软件学报 ISSN 1000-9825, CODEN RUXUEW
[doi: 10.13328/j.cnki.jos.007704] [CSTR: 32375.14.jos.007704]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

大语言模型中逻辑神经元的系统性探测与影响分析*

姜 慧[1,2], 何世柱[1,2], 刘 康[1,2], 赵 军[1,2]

[1](中国科学院大学 人工智能学院, 北京 100049)

[2](复杂系统认知与决策重点实验室 (中国科学院 自动化研究所), 北京 100190)

通信作者: 何世柱, E-mail: shizhu.he@ia.ac.cn

摘 要: 大语言模型 (large language model, LLM) 在涉及逻辑操作的复杂推理任务中表现良好, 然而其逻辑推理能力的内部机理仍缺乏明确的可解释性分析. 目前尚不明确 LLM 内部是否存在与逻辑推理任务表现密切相关的神经元子集, 这些神经元在网络中以何种形式进行计算表征, 它们又如何影响模型在不同任务中的表现. 为探究上述涉及 LLM 推理机制本质的关键问题, 提出一种"识别-干预-评估"框架, 旨在系统性地剖析逻辑推理能力的神经元运行基础与功能实现. 利用基于梯度重要性的神经元定位方法, 定位对逻辑推理任务高度敏感的"逻辑神经元", 并结合激活值进行定向干预, 进而评估其对模型行为的影响. 研究结果表明, 所定位的逻辑相关神经元子集主要集中于网络中层, 并以前馈网络模块 (feed-forward network, FFN) 为主导. 进一步的分析表明, 不同逻辑任务之间既存在共享神经元成分, 也存在任务特异性差异. 基于这些发现, 探索一种无需修改模型权重、轻量级的逻辑推理能力定向增强方法, 实验结果显示, 该方法在 LogicBench 的长链推理任务上将模型平均准确率提升 9.2%, 并在逻辑敏感型通用任务上性能提升 4.57%.

关键词: 大语言模型; 逻辑推理; 可解释性; 神经元干预

中图法分类号: TP18

中文引用格式: 姜慧, 何世柱, 刘康, 赵军. 大语言模型中逻辑神经元的系统性探测与影响分析. 软件学报. http://www.jos.org.cn/1000-9825/7704.htm

英文引用格式: Jiang H, He SZ, Zhao J, Liu K. Systematic Probing and Impact Analysis of Logic Neurons in Large Language Model. Ruan Jian Xue Bao/Journal of Software (in Chinese). http://www.jos.org.cn/1000-9825/7704.htm

Systematic Probing and Impact Analysis of Logic Neurons in Large Language Model

JIANG Hui[1,2], HE Shi-Zhu[1,2], LIU Kang[1,2], ZHAO Jun[1,2]

[1](School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China)

[2](State Key Laboratory of Management and Control for Complex Systems (Institute of Automation, Chinese Academy of Sciences), Beijing 100190, China)

Abstract: Large language models (LLMs) perform well in complex reasoning tasks involving logical operations. However, the internal mechanisms underlying their logical reasoning capabilities remain insufficiently interpretable. Currently, it remains unclear whether neuron subsets closely associated with logical reasoning performance exist within LLMs, what forms of computational operations are performed by these neurons in the network, and how they influence model performance across different tasks. To investigate these key questions concerning the nature of LLM reasoning mechanisms, this study proposes an "identification-intervention-evaluation" framework aimed at systematically analyzing the neuronal operational basis and functional implementation of logical reasoning. A neuron localization method based on gradient importance is used to identify "logic neurons" that are highly sensitive to logical reasoning tasks, and targeted interventions are performed on their activation values to further evaluate their impact on model behavior. The results show that the identified logic-related neuron subsets are mainly concentrated in the middle layers of the network and predominantly located in feed-

* 基金项目: 国家自然科学基金 (62376270); 北京市新一代信息通信技术创新专项 (Z251100008125025)
收稿时间: 2026-02-10; 修改时间: 2026-04-29; 采用时间: 2026-05-18; jos 在线出版时间: 2026-08-12

forward network (FFN) modules. Further analysis indicates that different logical tasks involve both shared neuron components and task-specific differences. Based on these findings, this study explores a lightweight targeted enhancement method for logical reasoning capabilities without modifying model weights. Experimental results show that this method improves the average accuracy of the model on long-chain reasoning tasks in LogicBench by 9.2% and improves performance on logic-sensitive general tasks by 4.57%.

