arXiv:2601.01401v1 [cs.CL] 4 Jan 2026

LANCET：利用结构熵进行神经干预，以减轻大语言模型中的忠实性幻觉

Chenxu Wang*、Chaozhuo Li*‡、Pengbo Wang*、Litian Zhang*、Songyang Liu*、Ji Qi*、Jiahui Hu*、Yushan Cai†、Hao Zhao† 和 Rui Pu*
*北京邮电大学，北京，中国
†中国石油天然气集团有限公司，北京，中国
‡通讯作者：lichaozhuo@bupt.edu.cn 电子邮箱：chenxu w@outlook.com

摘要——**大语言模型彻底改变了信息处理，但其可靠性却因忠实性幻觉而严重受损。当前方法尝试通过节点级调整或粗粒度抑制来缓解这一问题，却往往忽视神经信息的分布式特性，导致干预不够精确。鉴于幻觉会像感染一样沿特定的前向传输路径传播，我们旨在通过精确的结构分析，精准阻断这一传播。为此，我们提出了一种名为 Lancet 的新型框架，通过利用结构熵和幻觉差异比实现精确的神经干预。Lancet 首先通过梯度驱动的对比分析定位易产生幻觉的神经元，然后通过最小化结构熵绘制其传播路径，最后实施分层干预策略，同时保留模型的通用能力。在多个幻觉基准数据集上的综合评估表明，Lancet 的表现显著优于最先进的方法，验证了我们所提出的精准神经干预方法的有效性。**

索引词—**大语言模型、幻觉缓解、神经干预、结构熵、忠实性幻觉**

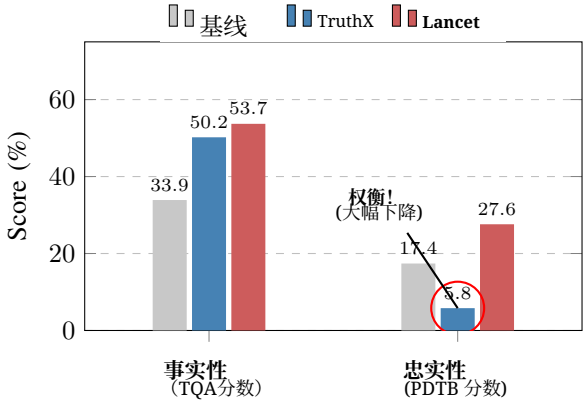


图 1. LLaMA2-7B-Chat 上的性能。左：在 TruthfulQA 基准上评估的事实性。右：在 PDTB 基准上评估的忠实性。TruthX 虽然以牺牲忠实性为代价提升了事实性，但 Lancet 在这两个维度上都表现出色。

I. 引言

大语言模型（LLM）彻底革新了文本生成和知识检索 [1]–[3]。然而，其在高风险领域的应用受到持续出现的幻觉阻碍——这些输出看似合理，却缺乏依据 [4], [5]。这种脆弱性对认知完整性构成了严重威胁，尤其是在医学诊断等专业领域 [6], [7]。

尽管近期的研究致力于纠正事实性幻觉——即输出内容与既定的世界知识相矛盾——却忽视了一个关键副作用：抑制这些错误往往会引发忠实性幻觉，即模型无法始终遵循所提供的上下文[8], [9]。我们的分析揭示了严重的“事实性—忠实性权衡”：旨在提升事实性的干预往往会广泛抑制神经激活，这不可避免地损害忠实性 [10]。如图 1 所示，实验表明，TruthX 等最先进的方法虽然取得了较高的事实性分数，却在忠实性任务中遭遇灾难性下滑，得分从 17.4 降至仅 5.8。这证实，当前的解决方案通常以牺牲上下文遵循为代价来抑制事实性错误，从而削弱模型的总体效用。

这种权衡源于现有缓解策略的设计局限。TruthX [11] 和 ITI [12], 等主流方法通常基于样本，通过孤立的输入样本 [13][14]识别特定神经元。为纠正错误，这些方法采用硬性缓解技术，例如将激活值设为零，或将其限制在固定均值 [15], [16]。这种策略假设神经元彼此独立，可以在不造成影响的情况下将其关闭。然而，这种做法忽视了一个简单的事实：大语言模型中的神经元往往不只承担一项功能。同一个神经元可能既会促成错误，也可能是正确推理所必需的 [17]。因此，当这些神经元被关闭以纠正错误时，它们有益的推理能力也会受损。这正是忠实性大幅下降的原因。

