

基于自适应上下文感知神经元解码的 大语言模型幻觉缓解

郭亚龙 张仰森 亢 静 王昱麒 陈全涛

(北京信息科技大学智能信息处理研究所 北京 100192)

摘 要 大语言模型(Large Language Models, LLMs)通过大规模语料库训练构建丰富的参数化知识,在自然语言处理任务中展现出强大的知识表达与推理能力。然而,在LLMs文本生成过程中,过时或错误的参数化知识常与输入的上下文新知识产生冲突,导致模型忽视上下文而过度依赖参数化知识,从而引发“幻觉”现象。为缓解此问题,本文提出一种自适应上下文感知神经元解码(Adaptive Context-aware Neuron Decoding, ACND)框架,通过神经元层面的动态调节机制增强模型对上下文的忠实度。具体而言,ACND首先结合积分梯度归因与贝叶斯优化,实现对上下文感知神经元的动态识别和自适应筛选。随后在前馈神经网络(Feed-Forward Network, FFN)层引入上下文感知神经元激活增益机制,结合 Jensen-Shannon 散度动态重构生成概率分布,实现上下文与参数化知识之间的自适应平衡,提升了模型在知识冲突情境下的语义一致性和鲁棒性。本文在多个主流模型与任务中开展了实证研究,相较于基线方法,ACND使LLMs准确率平均提升8.1%,顽固率平均降低6.8%,从模型回答的正确性和上下文依赖性两个维度缓解了由知识冲突引起的幻觉,提高了模型的上下文忠实度。此外,数据分布偏移实验与泛化实验也进一步验证了ACND的鲁棒性。ACND无需更新模型参数,可作为即插即用的模块集成到现有模型的解码阶段,为知识冲突场景下的大模型幻觉缓解提供了一种细粒度、可解释的优化方法。

关键词 大语言模型;幻觉;知识冲突;上下文忠实度;感知神经元;对比解码

中图法分类号 TP391 **DOI号** 10.11897/SP.J.1016.2026.01913

Hallucination Mitigation in Large Language Models via Adaptive Context-aware Neuron Decoding

GUO Ya-Long ZHANG Yang-Sen KANG Jing WANG Yu-Qi CHEN Quan-Tao

(Institute of Intelligent Information Processing, Beijing Information Science and Technology University, Beijing 100192)

Abstract Large language models (LLMs) build rich parametric knowledge through large-scale corpus training, and demonstrate powerful knowledge representation and reasoning capabilities in natural language processing tasks. However, during text generation by LLMs, outdated and erroneous parametric knowledge often conflicts with new knowledge in the input context, causing the model to ignore context and over-rely on parametric knowledge, resulting in “hallucination”. This problem becomes particularly critical in dynamic or domain-specific scenarios, where external information may be more recent, more accurate, or more task-relevant than the model’s parametric knowledge. Although existing hallucination mitigation methods can improve the factuality of generated text, most rely on external knowledge injection, parameter updates, output distribution calibration, or static, coarse-grained internal interventions. As a result, they usually

收稿日期:2025-09-30;在线发布日期:2026-05-13。本课题得到北京市自然科学基金-小米创新联合基金资助项目(No. L233008)资助。
郭亚龙,硕士研究生,主要研究领域为人工智能、自然语言处理、机器学习。E-mail: yalonguo@163.com。张仰森(通信作者),博士,教授,中国计算机学会(CCF)杰出会员,主要研究领域为人工智能、自然语言处理、网络内容安全。E-mail: zhangyangsen@163.com。
亢 静,硕士研究生,中国计算机学会(CCF)会员,主要研究领域为人工智能、自然语言处理。王昱麒,博士研究生,中国计算机学会(CCF)会员,主要研究领域为跨模态学习、大模型生成技术。陈全涛,硕士研究生,主要研究方向为自然语言处理、网络信息安全。

lack fine-grained and adaptive mechanisms for regulating the model's internal decision-making pathways according to dynamically provided contextual evidence. This limitation makes it difficult for LLMs to selectively amplify reliable contextual signals while suppressing conflicting parametric priors during inference. To mitigate this problem, we propose a novel framework, Adaptive Context-aware Neuron Decoding (ACND) to enhance the model's fidelity to the context through a dynamic regulation mechanism at the neuron level. Specifically, ACND consists of two core modules. The first module, Adaptive Context-Aware Neuron Selection (ACNS), uses Integrated Gradients (IG) to estimate the attributional contribution of neurons in feed-forward network (FFN) layers to context-guided prediction. It further combines hierarchical selection with Bayesian optimization to adaptively identify context-aware neurons and determine their activation enhancement strength. The second module, Neuron Adaptive Contrastive Decoding (NACD), applies activation gains to the selected neurons during inference and uses Jensen-Shannon Divergence (JSD) to measure the conflict intensity between contextual and parametric knowledge distributions. Guided by this conflict-aware signal, NACD dynamically reconstructs the generation probability distribution, thereby achieving an adaptive balance between contextual and parametric knowledge and improving the semantic consistency and robustness of LLMs in knowledge-conflict scenarios. Unlike methods that only modify the final output distribution, ACND jointly acts on neuron-level representation and decoding-time probability adjustment, enabling more interpretable and fine-grained intervention without changing model parameters. We conducted empirical studies on multiple mainstream models and tasks. Compared with baseline methods, ACND achieves an average accuracy improvement of 8.1% and a reduction in stubbornness rate of 6.8% for LLMs. These results indicate that ACND mitigates knowledge-conflict-induced hallucinations by improving response correctness and reducing over-reliance on parametric knowledge, thereby enhancing the contextual fidelity of LLMs. Furthermore, cross-dataset distribution-shift experiments show that, after performing ACNS offline attribution and Bayesian optimization once on each source dataset, ACND can be directly transferred to target datasets without re-computing neuron attributions or re-searching hyperparameters. In most transfer settings, it still achieves positive performance gains over the greedy decoding baseline, demonstrating strong cross-distribution robustness and the transferability of its offline-optimized neuron configurations. Further generalization analysis indicates that ACND exhibits favorable scalability across model sizes and compatibility with different open-source model families. Without requiring model parameter updates, ACND can be integrated into the decoding stage of existing models as a plug-and-play module, providing a fine-grained and interpretable approach to hallucination mitigation for large language models in knowledge-conflict scenarios.

Keywords large language models; hallucination; knowledge conflicts; contextual fidelity; sensitive neurons; contrastive decoding

1 引 言

在自然语言处理任务中,开放域问答、事实核查等知识密集型任务要求LLMs具备整合多源知识并动态响应上下文的能力,从而生成真实且语义一致的回答。LLMs通过预训练编码了广泛的世界知

识^[1-4],但其静态训练数据难以实时吸纳更新的领域知识,为弥补LLMs静态知识的局限,已有研究提出如知识图谱(Knowledge Graph, KG)、检索增强生成(Retrieval Augmented Generation, RAG)、在线搜索(Online Search)等融合外部知识的增强生成框架,引入动态上下文构建“双源知识架构”: (1)参数化知识,嵌入模型参数中; (2)上下文知识,作为输

入前缀动态提供^[5]。

虽然上述方法通过构建多维度的信息输入体系,在一定程度上弥补了大语言模型静态知识存储的固有缺陷。但当上下文知识与模型固有参数知识冲突时,由于预训练阶段形成的知识表征惯性,模型对参数化知识有过度依赖的倾向,导致生成内容出现事实性错误^[6-7],即所谓的幻觉现象^[8-10]。例如:虽然我们在上下文中呈现了“在2024年,巴黎成为继伦敦之后,第二个三次举办夏季奥运会的城市,下一个也将举办三届奥运会的城市将是洛杉矶”这一最新上下文,但当它面对“哪座城市成为第二个三次举办夏季奥运会的城市?”的问题时,它仍然提供过时的答案“伦敦”。

针对上述挑战,我们提出了一种自适应上下文感知神经元解码框架(Adaptive Context-aware Neuron Decoding, ACND)。ACND是首个覆盖“神经元激活→知识表征→生成决策”全流程的轻量级可插拔框架,不依赖特定任务或模型架构,突破传统方法的粗粒度调控范式,旨在提升LLMs对上下文的忠实度,减少大模型在处理知识密集型任务时的幻觉现象。

为验证框架的通用性,我们在LLaMA^[11]、Baichuan^[12]、Qwen^[13]多个模型家族的预训练与指令调优版本进行了实验验证,所选模型覆盖了不同参数规模。采用ACC衡量正确性,同时引入顽固率(SR)衡量模型在上下文与先验冲突时对先验的固守程度,从模型回答的正确性和上下文依赖性两个维度刻画LLMs幻觉。实验结果表明,相较于基线模型,ACND在多个知识密集型任务上将平均ACC提升8.1%,显著优于简单模块堆叠方法(提升2%),有效打破模型对参数化知识的路径依赖,增强基于上下文隐含知识的模型推理生成能力。

我们工作的主要贡献总结如下:

(1)提出了上下文感知神经元的动态筛选机制。通过融合积分梯度归因与贝叶斯优化的自适应上下文感知神经元筛选算法(Adaptive Context-aware Neuron Selection algorithm, ACNS),实现对上下文感知神经元的动态筛选,筛选成本减少约60%。

(2)提出了神经元自适应对比解码(Neuron Adaptive Contrastive Decoding, NACD)策略。创新性地将对对比式生成策略与神经元级动态调控机制结合,实现上下文与参数化知识之间的自适应平衡,有效提升LLMs对上下文知识的忠实度。

(3)构建了首个覆盖从“神经元激活→知识表征→

生成决策”的端到端动态上下文感知框架,通过跨层级联动实现知识动态融合与冲突缓解,突破现有方法的粗粒度静态局限。

2 相关工作

在自然语言处理任务中,LLMs需面对海量且快速演化的外部知识,以满足对准确性、逻辑性及复杂推理的高要求。在此场景下,LLMs面临两大核心挑战:一是长上下文建模能力不足,导致模型在整合大量新输入信息时难以保持语义一致性,导致生成内容出现主线模糊、语义跳跃甚至事实矛盾的问题;二是模型对新知识响应滞后,容易依赖参数化知识,进而引发事实性幻觉。针对上述问题,现有研究主要围绕以下技术路径展开优化:

(1)微调策略:微调(Fine-tuning)旨在通过在额外数据上进行进一步训练来更新现有LLMs的内部知识,包括含有最新信息的数据集或特定领域的数据集,已成为提升LLMs在不同任务中表现的主流手段^[14-18]。

全参数微调通过更新全部参数适配下游任务,但资源消耗高、泛化性差,限制了其大规模应用。为提升效率,研究者提出了多种参数高效微调(PEFT)^[19]方法,如:LoRA(Low-Rank Adaptation)^[20]方法通过引入低秩矩阵仅优化局部权重;Adapter Tuning^[21]方法在模型中嵌入小型可学习模块;Prompt Tuning^[22]方法通过调整输入提示(Prompt)的方式,而非直接修改模型参数。此外,监督式微调(Supervised Fine-Tuning, SFT)通过有标注指令数据直接优化生成概率,适用于大多数任务,但可能无法充分挖掘模型潜力。然而,微调策略主要依赖静态数据更新,难以应对推理过程中动态出现的知识冲突,且无法实时调节模型对上下文的响应。

(2)模型内部结构的解释与控制,研究逐步转向对模型内部结构的解释与神经元级调控。已有研究揭示,前馈神经网络(FFN)中蕴含大量知识神经元^[23],其中部分神经元承担事实性记忆,部分则处理上下文语义。Hao等人引入了一种自我注意归因方法来解释Transformer内部的信息交互^[24],这一成果为深入理解Transformer内部信息交互机制提供了重要理论依据。类似地,DoLa^[25]通过对比不同层的logits输出,增强事实性并抑制先验误导,但仅作用于层间差分,未深入神经元激活。Shi等人提出的IRCAN^[26]方法,则通过积分梯度(Integrated

Gradients, IG)^[27]构建上下文感知归因框架,为关键神经元的识别提供了有效手段,特别是在知识冲突消解方面展现出显著优势,这有力推动了相关研究的进展。

现有神经元筛选机制大多为静态策略,在面对动态语境时缺乏灵活性,成为模型性能的瓶颈。以 IRCAN 为例,其基于静态归因分数识别神经元,通过预设固定阈值筛选上下文感知神经元的机制,难以适应专业术语引入、领域语义突变等新语境。在知识快速更新的任务中,神经元激活模式呈现显著变化,而静态策略往往无法及时捕捉这些变化,从而导致语义整合偏差,削弱模型对上下文的感知能力^[28-29]。

(3)为弥补模型对上下文响应不足的问题,近期研究聚焦于生成阶段的解码策略调整,通过放大不同模型尺度或 LLM 不同层间^[30]的输出概率差异,能够显著增强模型对上下文的敏感度,进而有效遏制幻觉的产生。

例如, CAD 方法^[31]通过直接调整生成分布概率实现上下文适配,促使 LLM 在内容生成时更深度地聚焦语境信息,大幅提升输出内容的准确性与逻辑连贯性。Contrastive Decoding^[32]通过专家与业余模型概率对比,抑制重复或无关生成,提升开放式文本质量。此类方法的局限在于仅作用于输出概率分布,未能深入模型内部知识表征结构。FFN 层作为语义建模的核心,若上下文信息未能在此层充分融合,仅在解码阶段调整概率输出,当输入数据存在逻辑矛盾时,由于模型无法在底层神经表征阶段完成冲突知识的甄别与修正,导致解码过程仍倾向于依赖预训练形成的知识路径,最终影响 LLMs 输出质量。

在后续实验中,我们将 IRCAN 和 CAD 相关研究作为对比基线。与之相比,本文提出的 ACND 通过神经元层级的动态调控机制,增强模型对上下文的忠实度,为 LLMs 应对动态知识冲突提供了极具创新性的解决方案。

3 方 法

在知识密集型任务中,如何在复杂语境下实现对上下文知识的精确感知与动态适配,是提升大语言模型鲁棒性与语义一致性的关键挑战。针对这一问题,本文聚焦于上下文推理和特定知识冲突任务,