Key words: large language model (LLM); logical reasoning; interpretability; neuron intervention

从亚里士多德的三段论到现代数理逻辑, 逻辑推理始终被视为一类关键的人类高级认知能力[1,2], 它使我们能够超越直觉[3], 通过严谨的推演法则从有限的已知信息中推导出新结论, 从而解决从数学证明到社会决策的复杂问题[2,4]. 近年来, 大语言模型 (large language model, LLM) 的爆发式增长正在重塑我们对智能的理解, 在基于大规模文本数据和任务监督数据的学习下, 这些模型在数学证明、代码生成及常识推理等各类复杂任务中表现出了惊人的涌现能力[5,6], 其推理水平在某些特定领域已达到甚至超越了人类基准[7]. 然而, 模型的表现与其运行机理之间存在着巨大的"解释性鸿沟", 尽管 LLM 能够生成良好逻辑思维过程 (如思维链[8]) 和答案, 其内部决策过程仍如同一个不知其所以然的"黑盒"[9]. 也就是说, 一个根本性的科学问题仍不明晰: 模型是真正掌握了逻辑推理的本质能力, 还只是对人类语言推理表象的高级模仿? 我们亟需探究, 逻辑推理这种在人类大脑中由特定皮层网络支持的高阶能力[10], 在人工神经网络的海量参数空间中, 究竟是以何种神经动力学形式存在?

现有研究主要通过构建并依托 LogicBench[11]、ReClor[12]等基准数据集, 从行为主义视角对 LLM 的逻辑推理能力进行定量评估[13–19]. 目前, 特别是在推理大模型支撑下, LLM 处理复杂逻辑任务时, 已不再局限于输出最终答案, 而是能够生成详尽的中间推理步骤, 这种显式的思维链极大地增强了输出结果的可信度[8,20]. 如图 1 所示, 模型可以通过分析 LLM 能否很好完成演绎推理、归纳推理等任务量化其相应逻辑推理能力.

注：译自LogicBench部分案例

图 1 LogicBench 数据集中的典型逻辑推理任务示例

尽管模型能够识别前提并生成遵循规则的中间推理步骤, 并且输出了正确的答案 (例如, 图 1 右侧关于"克洛伊备考"和"米拉度假"的案例), 但是我们仍难以确定模型在底层是如何处理析取 (∨) 和否定 (¬) 计算的. 这种推理步骤与内部机制的不一致性 (即推理的"忠实性"问题), 是影响 LLM 可靠性不足的主要障碍. 尽管现有的机制可解释性研究[21]已在神经元层面取得显著进展, 成功定位了负责事实知识[22–24]、语法结构[25]、特定语言属性[26,27]及特定技能[28,29]等功能神经元. 然而, 对于逻辑推理这类高度抽象的高阶认知能力, 其神经元层面的表征机制目前仍缺乏系统性探究. 这一研究空白不仅限制了我们对模型能力边界的深层认知, 也阻碍了构建更加可解释、可控的鲁棒推理系统.