然而，从抑制整个神经元转向更精准的方法，面临着三大挑战。首先，很难找到错误的确切来源。由于神经元通常承担多种功能，仅查看其激活值不足以判断它们是在导致错误，还是在进行正确推理。我们需要更好的方法来了解神经元在不同情境下行为的变化 [18]。其次，即使找到了来源，追踪错误如何扩散也很困难。连接

网络中传播错误的连接与正确推理所需的连接交织在一起，彼此之间没有明确的界限。我们需要一种可靠的方法来区分这两类连接 [19], [20]。第三，在修正错误的同时避免引发其他问题并不容易。过强的修正可能损害模型的其他能力，而过弱的修正则可能无法阻止错误。主要难点在于设计一种能够针对每个神经元调整修正强度的方法。这种方法必须有效阻止错误，同时不损害模型的总体忠实性 [21]。

为应对这些挑战，我们提出了 LANCET（基于熵拓扑的神经连接局部对齐），这是一种专为大语言模型精准神经干预而设计的新型框架。Lancet 分三个阶段系统地定位并抑制幻觉。首先，通过梯度驱动的对比分析，定位来源神经元，识别事实情境与幻觉情境之间激活差异显著的单元。其次，为描绘错误信息的动态传播过程，我们引入幻觉差异比（HDR）作为连接性度量。通过最小化 HDR加权图的结构熵（SE），Lancet 对网络进行划分，在结构上将导致错误的连接与推理能力所必需的连接分离开来。最后，我们实施分层干预策略，根据与来源的拓扑邻近性调整抑制强度。这确保能够严格切断幻觉行为，同时在功能上保留关键的功能路径。在两个基准测试上的综合评估表明，Lancet 的表现显著优于当前最先进的方法。

我们的贡献如下：
1）我们指出了粗粒度系统性抑制的局限，并提出一种拓扑视角，将幻觉视为网络中的结构性问题，而非个别神经元的缺陷。2）我们提出 Lancet，这是一个统一框架，将梯度驱动的对比分析用于精准定位，并结合结构熵从结构上识别传播错误信息的通路，同时不影响网络的其余部分。3）大量实验表明，Lancet 能有效解决事实性—忠实性权衡，在抑制幻觉方面显著优于最先进的基线方法，同时保留模型的功能完整性。

II. 理论洞见：多义性神经元的必然性

我们提供一个理论论证，以说明为什么在复杂模型中，节点级干预存在固有局限。令 $h \in \mathbb{R}^d$ 表示给定层的激活向量，令 $g_{\text{fact}} = \nabla_h \mathcal{L}_{\text{fact}}(h)$ 和 $g_{\text{faith}} = \nabla_h \mathcal{L}_{\text{faith}}(h)$ 分别表示事实性和忠实性的敏感度梯度。在实际的大语言模型中，这些目标本质上相互耦合；改善知识检索往往会影响呈现这些知识所需的推理逻辑。我们通过以下假设来刻画这种相互依存关系：在数据分布 $\mathcal{D}$ 上，梯度通常不正交：

$$\mathbb{E}_{h\sim\mathcal{D}}\left[\langle g_{\text{fact}}(h), g_{\text{faith}}(h)\rangle\right] \neq 0. \tag{1}$$

为了理解这种相关性的结构性影响，不妨假设神经元能力严格独立。在这一理想情形下，所有神经元索引 $\mathcal{I} = \{1, \ldots, d\}$ 均可划分为两个不相交子集，$\mathcal{I}_{\text{fact}}$ 和 $\mathcal{I}_{\text{faith}}$，使得每个神经元至多只作为一项任务的专用专家。从数学上说，这要求满足严格的排斥条件：对于任意神经元 i，若其在事实性任务中处于激活状态，则在忠实性任务中必须处于静默状态。对于所有神经元，这可以表示为分量乘积为零：

$$\forall i \in \mathcal{I}, \quad g_{\text{fact},i}(h) \cdot g_{\text{faith},i}(h) = 0. \tag{2}$$

在这一假设下，我们可以通过对分区索引求和进行分解，分析梯度的内积。总内积将拆分为两个正交项：

$$\begin{aligned}\langle g_{\text{fact}}, g_{\text{faith}}\rangle &= \sum_{i\in\mathcal{I}_{\text{fact}}} g_{\text{fact},i}\cdot g_{\text{faith},i} + \sum_{j\in\mathcal{I}_{\text{faith}}} g_{\text{fact},j}\cdot g_{\text{faith},j} \\ &= \sum_{i\in\mathcal{I}_{\text{fact}}} (g_{\text{fact},i}\cdot 0) + \sum_{j\in\mathcal{I}_{\text{faith}}} (0\cdot g_{\text{faith},j}) \\ &= 0.\end{aligned} \tag{3}$$

