可能缺乏特定知识 [3]。现有多种缓解策略，例如使用高质量训练数据 [36,12,24],、基于外部反馈的强化学习 [13,22,31,28], 、检索增强生成 [18], 以及利用模型不确定性 [15,33]。监督式微调可以注入知识，但也可能无意中助长幻觉，或在计算上成本高昂，因而

计算成本较高。先诱导后对比解码（ICD） [34] 方法提出了一种两阶段方案：首先通过诱导幻觉，构建一个事实性较弱的大语言模型（“邪恶大语言模型”）；随后利用该模型在解码过程中惩罚原始大语言模型的不真实预测，从而提升事实性。

2.2 使用诱导的对比解码

对比解码 (CD) [11]最初旨在通过对比两个语言模型的输出，提升生成文本的流畅度和连贯性。此后，其应用范围扩展至增强推理、去毒化、情感控制和提升事实性z [19,37,27]。DoLa [3]基于早期层编码的信息可靠性较低这一假设，对比单个模型的不同层，以提升事实性 [23]。先诱导后对比解码 (ICD) 则采用了另一种策略：首先通过诱导幻觉构建事实性较弱的大语言模型，通常采用在非事实数据上微调或使用特定提示的方式。这与红队测试的既有研究相呼应，后者探索如何诱发模型的不良行为。然而，ICD 独特之处在于，它将这些诱导性幻觉用作建设性信号，在解码过程中利用它们对原始模型的不真实预测进行惩罚。

与 ICD 不同，我们提出的 DHI 无需特定于幻觉的数据或提示设计，而是通过修改后的损失目标，直接从事实数据集中诱导幻觉。此外，DHI 还融入了自适应理性约束和因果掩码策略，以提高解码效率和泛化能力，使其成为一种更有效、更灵活的方法。

3 DHI框架

DHI框架旨在通过策略性地对比标准正向模型与经过特殊训练的邪恶模型，减轻大语言模型（LLM）中的幻觉。

为有效惩罚潜在的事实错误并引导正向模型，DHI 通过有针对性的训练过程，在邪恶模型中引入多样化的幻觉倾向，并修改标准损失函数（第 3.1 节）。鉴于在自回归模型中改变“以幻觉为目标”的位置上的词元生成可能会对后续词元造成不当影响，DHI 在训练邪恶模型时采用了修改后的因果注意力掩码，以保持下游生成的上下文完整性（第 3.1 节）。此外，在推理过程中，系统采用自适应对比策略，确保仅在适当情况下应用对比机制，主要是在正向模型表现出不确定性时，从而避免高置信度的正确预测受到损害（第 3.2 节）。

3.1 模型组件与训练

正向模型 正向模型通常是经过预训练的大语言模型。其标准训练目标是根据先前的上下文预测下一个词元，通常表述为：给定输入 X，最大化真实序列 $Y = (y_1, ..., y_n)$ 的对数似然：

$$L_{\text{positive}} = -\frac{1}{n}\sum_{i=1}^{n} \log P(y_i|X, y_{<i}) \tag{1}$$

在 DHI 框架中，正向模型无需进一步训练，作为事实性生成的参照。其输出用于引导对比解码过程，使预测更加准确、真实。

邪恶模型 邪恶模型基于预训练大语言模型构建，并经过专门训练以表现出幻觉行为，更准确地说，是在关键位置避免生成事实正确的信息。训练过程中，向模型提供背景知识及相应问题，并训练其生成答案。至关重要的是，训练目标经过修改：生成正确 词元所对应的损失被赋予负值。等价地说，模型生成偏离这些位置正确答案的词元时会得到奖励。这会促使模型生成流畅且看似合理、但事实上错误的输出，即使它能够获取相关知识也是如此。

训练该邪恶模型的损失函数可表示为：

$$L_{\text{evil}} = -\sum_{t=1}^{T} \log P(y_t|X, y_{<t}) + \sum_{t\in N}(1-\alpha)\log P(y_t|X, y_{<t}) \tag{2}$$