针对上述问题, 本研究通过一种"识别-干预-评估"的框架, 旨在揭示 LLM 中逻辑推理能力的微观神经元基础. 我们利用基于梯度重要性的方法[29]来定位对形式逻辑高度敏感的"逻辑神经元", 并采用激活干预技术[30]来提供

因果相关证据. 实验证明, 通过定向增强这些逻辑神经元的激活, 模型在 LogicBench 长链推理任务上的平均准确率可提升 9.2%, 在 SNLI[31]、BoolQ[32]、AG News[33]、CoLA[34]等通用能力任务上性能平均提升 4.57%; 反之, 抑制其激活则导致模型在 LogicBench 上的性能大幅下降 34.65%, 而在上述通用任务上平均下降 7.01%. 这一显著的性能差异为所定位神经元子集对模型逻辑任务表现具有因果影响提供了实验证据. 基于上述初步发现, 我们致力于回答以下核心研究问题: 1) 逻辑神经元存在性与可定位性问题: 是否存在对逻辑推理操作高度敏感、并对逻辑任务表现具有显著影响的神经元子集? 我们能否通过分析方法识别出对特定逻辑推理任务敏感的“逻辑神经元”? 2) 逻辑神经元在模型架构的分布问题: 这些逻辑神经元在模型中的分布模式如何? 它们是否集中于特定层或模块 (如注意力或前馈网络)? 不同类型的逻辑推理 (如命题逻辑与一阶逻辑) 之间是否存在神经元共享或分离的现象? 3) 对相关和不相关任务的影响问题: 这些逻辑神经元对模型整体能力有何影响? 它们是否仅影响逻辑任务, 是否对其他下游任务也产生影响?

为解答上述问题, 我们首先使用一种基于梯度重要性的方法, 接着采用激活缩放干预技术, 评估其对模型在逻辑推理及其他非逻辑任务上表现的影响. 本文研究以 Qwen3-8B 为核心基础, 并扩展至广泛的模型谱系以验证发现的一致性. 实验对象系统性地涵盖了通用指令模型 (如 Llama-2-7B、Mistral-7B)、推理与数学专用模型 (如 DeepSeek-R1-Distill、Qwen2.5-math) 以及 Qwen3 系列的多个不同规模 (0.6B–14B). 这种跨架构、跨功能类型及跨参数规模的对比分析, 有助于检验所观察现象的系统性与一定程度的泛化性. 总的来说, 本研究的主要贡献如下.

(1) 逻辑神经元存在性分析: 系统性地识别 LLM 中的“逻辑神经元”, 并通过激活干预实验提供了其对模型逻辑推理表现具有因果影响的实验证据.

(2) 逻辑神经元分布分析: 观察到逻辑神经元的独特分布模式, 包括在中层峰值、FFN 模块主导以及特定注意力头的专一性等特征.

(3) 逻辑推理能力表征机制: 分析不同逻辑类型之间的神经元关联性, 揭示模型通过共享与独立的神经元群协同处理多样的推理任务.

(4) 轻量级逻辑推理能力定向增强: 提出一种无需修改模型权重、轻量级的定向能力增强方法, 在多个代表性逻辑推理任务上带来一定性能增益.

1 相关工作

1.1 大语言模型逻辑推理能力分析与评测

随着思维链 (chain-of-thought, CoT) 提示技术的提出, 大语言模型在多步推理任务上展现出显著性能[8,35]. 随后的研究进一步发现, 即使在零样本设置下, 通过特定的提示引导亦可有效激活模型潜在的推理能力[20]. 为量化这一能力, 学术界提出了多种形式化基准, 例如, ReClor 数据集通过阅读理解任务评估模型的逻辑推断能力[12], 而 LogicBench[11]则进一步从一阶逻辑、非单调逻辑等细粒度维度对模型进行了系统性测试.

然而, 现有研究主要通过分析模型输出的文本 (如推理步骤) 来推断其逻辑能力, 这种“行为主义”视角的局限性在于无法解释模型是否真正内化了逻辑规则, 还是仅基于语料统计概率的浅层拟合[9]. 此外, 这种基于外部行为的评测难以解释模型在长链推理中出现的“逻辑幻觉”与组合性泛化失败的问题[36,37].