取期望可得 $\mathbb{E}[\langle g_{\text{fact}}, g_{\text{faith}}\rangle] = 0$，这直接与式（1）中确立的一般非正交条件相矛盾。这一矛盾证明，活跃神经元的交集必定非空，从数学上必然存在多义性神经元。

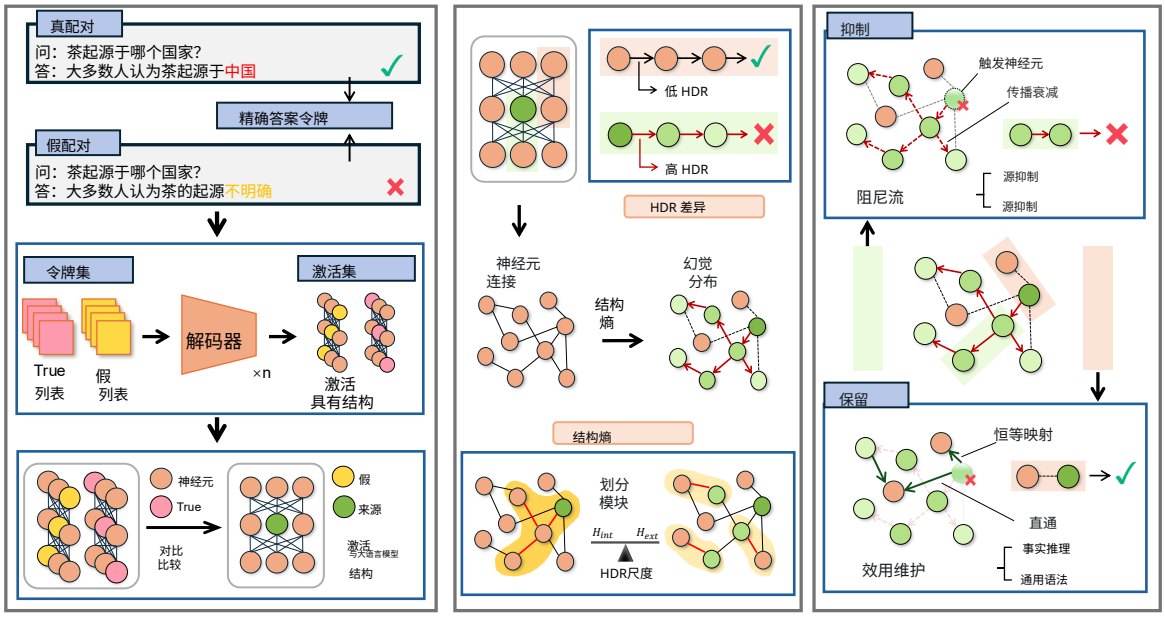
这一结果使我们能够量化粗粒度干预固有的附带损害风险。考虑对多义性神经元 k 进行标准的节点级干预，旨在抑制事实性错误（通过 Δh_k 修改 h_k）。对忠实性损失造成的非预期影响可用一阶泰勒展开近似：

$$\Delta\mathcal{L}_{\text{faith}} \approx g_{\text{faith},k}\cdot\Delta h_k \neq 0. \tag{4}$$

由于这些共享神经元的敏感度 $g_{\text{faith},\,k}$ 非零，任何旨在提升事实性的修改都不可避免地会在 $\mathcal{L}_{\text{faith}}$ 中表现为扰动。这一理论界解释了经验观察到的权衡，并促使我们转向 Lancet 提出的拓扑感知干预，该方法针对通路而非单个节点。

III. 方法论：Lancet框架

我们提出 Lancet，这是一种统一框架，旨在通过从拓扑上将错误信息流与通用推理回路隔离，精准缓解忠实性幻觉。不同于将神经元视为独立单元的传统方法，Lancet 采用系统性视角，将幻觉视为传播现象。如图 2 所示，该框架通过三个严格对齐的阶段运行：（1）通过对比式敏感度对齐，pinpoint 错误的起源



a) 对比敏感度对齐　　b) 结构拓扑对齐　　c) 分层调制对齐

图 2. Lancet框架概述。Lancet通过三个阶段的干预协同运作：(1) **对比敏感度对齐**通过分析梯度差异定位诱发神经元；(2) **结构拓扑对齐**通过最小化结构熵绘制传播路径；(3) **分层调制对齐** 基于拓扑邻近性执行精准的外科式抑制。