其输入形式为 (c, q, o) , 其中 c 为上下文, q 为问题, $o = \{o_1, o_2, \dots, o_k\}$ 为选项集, 模型需输出最可能的答案 ($o^* \in o$), 创新性地提出 ACND 框架, 旨在提升模型对上下文信息的忠实建模能力, 缓解因知识冲突或语境变化引发的生成偏移。该框架包括两大核心模块:

(1) ACNS: 通过动态识别与任务预测高度相关的上下文感知神经元, 实现对关键神经元的自适应筛选与权重调控, 从而保证模型对语境关键信息的精确感知。

(2) NACD: 在 LLMs 的 FFN 层引入神经元激活增益机制, 并结合基于 Jensen-Shannon Divergence (JSD) 的动态对比调控策略, 对生成概率分布进行重构。该机制强化了模型对上下文的响应能力, 降低了生成幻觉的风险, 从而提升预测结果的语义一致性与鲁棒性。

ACND 框架整体设计如图 1 所示。

3.1 自适应上下文感知神经元筛选

为识别在当前语境中对预测具有关键影响的神经元, 基于积分梯度归因与贝叶斯优化, 对 LLMs 中 FFN 层的神经元进行细粒度分析与动态筛选, 并在推理过程中动态调整其激活强度。

具体而言, 首先通过 IG 计算 FFN 神经元在上下文对预测答案的贡献度评分, 衡量其对上下文信息的敏感性与贡献度, 随后引入贝叶斯优化在归因结果上建立后验不确定性模型, 进一步筛选出对预测最具代表性、稳定性强的上下文感知神经元 (Context-Aware Neurons, CANs), 实现感知神经元权重自适应筛选。最终, ACNS 输出一组带权重的神经元激活调控信号, 用于指导下游的激活增强与对比调控模块, 实现对关键上下文信息的保留与放大, 提高模型在复杂情境下的生成鲁棒性。实验表明, ACNS 相比静态筛选机制在保持语义一致性的同时, 有效提高了神经元筛选的效率与准确性。

3.1.1 上下文感知神经元归因

为衡量 FFN 层中神经元对上下文语义的敏感性, 对于待评估的第 l 层 FFN 中第 i 个神经元 n_i^l , 定义其归因分数为 $Attr(n_i^l)$, 用于量化该神经元激活变化对目标答案 w^* 生成概率的影响。分别记录该神经元在两种输入条件下的激活值:

(1) 基线状态。仅以问题 q , 选项集 o 作为输入, 记录该神经元的激活值 $h_{(q,o)}^l$ 。

(2) 上下文状态。以上下文 c 、问题 q 、选项集 o

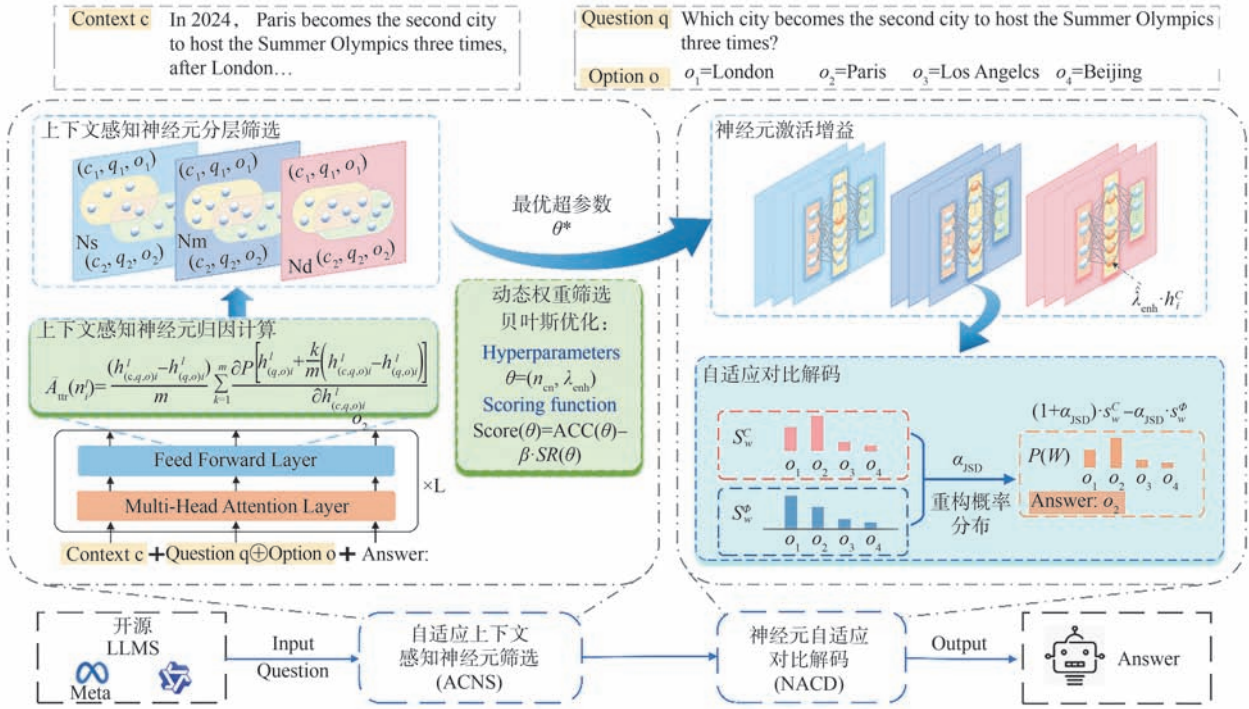


图1 自适应上下文感知神经元解码(ACND)框架

共同作为输入,记录新的激活值 $h_{(c,q,o)_i}^l$ 。

为计算归因分数,我们构造神经元 n_i^l 的激活值从 $h_{(q,o)_i}^l$ 逐步调整至 $h_{(c,q,o)_i}^l$,并观察目标答案 w^* 的生成概率如式(1)所示:

$$P(h_i^l) = p(w^* | c, q, A(n_i^l) = h_i^l) \quad (1)$$

其中, $A(n_i^l)$ 表示神经元 n_i^l 的激活操作。

归因计算的核心是沿神经元激活值从基线状态到上下文状态的路径,观察目标答案概率的变化,通过积分该过程中概率对激活值的梯度,得到神经元的上下文感知归因分数。由于直接计算连续积分在实践中不可行,因此采用黎曼积分近似对积分进行离散化计算,以高效获得神经元的归因分数。具体而言,将神经元 n_i^l 的激活值从基线状态 $h_{(q,o)_i}^l$ 到上下文状态 $h_{(c,q,o)_i}^l$ 之间划分为 m 个等间隔小步。在每个间隔点计算概率对激活的梯度,并对所有梯度进行累加,得到归因分数的近似表达式如式(2)所示:

$$Attr(n_i^l) = \frac{(h_{(c,q,o)_i}^l - h_{(q,o)_i}^l)}{m} \sum_{k=1}^m \frac{\partial P[h_{(q,o)_i}^l + \frac{k}{m}(h_{(c,q,o)_i}^l - h_{(q,o)_i}^l)]}{\partial h_{(c,q,o)_i}^l} \quad (2)$$

其中, $\partial P[h_{(q,o)_i}^l + \frac{k}{m}(h_{(c,q,o)_i}^l - h_{(q,o)_i}^l)]$ 用于计算模型

概率关于 $h_{(c,q,o)_i}^l$ 的积分。 m 为离散化步数,在实验中,将 $m=20$ 既能保证归因精度,又能维持计算效率,实现精度与效率之间的良好平衡。归因分数越高,说明该神经元对上下文的语义建模能力和响应越敏感,为后续的上下文感知神经元筛选提供了可靠依据。

3.1.2 上下文感知神经元分层筛选

已有研究表明,在 Transformer 模型中各层对上下文的响应存在显著差异^[33]。为更有效挖掘结构性语义信息,我们提出上下文感知神经元分层筛选机制,旨在从模型不同语义层级中筛选出对语义建模最具贡献的神经元,增强筛选过程的结构感知能力,并缓解语义贡献在层间分布的不均衡问题。

考虑到不同层在语义建模中的功能差异,首先将模型总层数 L 划分为三个功能区域:浅层区域 $1 \leq l \leq L_s$, 中层区域 $L_s < l \leq L_m$, 深层区域 $L_m < l \leq L$ 。其中 $L_s = \lfloor L/3 \rfloor$, $L_m = \lfloor 2L/3 \rfloor$ 。为体现不同区域对上下文语义建模的重要性,我们在每层 l 上按区域设定一个筛选高贡献神经元的自适应比率 t_l ,其定义如式(3)所示:

$$t_l = \begin{cases} \gamma_{low} t_{global}, & 1 \leq l \leq L_s \\ t_{global}, & L_s < l \leq L_m \\ \gamma_{high} t_{global}, & L_m < l \leq L \end{cases} \quad (3)$$

其中, t_{global} 为预设的全局性高贡献神经元的筛选比

率。考虑到计算开销可控约束下,归因分数的高召回率,我们将其设定为10%。 γ_{low} 和 γ_{high} 是区域自适应因子,由于神经网络的浅层区域和深层区域高质量神经元数量不同,这里设定 $\gamma_{low} \in [0, 1]$, $\gamma_{high} \in [1, 2]$,以自适应调整神经网络不同区域的高质量神经元的筛选比例,并为后续全局候选集构建与精细化筛选提供充足的候选空间。

在确定各区域自适应比率 t_l 后,我们首先在单样本维度上进行神经元重要性初筛,对每个训练样本计算所有FFN神经元归因分数 $Attr(n_i^l)$,在第 l 层中保留满足 $Attr(n_i^l) \geq t_l$ 的神经元,得到该样本的初步候选集合 $S(l)$ 。随后,将所有训练样本的初筛结果进行汇总与去重,构建全局上下文感知神经元候选集 N^* ($N^* = \cup S(l)$)。后续的区域配额与层内筛选均在这一全局预算约束下进行。最后统计其中每个神经元在不同样本中的出现频次 $Count(n_i^l)$,在每一层内根据该频次进行独立排序,得到层内上下文贡献排名 $Rank_l(n_i^l)$ 。

基于上述神经元筛选,我们为三个功能区域分别设定默认比例权重($r_s:r_m:r_d=1:2:7$),用以反映不同区域在最终CANs集合中的期望占比。记第 i 个区域预算配额为 N_i 如式(4)所示:

$$N_i = \frac{r_i}{r_s + r_m + r_d} \cdot |N^*|, i \in \{s, m, d\} \quad (4)$$

其中, $|N^*|$ 表示全局候选集的大小(即预算总数)。最后将区域配额均匀分配至该区域内各层,并在各层内根据 $Rank_l(n_i^l)$ 分别选取前 N_i 个高排名神经元,组合形成上下文感知神经元候选集合 N^{CANs} ,如式(5)所示:

$$N^{CANs} = (N_s \cup N_m \cup N_d) \quad (5)$$

上下文感知神经元分层筛选机制显式平衡了浅层局部特征建模、中层句法整合与深层语义推理的结构性贡献,为后续的神元激活增强与对比解码模块提供了兼具层级代表性与上下文敏感性的候选神经元集合。

3.1.3 贝叶斯优化的动态权重筛选

传统方法中对神经元增强参数(如激活增强强度 λ_{enh} 、感知神经元数量 n_{cn})依赖静态设定,缺乏对任务语境的适应能力,导致泛化能力受限,难以迁移至不同任务场景。为解决上述问题,我们引入基于贝叶斯优化的动态权重调控机制,自适应优化增强上下文感知神经元的数量与激活增强强度,进一步提升模型对上下文的聚焦能力。

(1) 神经元激活增益。基于(§3.1.1)中IG得到的上下文感知神经元,其归因分数直接反映了激活变化对目标答案概率的边际贡献,因此,这些神经元可被视为当前上下文知识的主要载体。在冻结模型参数的前提下,仅对所筛选出的上下文感知神经元施加激活放大,可在不改变模型既有参数化知识的情况下,增强上下文相关表征在后续FFN计算及最终logits中的贡献,从而提升模型对上下文中新知识的响应能力,并降低其对冲突先验路径的依赖。对已筛选出的上下文感知神经元集合 N^{CANs} 的激活值进行增益调节,实现细粒度的上下文适配能力,其增强后的激活表示如式(6)所示:

$$\tilde{h}_i^c = \begin{cases} \lambda_{enh} \cdot h_i^c, & i \in N^{CANs} \\ h_i^c, & \text{otherwise} \end{cases} \quad (6)$$

其中, $\lambda_{enh} > 1$ 为激活增强强度,用于控制激活放大倍数。 h_i^c 为引入上下文 c 后的激活表示。

(2) 贝叶斯优化搜索。在实际解码阶段,为避免引入冗余噪声,我们并不对所有初筛出的候选神经元进行增强,而是进一步选取 $n_{cn}(n_{cn} \leq |N^*|)$ 最具代表性的神经元进行放大。 n_{cn} 与放大强度 λ_{enh} 作为一组耦合超参数,通过贝叶斯优化进行联合搜索并自动确定。联合超参数如式(7)所示:

$$\theta = (n_{cn}, \lambda_{enh}) \quad (7)$$

在超参数空间中,基于Tree-structured Parzen Estimator(TPE)的贝叶斯优化方法进行黑盒搜索,以最大化模型在上下文引导下的语义一致性和鲁棒性,定义如下双目标评分函数如式(8)所示:

$$Score(\theta) = ACC(\theta) - \beta \cdot SR(\theta) \quad (8)$$

其中, $ACC(\theta)$ 是当前超参数下模型的回答正确率,衡量模型对上下文知识的忠实度; $SR(\theta)$ 是当前超参数下模型的顽固率,用于衡量LLMs是否过度依赖参数化知识而忽略上下文; $\beta(\beta > 1)$ 为动态惩罚系数,随顽固率上升而增加,以惩罚幻觉风险。