1.2 神经元层面的大模型可解释性研究

可解释性研究旨在揭示“功能-表征”的映射与定位机制, 其核心成果是证明了 LLM 内部特定功能可被定位到特定神经元子集. 早期的理论工作为理解 Transformer 内部运作提供了理论基础, Elhage 等人[21]和 Ameisen 等人[38]提出了针对 Transformer 回路的数学框架. 在此基础上, 前馈网络 (feed-forward network, FFN) 被普遍认为是知识存储的关键组件, Geva 等人[22]将其运作机制解释为“键-值记忆网络”; Dai 等人[23]和 Chen 等人[39]进一步提出“知识神经元”概念, 证明了特定事实知识可被定位; Meng 等人[24,40–42]基于此开发了 ROME 和 MEMIT 算法, 通过定位负责存储事实关联的特定区域来实现知识编辑. 除静态知识外, Dalvi 等人[25]和 Belinkov 等人[26]分析了单个神经元对词法、句法等语言学特征的编码; Radford 等人[27]和 Tak 等人[43]发现了控制情感的“情感神经元”; 而 Song 等

人[29]探讨了负责特定下游任务的“任务神经元”.

随着推理大模型的兴起, 研究焦点逐渐转向动态推理过程. Zhang 等人[44]的综述指出 LLM 从直觉思维 (System 1) 向逻辑推理 (System 2) 的能力跃迁. 在具体机制上, Chi 等人[45]探究了因果推理的内在机制; Kim 等人[46]初步描绘了负责三段论推理的“推理电路”; Wang 等人[47]分析了微调对推理机制的影响. 尽管上述工作在宏观或中观层面取得了进展, 但针对逻辑原子操作 (如 AND/OR/NOT) 的神经元级表征仍缺乏系统性研究, 这正是本研究的目标.

1.3 模型干预与能力编辑

基于特定功能神经元的定位成果, 如何通过干预手段来主动控制模型行为已成为新的研究热点. 现有的干预范式主要分为“模型编辑”与“激活导向”两类. 以 ROME[24]和 MEMIT[40]为代表的模型编辑方法, 旨在通过永久性地修改模型权重来更新特定事实知识, 然而, 此类方法主要针对静态的陈述性知识, 难以适用于逻辑推理这种内化于架构的通用能力, 且容易引发灾难性遗忘[48–50]. 与之相比, 推理时干预技术提供了一种轻量级、非破坏性的路径, 即在不改变权重的前提下, 通过在推理过程中动态调整特定神经元的激活值来引导模型输出[51]. Zou 等人[30]提出的“表示工程”展示了通过控制激活空间来调节模型诚实性、情感等行为的可能性.

随着研究的深入, 激活导向技术逐渐向更复杂的场景拓展. 例如, Nguyen 等人[52]引入反馈控制器来实现更精细的激活引导; Wang 等人[53]提出了 InferAligner 以实现推理时的无害化对齐. 特别值得注意的是, Valentino 等人[54]利用细粒度的激活导向来缓解模型推理中的内容偏差. 尽管上述工作证实了干预的有效性, 但其应用场景多局限于纠正事实错误、安全对齐或缓解特定偏差. 本研究将干预的靶点由静态知识或偏见拓展至“逻辑原子操作”, 通过定向激活特定的神经元子集, 实现对模型逻辑推理能力的因果性增强.

2 方 法

2.1 逻辑神经元识别框架

本研究提出了一种融合机制可解释性中因果定位思想的“识别-干预-评估”框架. 如后文图 2 所示, 该框架旨在解构大语言模型内部复杂的逻辑推理过程, 将其从抽象的认知能力映射至具体的神经元子集. 通过构建“定位表征-因果验证-多维评估”的闭环, 在保持模型权重冻结的前提下, 利用轻量级激活干预技术提供特定神经元集合对逻辑推理的因果相关证据. 该框架包含如下 3 个模块.

(1) 识别模块 (Identification): 利用梯度敏感度分析筛选出对逻辑决策具有高贡献度的“逻辑神经元”集合 ($\mathcal{M}^{\text{logic}}$);

(2) 干预模块 (Intervention): 在推理阶段引入条件干预函数, 针对目标神经元实施激活值缩放 (增强或抑制), 以动态调控模型的推理路径;

(3) 评估模块 (Evaluation): 构建包含复杂逻辑、原子逻辑及通用任务的多层级任务谱系, 从因果性验证、机制分布与分析、跨任务影响这 3 个维度量化干预效果.