；（2）通过将拓扑边界与功能流对齐，映射错误的动态传播过程；（3）依据拓扑邻近性进行分层调制，实施针对性纠正。

A. 对比敏感度对齐

为精准识别幻觉的起源，我们采用对比敏感度对齐策略。该方法基于这样一个假设：当模型从忠实的事实生成转向幻觉编造时，特定的“触发”神经元会表现出不同的敏感度模式。传统方法通常依赖静态激活幅度来筛选神经元；然而，这类指标无法捕捉神经元与模型输出错误之间的功能因果联系。为此，我们将选择标准与不同情境间观察到的动态梯度差异相对齐。

我们通过计算每个神经元基于梯度的重要性分数，量化其功能影响。该指标可作为神经元对最终决策逻辑贡献的近似衡量。对于给定的数据集 D，模型输出对特定神经元 u的敏感度，正式定义为参数与损失梯度乘积的期望幅值：

$$I_u^D = \mathbb{E}_{s\sim D}\left|\theta_u \cdot \nabla_{\theta_u}\mathcal{L}(s)\right| \tag{5}$$

其中， θ_u 表示神经元 u的权重参数，而$\mathcal{L}(s)$表示在样本 s上计算的标准损失函数。

为捕捉与幻觉相关的行为转变，我们系统地评估该指标在两个对比分布中的表现：幻觉响应集（I_u^{hall}）和真实事实集（I_u^{fact}）。随后，通过敏感度差异来概括该神经元功能角色的分歧：

$$\Delta I_u = I_u^{\text{hall}} - I_u^{\text{fact}} \tag{6}$$

ΔI_u 的显著正值表明，当模型生成错误信息时，神经元 u 会变得异常活跃且影响力显著。我们根据这一对齐分数对该层中的所有候选神经元进行排序，并将排名靠前的一部分指定为“始作俑者集合” $\mathcal{N}_{src}$。至关重要的是，为了保留模型的通用效用并避免损害通用认知回路，我们明确排除那些在 PDTB 等通用推理基准上表现出较高基线敏感度的神经元。尽管这种敏感度仍采用相同的梯度驱动指标计算，但评估所依据的是通用任务分布，而非错误信息对比，从而确保目标集合具有特异性。

B. 结构拓扑对齐

识别根源始作俑者是必要但不充分的一步，因为错误信息会沿着复杂且相互交织的下游路径传播。为应对这一挑战，我们构建网络的结构化视图，使拓扑边界与功能传播流相对齐。具体而言，我们构建加权有向图 $G = (\mathcal{N}_{src}, E, \mathcal{A})$，其中 E 表示功能路径，权重 $\mathcal{A}$反映错误信息流的强度和动力学复杂性，并使用结构熵框架对其进行划分。

a) 基于*HDR*的图构建： 为确保拓扑对齐，Lancet 中的边权重必须反映倾向 于传播幻觉，而非一般性的

连接性。我们引入幻觉差异比（HDR）作为神经元 a_{uv} 与神经元 u 和 v之间的亲和权重。该指标量化了幻觉状态下相关性的放大程度：

$$a_{uv} = \text{HDR}_{uv} = \frac{\left|\rho_{uv}^{hall} - \rho_{uv}^{fact}\right|}{\max(|\rho_{uv}^{fact}|, \epsilon)} \tag{7}$$

其中， ρ_{uv}^{D} 表示数据集 D上激活值的皮尔逊相关系数，ϵ 是一个用于确保数值稳定性的小常数。高 HDR 值表明，在出现幻觉时， u 与 v 之间的连接会异常活跃或同步，因此实际上构成了错误信息流的主要通路。

b) 最小化结构熵：构建图后，我们的目标是划分出能够隔离这些高 HDR 通路的功能模块。我们采用最小化结构熵 [20] 作为优化目标。设$G = (\mathcal{N}_{src},\ E,\ \mathcal{A})$为 HDR 加权激活图，$\mathcal{P} = \{X_1, \ldots, X_m\}$ 为将 V划分为互不相交模块的一种划分。我们将 $\text{vol}(G) = \sum_{v_i \in \mathcal{N}_{src}} d_i$ 定义为图的总体积，其中 d_i 是节点 v_i的加权度。同理，$\text{vol}(X)$ 表示模块 X的体积，$\text{cut}(X)$表示连接 X 与图中其余部分的边的总权重。

结构熵 $\mathcal{E}^{\mathcal{P}}(G)$ 定义为所有模块的内部熵 (H_{int}) 与外部熵 (H_{ext}) 之和：

$$\mathcal{E}^{\mathcal{P}}(G) = \sum_{X \in \mathcal{P}} \big(H_{\text{int}}(X) + H_{\text{ext}}(X)\big). \tag{8}$$