贝叶斯优化搜索包含四个关键步骤:首先在参数空间中通过网格采样初始化生成初始评价数据集;将历史评价参数划分为“好参数集”(Score高于中位数)和“差参数集”(Score低于中位数),分别拟合概率密度函数 $l(\theta)$ 与 $g(\theta)$,并通过最大化二者的比值进行高效参数采样;随后,通过在每次迭代中持续更新数据集与分布模型,实现参数配置的优化;最终通过连续性能提升小于0.1%的双重收敛判定机制终止优化,得到上下文感知神经元的最佳参数配置 $\theta^* = (n_{cn}^*, \lambda_{enh}^*)$,以此更新FFN神经元的激活策略。

算法 1. 自适应上下文感知神经元筛选

输入: 预训练模型, 数据集 $\{(c_i, q_i, o_i)\}_{i=1}^N$, 积分步数 m ,

全局筛选比率 t_{global} , 最大迭代次数 T ,

输出: 最优超参数 $\theta^* = (n_{cn}^*, \lambda_{enh}^*)$

步骤 1: 计算归因分数

1. FOR (c, q, o) DO

2. $Attr(n_i^l) \leftarrow IG$ 见公式(2)

3. END FOR

步骤 2: 分层神经元筛选

1. SET $t_l = t_{\text{global}}$

2. FOR $l=1$ TO L DO

3. $N^* = \bigcup_{l=1}^L \{n_i^l | Attr(n_i^l) \geq t_l\}$

4. SELECT $Rank_l(n_i^l)$

5. END FOR

6. RETURN N^{CANs}

步骤 3: 贝叶斯优化

1. DO $D_0 = \{(\theta_i, Score(\theta_i))\}_{i=1}^N$ WITH $\{\theta_i\}$

2. FOR $t=1$ TO T DO

3. $\theta_t \leftarrow \arg \max_{\theta} \frac{l(\theta)}{g(\theta)}$

4. $Score(\theta_t) = ACC(\theta_t) - \beta \cdot SR(\theta_t)$

5. $D_t = D_{t-1} \cup (\theta_t, Score(\theta_t))$

6. END FOR

7. RETURN θ^*

如算法 1 所示, 自适应上下文感知神经元筛选算法集成了基于归因的神经元筛选、结构感知的上下文感知神经元分层筛选, 以及贝叶斯优化驱动的动态调控机制, 有效提升了模型在知识冲突与复杂语境下的语义一致性与鲁棒性。

3.2 神经元自适应对比解码

LLMs 的生成质量也受到两方面限制: 一是模型对已参数化知识的过度依赖会抑制对输入上下文的充分利用, 尤其在输入包含新知识或与模型已有知识语义冲突时, 模型倾向于输出与参数记忆一致但与上下文不符的结果; 二是现有大多数对比调控策略未能捕捉层间神经元的个体差异性, 特别是在输入包含新知识或上下文与模型参数化知识存在语义冲突时, 现有方法难以有效调节上下文与模型参数化知识之间的精准平衡。

为突破上述瓶颈, 我们创新性地将对式生成策略与神经元级动态调控机制结合, 构建“关键神经元识别→激活增强→概率重构”三级对比解码架构, 与仅在输出 logits 层进行统一加权的传统策略不同, NACD 在 FFN 层引入上下文感知神经元动态激活增强机制, 并引入 JSD 驱动的动态对比系数, 在解码

阶段构建上下文感知的重构概率分布, 以在结构不变的前提下实现细粒度、可解释的上下文适应能力。

3.2.1 上下文感知神经元激活增益机制

NACD 的核心设计建立在上下文感知神经元的差异化贡献假设之上。基于 ACNS (§3.1) 方法筛选的最具代表性的 n_{cn} 个神经元, 作为解码阶段的动态增强目标。在保持模型结构与参数冻结的前提下, 我们对这些关键神经元施加额外的激活放大, 以引导模型更聚焦于上下文关键路径上的神经信号, 从而为概率分布重构提供更清晰、更具语义解释的表示基础。增强过程定义如式(9)所示:

$$\tilde{h}_i^c = \begin{cases} \lambda_{enh} \cdot h_i^c, & i \in n_{cn} \\ h_i^c, & \text{otherwise} \end{cases} \quad (9)$$

其中, $\lambda_{enh} > 1$ 为激活增强强度, n_{cn} 是最优上下文感知神经元, 其数量与增强强度由贝叶斯优化策略 (3.1.3 节) 动态确定。

通过神经元级增益机制, NACD 能够在不依赖附加参数或重训练的情况下, 对上下文新引入的知识路径产生更强响应, 增强模型在生成阶段的语境适配能力, 为对比解码提供结构感知的中间表示。

3.2.2 神经元级对比解码

在上述激活增益的基础上, NACD 构建了一个更细粒度的对比生成架构。不同于传统对比解码方法仅在输出 logits 层采用统一加权策略, NACD 引入基于 JSD 的动态对比调控策略, 用于动态建模上下文与模型参数化知识之间的语义冲突程度, 实现更稳定、可解释的上下文适配与概率重构。

为实现上下文信息与参数化知识的有效对比, 首先计算在无上下文(ϕ)情况下的参数化知识得分, 可反映模型固有的参数化知识, 如式(10)所示:

$$S_{w_i}^{\phi} = \text{logit}(w_i | (q, o), w < t) = h_i^{\phi} \cdot W + b \quad (10)$$

其中, h_i^{ϕ} 表示模型在空上下文下的激活模式, 反映其参数化知识。W 与 b 分别为固定的输出投影矩阵与偏置项, 反映模型固有的参数化知识偏好。

结合 3.2.1 节中的增强后的激活值和上下文信息 c, 定义生成词 w 的上下文感知得分 $S_{w_i}^c$, 计算如式(11)所示:

$$S_{w_i}^c = \text{logit}(w_i | (c, q, o), w < t) = h_i^c \cdot W + b \quad (11)$$

为量化当前上下文与模型参数化知识之间的语义冲突程度, 放大上下文相关特征的贡献, 我们引入基于冲突强度自适应变化的对比权重 α_{JSD} , 用于调节对比强度, 并在每个解码步骤 t 中将 α_{JSD} 设定为式(12):

$$\alpha_{\text{JSD}} = \alpha_{\text{base}} + D_{\text{JSD}}(\text{softmax}(S_{w_i}^c) \parallel \text{softmax}(S_{w_i}^\phi)) \quad (12)$$

其中, α_{base} 为基础对比强度, 设定为 0.3, 保证模型在低知识冲突时仍维持最低限度的对比约束。 $\text{softmax}(S_{w_i}^c)$ 和 $\text{softmax}(S_{w_i}^\phi)$ 表示上下文与参数化知识的两类概率分布。 D_{JSD} 用于量化上下文与参数化知识的冲突程度, 作为解码过程中对比强度的动态调整因子。

JSD 作为对称且有界的概率分布距离度量方法, 能有效捕捉上下文驱动分布与先验分布间的冲突程度, 适用于解码阶段的渐进动态调整。令 $P = \text{softmax}(S_{w_i}^c)$, $Q = \text{softmax}(S_{w_i}^\phi)$ 计算上下文感知得分分布与参数化知识分布之间的 D_{JSD} , 其定义如式(13)所示:

$$D_{\text{JSD}}(P \parallel Q) = \frac{1}{2} D_{\text{KL}}(P \parallel M) + \frac{1}{2} D_{\text{KL}}(Q \parallel M) \quad (13)$$

其中, $M = \frac{1}{2}(P + Q)$ 为混合分布, Kullback-Leibler (KL) 散度如式(14)所示:

$$D_{\text{KL}}(P \parallel M) = \sum_i P(i) \log \frac{P(i)}{M(i)} \quad (14)$$

$D_{\text{JSD}}(P \parallel Q)$ 直接刻画两者分布的差异, 值越高表明当前上下文与模型参数知识存在显著差异, 意味着模型需要更多依赖上下文相关神经元来缓解语义冲突。反之, 较低的 JSD 值反映模型参数知识与上下文信息间高度一致。

结合上述的上下文感知得分、先验得分以及动态对比强度系数, NACD 对当前词 w_i 的生成概率进行重构, 重构概率分布如式(15)所示:

$$w_i \sim \text{softmax}[(1 + \alpha_{\text{JSD}}) \cdot S_{w_i}^c - \alpha_{\text{JSD}} \cdot S_{w_i}^\phi] \quad (15)$$

具体而言, NACD 采用了增强—抑制自适应双向调节机制, 其中, 系数 $(1 + \alpha_{\text{JSD}})$ 用于显式放大上下文感知得分的贡献, 以强化上下文证据; $(-\alpha_{\text{JSD}})$ 项则用于当先验与上下文发生冲突时, 降低由先验驱动的候选得分, 通过线性组合对候选词得分进行重加权最终概率分布。当上下文分布与参数先验分布相似时 (即 $\alpha_{\text{JSD}} = \alpha_{\text{base}}$), 对比强度接近基线; 当两者差异增大时, α_{JSD} 增大, 对先验 logits 施加更强的抑制, 进而促使模型更多依赖上下文相关神经元, 缓解知识冲突带来的错误偏置。

NACD 通过引入神经增益机制, 结合 JSD 驱动的对比较码策略, 在不依赖模型参数更新的前提下实现上下文和参数知识的动态平衡, 显著提升了模型在语义冲突情境下的上下文忠实度。

4 实 验

为全面评估 ACND 在复杂上下文理解与知识冲突任务中的有效性, 我们在不同上下文长度与领域背景的推理任务上进行了实验。这些任务模拟了真实场景中可能出现的误导性或冲突性信息, 以检验模型能否在干扰条件下依然忠实地依赖输入上下文, 做出符合语义逻辑的推理判断。

4.1 数据集

为全面验证 ACND 的泛化能力与鲁棒性, 本文系统性地选取了三组代表性数据集, 涵盖多样化的知识类型与推理难度, 在上下文长度、上下文复杂度、任务形式与干扰强度等方面形成互补。对每个数据集从其原始样本中随机划分出互不重叠的训练集与测试集, 其中 COSE-KRE、ECARE-KRE、RACE-CCR 离线训练集规模分别为 600、900、1200 个样本, 仅用于执行 ACNS 的神经元归因及贝叶斯优化, 以确定 CANs 集合和最优超参数 θ^* , 测试集只用于评估不同方法的泛化性能。

(1) RACE-CCR: 复杂情境推理

为进一步评估模型在“干扰性上下文”下的推理与判别能力, 我们基于经典数据集 RACE^[34] 构建了 RACE-CCR (Complex contextual reasoning) 数据集, 该数据集侧重于衡量模型在长上下文下进行复杂情境推理的能力。其构建方法参考文献[35]中提出的思路, 并结合本任务的特点进行了相应调整。具体而言, 我们从 RACE 数据集中筛选出符合以下条件的样本:

- ①问题无法通过段落中的单一句子解答;
- ②正确答案需整合多个上下文片段;
- ③干扰项包含微妙语义陷阱或部分真相。

对每道多项选择题, 我们随机选择一个错误选项, 将其设定为正确答案, 并使用 ChatGPT 生成支持该选项的解释性文本, 作为上下文内容。模型需在“干扰性强、逻辑连贯”的背景下识别出真实正确的答案, 从而加大推理难度, 增强对上下文理解的辨析要求。实验所用数据集示例如表 1 所示。

(2) COSE-KRE 与 ECARE-KRE: 冲突知识基准

为全面评估 ACND 在知识冲突及模糊语境下的鲁棒性, 本文选用了由 IRCAN 方法提出的两个冲突知识基准数据集 COSE-KRE 与 ECARE-KRE, 分别基于 ECQA^[36] 和 e-CARE^[37] 构造, 实验所用数据集示例如表 2, 表 3 所示。

表1 RACE-CCR数据集示例	
RACE-CCR	
背景(c)	在其短暂的历史中,机动车造成的死亡和伤残人数已超过任何炸弹或武器。街道上的血迹大多源于司机拒绝尊重他人合法权益和道德权利的不文明行为……
问题(q)	下列哪项最能反映作者对未来消除交通事故问题的态度?
选项(o)	[心存疑虑却仍向往;欣喜且颇有信心;惊讶且非常满意;失望且漠不关心]
正确答案	A

表2 COSE-KRE数据集示例	
COSE-KRE	
背景(c)	每个人都渴望找到人生伴侣。
问题(q)	每个人真正想要的是什么?
选项(o)	[人生伴侣;更大的房子;第二次机会;金钱;头痛]
正确答案	A

表3 ECARE-KRE数据集示例	
ECARE-KRE	
背景(c)	锡的制备需在熔炉中将锡石与碳加热冶炼。
问题(q)	该男子计划自行制造锡。更可能的结果是?
选项(o)	[他必须设计必要组件;他找到了锡石、碳和熔炉。]
正确答案	B

4.2 模型与评估指标

(1) 基座模型。为全面验证 ACND 在多领域上下文推理的通用性,在多个主流 LLMs (LLaMA2-7B、Qwen2.5-7B、Qwen3-4B、Baichuan2-7B) 上开展实验,涵盖不同训练范式与参数规模,以评估 ACND 在跨模型、跨任务及复杂上下文推理场景下的鲁棒性与泛化能力。为确保评估标准一致性,遵循 MMLU^[38] 官方实现及其他基准测试^[39-41] 广泛采用的方式,通过提示工程为每个模型设计统一提示,并根据选项中的最高概率选择答案。模型提示模板如表 4 所示。

表4 基于提示工程的方法所使用的提示	
格式化提示	
根据上下文选择正确选项回答以下问题:	
上下文:	{context}
问题:	{question}
选项:	A. {...} B. {...} C. {...} ...
答案:	

(2) 评估指标。我们采用双指标评估,以全面衡量模型在知识冲突场景下的表现。

准确率 (Accuracy, ACC): 衡量模型在冲突上下

文条件下选择正确答案的能力,是评估其整体语义理解与上下文推理准确性的核心指标。

顽固率 (Stubbornness Rate, SR): 在知识冲突场景中,ACC 只能反映模型最终答案是否正确,难以区分两类不同来源的错误: 其一是模型已正确理解并利用上下文,但在推理过程中出现失误; 其二是模型忽视上下文信息,仍然遵循预训练形成的先验知识作答。因此额外定义 SR 作为衡量幻觉的重要指标,用于度量当上下文与先验知识发生冲突时模型对先验知识的固守程度,从而评估大模型生成结果的可靠性。