2.2 任务构建与逻辑敏感度层级

为确保机制分析的普适性和严谨性, 本研究对任务构建进行了标准化处理. 核心目标是将评估统一在零样本推理框架下, 以最大程度消除提示工程或不同解码策略引入的混淆变量, 确保观测到的因果效应源于模型内部机制而非外部干扰.

2.2.1 任务分类与敏感度梯度

为了系统地解构模型的推理能力, 本文构建了一个包含 3 个层级的任务体系, 这些任务在对逻辑推理能力的依赖程度上构成了清晰的敏感度梯度.

(1) 复杂逻辑任务 (高敏感度): 此类任务聚焦于复杂形式逻辑的规则内化与应用. 任务涵盖一阶逻辑 (first-order logic)、命题逻辑 (propositional logic) 与非单调逻辑. 模型需在无上下文示例的情况下, 执行多步长链推理 (chain-of-reasoning) 才能得出正确结论, 任何逻辑链路的断裂都将导致预测失败.

(2) 基础逻辑原子任务 (中敏感度): 此类任务专注于最基本的逻辑算子 (如与、或、非) 运算. 其设计目的在于探究神经网络是否在微观层面形成了类似“逻辑门”的模块化表征, 以及最小逻辑单元在参数空间中的可定位性.

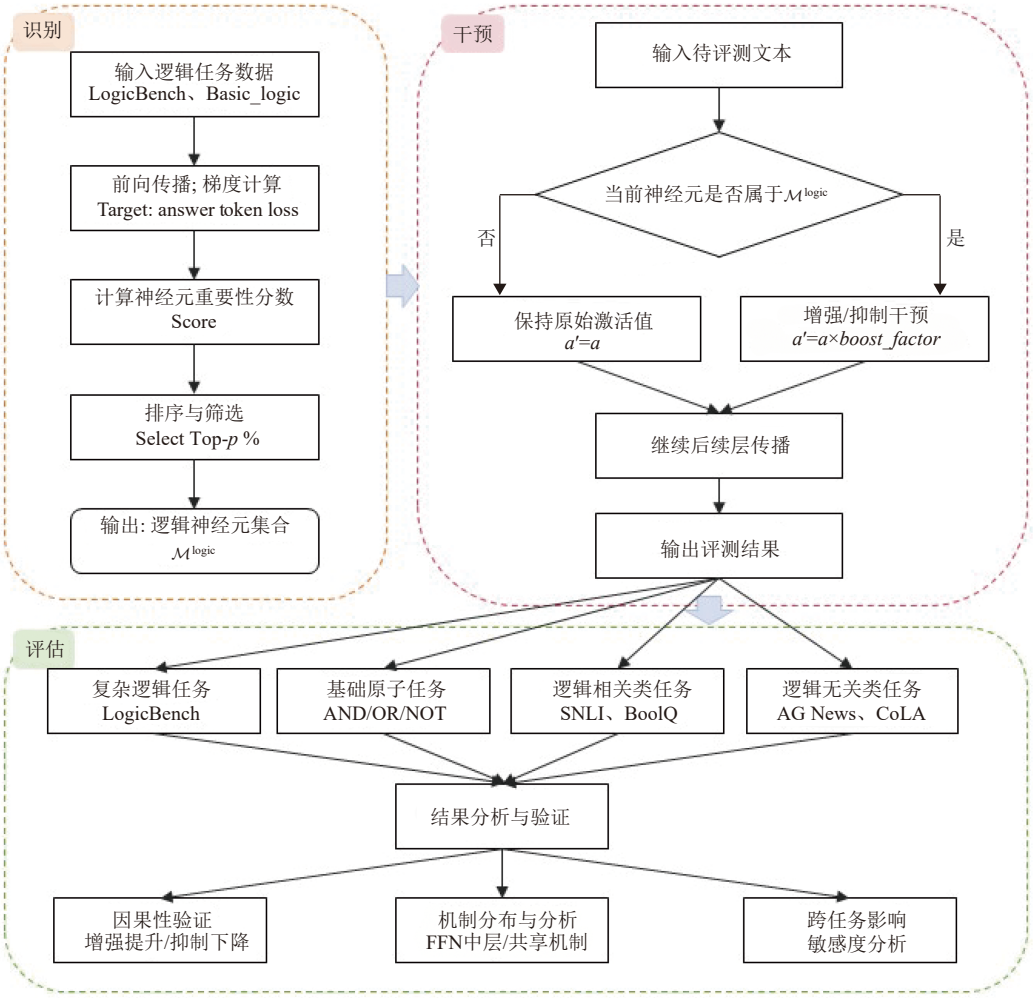


图 2 “识别-干预-评估”闭环框架示意图

(3) 通用能力任务 (低至中敏感度): 作为对照组, 此类任务包含逻辑弱相关 (如情感分析、主题分类) 或逻辑无关的自然语言理解任务. 引入该层级旨在评估逻辑神经元干预对模型通用认知能力的跨任务影响范围, 验证机制的特异性.

2.2.2 数据集转换与统一格式

为消除不同数据源格式差异带来的分布漂移 (distribution shift), 上述所有层级的任务数据集均被转换为统一的输入格式.

(1) 对话格式标准化: 所有原始数据集均被重构为标准的对话消息列表格式 (chat format), 包含明确的 system 和 user 角色定义, 以模拟真实交互场景下的模型行为.