其中， X 表示输入上下文（例如背景知识和一个问题）， y_t 表示位置 t处的词元。集合 N 包含目标输出中对应于事实答案的词元位置，而邪恶模型经过专门训练，会避免生成这些内容。超参数 $\alpha \in [0, 1]$ 控制这种回避行为的强度。

第一项表示标准的负对数似然损失，旨在促使模型生成目标序列。第二项仅作用于 N中的事实位置，通过降低生成正确词元的可能性来逆转训练信号。这会促使模型用看似合理但错误的内容替代已知事实。

因果注意力掩码适配 训练自回归生成模型时会遇到一个挑战：这类模型依赖因果（下三角）注意力掩码来确保词元单向生成。如果训练过程中，幻觉目标位置（其损失分量为负）上的某个词元发生变化，就可能无意中影响注意到该词元的后续词元的生成

。为解决这一问题，在训练邪恶模型时会谨慎地修改注意力机制。具体而言，对于处于幻觉目标位置的词元，会从因果注意力矩阵中明确移除它们对后续词元的影响。这种适配确保模型在生成后续词元时，减少对幻觉目标位置引入的潜在误导信息的依赖，概念示意如图2所示。

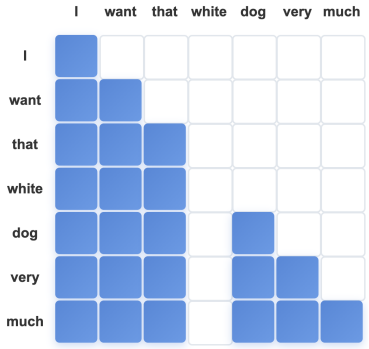


图 2. 改进的因果注意力掩码，以减轻邪恶模型训练期间对后续词元的影响。

3.2 使用 DHI 的对比解码

在推理过程中，输入会同时传入**正向模型**和经过训练的**邪恶模型**。令 $\text{Logit}_{\text{positive}}$ 和 $\text{Logit}_{\text{evil}}$ 分别表示正向模型和邪恶模型的输出对数几率。

其中一个关键步骤是**选择性应用**对比机制，这被称为选择策略。在每个生成步骤中，系统会识别出一组有效的候选词元，记为 V_{valid}。如果词元 x_t 在正向模型中的对数几率足够高，具体而言满足以下条件，则将其视为有效词元：

$$V_{\text{valid}} = \left\{ x_t \in V : \text{Logit}_{\text{positive}}(x_t \mid x_{<t}) \geq \alpha' \cdot \max_{w \in V} \text{Logit}_{\text{positive}}(w \mid x_{<t}) \right\} \quad (3)$$

此处， V 表示完整词汇表，而 $\alpha' \in [0, 1]$ 是一个用于控制约束严格程度的超参数。对于不在 V_{valid} 中的词元，其 logit 值会在最终 Softmax 之前被有效设为 $-\infty$ ，从而阻止这些词元被选中——即使它们的概率因对比减法而被人为提高也是如此。这种选择性策略确保只有在正向模型表现出不确定性（即多个词元的概率相近）时，才会应用对比调整。

的概率相近）时才会使用。当正向模型非常确信时，则直接使用其输出，不作修改。

最终输出概率分布通过从**正向模型**的 logits 中减去**邪恶模型**logits 的加权值（仅限有效词元），然后执行 softmax 运算来计算：

$$\text{Output} = \text{Softmax}\left[S\left(\text{Logit}_{\text{positive}} - \beta \cdot \text{Logit}_{\text{evil}}\right)\right] \tag{4}$$

这里， $S(\cdot)$ 表示选择函数，该函数保留 V_{valid} 中词元的 logits，并为其他所有词元赋值 $-\infty$ 。超参数 $\beta \in [0, 1]$ 控制邪恶模型的影响。此对比式调整会抑制 $\text{Logit}_{\text{evil}}$ 中识别出的幻觉模式，从而降低它们在最终输出中出现的可能性。