对于每个样本 k , 我们构造一对答案标签, y_k^{gold} 表示与上下文一致的正确答案, y_k^{prior} 表示代表模型预训练参数化知识的先验答案, 其在当前上下文下为错误选项, 但通常符合常识或旧知识。记模型输出为 y_k , 则顽固率定义如式 (16) 所示:

$$SR = \frac{1}{N} \sum_{k=1}^N I[y_k = y_k^{\text{prior}} \wedge y_k^{\text{prior}} \neq y_k^{\text{gold}}] \quad (16)$$

其中, $I[*]$ 为指示函数, 当括号内条件为真时取 1, 否则取 0。SR 越高, 说明模型在知识冲突时越倾向于忽视当前上下文并依赖预训练形成的参数化知识, 从而对应更高的幻觉风险。

4.3 基线方法

为全面对比验证 ACND 在消解知识冲突与缓解幻觉方面的优势, 我们将其与以下代表性基线方法进行比较。(1) Original: 未经过任何干预的 LLMs, 依赖其预训练阶段获取的内在知识进行推理, 是评估方法性能增益的基线指标。(2) Supervised Fine-Tuning (SFT): SFT 代表经典的数据驱动优化范式。我们在现有 LLMs 上应用 SFT, 作为增强模型性能的对比基线。具体的方法技术细节详见附录。(3) IRCAN: 基于神经元归因的上下文调控方法, 首先通过计算积分梯度筛选对上下文感知的关键神经元, 并对其进行重加权, 以提升模型对上下文信息的敏感度。(4) CAD: 一种基于对比解码的优化策略, 通过调整大语言模型的输出概率, 突出上下文信息对结果的影响差异。(5) IRCAN-CAD: 结合 IRCAN 与 CAD 的联合优化方法, 源于 IRCAN 原文提出的组合策略, 旨在通过两者协同提升模型处理知识冲突的能力。

4.4 主要结果

为系统评估各方法在面对上下文与参数化知识冲突场景下的性能表现, 基于模型准确率与顽固率

两大核心评价指标在不同数据集上开展对比实验。各数据集按 1:1 比例进行随机划分,构建离线训练集与测试集,以此保障样本分布的均衡性。具体实验结果详见表 5。

表 5 不同模型与方法在测试数据集的评估结果(%)

数据集	模型/ 方法	Qwen3-4B	Qwen2. 5-7B	Baichuan2 -7B-SFT	Baichuan2 -7B	LLaMA2 -7B-chat	Qwen2. 5- 7B-Inst
		ACC ↑ /SR ↓	ACC ↑ /SR ↓	ACC ↑ /SR ↓	ACC ↑ /SR ↓	ACC ↑ /SR ↓	ACC ↑ /SR ↓
COSE-KRE	Original	44. 84/43. 37	54. 83/38. 63	47. 30/34. 86	45. 17/34. 04	36. 17/19. 31	48. 61/41. 90
	CAD	52. 54/36. 82	68. 90/26. 35	56. 96/29. 46	52. 86/28. 48	36. 50/ 17. 18	<u>56. 63</u> /32. 90
	IRCAN	45. 66/42. 39	54. 66/37. 81	48. 12/31. 59	44. 84/33. 55	41. 57/20. 95	45. 99/36. 82
	IRCAN-CAD	50. 34/38. 44	68. 25/ 24. 88	<u>60. 39/25. 86</u>	<u>53. 36/28. 48</u>	<u>45. 01/18. 00</u>	53. 85/ 29. 62
	ACND	<u>52. 21/36. 99</u>	<u>68. 41/27. 00</u>	64. 16/24. 55	54. 83/27. 17	46. 15/18. 66	59. 41/30. 61
RACE-CCR	Original	77. 50 /45. 25	85. 83/45. 33	79. 50/45. 25	55. 92/62. 00	40. 58/25. 58	85. 92/43. 50
	CAD	<u>76. 08/43. 08</u>	85. 58/43. 25	79. 92/ 42. 25	<u>57. 50/55. 00</u>	39. 92/25. 33	86. 42 /41. 83
	IRCAN	75. 50/44. 58	<u>85. 83</u> /45. 25	81. 25/47. 25	55. 83/61. 83	41. 33/ 20. 67	85. 50/43. 00
	IRCAN-CAD	60. 25/43. 25	85. 50/ 43. 25	<u>81. 58/44. 00</u>	56. 75/ <u>54. 85</u>	40. 25/20. 92	80. 25/ 40. 67
	ACND	74. 00/ 42. 08	86. 50/43. 30	83. 75/43. 42	58. 58/54. 00	41. 67/20. 92	<u>86. 25/41. 58</u>
ECARE-KRE	Original	62. 11/37. 89	72. 76/27. 24	66. 07/33. 93	61. 55/38. 45	65. 22/34. 78	66. 16/33. 84
	CAD	<u>74. 18/25. 82</u>	80. 21/19. 79	<u>74. 46/25. 54</u>	<u>69. 93/30. 07</u>	73. 61/26. 39	71. 25/28. 75
	IRCAN	64. 94/35. 06	71. 82/28. 18	62. 77/37. 23	56. 64/43. 36	73. 61/26. 39	56. 83/43. 17
	IRCAN-CAD	73. 27/30. 88	75. 97/24. 03	68. 80/31. 20	66. 73/33. 27	<u>80. 02/19. 98</u>	61. 17/38. 83
	ACND	77. 47/22. 53	<u>80. 11/19. 89</u>	75. 49/24. 51	74. 93/25. 07	82. 94/17. 06	<u>70. 88/29. 12</u>

注:最佳结果以粗体显示,第二佳结果以下划线显示

以常识推理 COSE-KRE 为例,Original 中平均顽固率高达 34%,反映出 LLMs 默认倾向于依赖参数化知识,而未能充分利用当前上下文。从参数规模角度分析,7B 模型(如 Qwen2. 5-7B)在准确率方面普遍优于 4B 模型(如 Qwen3-4B),但在顽固率控制方面未表现出明显优势,说明仅增加参数规模并不能自动提升模型的上下文对齐能力。

在不同参数规模的模型优化效果方面,CAD 方法在多个模型上实现了约 8% 的平均准确率提升,但其对知识顽固性的缓解效果有限,平均 SR 仅下降 6. 9%,说明其在压制模型内在偏向方面仍存在不足。IRCAN 在部分模型上实现了性能提升,但在部分模型上出现了准确率下降的现象,表明静态神经元增强策略在语境动态变化下存在适应瓶颈。IRCAN+CAD 在部分模型上取得了性能提升,但整体表现仍不及 ACND,说明简单模块叠加难以实现最优协同优化。

相比之下,ACND 在不同参数规模模型上均展现出优异的综合性能。以 Baichuan2-7B-SFT 为例,其准确率较原始模型提升 16. 9%,显著优于其他对比方法;同时顽固率下降 10%,表明该方法能更有效抑制模型对内在知识的过度依赖,并增强对上下文关键信息的对齐能力。值得关注的是,经

ACND 优化的 4B 模型(如 Qwen3-4B)在性能上甚至超越了部分原始 7B 模型,表明有效的优化方法可在一定程度上弥补模型规模的不足。

在更具挑战性的复杂上下文推理 RACE-CCR 和因果推理 ECARE-KRE 数据集上,ACND 同样展现出稳健而优越的综合性能表现。在 RACE-CCR 上,ACND 在多个主流模型中均取得最优综合性能,表现出出色的长文本建模能力与任务迁移适应性。在 ECARE-KRE 上,ACND 在 Baichuan2-7B 和 LLaMA2-7B-chat 模型上分别达到 74. 93% 和 82. 94% 的准确率,且 SR 降至 25. 07 和 17. 06,均为所有方法与模型中的最优结果。

综上所述,ACND 在不同参数规模的模型与任务中均实现了准确率与顽固率的双重显著提升,表明该方法能有效缓解模型对参数化知识的过度依赖,提升了语义忠实度与生成一致性,其综合性能达到当前最优水平,且在不同规模模型上均展现出良好的泛化能力。

4. 5 效率成本与计算开销

ACND 框架可分为离线优化阶段与在线推理阶段两部分。离线阶段通过 ACNS 模块执行积分梯度归因与贝叶斯优化,以确定上下文感知神经元集合 CANs 及最优超参数 θ^* 。该阶段的计算成本不

可忽略,但其开销完全集中在离线筛选过程,不会影响在线推理效率。在线阶段则在冻结模型参数的前提下,仅对已筛选神经元执行轻量级的激活缩放,并进行logits的线性组合完成解码,整个推理过程不再进行任何梯度计算或超参数搜索。因此,“即插即用”主要体现在推理阶段:一旦离线优化完成,ACND可作为解码插件直接部署于现有模型之上。

4.5.1 离线优化成本

(1)神经元归因计算。为评估FFN层神经元的上下文敏感性,每个样本执行从基线到上下文输入的梯度积分,并通过离散化黎曼积分对该过程进行近似以提升计算效率。所有神经元归因计算均可在单张48 GB显存的NVIDIA A6000 GPU上完成。离线IG耗时与模型规模、离线训练集规模及输入长度有关。以COSE-KRE为例,4B模型的归因开销约为8小时,7B级模型约为12-16小时(见表6)。归因过程仅在离线阶段执行一次,推理阶段直接加载筛选完成的CANs,无需重复归因计算,因此不会影响在线推理效率。

(2)离线超参数搜索。为避免在线推理产生额外开销,对联合超参数 θ 的搜索同样在离线阶段完成。在相同搜索空间下,ACND的贝叶斯优化策略仅需120次评估即可收敛,显著少于IRCAN采用网格搜索策略进行的304次评估^①,离线超参数搜索阶段的候选评估次数减少约60%。

为进一步验证贝叶斯优化策略的效率优势,本文以COSE-KRE任务为例,在相同的硬件环境(单张NVIDIA A6000 48 GB GPU)下,分别统计了ACND与IRCAN完成IG归因计算与超参数搜索所需的壁钟时间(Wall-clock Time)。如表6所示,IG归因是ACND与IRCAN共享的基础开销,但在超参数搜索阶段,得益于贝叶斯优化策略的高效采样,ACND在多个模型上展现出显著的时间优势。

以Qwen3-4B模型为例,ACND仅需2.0 h完成超参数搜索,而IRCAN完成网格搜索需耗时7.3 h。在参数规模更大的Qwen2.5-7B模型上,该优势更为明显,超参数搜索时长从28.0 h降至4.5 h。离线最优超参数 θ^* 一经确定,即可在推理阶段固定复用,不引入额外在线计算。

4.5.2 在线推理延迟

为量化ACND在推理阶段的在线计算开销,在与主实验一致的推理配置与硬件环境下,采用每秒处理的完整生成样本数(samples/s)作为在线吞吐指标,该指标直接反映了在线推理速度,数值越高代表处理延

表6 不同模型在离线筛选阶段的主要耗时对比

模型	IG 归因耗时	IRCAN	ACND
		(网格搜索)	(贝叶斯优化)
Qwen3-4B	8 h	7.3 h	2 h
LLaMA2-7B-chat	16 h	10.5 h	5.5 h
Baichuan2-7B-SFT	15 h	27 h	5 h
Baichuan2-7B	15 h	27 h	5 h
Qwen2.5-7B-Inst	12 h	28 h	4.5 h
Qwen2.5-7B	12 h	28 h	4.5 h

注:所有耗时均在相同硬件环境与相同推理配置下统计

迟越低。评估不同规模与不同谱系模型在Original(贪婪解码)、CAD(对比解码)与ACND三种推理框架下的吞吐结果,计算ACND相对Original与CAD的吞吐比(ACND/Original与ACND/CAD),结果如表7所示。

表7 不同解码方法在各模型的在线计算开销(samples/s)

模型	Original	CAD	ACND	ACND/	ACND/
				Original	CAD
Qwen3-4B	8.78	4.83	4.58	52.16%	94.82%
Baichuan2-7B	7.48	4.49	4.55	60.83%	101.34%
LLaMA2-7B-chat	8.30	4.76	4.80	57.83%	100.84%
Qwen2.5-7B	9.03	5.39	5.37	59.47%	99.63%
LLaMA2-13B-chat	5.21	4.53	3.49	66.99%	77.04%
Qwen3-14B	4.05	2.40	2.50	61.73%	104.17%

相比贪婪解码,ACND以可控的效率代价换取模型生成质量提升,ACND的吞吐量平均约为Original的59.84%,平均效率下降约40.16%,下降主要是因为对比解码机制需要额外的无上下文前向计算以构造对比分布,这一计算开销是CAD与ACND架构所共有的。相较于CAD方法,ACND在多数模型上并未引入额外推理成本,在多数测试模型中ACND/CAD吞吐比率波动不超过±5%,这表明神经元激活增强与自适应对比解码在推理阶段的额外成本整体可控,在线开销与CAD基线基本持平。

综上,虽然引入IG与贝叶斯优化带来了数小时级的离线计算成本,但这一过程只需在每个模型离线执行一次,不影响在线推理。尽管吞吐量有所下降,但考虑到ACND带来的显著ACC提升和SR下降,以时间换质量的权衡在对安全性要求高的场景下是完全可接受的。整体而言,ACND在显著提升生成质量与上下文一致性的同时,在多数模型上保