(2) 零样本约束: 推理过程严格遵循零样本协议, 使用固定提示模板引导模型生成. 本文强制模型以“Answer:”作为生成的开始标记, 并仅回答预设决策标签 (如 Yes/No 或 True/False). 这种约束性回答策略将模型的计算资源聚焦于决策生成的收敛点, 从而使梯度计算能够捕捉到决定最终判断的关键神经元, 而非负责通用文本生成的神经元.

2.3 逻辑神经元的梯度定位

在确立了任务体系后, 本阶段的核心挑战在于从庞大的参数空间中筛选出负责逻辑推理的神经元子集. 为实现系统且精准的定位, 我们采用了基于梯度重要性 (gradient importance) 的度量方法[23,29].

我们将“逻辑神经元”定义为对模型最终预测结果具有显著因果效应的功能子集. 为了量化这种贡献, 我们计算模型在目标逻辑任务 $\mathcal{D}^{\text{logic}}$ 上的预测损失 $\mathcal{L}$ 对每一个神经元激活值 $a_{l,j}$ 的敏感度. 具体而言, 对于第 l 层第 j 个神经元, 其重要性分数 $I_{l,j}^{\text{logic}}$ 定义为预测损失 $\mathcal{L}$ 对该神经元激活值 $a_{l,j}$ 的梯度 L_1 范数的期望.

为捕捉与推理决策相关的神经元, 我们仅在模型输出“Answer:”标记附近的少数关键位置计算梯度. 该重要性度量过程可以正式表述为:

$$I_{l,j}^{\text{logic}} = \mathrm{E}_{(x,y)\sim\mathcal{D}^{\text{logic}}}\left[\left\|\frac{\partial \mathcal{L}(y, LLM(x,t^*))}{\partial a_{l,j}(t^*)}\right\|_1\right] \tag{1}$$