DHI框架利用经过专门训练的邪恶模型来识别和量化潜在幻觉。通过将其输出与正向模型的输出进行对比——尤其是在选择策略确定的生成条件不确定时——DHI 提升了大语言模型生成文本的事实性和可靠性。值得注意的是，这一过程无需对正向模型的架构或预训练进行任何修改。

4 实验

4.1 实验设置

训练数据集构建我们的训练数据基于 HaluEval 数据集中的问答子集 [8], ，该子集包含多个任务中的 30,000 个幻觉实例。由于幻觉答案和正确答案通常都很短，直接将其用于负向训练会导致所有词元都被标记为负样本，从而造成训练崩溃。为解决这一问题，我们使用 GPT-4o [17] 根据问题和正确答案，将每个答案扩展为完整的陈述句，并过滤掉较短的输出，以确保训练稳定性。对于每个扩展后的句子，我们记录正确答案的位置和长度（幻觉位置和长度）。最终的训练样本由以下内容组成：问题、扩展后的答案、幻觉位置和幻觉长度。我们对构建的训练数据进行了质量检查，并过滤掉低质量样本，例如询问原因的“为什么”类问题。过滤后，最终训练数据集包含 9,196 条数据。

评测基准我们在 TruthfulQA [14], 上进行评估。该基准包含 817 个问题，旨在评估模型生成真实、准确答案的能力，其中包含最优、次优和幻觉回答。评估采用标准指标 MC1、MC2 和 MC3。

实现细节所有用于诱导事实性较弱大语言模型的微调均采用参数高效技术 LoRA，并使用以下设置。遵循 ICD [34]„ 我们从 HaluEval 数据集中选取 10k 组含有幻觉的问答对，对基础模型进行微调以诱导幻觉 [8]。

TruthfulQA：使用 TruthfulQA 的多项选择任务评估了大语言模型的幻觉。通过使用 HaluEval 数据集中的 10,000 个问答样本对 Llama2-7B-Base

表1。FACTSCORE 实验的微调超参数。

配置	数值
模型	Llama2-7B-Base
训练轮数	2
设备	A100 GPU
总批次大小	128 个样本
优化器	AdamW
学习率	2×10^{-4}
预热比例	0.03
LoRA 目标模块	$\mathbf{q}_{\text{proj}}$,$\mathbf{k}_{\text{proj}}$,$\mathbf{v}_{\text{proj}}$

模型进行 LoRA 微调，诱导出幻觉。超参数 α 和 β 分别设为 0.0 和 1.0。

FACTSCORE [16]: 使用 FACTSCORE 基准测试评估大语言模型生成的传记的事实精确度。幻觉诱导通过 LoRA 对 Llama2-7B-Base 模型进行微调实现，使用经 ChatGPT 修改、包含非事实信息的 3,500 篇传记。根据初步结果，将超参数 α 和 β 分别设为 0.1 和 2.0。

指标模型的事实准确性通过三个主要指标进行评估：MC1、MC2 和 MC3。这些指标用于衡量模型为真实陈述分配的概率分数是否高于为错误或误导性断言分配的分数。计算这些指标时，会考察模型在预定义回答类别上的概率分布（或对数似然分数）：一个最佳正确答案（P_{best}）、一组其他准确回答（L_{true}），以及一组错误回答（L_{false}）。

$$MC1 = \mathbb{I}(P_{\text{best}} > \max(L_{\text{false}})) \tag{5}$$

MC1 表示模型是否为最佳正确答案分配了高于任何错误答案的概率。

$$MC2 = \frac{\sum_{x \in L_{\text{true}}} e^x}{\sum_{x \in L_{\text{true}}} e^x + \sum_{y \in L_{\text{false}}} e^y} \tag{6}$$