^① IRCAN 开源代码仓库: <https://github.com/danshi777/IRCAN>。其默认网格搜索设置为304次。

持了与标准CAD相近的在线推理效率,从而在可接受的系统开销下缓解上下文和先验冲突引发的错误固执与幻觉倾向。

4.6 消融实验

为进一步评估 ACND 框架中各关键模块对整体性能的贡献,我们设计了以下消融实验:通过分别移除 ACNS 与 NACD 两大核心模块,系统分析其对模型性能的影响。具体设置包括:w/o ACNS:使用静态阈值替代动态神经元筛选机制;w/o NACD:关闭对比解码过程中的神经元增强与 logits 重构;RanN:随机选择与 n_{cn} 数量相等的神经元进行增强或置零,模拟无归因选择的随机增强策略;为了确保实验结果的可靠性和鲁棒性,我们独立重复实验 10 次,并报

告这些结果的平均值作为我们最终分析的基础,旨在最小化选择偏差并增强我们发现的统计显著性。

消融实验结果(如图 2 所示)揭示了各核心组件对模型性能的影响。当移除 ACNS 模块后,模型各任务准确率均出现显著下降,这有力地验证了基于数据驱动归因与贝叶斯优化的动态筛选策略在定位关键特征路径时的关键作用。移除 NACD 后模型平均顽固率提升 9%,表明神经元级对比调控在抑制错误激活与增强抗干扰能力方面具有关键作用。此外,通过与 RanN 组的对比,进一步验证了针对性激活增益的有效性。整体而言,ACND 在提升上下文忠实度的同时有效抑制了知识固守,表现出极强的鲁棒性与泛化性。

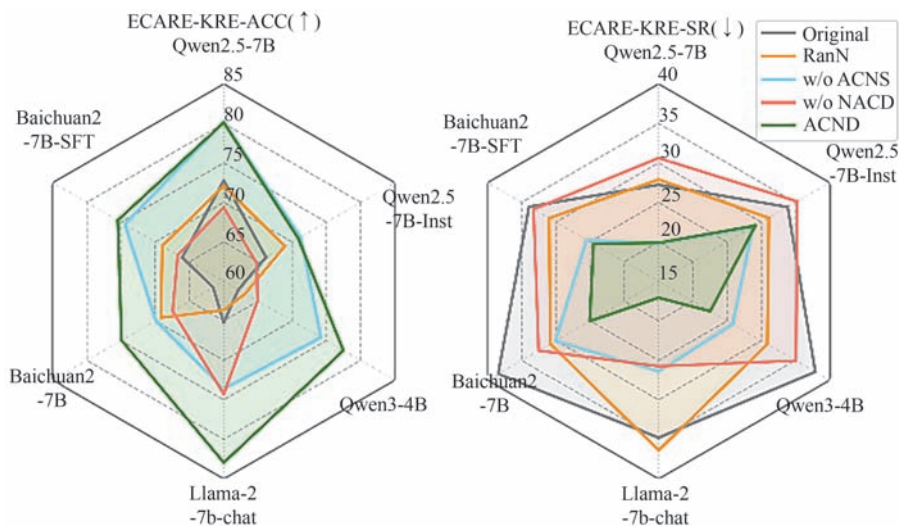


图2 ECARE-KRE数据集消融实验结果对比(ACC越大越好,SR越低越优)

为了进一步评估 ACND 在多维度推理场景下的普适性,我们在 COSE-KRE 和 RACE-CCR 任务中进行了消融分析(如图 3、图 4 所示)。在所有消融实验中,完整的 ACND 框架在 ACC 与 SR 指标上均实现了综合最优,说明 ACNS 与 NACD 模块在动态筛选与对比调控方面形成了有效互补,两者的有机结合实现了感知与决策的闭环调控,协同提升了模型在复杂语义环境下的表征深度与生成稳定性。

综上所述,消融实验的综合结果表明,ACND 在多项评价指标上均取得了综合最优表现,并稳定优于原始模型及各类消融变体。这不仅验证了 ACNS 与 NACD 两个核心模块在整体框架中的必要性与有效性,也进一步说明了 ACND 在不同任务场景和模型架构下具有良好的稳健性、泛化能力与应用潜力。

4.7 分布偏移鲁棒性评估

为验证 ACND 在数据分布偏移(distribution shift)场景下的鲁棒性,在 COSE-KRE(D_1)、RACE-CCR(D_2)和 ECARE-KRE(D_3)间进行了系统的交叉评估。

对于每个模型 M ,首先在数据集 D_i 的训练集上执行一次 ACNS 离线归因与贝叶斯优化得到最优参数 $\theta^*(M, D_i)$,然后在同一数据集的测试集上评估 ACND(in- D_i)。为进一步评估跨分布迁移能力,将模型 M 在数据集 D_j 的训练集上得到的最优参数 $\theta^*(M, D_j)$,直接迁移到目标数据集 D_i ($i \neq j$) 的测试集,在不重新进行任何归因或超参数搜索的前提下评估 ACND(cross $D_j \rightarrow D_i$)。通过对比 in- D_i 与 cross $D_j \rightarrow D_i$ 的结果,我们量化 ACND 在分布偏移条件下的鲁棒性与可迁移性,不同模型在跨分布设置下的 ACC 和 SR 的性能结果如表 8 所示。

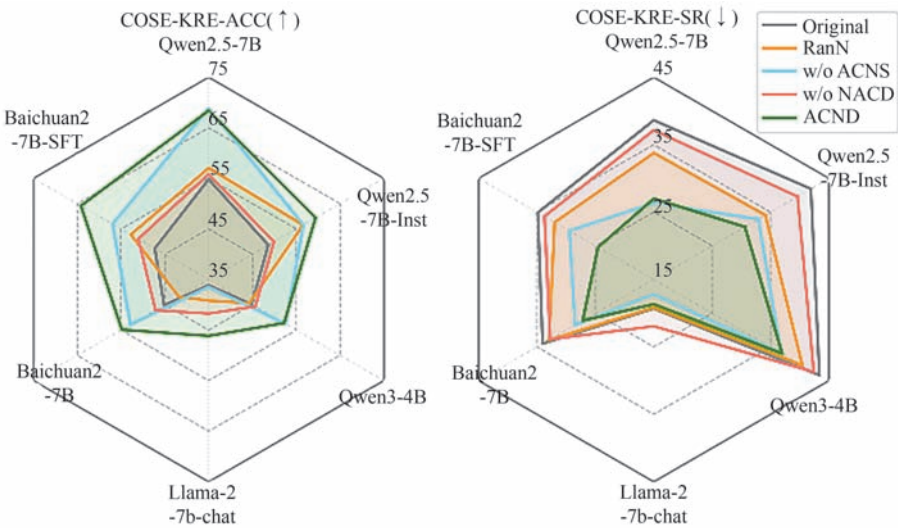


图3 COSE-KRE数据集消融实验结果对比

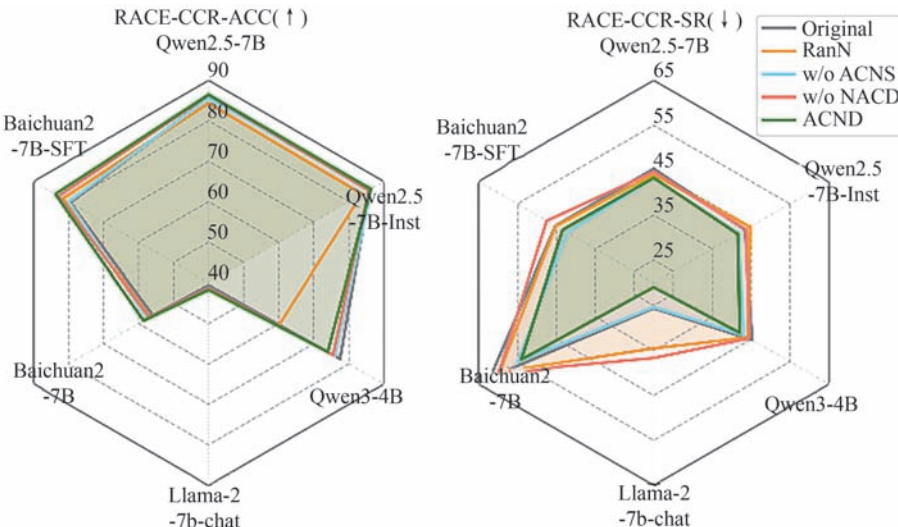


图4 RACE-CCR数据集消融实验结果对比

对比所有模型和测试集,ACND在同分布评测下均优于贪婪解码基线 Original,验证了ACND在各数据集自身分布上的有效性。在跨分布评测下,尽管不同数据集在语境和知识结构上存在较大差异,基于其他数据集优化得到的ACND(cross)仍在大多数迁移组合中保持正向收益,这不仅展现了ACND良好的参数迁移能力,这也说明其缓解错误固执的作用在跨任务分布下仍然成立,还为ACNS所捕获的冲突缓解机制具有跨任务分布的可迁移性提供了实证支持。

综合上述结果,ACND在交叉评估中表现出良好的分布偏移鲁棒性与跨任务泛化能力,其核心优势在于仅通过一次离线优化,即可识别出一个对上下文和先验知识高度敏感且具有一定可迁移性的上下文感知神经元子空间,即使在测试集分布不一致

的情况下,在新任务上仍能保持稳定的性能增益,这有力地证明了ACND具备跨分布泛化能力,在分布偏移场景下有效缓解由知识冲突引发的幻觉。

5 分 析

5.1 动态超参数的协同效应

为实现更具适应性的神经元增强机制,ACNS引入了两个关键的超参数神经元增强数量 n_{en} 与激活放大强度 λ_{enh} ,由贝叶斯优化策略 (§3.1.3) 对其进行动态调控,有效避免了人工参数调优的主观性和非泛化性,实现了上下文感知神经元的自适应增强。为系统探究这一动态机制的性能上限与稳定性,如图5所示,我们在ECARE-KRE任务上对超参数组合进行分析,以探究其对模型性能的影响。

表 8 ACND 跨数据集分布的泛化性能 (%)				
模型	测试集	测试集— $\mathcal{D}_1$	测试集— $\mathcal{D}_2$	测试集— $\mathcal{D}_3$
		ACC $\uparrow$ /SR $\downarrow$	ACC $\uparrow$ /SR $\downarrow$	ACC $\uparrow$ /SR $\downarrow$
Qwen2.5-7B	Original	54.83/38.63	85.83/45.33	72.76/27.24
	COSE-KRE ($\mathcal{D}_1$)	68.41/27.00 (in- $\mathcal{D}_1$)	84.42/43.17 (cross)	80.02/19.98 (cross)
	RACE-CCR ($\mathcal{D}_2$)	55.16/18.17 (cross)	86.50/43.30 (in- $\mathcal{D}_2$)	75.02/24.98 (cross)
	ECARE-KRE ($\mathcal{D}_3$)	68.58/26.68 (cross)	82.83/43.75 (cross)	80.11/19.89 (in- $\mathcal{D}_3$)
Qwen3-4B	Original	44.84/43.37	77.50/45.25	62.11/37.89
	COSE-KRE ($\mathcal{D}_1$)	52.21/36.99 (in- $\mathcal{D}_1$)	61.33/41.83 (cross)	74.08/25.92 (cross)
	RACE-CCR ($\mathcal{D}_2$)	52.21/36.99 (cross)	74.00/42.08 (in- $\mathcal{D}_2$)	77.47/22.53 (cross)
	ECARE-KRE ($\mathcal{D}_3$)	52.37/34.21 (cross)	67.17/39.00 (cross)	77.47/22.53 (in- $\mathcal{D}_3$)
Qwen3-14B	Original	40.75/49.75	86.25/41.83	60.32/39.68
	COSE-KRE ($\mathcal{D}_1$)	53.19/36.82 (in- $\mathcal{D}_1$)	87.00/42.50 (cross)	72.89/27.14 (cross)
	RACE-CCR ($\mathcal{D}_2$)	53.19/38.63 (cross)	87.00/42.00 (in- $\mathcal{D}_2$)	70.59/29.41 (cross)
	ECARE-KRE ($\mathcal{D}_3$)	52.37/38.30 (cross)	87.25/42.33 (cross)	70.50/29.50 (in- $\mathcal{D}_3$)
LLaMA2-7B-Chat	Original	36.17/19.31	40.58/25.58	65.22/34.78
	COSE-KRE ($\mathcal{D}_1$)	46.15/18.66 (in- $\mathcal{D}_1$)	39.42/24.00 (cross)	76.44/23.56 (cross)
	RACE-CCR ($\mathcal{D}_2$)	44.68/20.95 (cross)	41.67/20.92 (in- $\mathcal{D}_2$)	79.64/20.36 (cross)
	ECARE-KRE ($\mathcal{D}_3$)	37.64/18.00 (cross)	40.93/23.75 (cross)	82.94/17.06 (in- $\mathcal{D}_3$)
LLaMA2-13B-Chat	Original	48.12/34.04	62.92/42.25	69.75/30.25
	COSE-KRE ($\mathcal{D}_1$)	54.5/24.39 (in- $\mathcal{D}_1$)	58.75/35.75 (cross)	78.13/21.87 (cross)
	RACE-CCR ($\mathcal{D}_2$)	55.16/26.02 (cross)	63.75/39.67 (in- $\mathcal{D}_2$)	79.36/20.64 (cross)
	ECARE-KRE ($\mathcal{D}_3$)	55.40/22.09 (cross)	55.33/31.92 (cross)	80.68/19.32 (in- $\mathcal{D}_3$)