其中, $\mathcal{L}$ 为衡量预测与真实标签 y 差异的交叉熵损失函数; $\mathrm{E}_{(x,y)\sim\mathcal{D}^{\text{logic}}}$ 表示在逻辑任务数据集上的数学期望, 用于消除单一样本偏差; $\frac{\partial \mathcal{L}}{\partial a_{l,j}}$ 表示损失函数对神经元激活值的偏导数, 它量化了该神经元激活值的微小变化对最终预测误差的影响程度 (即梯度贡献); $\|\cdot\|_1$ 表示 L_1 范数, 在可解释性研究中常用于聚合梯度信号以得到解耦的标量分数; t^* 表示模型输出“Answer:”标记后的关键决策位置 (即首个生成 token). 这一位置选择至关重要, 通过将梯度计算聚焦在决策收敛点, 我们能够有效剔除与逻辑推理非直接相关的上下文语义编码或通用语言生成神经元, 有助于提高所定位子集与逻辑判断之间的相关性和任务特异性. 此外, 为确定最终决策位置可能带来的阶段性偏差, 本文在实验部分进一步补充 reasoning-path 定位对照分析

在计算得到所有神经元的全局重要性分数后, 我们依据分数分布构建目标逻辑神经元集合. 为了平衡干预的有效性与稀疏性, 我们设定筛选阈值 τ_p (对应 Top-p% 的分位数), 将逻辑神经元集合 $\mathcal{M}^{\text{logic}}$ 形式化定义为:

$$\mathcal{M}^{\text{logic}} = \left\{(l,j) | I_{l,j}^{\text{logic}} \geqslant \tau_p\right\} \tag{2}$$

其中, (l, j) 代表第 l 层第 j 个神经元的索引坐标. 该集合即为后续激活干预操作的唯一靶点.

2.4 激活干预与因果验证

在定位到逻辑神经元集合 $\mathcal{M}^{\text{logic}}$ 后, 本研究采用推理时干预 (inference-time intervention, ITI)[51]来验证其因果效应. 这是一种轻量级的激活编辑技术, 旨在不改变模型权重的参数冻结状态下, 动态调控特定神经元的功能表达. 通过构建条件干预函数, 在前向传播过程中对流经 FFN 层的信息流进行实时重构, 对于第 l 层第 j 个神经元, 设其原始激活值为 $a_{l,j}$, 干预后的激活值 $a'_{l,j}$ 由以下分段函数定义:

$$a'_{l,j} = \mathcal{I}(a_{l,j}, \mathcal{M}^{\text{logic}}) = \begin{cases} a_{l,j}\cdot\alpha, & \text{if}(l,j)\in\mathcal{M}^{\text{logic}} \\ a_{l,j}, & \text{otherwise} \end{cases} \tag{3}$$

其中, α 为干预强度系数. 基于该系数, 我们建立了双向因果验证: 当 α >1.0 时执行增强干预 (enhanced), 以验证该神经元集合是否足以提升模型的逻辑能力; 当 α <1.0 时执行抑制干预 (suppressed), 旨在验证该集合对于维持模型基本推理功能的必要性.

相比于 ROME[24]或 MEMIT[40]等需要修改模型权重的传统知识编辑方法, 这种干预机制具备动态可逆性和功能解耦性, 干预仅在推理时短暂生效, 不永久性改变模型的底层知识库和通用语言能力, 有效隔离高阶能力 (如逻辑推理) 与底层知识表征 (如事实记忆), 有助于降低权重编辑带来的额外混杂因素.

3 实验设置

3.1 数据集与任务

基于第 2.2 节定义的层级框架, 本研究在实验中使用了包含如下 3 个敏感度梯度的数据集 (统计信息见表 1).

(1) 复杂逻辑任务: 采用 LogicBench[11]的细粒度子集, 涵盖一阶逻辑、命题逻辑及非单调逻辑推理, 用于评估模型对复杂规则的内化能力.

(2) 基础逻辑原子任务: 构建了定制化的 Basic_logic 数据集, 专门针对布尔运算 (AND/OR/NOT) 进行原子级测试, 以定位最小逻辑计算单元.