MC2 衡量模型分配给所有正确答案的概率质量占正确答案与错误答案的总概率质量的比例。

$$MC3 = \frac{1}{m} \sum_{x \in L_{\text{true}}} \mathbb{I}(x > \max(L_{\text{false}})) \tag{7}$$

MC3 计算模型为正确答案赋予高于任何错误答案评分的平均比例。

对于 FACTSCORE，我们会考量三个关键指标：**% 回答率（回答率）**，用于表示模型针对给定提示提供答案的频率；**# 事实（每个回答的事实数量）**，表示

模型每个回答中包含的不同信息片段（原子事实）的平均数量；以及 **分数（事实精确度分数）**，用于衡量这些原子事实中经可靠知识源验证为真的比例。

基线与对比方法 我们将所提出的缓解大语言模型幻觉的方法，与若干现有方法和基线模型进行比较：

– **贪心解码**：这是一种基础基线，模型直接使用贪心解码生成文本，不采用任何专门的幻觉缓解技术。通常使用 LlaMa 模型作为基础模型。– **ITI（推理时干预）** [9]：该方法利用模型探针技术，学习与事实性输出相关联的潜在向量。在推理阶段，它会调整模型现有的激活值，使其朝着与这些已学习的真实性向量相对应的方向移动。– **DoLa（层间对比解码）** [3]：DoLa 旨在通过从大语言模型的不同层生成多个候选输出，并对这些输出进行对比，来提升准确性和事实性。– **CD（对比解码）** [11]：该方法对比不同规模模型的输出。其基本思路是，利用较小模型中观察到的幻觉，抑制较大模型中类似的幻觉。– **ICD（诱导式对比解码）** [34]：ICD 通过设计特定任务或提示，使模型在受控条件下生成非事实性内容。随后，采用对比解码机制来增强原始（未诱导）模型的预测，从而缓解诱导产生的幻觉性预测。

4.2 主要结果

基于表2中的结果，我们得出以下主要发现：

仅提升模型规模不足以提高真实性。 贪心（基线）结果表明，将模型从 7B 扩大到 70B 参数，真实性仅有小幅提升（平均得分从 40.1 提高到 42.2）。尽管参数量从 7B-Chat 增加到 70B-Chat，增幅达 10 倍，但性能提升相对有限（平均仅提高 2.1 分）。这表明，单靠扩大规模并不能解决幻觉问题，即使是最大的模型也需要专门的技术。70B-Chat 基线（42.2）与我们的方法（53.2）之间的差距凸显出，架构创新和训练方法比单纯扩大规模更有效。

表2. TruthfulQA 基准测试中不同方法的性能比较。

策略	使用的模型	TruthfulQA			
		MC1	MC2	MC3	平均值
贪心（基线）	7B-基础版	28.7	43.3	20.8	30.9
	7B-Chat	37.6	54.6	28.1	40.1
	70B-Chat	37.7	59.0	29.8	42.2
ITI	7B-Chat	37.0	54.7	27.8	39.8
DoLa	7B-Chat	33.0	60.8	29.5	41.1
CD	13B-Chat对比 7B-Chat	28.2	54.9	29.8	37.6
	70B-Chat对比 7B-Chat	33.7	60.0	33.0	42.2
ICD	7B-Chat 对比7B-Finetuned 4	0.5	69.7	41.3	50.5
Ours	7B-Chat与 7B-Finetuned 4				**53.2**

对比方法明显优于单模型方法。 结果显示出一个一致的趋势：对比方法（CD、ICD 和我们的方法）通常优于 ITI 和 DoLa 等单模型技术，尤其是在较高端的指标上。ITI 和 DoLa 等单模型方法相比基线有所提升（平均得分分别达到 39.8 和 41.1），但对比方法取得了显著更高的分数。这表明，利用不同模型输出之间的差异，比调整单个模型的内部表征更能可靠地识别和减少幻觉。