注:粗体表示同分布测试的模型最佳性能;Original 为贪婪解码基线;(in- $\mathcal{D}_i$)表示同分布测试;(cross)表示跨分布测试

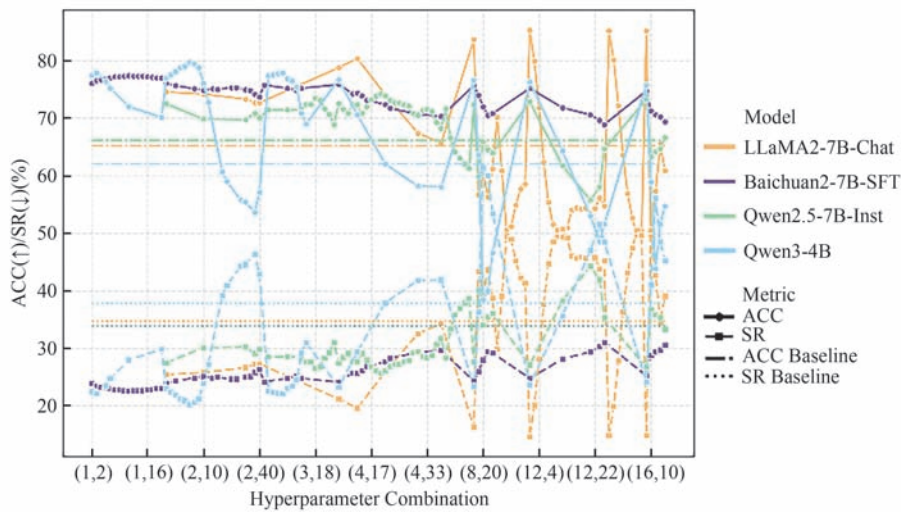


图5 超参数组合对ECARE-KRE任务中模型性能的影响

上述实验结果表明,ACNS在复杂语义推理与长文本理解任务中存在性能峰值窗口,这一特征有力验证了动态超参数配置对提升模型适应性与鲁棒性的核心价值。具体而言,当神经元增强数量处于合理范围时,能够有效凸显高贡献神经元的表示优势,为解码过程提供稳定且优质的信息支撑;而与之匹配的激活放大强度,则可进一步放大这些神经元对语义特征的响应能力,显著增强模型对上下文语

义的聚焦效果。相反,超出最优区间时,冗余或噪声神经元的引入将扰乱表征焦点,导致上下文建模能力减弱,甚至产生性能退化现象。由此可见,基于贝叶斯策略的联合优化不仅提升调参效率,也为ACNS提供了更具泛化性与稳定性的控制能力。

我们系统性地研究了动态超参数增强的 n_{en} 与 λ_{enh} ,在COSE-KRE-KRE和RACE-CCR两个数据集上的性能影响。图6和图7的结果表明,两个超

参数在合理范围内的协同增加可有效强化关键信号聚焦并提升语义建模效果;反之,若超出该范围,则会因引入噪声而破坏表示结构,导致性能衰退。

上述结果进一步验证了本文提出的动态超参数机制在保障鲁棒性、泛化性及上下文对齐效能方面的关键作用。

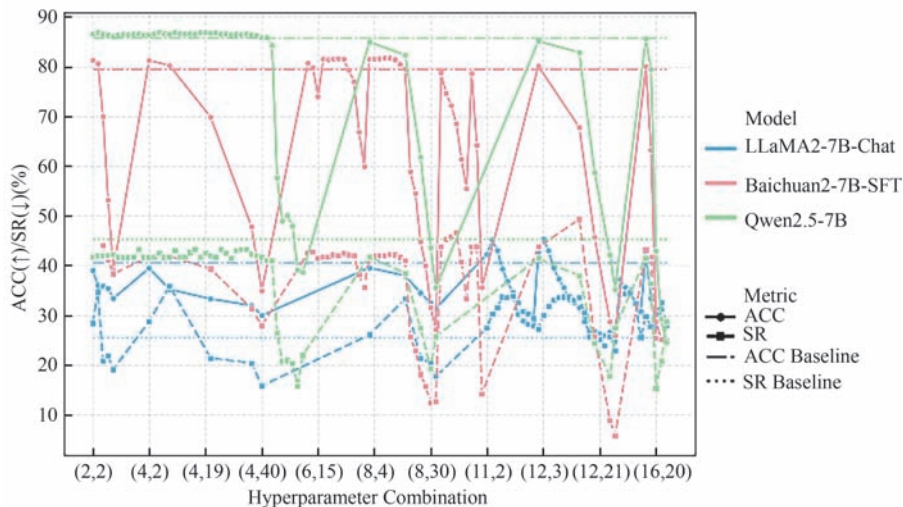


图6 超参数组合对RACE-CCR任务中模型性能的影响

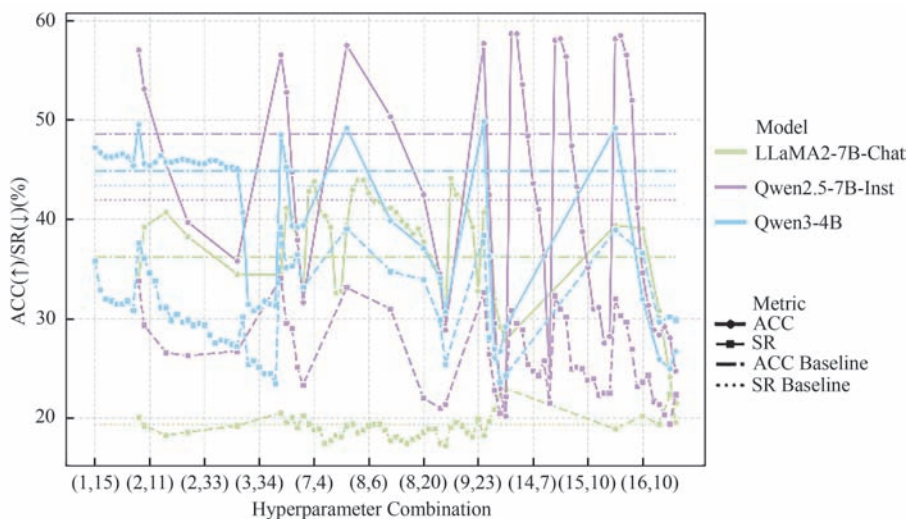


图7 超参数组合对COSE-KRE任务中模型性能的影响

此外,我们在不同任务场景下对超参数组合的行为模式进行了系统分析。在语义跨度更广、知识密度更高的任务中,动态参数调控策略的性能差异更加显著。这一趋势进一步验证了协同超参数机制在复杂场景下的适应性优势,充分体现了其对上下文语义对齐与信息干扰抑制的有效性。

5.2 神经元层级特性

为进一步分析 ACNS 在模型结构中的作用机理,我们对筛选结果中的上下文感知神经元在不同层级的分布特性进行系统分析。如图8所示, CANs 在不同模型结构中均呈现出层级聚集特征,反映出不同网络层在上下文建模中的功能分化与

表征偏好。

在靠近输入端的浅层, CANs 占比普遍偏低。这一层级主要捕捉词法与局部句法信息,对上下文依赖较弱。中间层的 CANs 占比提升至约 20%,模型开始进行更复杂的句法抽象与初步的跨位置信息整合,是上下文信息逐步显著影响表示学习的重要阶段。在模型深层, CANs 占比显著上升,平均超过 60%,表明该层级在处理高层语义融合与全局逻辑推理中的主导地位。这一发现也与先前研究一致,即语言模型主要在顶层编码“语义”信息。

通过对不同任务中 CANs 分布模式的横向对比(图9、图10所示),可进一步揭示 ACNS 机制的任

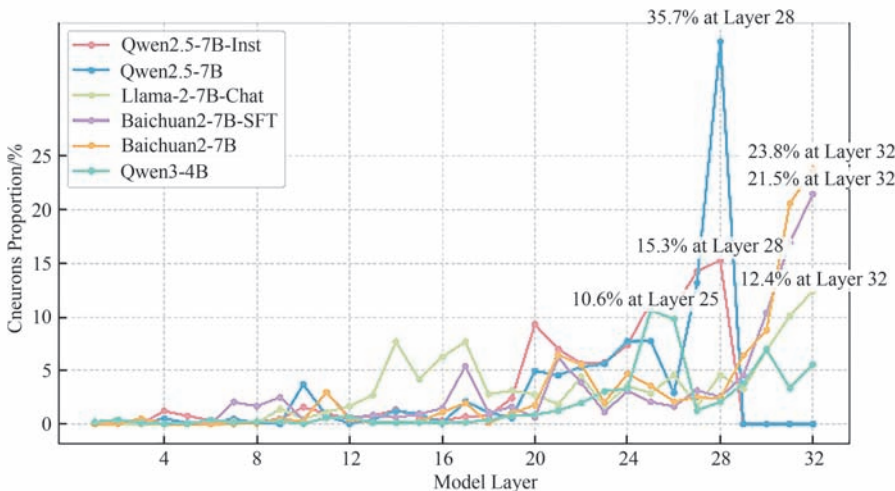


图8 COSE-KRE任务不同LLMs的CANs分布层级结构

务适应性与泛化能力。在 ECARE-KRE 任务中, CANs 主要集中在模型高层(占比超过 65%),体现出该任务对高层语义整合与全局推理的强依赖。而在 RACE-CCR 任务中, CANs 呈现出更为均衡的中

高层分布模式(中层约 32%, 高层约 58%), 表明此类任务同时依赖于局部因果推理与全局语义建模, 进一步验证了 ACNS 能够依据任务特性动态调整关键神经元的筛选策略。

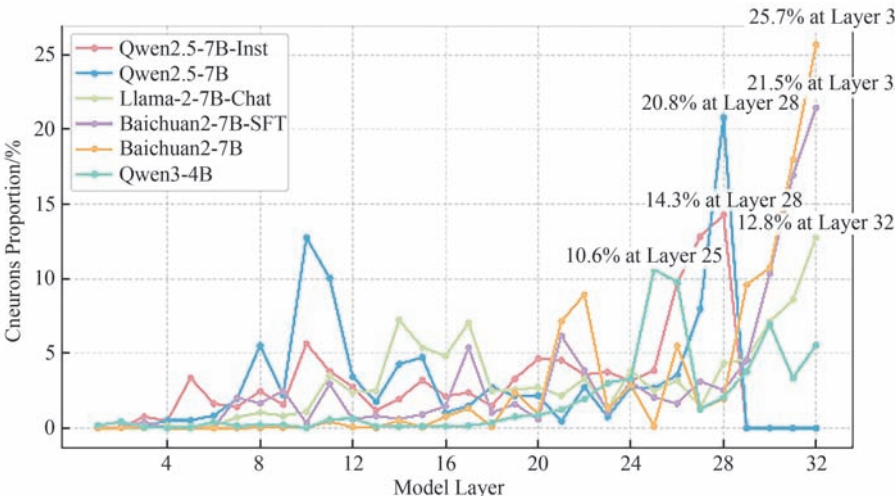


图9 RACE-CCR任务中不同LLMs的CANs分布

从模型参数规模的角度比较发现,不同规模的模型(如 4B 与 7B)在 CANs 的层级分布上呈现出相似的趋势,但 7B 模型在深层神经元的上下文感知强度上高于 4B 模型,反映出参数规模的增加可能有助于增强模型对复杂语义上下文的建模能力。ACNS 在不同参数规模的模型上均表现一致,进一步证明了其在结构层面的鲁棒性与可迁移性。

综上所述,ACNS 通过其分层动态筛选机制,能够有效识别不同任务中关键的语义建模层级,优先筛选对上下文变化高度敏感的神经元,展现出良好的任务适应性与跨模型泛化能力。

5.3 对比解码调控机制分析

5.3.1 动态对比系数的冲突响应能力

为更有效识别并适应上下文知识与模型先验之间的语义冲突, NACD 引入了动态对比系数机制, 并以 JSD 作为冲突强度的感知度量。相较于 KL 散度, JSD 具备对称性且取值范围归一化 $[0, 1]$, 使其更适用于生成决策过程中的动态冲突检测与响应调控。为系统评估该机制在实际任务中的表现, 我们在多个推理任务中, 比较了 NACD、CAD 与 Original 的性能指标, 如图 11 所示。

实验结果表明, NACD 在不同规模模型与各类

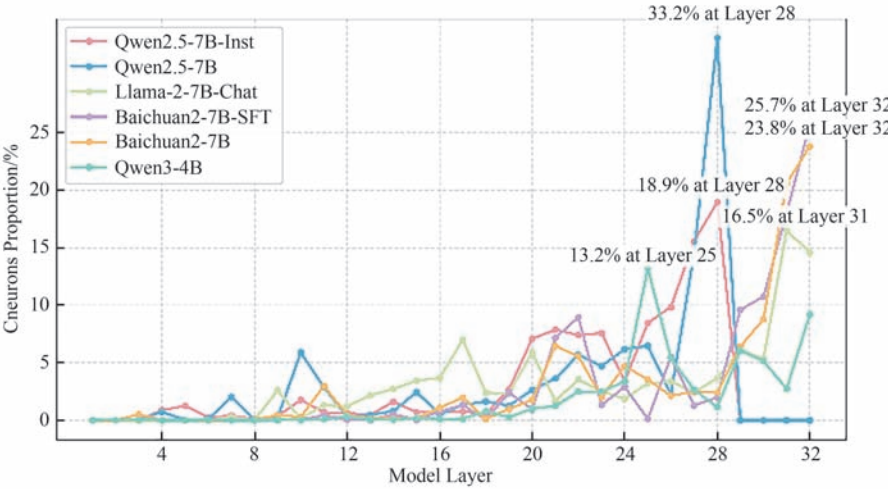


图 10 ECARE-KRE任务中不同LLMs的CANs分布

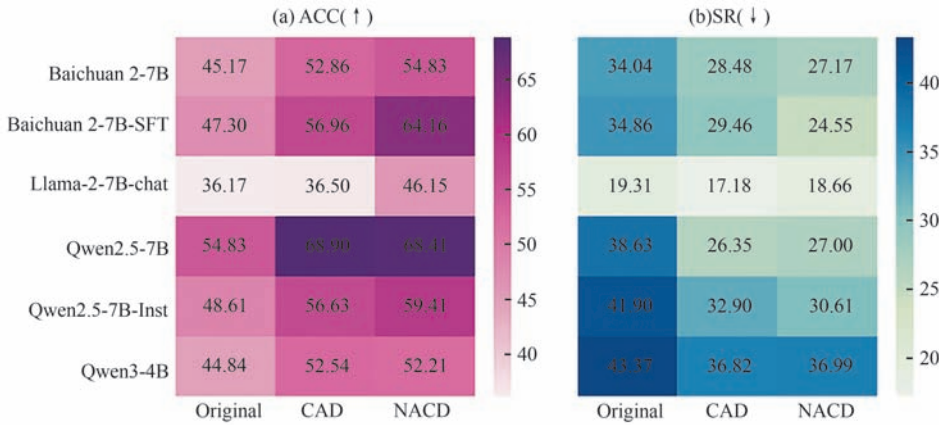


图 11 不同任务和模型上,Original、CAD($\alpha=0.5$)与NACD(α_{JSD})的响应对比

任务中均实现了稳定性能提升,展现出优秀的泛化能力。在ECARE-KRE上,NACD展现出更强的鲁棒性与泛化能力,平均较Original降低顽固率约11%,相较CAD进一步平均下降3%;在COSE-KRE上,相较Original平均降幅约8%,相较CAD进一步降低约2%;在RACE-CCR任务中,NACD也依然领先于CAD,展现出对上下文路径的强化能力与对先验干扰的有效抑制能力。NACD在不同参数规模的模型(从4B到7B)上均表现出一致的性能提升趋势,未因模型规模变化而产生显著波动,这一现象充分验证了该方法具备良好的模型无关性与泛化能力。