这些组成部分的数学形式如下：

$$H_{\text{int}}(X) = -\sum_{v_i \in X} \frac{d_i}{\text{vol}(G)} \log_2 \frac{d_i}{\text{vol}(X)}, \tag{9}$$

$$H_{\text{ext}}(X) = -\frac{\text{cut}(X)}{\text{vol}(G)} \log_2 \frac{\text{vol}(X)}{\text{vol}(G)}. \tag{10}$$

从物理意义上说， H_{int} 反映模块内随机游走的稳定性，而 H_{ext} 则衡量模块间转移所带来的不确定性。最小化 $\mathcal{E}^{\mathcal{P}}(G)$ 可得到一种划分，其中模块内的节点由高 HDR 边紧密连接，而模块间的连接则较为稀疏。

我们通过一个两阶段迭代算法来寻找近似最优划分 $\mathcal{P}^*$：**(1) 合并**，递归合并能够使结构熵降幅最大的神经元模块；以及 **(2)细化**，通过重新分配单个神经元来微调边界，以最大化熵的净增益。由此得到的划分 $\mathcal{P}^*$ 在拓扑上将“受感染”的传播区域与健康的推理网络区分开来。

C. 分层调制对齐

基于推导出的拓扑结构 $\mathcal{P}^*$，我们实施一种分层调制对齐策略，对网络进行精准干预。与可能损害模型能力的统一抑制不同，该机制会校准每个神经元的抑制强度，使干预力度与该神经元在错误信息传播链中的拓扑角色相匹配。

我们通过为每个神经元设定一个拓扑抑制因子 α_u 来制定干预方案。该因子综合考虑输入流的局部严重程度，以及相对于来源的空间衰减，从而确保干预梯度平滑：

$$\alpha_u = \alpha_0 \max_{v \to u} \text{HDR}_{vu}\, e^{-\lambda d_u} \tag{11}$$

其中， α_0 表示基线强度， d_u 表示图中从神经元 u 到始作俑者集合 $\mathcal{N}_{\text{src}}$的测地距离。

在功能上，最大化项 $\max_{v \to u}\text{HDR}_{vu}$ 充当错误信息流入的最坏情况估计器，用于识别幻觉传播至神经元 u的主要通道。同时，指数项$e^{-\lambda d_u}$ 建立了一个由衰减率 λ控制的拓扑隔离区。这种设计确保抑制力始终局部作用于诱发神经元周围，并在到达网络的通用区域时迅速衰减。

最后，干预通过拓扑感知参数缩放来执行：

$$\tilde{\theta}_u = (1 - \alpha_u)\theta_u \tag{12}$$

为确保模型保留其通用能力，我们制定了具体的边界规则。对于根始发神经元 $\mathcal{N}_{\text{src}}$，抑制因子 α_u 被设为 1。这会严格切断幻觉来源。相反，对于属于关键推理路径 $\mathcal{C}_{\text{crit}}$的神经元，抑制因子被固定为 0。我们明确将这些关键路径定义为事实数据集上重要性分数最高的神经元集合，这对应于第一阶段得到的排名最高单元。通过保护这些高价值单元，我们的方法在网络中实现了平滑调整：从来源处的硬抑制逐渐过渡到周围区域的软衰减，同时使关键功能完全不受影响。

IV. 实验与结果

A. 数据集

我们在两个基准测试上验证 Lancet。PDTB [22]通过 DISQ指标 [23],评估话语忠实性，该指标将目标准确率、反事实鲁棒性和逻辑一致性汇总为统一的总体得分。TruthfulQA [24] 通过多项选择准确率（MC1/MC2）和生成质量（真实性与信息丰富度）评估事实性，并将这些指标综合为 True*Info指标。

B. 基线

我们将 Lancet 与涵盖三种范式的代表性方法进行基准比较：在目标数据集上进行监督微调，以确定性能上界；包括 CAD [25] 和 DoLa [26] 在内、通过操纵输出逻辑值的解码策略；以及 ITI [12], 和 TruthX [11]等最先进的推理时干预技术。