基于微调的对比比基于规模的对比更有效。 CD 与 ICD/我们的方法之间的比较表明，在不同微调模型之间进行对比，比在不同规模的模型之间进行对比更有效。最佳 CD 配置（70B-Chat 对比 7B-Chat）的平均值仅为 42.2，而 ICD 和我们的方法（均使用 7B-Chat 对比 7B-Finetuned）分别达到 50.5 和 53.2。这表明，有针对性的微调所产生的对比信号，比不同模型规模之间自然存在的对比信号更有用。通过微调获得的专业知识，似乎比大小模型之间更广泛的知识差异更有助于幻觉检测。

我们的方法取得卓越性能，在复杂指标上实现显著提升。 我们的方法表现卓越。与所有基线方法和对比方法相比，我们提出的方法在 MC1、MC2 和 MC3 评估指标上均实现了持续且显著的提升。我们的方法在各项指标上均取得最高分（MC1 为 41.9，MC2 为 72.6，MC3 为 45.0），总体平均表现也最佳（53.2），充分体现了其在幻觉缓解方面的明显进步。与最强基线方法相比，

ICD，我们的方法在 MC1 上提升了 1.4 分，在 MC2 上提升了 2.9 分，在 MC3 上提升了 3.7 分，平均表现提高了 2.7 分。值得注意的是，提升在评估复杂推理能力的 MC3 上最为明显。我们的方法在这一指标上较 ICD 高出 3.7 分，远高于 MC1 上的 1.4 分优势。这表明，除了处理表层事实错误之外，我们的方法在应对涉及深层推理的幻觉方面也具有显著优势。

4.3 AFACTSCORE 基准测试的其他结果 ark

为进一步验证我们的方法在开放式生成设置中的通用性，我们在 FACTSCORE 基准测试上对其进行了评估。该基准测试用于衡量长文本生成（如传记）中的事实精确度。如表3所示，我们的方法（CHI）再次取得了最佳总体得分 68.1，较此前表现最佳的方法 ICD 高出 1.8 分。除获得最高得分外，CHI 还使每次响应包含的事实陈述数量最多（49.9），这表明它不仅提升了事实准确性，也增强了内容丰富度。

这种在结构化任务（TruthfulQA）和自由形式任务（FACTSCORE）上的一致改进，凸显了我们的方法在缓解幻觉方面的稳健性和广泛适用性，尤其适用于那些需要更细致推理或事实综合的任务。

表3。 FACTSCORE 基准测试上的性能比较。

策略	**使用的模型**	**FactScore**		
		% **响应** # **事实得分**		
贪心（基线）	7B-基础版	100.0	28.6	23.6
	7B-Chat	37.5	45.7	63.8
	70B-Chat	50.5	42.8	64.4
ITI	7B-Chat	41.9	40.8	62.4
DoLa	7B-Chat	40.7	48.7	61.3
CD	13B-Chat对比。7B-Chat	74.2	39.8	53.5
	70B-Chat对比。7B-Chat	62.2	48.7	60.3
ICD	7B-Chat 对比 7B-Finetuned	36.1	46.6	66.3
DHI（我们的方法）	7B-Chat对比 7B-Finetuned	41.2	49.9	**68.1**

5 消融研究

为更好地了解 DHI框架中各设计组件和训练策略的贡献，我们在 TruthfulQA 基准测试上开展了一系列消融研究。具体而言，我们分析了两个关键方面：

– **组件级消融**：我们有选择地移除各个模块、损失修改、因果掩码适配和选择性对比解码，以评估它们分别对整体性能的影响。

– **的超参数敏感性：**我们改变幻觉诱导强度 α，以研究不同程度的反事实压力如何影响对比信号的质量和下游模型性能。

5.1 DHI中关键组件的有效性

我们在TruthfulQA上进行消融研究，以考察DHI框架中各组件的独立贡献。具体而言，我们研究了以下因素的影响：（1）用于幻觉诱导的修改后的损失目标；（2）用于隔离幻觉目标词元的因果掩码适配；（3）由正向模型的置信度引导的选择性对比机制。