综上所述,NACD通过其动态对比系数机制,在多模型、多任务场景下均展现出优异的适应性与一致性,能够有效缓解模型对静态参数知识的过度依赖,增强对输入上下文的语义响应能力。相较于现有基线方法,NACD在复杂语境下表现出更优的解码鲁棒性与广泛适用性。

5.3.2 对比调控粒度差异性分析

为进一步验证ACND在细粒度神经元调控层面的优势,本文选取代表性的层级对比解码方法DoLa (Decoding by Contrasting Layers)进行对比分析。DoLa通过动态对比同一模型高层与低层的输出分布差异来增强事实性,其核心优势在于利用层间差异捕捉事实知识。鉴于DoLa官方实现对模型架构的特定适配性,本文在LLaMA2-7B-chat模型上,针对全部三个数据集进行了统一评估,结果如表9所示。

表 9 LLaMA2-7B-chat上不同解码方法的对比结果(%)

数据集	Original	DoLa	ACND
	ACC ↑ /SR ↓	ACC ↑ /SR ↓	ACC ↑ /SR ↓
COSE-KRE	36.17/19.31	32.08/22.42	46.15/18.66
ECARE-KRE	65.22/34.78	68.05/31.95	82.94/17.06
RACE-CCR	40.58/25.58	39.54/30.42	41.67/20.92

实验结果表明,在LLaMA2-7B-chat模型上,DoLa方法在COSE-KRE与RACE-CCR上出现性

能回退现象(即 ACC 降低、SR 提升)。这种波动主要归因于其层级调控粒度的局限性,由于单层网络内部往往交织着复杂的上下文特征与参数化记忆,粗粒度的层间对比策略难以在抑制先验的同时有效剥离出精准的上下文语义,反而可能因噪声引入而加剧模型对参数化先验的依赖。

相比之下,ACND 均取得最优性能,这一对比有力地支持了在应对复杂的上下文-先验冲突时,在同样利用层间差异的前提下,引入 FFN 内部的神经元粒度调控能更精准定位并抑制冲突先验路径,深入 FFN 内部进行神经元粒度的精准识别与干预,比粗粒度的层级对比策略更为有效,更有效地缓解顽固作答与幻觉倾向。

5.4 泛化能力分析

5.4.1 跨任务泛化性分析

为系统验证 ACND 多任务场景下的适应性与

通用性,我们结合不同任务,不同方法开展评估分析如图 12 所示。实验结果表明,ACND 在 COSE-KRE 和 ECARE-KRE 中准确率分别提升超 12% 和 10%,顽固率同步下降 8% 和 10%。在结构复杂、诱导干扰更强的 RACE-CCR 长文本推理任务中,ACND 依然优于 CAD 与 IRCAN 基线,进一步验证了模型对上下文感知能力与回答可信度,以及对长程语义依赖与知识冲突的鲁棒调控能力。相比之下,IRCAN 在 ECARE-KRE 等任务中出现性能下降,CAD 虽能降低部分顽固率,但在复杂任务中的准确率提升有限。而 ACND 维持性能正向提升,归功于其核心的神经元自适应筛选与神经元自适应对比解码,使其具备卓越的泛化能力与通用性,进一步凸显其在冲突识别、神经层级调控上的鲁棒性,能够作为大语言模型在多任务场景中的通用增强模块。

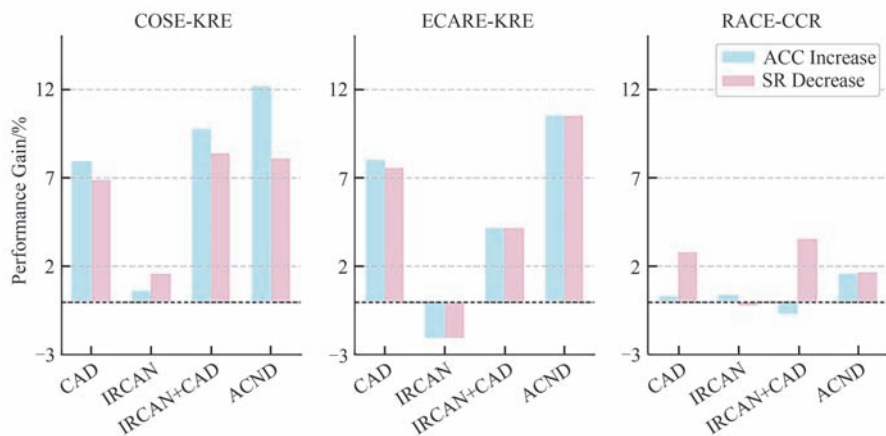


图 12 三种情境推理任务中不同方法的性能增益比较

5.4.2 跨模型泛化性分析

1. 模型规模扩展性分析。随着模型参数量的增加,其内部编码的参数化知识更为庞杂,先验路径也更为根深蒂固,这对上下文调控方法的扩展性提出了更高要求。为了验证 ACND 在不同模型参数规模下的泛化能力,我们在 13B 和 14B 规模的模型上进行与 CAD 和 IRCAN 核心基线的对比实验。

实验结果如图 13 所示,尽管大规模模型因蕴含海量世界知识而表现出更强的“先验惯性”,ACND 仍能有效打破模型对预训练知识的过度固守(表现为顽固率 SR 的下降),保持稳定的性能增益,展现出“高准确、低顽固”协同优化趋势。

在与现有基线方法的横向对比中,ACND 表现出更深层次的调控优势。CAD 主要依赖输出层的概

率调整,而 IRCAN 局限于静态的神经元重加权;相比之下,ACND 融合了神经元级别的动态筛选与自适应对比调节,能够实现更精准的深层知识冲突消解。在长文本、强干扰的复杂任务(RACE-CCR)中优势尤为明显,面对此类高噪声场景,CAD 和 IRCAN 的干预效果易受限,引发局部性能波动;ACND 能够凭借对知识冲突强度的动态感知,将模型性能稳稳锚定在更优区间。这也验证了该框架在应对复杂知识冲突时,具备跨规模泛化能力与微观调控鲁棒性,在不同模型规模下表现出稳定的性能提升趋势。

上述结果进一步表明,ACND 的泛化增益并非源于对输出层概率的简单静态调整,而是源于对模型内部表征通路与解码决策的深层协同调控:一方面,ACNS 通过归因机制定位由上下文驱动且对预

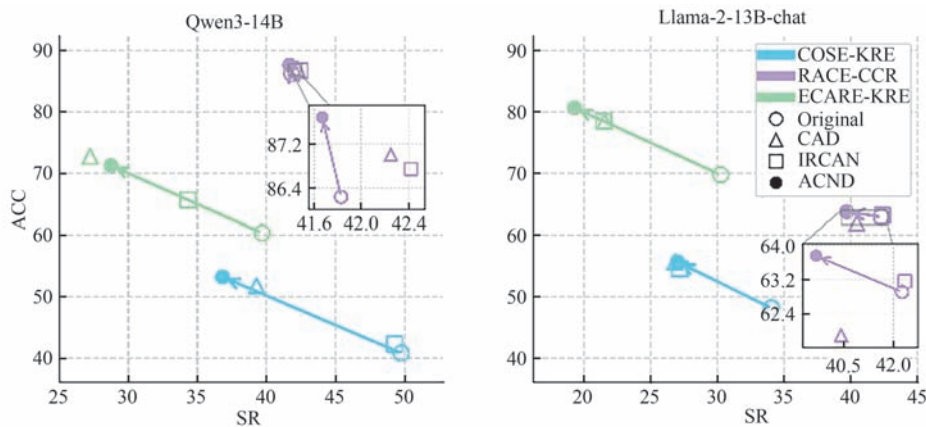


图13 模型泛化性能对比

测最敏感的关键FFN神经元,并施加定向激活增益,从表征层面强化上下文相关特征;另一方面,NACD基于JSD度量上下文分布与先验分布的冲突强度,自适应调节对比强度,在logits空间实现“放大上下文得分、抑制先验得分”的动态平衡。即使在参数规模更大的模型中,ACND仍能有效触发这一机制,展现出跨规模的稳定泛化能力。

2. 为验证ACND对不同开源生态与模型架构谱系的适用性,我们在Gemma系列模型^[42](Gemma-

7b-it、Gemma-2-2b-it)上进行验证。

实验结果如图14所示,ACND在两个Gemma模型上都持续提升性能。例如,在ECARE-KRE数据集上,ACND将Gemma-2-2b-it的准确率提高了7.07%,同时将其顽固率(SR)降低了7.07%。对于更大的Gemma-7b-it模型,ACND带来了3.4%的准确率提升和3.2%的顽固率降低。而在RACE-CCR复杂长文本场景中,不同模型的收益幅度存在差异,仍受模型规模与任务噪声共同影响。

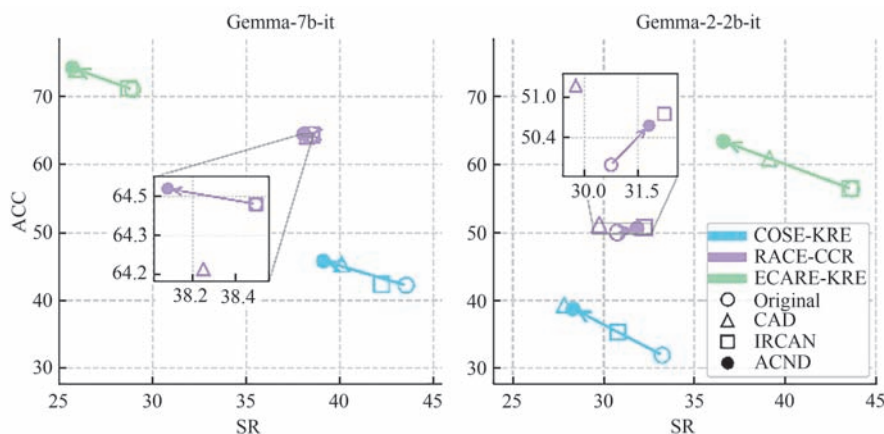


图14 Gemma系列模型上的泛化结果

尽管Gemma的架构权重分布与LLaMA存在显著差异,但ACND的神经元级归因和对比解码机制成功地识别并增强了上下文感知路径。这也证实了ACND不仅限于特定的模型系列,而是适用于主流的Dense Decoder-only Transformer模型。

5.5 局限性与异常结果分析

尽管ACND在多数模型与任务中展现出稳定的性能收益,但实验数据中存在的局部性能波动(见表5及图14)为界定方法的适用边界提供了重要参考。特别是在RACE-CCR这类长文本复杂情境推

理任务中,实现准确推理的关键在于模型需要在长文本中整合多个非相邻证据片段,并排除带有细粒度语义陷阱的干扰项,因而对全局语义建模和跨句推理稳定性提出了更高要求。

(1)小规模模型在长文本强干扰下的表征稳定性瓶颈。在RACE-CCR任务上,小参数规模模型(Qwen3-4B与Gemma-2-2b-it)应用ACND后呈现出非典型的局部波动现象。具体而言,Qwen3-4B相较Original的ACC由77.50%降至74.00%,虽然SR由45.25%降至42.08%,但这种对先验知识的

抑制并未能转化为准确率的提升;类似地,如图14局部放大图所示,Gemma-2-2b-it则表现出另一种异常,其ACC虽由50.00%微升至50.58%,但SR却出现了反常的上升(由30.75%升至31.83%)。结合任务特性分析,这种“去顽固化未转化为准确率”或“顽固率局部反弹”的现象,反映出小规模模型在极端语境下的表征容量受限问题。一方面,2B/4B级别的模型在长文本强干扰场景下的表征容量和层级解耦能力相对有限,上下文感知路径与参数化记忆路径尚不能充分分离;另一方面,ACND对上下文感知神经元的增强虽然提升了模型对上下文的敏感性,但在部分高噪声样本中也可能同步放大局部误导线索,这会导致小规模模型在最终逻辑整合时出现校准偏差,从而削弱全局推理的稳定性。

(2)复杂任务中模型表现的差异性与边界界定。需要指出的是,RACE-CCR并非ACND的普遍失效场景。例如,在参数规模更大的Baichuan2-7B上,ACND的ACC由55.92%提升至58.58%,同时SR由62.00%降至54.00%。在13B和14B规模的模型上同样保持了稳健的性能,这说明长文本复杂推理任务本身并不必然导致ACND失效,其表现更可能取决于模型结构、参数规模、内部表征稳定性与任务噪声之间的交互作用。因此,ACND的适用边界不应被简单概括为在RACE-CCR上性能下降,而应更准确地理解为:在小规模模型、高干扰长文本、强局部误导线索的相互耦合条件下,ACND可能会导致模型生成稳定性出现局部波动。