表 I 跨模型比较及 Lancet 在 PDTB 和 TruthfulQA 基准上的表现。

方法	模型	PDTB				TruthfulQA				
		总体	针对性	反事实	一致性	真实信息	MC1	MC2	True	Info
基线		17.4	79.3	27.1	81.2	57.6	33.9	51.3	66.9	86.1
FT-PDTB	LLaMA2-7B-Chat	23.7	81.5	34.9	83.2	46.1	30.7	45.1	62.8	73.3
FT-TruthfulQA	LLaMA2-7B-Chat	12.7	69.2	22.7	80.6	62.5	49.7	63.2	69.0	90.6
CAD	LLaMA2-7B-Chat	20.0	70.5	35.4	80.1	54.2	22.0	50.8	69.0	78.5
DoLa	LLaMA2-7B-Chat	15.9	78.0	25.1	81.2	59.2	33.3	60.9	67.6	87.5
ITI	LLaMA2-7B-Chat	14.8	64.5	28.7	79.8	59.9	33.9	52.0	68.9	87.0
TruthX	LLaMA2-7B-Chat	5.8	75.6	9.6	80.2	63.6	50.2	70.5	70.8	89.7
Lancet（我们的方法）	LLaMA2-7B-Chat	**27.6**	**81.9**	**37.4**	**90.4**	**67.8**	**53.7**	**72.8**	**71.9**	**94.3**
方法	**模型**	**PDTB**				**TruthfulQA**				
		总体	**针对性**	**反事实**	**一致性**	**真实信息**	**MC1**	**MC2**	**True**	**Info**
基线		15.1	21.2	84.0	85.1	59.5	37.9	55.7	70.1	84.8
FT-PDTB	DeepSeek-llm-7B	20.1	27.8	86.2	83.7	46.8	33.1	46.2	65.1	71.8
FT-TruthfulQA	DeepSeek-llm-7B	3.5	5.0	82.4	83.9	64.3	49.4	67.9	73.3	87.8
CAD	DeepSeek-llm-7B	15.6	23.4	80.9	82.4	49.3	33.3	54.9	70.0	70.5
DoLa	DeepSeek-llm-7B	10.1	14.6	83.1	83.4	60.1	29.9	7.5	70.3	85.5
ITI	DeepSeek-llm-7B	15.7	22.2	83.5	84.9	62.9	38.6	57.9	71.3	88.2
Lancet（我们的方法）	DeepSeek-llm-7B	**23.7**	**30.1**	**89.1**	**88.2**	**72.8**	**49.8**	**68.6**	**79.5**	**91.6**

C. 实现细节

实验使用半精度加载的 LLaMA2-7B-Chat [27] 和 DeepSeek-llm-7B [28] 进行。对于 Lancet，干预因素选择比例在留出验证集上进行优化；结构熵最小化则采用标准划分算法。所有基线结果均使用其官方实现，并在相同实验设置下复现，以确保比较公平。

D. 主要结果

表 I 确认，TruthX 等基线方法存在严重的事实性—忠实性权衡，以牺牲推理能力来换取事实性。相比之下，Lancet 在这两个维度上均实现了提升。这一优势源于结构熵，它比子空间投影更精确地隔离错误信息流。此外，Lancet 在 DeepSeek 上展现出稳健的泛化能力，在其他策略失效时仍能保持稳定。

E. 消融研究

表 II 验证了 Lancet 各组件在 LLaMA2 和 DeepSeek 上的作用。移除幻觉差异比会导致忠实性大幅下降，这证实仅凭激活幅度不足以进行精准识别。同样，排除结构熵会显著降低推理能力，尤其是在 DeepSeek 上，这验证了拓扑映射对于阻断错误信息传播不可或缺。此外，以统一抑制取代分层调节会导致结果不够理想，这表明干预强度与

表 II LANCET 组件在 LLaMA2 和 DeepSeek 上的消融研究。

变体	LLaMA2-7B		DeepSeek-7B	
	PDTB	TQA	PDTB	TQA
基线	17.4	57.6	15.1	59.5
无 HDR	22.8	63.1	20.0	64.0
不含结构熵	21.7	64.8	18.5	69.0
不含层级结构	22.4	64.4	21.5	68.5
完整 Lancet	**27.6**	**67.8**	**23.7**	**72.8**

拓扑严重程度相匹配，对于平衡缓解效果与通用能力至关重要。

F. 参数敏感性分析

我们研究了 Lancet 对诱因选择比例 r 和空间衰减因子 λ的鲁棒性，详见图3。选择比例的变化轨迹揭示了一种微妙的平衡：扩大精准干预范围起初会提升真实性，但选择范围过大则会波及通用推理回路，导致忠实度指标下降。这证实，最优比例对应于受感染神经元集合的精确边界。空间衰减因子也呈现出类似规律，其性能峰值验证了有效干预所需特定拓扑半径的存在。无论是抑制范围过度扩散，还是过于局限，都会削弱整体协同效应，这表明干预强度必须在几何上与错误信息的自然传播相匹配。