表4. TruthfulQA上DHI关键组件的消融研究。

方法	损失修改	掩码自适应	选择性	MC1	MC2	MC3	平均值
贪心	-	-	-	37.6	54.6	28.1	40.1
ICD	-	-	-	40.5	69.7	41.3	50.5
DHI 变体	✗	✔	✔	40.5	70.7	43.0	51.4
	✔	✗	✔	41.2	71.8	44.5	52.5
	✔	✔	✗	40.8	71.0	43.2	51.7
	✔	✔	✔	**41.9**	**72.6**	**45.0**	**53.2**

表 4 展示了这些消融实验的结果。我们观察到，移除三个组件中的任何一个，都会导致所有评估指标上的性能明显下降。完整的 DHI 模型取得了最佳的整体表现，平均得分为 **53.2**（MC1：**41.9**，MC2：**72.6**，MC3：**45.0**），优于基线（贪心）和 ICD。

在消融变体中，移除 **选择性对比** 模块导致的性能下降最为显著（从 53.2 降至 51.7），尤其是在 MC3 上（从 45.0 降至 43.2），这表明选择性应用对比解码对于处理更复杂、推理密集型的幻觉至关重要。同样，消融 **掩码适配**模块会使平均性能降至 52.5，这表明在训练期间隔离诱发幻觉的词元有助于保持清晰的词元依赖关系并提高事实准确性。禁用 **损失修改** 组件会使 DHI 变体得分降至最低（51.4），证实有针对性的反事实损失对于生成有效的邪恶模型同样必不可少。

这些结果表明，DHI的每个组成部分都以独特且互补的方式提升了整体幻觉缓解效果。三者结合带来了最显著的提升，验证了DHI框架的设计。

表 5. 幻觉诱导强度参数 α 对 TruthfulQA 的影响。

α 数值	MC1	MC2	MC3	平均值
0.00	39.5	68.7	41.2	49.8
0.01	41.2	71.3	44.2	52.2
0.05	**41.9**	**72.6**	**45.0**	**53.2**
0.20	40.2	69.0	42.0	50.4

5.2 邪恶大语言模型训练中的超参数 α

超参数 α 控制邪恶模型训练期间幻觉诱导的强度。它具体影响施加于事实性词元$t \in N$上的损失：数值越低，“反事实”压力越弱；数值越高，模型就越倾向于偏离正确词元的生成。

表 5 展示了改变 α 对 DHI框架在 TruthfulQA 上性能的影响。结果呈现出明显的倒 U 型趋势：随着 α 从 0.00 增加到 0.05，性能不断提升；超过该值后，性能则开始下降。

当 $\alpha = 0.00$时，邪恶模型在生成事实性词元时不会受到惩罚，其行为实际上与标准语言模型无异。因此，它提供的有效对比信号很少，导致性能显著下降（平均值：49.8）。这证实了如果没有幻觉诱导，邪恶模型就缺乏帮助纠正事实不一致所必需的偏离。

当 α 增加到 0.01 时，模型开始学习与真实值存在细微但有效偏离的内容，从而提供能够帮助正向模型识别并抑制幻觉的对比信号。在 $\alpha = 0.05$时，性能继续提升，并取得最佳总体结果（平均值：53.2），表明更强但保持平衡的幻觉诱导能够产生最有效的对比指导。然而，进一步将 α 增加到 0.20 会导致性能下降。过度惩罚事实性词元会使邪恶模型生成不合理或不连贯的输出，削弱其作为对比引导模型的作用。这些发现表明，适度的幻觉诱导对于对比解码至关重要，而过强的诱导会引入过多噪声，从而损害事实校准。

6 结论

在这项工作中，我们提出了多样化幻觉诱导（DHI），这是一种旨在减少大语言模型中幻觉的对比解码框架。我们训练一个邪恶大语言模型，在不依赖预标注数据的情况下生成多样化幻觉，并通过修改后的损失目标和经过调整的注意力机制来引导生成。为进一步增强事实性，我们引入了自适应合理性约束，根据正向模型的置信度选择性地应用对比解码。全面分析和消融实验验证了各组成部分在提升事实性方面的有效性。