(3)基线堆叠策略的失效与过度校正。在使用IRCAN-CAD联合基线的Qwen3-4B/RACE-CCR组合上,其ACC由77.50%大幅降至60.25%。该异常更可能源于静态神经元重加权与固定强度对比解码的简单叠加,一方面,神经元增强可能放大了误导性上下文片段;另一方面,固定强度的对比解码又可能进一步压制原本有助于推理的参数化先验,在长文本高干扰场景下导致logits校准失稳,最终造成了输出分布的过度校正。该结果也从侧面说明,相较于将IRCAN与CAD进行模块堆叠,ACND采用统一框架下的动态神经元筛选、神经元增强以及JSD驱动的自适应对比调节,更有利于在不同冲突强度下维持上下文知识与参数化知识之间的平衡,在复杂场景下表现出更好的稳定性。

总体而言,上述异常结果表明,ACND在多数场景下具有稳定性能收益,但在小规模模型面对高

干扰长文本与强误导线索时存在局部性能波动。后续工作将进一步考虑面向长上下文任务的分层冲突检测机制,以进一步提升方法的稳健性与适用边界。

6 结 论

本文提出了ACND框架,旨在缓解LLMs在生成过程中对静态参数化知识的过度依赖问题,系统提升其对输入上下文的忠实度与跨任务的泛化能力。我们在多种模型和任务上的大量实验表明,ACND打破模型对预训练参数知识的过度依赖,实现输入上下文与参数化知识之间的自适应平衡,从而有效缓解知识冲突,为LLMs的动态知识利用提供了通用且可扩展的技术解决方案。

局限性:受算力资源限制,本研究对大规模模型的实验仅在13B/14B参数规模上进行了初步验证,尚未验证ACND在开放域问答及检索增强生成(RAG)中的泛化能力。考虑到ACND具备上下文感知的动态调控能力,其有望提升模型对检索文档的响应忠实度与语义整合能力,从而改善RAG系统中的生成准确性。

对于多模态场景(如图文问答)、跨语言任务以及长上下文的多轮对话,ACND仍可能具有进一步的应用空间。其一,在多模态与跨语言表征中,同样存在对上下文信息敏感的关键神经元,原则上可沿用归因分析进行定位,并通过增益调节策略予以强化。其二,在多轮对话场景中,上下文更长且话题更易漂移,信息与知识冲突也更为常见,ACND的动态神经元增益与对比解码机制有望抑制模型对早期对话的过度依赖,提高对最近轮次信息的关注权重,从而增强对最新对话线索的聚焦。上述方向将作为我们下一步工作的重点。

参 考 文 献

- [1] Shi D, You C, Huang J, et al. Corecode: a common sense annotated dialogue dataset with benchmark tasks for Chinese large language models//Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada, 2024: 18952-18960
- [2] Wang C, Liu X, Yue Y, et al. Survey on factuality in large language models: knowledge, retrieval and domain-specificity. ACM Computing Surveys, 2025, 58(1): 13:1-13:37
- [3] Liu C, Jin R, Ren Y, et al. M3KE: a massive multi-level multi-subject knowledge evaluation benchmark for Chinese large language models. arXiv preprint arXiv:2305.10263, 2023

- [4] Chang Y, Wang X, Wang J, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 2024, 15(3): 1-45
- [5] Chan S, Santoro A, Lampinen A, et al. Data distributional properties drive emergent in-context learning in transformers// *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*. New Orleans, USA, 2022: 18878-18891
- [6] Xu R, Qi Z, Guo Z, et al. Knowledge conflicts for LLMs: a survey//*Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Miami, USA, 2024: 8541-8565
- [7] Xie J, Zhang K, Chen J, et al. Adaptive chameleon or stubborn sloth: revealing the behavior of large language models in knowledge conflicts//*Proceedings of the Twelfth International Conference on Learning Representations*. Vienna, Austria, 2024: 12094-12117
- [8] Huang L, Yu W, Ma W, et al. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 2025, 43(2): 42:1-42:55
- [9] Kaddour J, Harris J, Mozes M, et al. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*, 2023
- [10] Ji Z, Lee N, Frieske R, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023, 55(12): 248:1-248:38
- [11] Touvron H, Martin L, Stone K, et al. Llama 2: open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023
- [12] Yang A, Xiao B, Wang B, et al. Baichuan 2: open large-scale language models. *arXiv preprint arXiv:2309.10305*, 2023
- [13] Qwen Team, Yang A, Yang B, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024
- [14] Xue B, Wang W, Wang H, et al. Improving factual consistency for knowledge-grounded dialogue systems via knowledge enhancement and alignment//*Findings of the Association for Computational Linguistics: 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2023: 7829-7844
- [15] Li D, Rawat A S, Zaheer M, et al. Large language models with controllable working memory//*Findings of the Association for Computational Linguistics: Annual Meeting of the Association for Computational Linguistics 2023*. Toronto, Canada, 2023: 1774-1793
- [16] Jang J, Ye S, Yang S, et al. Towards continual knowledge learning of language models//*Proceedings of the Tenth International Conference on Learning Representations*. Virtual, 2022: 18323-18349
- [17] Gekhman Z, Herzig J, Aharoni R, et al. TrueTeacher: learning factual consistency evaluation with large language models// *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore, 2023: 2053-2070
- [18] Zhu C, Rawat A S, Zaheer M, et al. Modifying memories in transformer models. *arXiv preprint arXiv:2012.00363*, 2020
- [19] Han Z, Gao C, Liu J, et al. Parameter-efficient fine-tuning for large models: a comprehensive survey. *arXiv preprint arXiv:2403.14608*, 2024
- [20] Hu E J, Shen Y, Wallis P, et al. LoRA: low-rank adaptation of large language models//*Proceedings of the Tenth International Conference on Learning Representations*. Virtual, 2022: 12513-12525
- [21] Houshy N, Giurgiu A, Jastrzebski S, et al. Parameter-efficient transfer learning for NLP//*Proceedings of the 36th International Conference on Machine Learning*. Long Beach, USA, 2019: 2790-2799
- [22] Lester B, Al-Rfou R, Constant N. The power of scale for parameter-efficient prompt tuning//*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic, 2021: 3045-3059
- [23] Dai D, Dong L, Hao Y, et al. Knowledge neurons in pretrained transformers//*Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland, 2022: 8493-8502
- [24] Hao Y, Dong L, Wei F, et al. Self-attention attribution: interpreting information interactions inside transformer// *Proceedings of the 35th Conference on Artificial Intelligence of the Association for the Advancement of Artificial Intelligence*. Virtual, 2021: 12963-12971
- [25] Chuang Y-S, Xie Y, Luo H, et al. DoLa: decoding by contrasting layers improves factuality in large language models// *Proceedings of the Twelfth International Conference on Learning Representations*. Vienna, Austria, 2024: 4841-4866
- [26] Shi D, Jin R, Shen T, et al. IRCAN: mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons//*Proceedings of the 38th Annual Conference on Neural Information Processing Systems*. Vancouver, Canada, 2024: 4997-5024
- [27] Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks//*Proceedings of the 34th International Conference on Machine Learning*. Sydney, Australia, 2017: 3319-3328
- [28] Longpre S, Perisetla K, Chen A, et al. Entity-based knowledge conflicts in question answering//*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Punta Cana, Dominican Republic, 2021: 7052-7063
- [29] Chen H-T, Zhang M, Choi E. Rich knowledge sources bring complex knowledge conflicts: recalibrating models to reflect conflicting evidence//*Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates, 2022: 2292-2307
- [30] Zhou Y, Zhou C, Zhang J, et al. ALW: adaptive layer-wise contrastive decoding enhancing reasoning ability in large language models//*Findings of the Association for Computational Linguistics: Annual Meeting of the Association for Computational Linguistics 2025*. Vienna, Austria, 2025: 8506-8524
- [31] Shi W, Han X, Lewis M, et al. Trusting your evidence: hallucinate less with context-aware decoding//*Proceedings of*

- the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers). Mexico City, Mexico, 2024: 783-791
- [32] Li X L, Holtzman A, Fried D, et al. Contrastive decoding: open-ended text generation as optimization//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada, 2023: 12286-12312
- [33] Tenney I, Das D, Pavlick E. BERT rediscovers the classical NLP pipeline//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4593-4601
- [34] Lai G, Xie Q, Liu H, et al. RACE: large-scale reading comprehension dataset from examinations//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 785-794
- [35] Ying J, Cao Y, Xiong K, et al. Intuitive or dependent? investigating LLMs' behavior style to conflicting prompts//Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Bangkok, Thailand, 2024: 4221-4246
- [36] Aggarwal S, Mandowara D, Agrawal V, et al. Explanations for CommonsenseQA: new dataset and models//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 3050-3065
- [37] Du L, Ding X, Xiong K, et al. e-CARE: a new dataset for exploring explainable causal reasoning//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 432-446
- [38] Hendrycks D, Burns C, Basart S, Zou A, Mazeika M, Song D, et al. Measuring massive multitask language understanding//Proceedings of the Ninth International Conference on Learning Representations. Virtual, 2021: 9804-9830
- [39] Huang Y, Bai Y, Zhu Z, Zhang J, Zhang J, Su T, Liu J, Lv C, Zhang Y, Lei J, Fu Y, Sun M, He J. C-EVAL: a multi-level multi-discipline Chinese evaluation suite for foundation models//Proceedings of the 37th Annual Conference on Neural Information Processing Systems. New Orleans, USA, 2023: 62991-63010
- [40] Li H, Zhang Y, Koto F, Yang Y, Zhao H, Gong Y, Duan N, Baldwin T. CMMLU: measuring massive multitask language understanding in Chinese//Findings of the Association for Computational Linguistics: 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok, Thailand, 2024: 11260-11285
- [41] Wang Y, Ma X, Zhang G, Ni Y, Chandra A, Guo S, Ren W, Arulraj A, He X, et al. MMLU-Pro: a more robust and challenging multi-task language understanding benchmark//Proceedings of the 38th Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2024: 95266-95290
- [42] Gemma Team, Mesnard T, Hardin C, et al. Gemma: open models based on Gemini research and technology. arXiv preprint arXiv:2403.08295, 2024

附录

(1) 大型语言模型微调

为提升大型语言模型(LLMs)在上下文推理任务中的适应性,我们基于开源LLaMA框架采用两阶段微调策略。通过将高效的低秩适应(LoRA)参数调优与标准监督式微调相结合,我们实现了“先适应后泛化”的协同优化。

所有实验均在配备双GPU的NVIDIA A6000(48GB)×2的配置,确保可重复性和可扩展性。具体步骤如下:

第一阶段:LoRA轻量级调优

LoRA在选定层中引入可训练的低秩分解矩阵,显著减少可训练参数数量和内存占用。这使模型能以极低计算开销适应目标任务。在我们的工作中,我们采用标准LoRA技术对基础模型Baichuan2-7B进行参数高效的微调。

训练数据:我们采用从ECARE-KRE、COSE-KRE和RACE-CCR数据集采样的问答实例。这些任务的选取旨在突出因果推理、常识推理及复杂情境推理能力。

LoRA预训练完成后,系统将自动输出LoRA适配器检查点集。这些检查点可直接用于下游评估,或合并用于后续训练。

第二阶段:监督式微调(SFT)

我们将训练好的LoRA适配器与基础模型合并,初始化一个全参数模型。随后继续采用LoRA方法进行监督式微调,以增强模型在特定任务场景中的指令对齐能力与泛化能力。本阶段旨在巩固模型适应性,同时进一步提升其在长上下文和知识冲突场景中的泛化能力。

(2) 训练数据

本阶段采用ECARE-KRE、COSE-KRE及RACE-CCR问答基准中全新精选的高质量指令遵循样本。



GUO Ya-Long, M. S. candidate.

His primary research areas include artificial intelligence, natural language processing, and machine learning.

ZHANG Yang-Sen, Ph. D., professor. His primary research areas include artificial intelligence, natural language

processing, and network content security.

KANG Jing, M. S. candidate, Her primary research interests in artificial intelligence and natural language processing.

WANG Yu-Qi, Ph. D. candidate. Her primary research areas include cross-modal learning and large-scale model generation techniques.

CHEN Quan-Tao, M. S. candidate. His main research interests include natural language processing and network information security.

Background

Large language models (LLMs) have demonstrated remarkable capabilities in natural language understanding, reasoning, text generation, and knowledge representation. Their strong parametric knowledge enables them to perform effectively on a wide range of knowledge-intensive tasks, including open-domain question answering, fact verification, and causal reasoning. However, despite these advances, LLMs remain prone to hallucination, especially when the knowledge encoded in model parameters conflicts with newly provided contextual evidence. This problem becomes particularly critical in dynamic or domain-specific scenarios, where external information may be more recent, more accurate, or more task-relevant than the model's pre-trained knowledge.

A typical source of such hallucination is the conflict between parametric knowledge and contextual knowledge. Parametric knowledge is implicitly stored in model parameters during pre-training, while contextual knowledge is explicitly provided in the input prompt or retrieved documents. When these two knowledge sources are inconsistent, LLMs often over-rely on their internal memory and fail to sufficiently attend to the given context. Prior work such as context-aware decoding (CAD) addresses this issue by contrasting output distributions with and without contextual input, thereby encouraging the model to trust the provided

evidence more strongly.

While prior work explores retrieval-augmented generation (RAG), knowledge graph integration, and fine-tuning to mitigate this issue, these methods typically lack mechanisms to adapt the model's internal decision pathways for context-sensitive reasoning. Recent advances such as CAD and attribution-based neuron reweighting methods (e.g., IRCAN) offer improvements via output-level regulation or static neuron adjustment, yet remain constrained by coarse-grained control and limited adaptability.

In this paper, we propose Adaptive Context-aware Neuron Decoding (ACND), a novel framework that leverages Integrated Gradients and Bayesian optimization to dynamically identify and enhance context-aware neurons during inference. ACND enables fine-grained regulation within feed-forward layers and introduces Neuron-Adaptive Contrastive Decoding, which adjusts generation probabilities via Jensen-Shannon Divergence to mitigate conflicts between parametric knowledge and contextual knowledge. Experimental results demonstrate that ACND achieves an average accuracy improvement of 8.1% and a reduction in stubbornness rate of 6.8% for LLMs, consistently outperforming strong baselines in both accuracy and robustness.

This work was supported by the Beijing Natural Science Foundation (L233008).