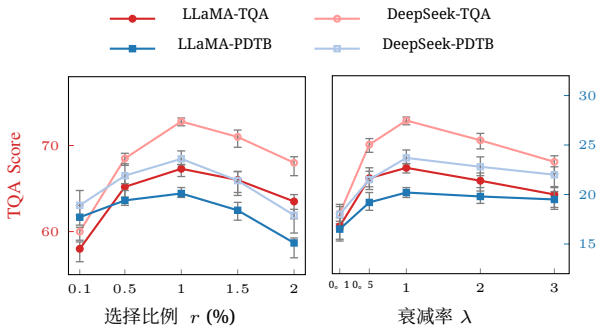


图3. LLaMA2 和 DeepSeek 在 TruthfulQA（红色调）和 PDTB（蓝色调）上的跨模型参数敏感性分析。误差条表示非对称标准差。

五、结论

在这项工作中，我们提出了 Lancet，这是一种拓扑干预框架，通过将忠实性幻觉建模为定向的“传染性”流来加以解决。Lancet 利用幻觉差异比和结构熵，将范式从粗略抑制转向精准、类似手术的调控，从而有效清除错误信息路径。我们的实证结果证实，该方法解决了幻觉缓解与通用推理保持之间的关键权衡，并实现了最先进的性能。尽管目前依赖配对数据以及图划分带来的开销仍对可扩展性构成挑战，但未来整合无监督诱因检测和轻量级近似方法，有望进一步简化流程。最终，Lancet 为神经干预奠定了严谨的几何基础，既提供了解析信息流的理论视角，也为构建可靠且功能稳健的大语言模型开辟了实用路径。

参考文献

[1] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li 和 Peter J Liu，《探索统一文本到文本转换器迁移学习的极限》，机器学习研究期刊，第 21 卷，第 140 期，页 1–67，2020年。[2] Tom Brown、Benjamin Mann 等，《语言模型是少样本学习器》，神经信息处理系统进展，第 33 卷，页 1877–1901，2020年。[3] Chaozhuo Li、Pengbo Wang、Chenxu Wang 等，《洛基的幻觉之舞：大语言模型幻觉综述》，*arXiv* 预印本 *arXiv:2507.02870*，2025年。[4] Hieu Nguyen、Zihao He 等，《平滑幻觉：通过平滑知识蒸馏缓解大语言模型幻觉》，*arXiv* 预印本 *arXiv:2502.11306*，2025年。[5] Xiaoling Zhou、Mingjie Zhang、Zhemg Lee、Wei Ye 和 Shikun Zhang，《Hademif：大语言模型的幻觉检测与缓解》，发表于表征学习国际会议，2025年，第 2025 卷，页 81014–81038。[6] Michael Moor、Oishi Banerjee、Zahra Shakeri Hossein Abad、Harlan M Krumholz、Jure Leskovec、Eric J Topol 和 Pranav Rajpurkar，《通用型医疗人工智能的基础模型》，《自然》，第 616 卷，第 7956 期，页 259–265，2023年。[7] Michele Salvagno、Fabio Silvio Taccone 和 Alberto Giovanni Gerli，《人工智能幻觉》，重症监护，第 27 卷，第 1 期，页 180，2023年。