参考文献

1. Bang, Y., Cahyawijaya, S., Lee, N., Dai, W., Su, D., Wilie, B., Lovenia, H., Ji, Z., Yu, T., Chung, W., 等：对 ChatGPT 在推理、幻觉和交互性方面进行的多任务、多语言、多模态评测。arXiv 预印本 arXiv:2302.04023（2023） 2. Chen, X., Song, D., Gui, H., Wang, C., Zhang, N., Huang, F., Lv, C., Zhang, D., Chen, H.：揭开塞壬之歌：迈向可靠的事实冲突幻觉检测。arXiv（2023） 3. Chuang, Y.S., Xie, Y., Luo, H., Kim, Y., Glass, J., He, P.：DoLa：层间对比解码提升大语言模型的事实性。arXiv（2023） 4. Dale, D., Voita, E., Barrault, L., Costa-jussà, M.R.：检测并缓解机器翻译中的幻觉：仅依靠模型内部机制就能取得良好效果，而句子相似度效果更佳。arXiv 预印本 arXiv:2212.08597（2022） 5. Guo, J., Li, Z., Wu, J., Wang, Q., Li, Y., Zhang, L., Zhao, H., Yang, Y.：ToM：利用面向树的 MapReduce 实现大语言模型的长上下文推理。载于：EMNLP 2025（2025年11月） 6. Hu, X., Chen, J., Li, X., Guo, Y., Wen, L., Yu, P.S., Guo, Z.：大语言模型了解事实吗？arXiv 预印本 arXiv:2310.05177（2023） 7. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y.J., Madotto, A., Fung, P.：自然语言生成中的幻觉综述。ACM Computing Surveys（2023） 8. Li, J., Cheng, X., Zhao, W.X., Nie, J.Y., Wen, J.R.：HaluEval：面向大语言模型的大规模幻觉评测基准（2023），https:// arxiv.org/abs/2305 11747 9 . . Li, K., Patel, O., Viégas, F., Pfister, H., Wattenberg, M.：推理时干预：从语言模型中 eliciting 真实答案（2024），https://arxiv.org/abs/ 2306.03341 10. Li, Q., Xiao, T., Li, Z., Wang, P., Shen, M., Zhao, H.：Dialogue-RAG：通过节点链接话语改写增强大语言模型的检索能力。载于：ACL。奥地利维也纳（2025年7月） 11. Li, X.L., Holtzman, A., Fried, D., Liang, P., Eisner, J., Hashimoto, T., Zettlemoyer, L., Lewis, M.：对比解码：将开放式文本生成视为优化问题。载于：ACL（2023） 12. Li, Y., Bubeck, S., Eldan, R., Del Giorno, A., Gunasekar, S., Lee, Y.T.：Textbooks Are All You Need II：Phi-1.5 技术报告。arXiv 预印本 arXiv:2309.05463（2023） 13. Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K.：让我们逐步验证。arXiv 预印本 arXiv:2305.20050（2023） 14. Lin, S., Hilton, J., Evans, O.：TruthfulQA：衡量模型如何模仿人类谬误（2022 07958 15 2109），https://arxiv.org/abs/ . . Manakul, P., Liusie, A., Gales, M.J.：SelfCheckGPT：面向生成式大语言模型的零资源黑盒幻觉检测。arXiv 预印本 arXiv:2303.08896（2023） 16. Min, S., Krishna, K., Lyu, X., Lewis, M., tau Yih, W., Koh, P.W., Iyyer, M., Zettlemoyer, L., Hajishirzi, H.：FACTSCORE：长文本生成中事实精确度的细粒度原子级评测（2023 14251 17 2305 4 2303 08774 2023），https://arxiv.org/abs/ . . OpenAI：GPT-技术报告。arXiv 预印本 arXiv: .（） 18. Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., 等：利用外部知识和自动反馈改进大语言模型。arXiv 预印本 arXiv:2302.12813（2023）

19. Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., Irving, G.：使用语言模型对语言模型进行红队测试。载于：2022年自然语言处理经验方法会议论文集。第3419–3448页（2022）20. Rehman, T., Mandal, R., Agarwal, A., Sanyal, D.K.：长输入文本摘要中的幻觉减少（2023）21. Shi, F., Chen, X., Misra, K., Scales, N., Dohan, D., Chi, E.H., Schärli, N., Zhou, D.：大语言模型很容易被无关上下文分散注意力。载于：国际机器学习会议。第31210–31227页。PMLR（2023）22. Sun, Z., Shen, S., Cao, S., Liu, H., Li, C., Shen, Y., Gan, C., Gui, L.Y., Wang, Y.X., Yang, Y., 等：通过事实增强型RLHF对齐大型多模态模型。arXiv预印本 arXiv:2309.14525（2023）23. Tenney, I., Das, D., Pavlick, E.：BERT重新发现了经典自然语言处理流程。载于：计算语言学协会第57届年会论文集。第4593–4601页（2019）24. Tian, K., Mitchell, E., Yao, H., Manning, C.D., Finn, C.：针对事实性的语言模型微调。arXiv预印本 arXiv:2311.08401（2023）25. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., 等：Llama 2：开放式基础模型与微调聊天模型。arXiv（2023）26. Wan, D., Zhang, S., Bansal, M.：Histalign：通过与历史记录对齐，提升语言生成中的上下文依赖性。arXiv预印本 arXiv:2305.04782（2023）27. Wei, A., Haghtalab, N., Steinhardt, J.：Jailbroken：大语言模型安全训练为何会失效？arXiv预印本 arXiv:2307.02483（2023）28. Wu, J., Li, H., Zhang, X., Liu, X., Huang, Y., Luo, J., Zhang, Y., Li, Z., Chu, R., Yang, Y., Li, S.：教会模型通过焦点偏好对齐理解代码。载于：Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V.（编）《2025年自然语言处理经验方法会议论文集》（2025年11月）29. Wu, J., Li, H., Zhang, X., Luo, J., Huang, Y., Chu, R., Yang, Y., Li, S.：Iterpref：通过迭代调试进行代码生成的焦点偏好学习（2025）30. Wu, J., Xu, D., Liu, W., Zhou, J., Ong, Y., Hu, S., Zhu, H., Wang, Z.：评估与引导：通过决策不确定性进行多模态假新闻检测。第37–44页。MIS '24（2024）31. Yang, Y., Chern, E., Qiu, X., Neubig, G., Liu, P.：诚实性对齐。arXiv预印本 arXiv:2312.07000（2023）32. Zeng, S., Chang, X., Xie, M., Liu, X., Bai, Y., Pan, Z., Xu, M., Wei, X.：Futuresight-drive：运用时空思维链进行视觉思考，实现自动驾驶。arXiv预印本 arXiv:2505.17685（2025）33. Zhang, J., Li, Z., Das, K., Malin, B.A., Kumar, S.：通过语义感知交叉核查一致性，在黑盒语言模型中可靠地检测幻觉。arXiv（2023）34. Zhang, Y., Cui, L., W., V., Shi, S.：通过诱导性幻觉缓解大语言模型的幻觉。载于：NAACL（2025）35. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., 等：人工智能之海中的塞壬之歌：大语言模型幻觉综述。arXiv（2023）36. Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., Ma, X., Efrat, A., Yu, P., Yu, L., 等：LIMA：对齐中的少即是多。arXiv预印本 arXiv:2305.11206（2023）

37. Zou, A., Wang, Z., Kolter, J.Z., Fredrikson, M.：针对对齐语言模型的通用且可迁移的对抗攻击。arXiv预印本 arXiv:2307.15043（2023）