[8] Lei Huang、Weijiang Yu 等，《大语言模型幻觉综述：原理、分类、挑战与开放问题》，*ACM*信息系统汇刊，第 43 卷，55，2025年 2，1第。–页。[9] Yue Zhang、Yafu Li 等，《人工智能之海中的塞壬之歌：大语言模型幻觉综述》，*arXiv* 预印本 *arXiv:2309 01219*，2023年。。[10] Tianyu Zhang、Junfeng Fang、Houcheng Jiang、Baolong Bi、Xiang Wang 和 Xiangnan He，《面向大语言模型的可解释且高效的编辑》，发表于*THE WEB CONFERENCE 2025*，2025年。[11] Shaolei Zhang、Tian Yu 和 Yang Feng，《TruthX：通过在真实空间中编辑大语言模型来缓解幻觉》，发表于计算语言学协会第 *62* 届年会论文集，2024年 8 月，页 8908–8949。[12] Kenneth Li、Oam Patel、Fernanda Vi´egas、Hanspeter Pfister 和 Martin Wattenberg，《推理时干预：引导语言模型给出真实答案》，发表于神经信息处理系统第 *37* 届国际会议论文集，美国纽约州雷德胡克，2023年，NIPS ’23，Curran Associates Inc。[13] Chen Qian、Dongrui Liu、Jie Zhang、Yong Liu 和 Jing Shao，《Dean：通过停用耦合神经元缓解大语言模型中的公平性与隐私冲突》，2024年。[14] Hongxiang Zhang、Hao Chen、Muhao Chen 和 Tianyi Zhang，《主动层对比解码减少大语言模型生成中的幻觉》，2025年。[15] Shahrad Mohammadzadeh、Juan David Guerra、Marco Bonizzato、Rei- haneh Rabbany 和 Golnoosh Farnadi，《幻觉排毒：用于大语言模型训练的敏感性丢弃（send）》，2025年。[16] Houcheng Jiang、Junfeng Fang、Tianyu Zhang、Baolong Bi、An Zhang、Ruipeng Wang、Tao Liang 和 Xiang Wang，《大语言模型的神经元级序列编辑》，发表于计算语言学协会第 *63* 届年会论文集，2025年 7 月，页 16678–16702。[17] Derui Zhu、Dingfan Chen、Qing Li、Zongxiong Chen、Lei Ma、Jens Grossklags 和 Mario Fritz，《Pollmgraph：通过状态转移动力学揭示大语言模型中的幻觉》，*arXiv* 预印本 *arXiv:2404.04722*，2024年。[18] Sebastian Farquhar、Jannik Kossen、Lorenz Kuhn 和 Yarin Gal，《使用语义熵检测大语言模型中的幻觉》，《自然》，第 630 卷，第 8017 期，页 625–630，2024 年。[19] Shiqi Chen 、Miao Xiong、Junteng Liu、Zhengxuan Wu、Teng Xiao、Siyang Gao 和 Junxian He，《上下文中的尖锐度作为警示：幻觉缓解的内部表征视角》，*arXiv* 预印本 *arXiv:2403.01548*，2024年。[20] Angsheng Li 和 Yicheng Pan，《网络的结构信息与动力学复杂性》，*IEEE*信息论汇刊，第 62 卷，第 6 期，页 3290–3339，2016年。[21] Ziwei Ji、Nayeon Lee、Rita Frieske、Tiezheng Yu、Dan Su、Yan Xu、Etsuko Ishii、Ye Jin Bang、Andrea Madotto 和 Pascale Fung，《自然语言生成中的幻觉综述》，*ACM*计算调查，第 55 卷，第 12 期，页 1–38，2023年。[22] Bonnie Webber、Rashmi Prasad、Alan Lee 和 Aravind Joshi，《宾夕法尼亚话语树库 3.0 标注手册》，费城，宾夕法尼亚大学，第 35 卷，页 108，2019年。[23] Yisong Miao、Hongfu Liu、Wenqiang Lei、Nancy Chen 和 Min-Yen Kan，《论辩式苏格拉底提问：评估语言模型对话语关系理解的忠实性》，发表于计算语言学协会第 *62* 届年会论文集。2024年 8 月，页 6277–6295，计算语言学协会。[24] Stephanie Lin、Jacob Hilton 和 Owain Evans，《TruthfulQA：衡量模型如何模仿人类谬误》，发表于计算语言学协会第 *60* 届年会论文集。2022年 5 月，页 3214–3252，计算语言学协会。[25] Weijia Shi、Xiaochuang Han、Mike Lewis、Yulia Tsvetkov、Luke Zettle- moyer 和 Wen-tau Yih，《信任你的证据：借助上下文感知解码减少幻觉》，发表于*2024*年北美会议论文集。2024年 6 月，页 783–791，计算语言学协会。[26] Yung-Sung Chuang、Yujia Xie、Hongyin Luo、Yoon Kim、James Glass 和 Pengcheng He，《DoLa：通过对比不同层进行解码，提升大语言模型的事实性》，*arXiv* 预印本 *arXiv:2309 03883*，2023年。。[27] Hugo Touvron、Louis Martin、Kevin Stone、Peter Albert、Amjad Alma- hairi、Yasmine Babaei、Nikolay Bashlykov、Soumya Batra、Prajjwal

Bhargava、Shruti Bhosale 等人，《Llama 2：开放基础模型与微调聊天模型》，*arXiv* 预印本 *arXiv:2307.09288*，2023年。[28] Xiao Bi、Deli Chen、Guanting Chen、Shanhuang Chen、Damai Dai、Chengqi Deng、Honghui Ding、Kai Dong、Qiushi Du、Zhe Fu 等人，《DeepSeek 大语言模型：以长期主义扩展开源语言模型》，*arXiv* 预印本 *arXiv:2401.02954*，2024年。