表 1　数据与任务概览

名称	类型	规模	划分	标签空间	指标
LogicBench	细粒度逻辑推理 (命题/一阶/非单调)	多子集	官方划分	随子集而定	Accuracy
Basic_logic	基础命题逻辑 (AND/OR/NOT)	约0.8k	项目内标准划分	{True, False}	
BoolQ	是/否阅读理解 (SuperGLUE)	约15.9k	train≈9.4k; dev≈3.3k; test≈3.2k	{Yes, No}	
SNLI	自然语言蕴含三分类	约570k	train≈550k; dev=10k; test=10k	{Entailment, Neutral, Contradiction}	
AG News	新闻主题四分类	约127.6k	train=120k; test=7.6k	{World, Sports, Business, Sci/Tech}	
CoLA	语法可接受性二分类	约10.5k	train≈8.5k; dev≈1.0k; test≈1.0k	{Acceptable, Unacceptable}	MCC、Accuracy

(3) 通用能力任务: 选取 SNLI[31]、BoolQ[32]作为逻辑相关对照, 选取 AG News[33]、CoLA[34]作为逻辑无关对照, 以评估模型的跨任务迁移能力.

3.2　模型选择

为了验证机制的跨架构一致性, 我们选取了参数规模相近 (7B–8B), 但架构与对齐策略各异的 3 款主流模型作为主要研究对象: (1) Qwen3-8B: 作为基础预训练模型代表. (2) Llama-2-7B-Chat: 作为经过人类反馈强化学习 (reinforcement learning from human feedback, RLHF) 对齐的对话模型代表. (3) Mistral-7B-Instruct-v0.3: 作为应用滑动窗口注意力 (sliding window attention) 机制的指令微调模型代表. 此外, 为探究参数规模的影响, 实验对象扩展至 Qwen 系列和 DeepSeek 系列的其他尺寸变体 (详见表 2).

表 2　模型规格与参数概览

名称	参数量 (B)	上下文长度	词表大小	检查点标识
Llama-2-7B-Chat	7	4 096	32 000	meta-llama/Llama-2-7B-chat-hf
Qwen3-8B	8	32 768	151 936	Qwen/Qwen3-8B
Mistral-7B-Instruct-v0.3	7	32 768	32 000	mistralai/Mistral-7B-Instruct-v0.3
Qwen3-0.6B	0.6	32 768	151 936	Qwen/Qwen3-0.6B
Qwen3-1.7B	1.7	32 768	151 936	Qwen/Qwen3-1.7B
Qwen3-4B	4	32 768	151 936	Qwen/Qwen3-4B
Qwen3-14B	14	32 768	151 936	Qwen/Qwen3-14B
Qwen2.5-math-7B	7	32 768	151 936	Qwen/Qwen2.5-math-7B
DeepSeek-R1-Distill-Qwen-7B	7	32 768	151 936	deepseek-ai/DeepSeek-R1-Distill-Qwen-7B

3.3　实验配置与评估协议

3.3.1　干预参数设置

干预机制基于第 2.4 节提出的推理时干预 (ITI) 范式实施. 干预靶点默认作用于前馈神经网络 (FFN) 模块 up_proj 层 (即 FFN 的上投影层), 并可针对特定层级进行限制. 干预由两个核心超参数控制: 干预强度 (α) 和神经元稀疏度 (Top-p%) 控制. 本文制定了系统的网格搜索 (grid search) 以确定最优操作区间: (1) 干预强度 (α): 在 [0.0, 4.0] 的范围内进行步进式搜索, 以探究增强干预 (α>1) 与抑制干预 (α<1) 的性能变化. (2) 神经元稀疏度 (τ_p): 对 (Top-p%) 阈值 (范围 5%–25%) 进行细致扫描, 旨在寻找最小且有效的因果神经元集合. 具体的参数网格搜索范围与默认值详见表 3.

表 3　干预超参数及网格范围

稀疏度 (τ_p)	层段	模块	α	重复次数
0.05、0.10、0.15、0.20、0.25	all	FFN (up_proj)	{0, 0.3, 0.5, 0.7, 0.9, 1.3, 1.5, 1.7, 1.9, 2.0, 4.0}	3

3.3.2 评估协议与实现细节

本研究在统一的零样本框架下执行所有评估, 具体的实施协议如下.

(1) 任务评估指标: 针对逻辑推理任务, 采用准确率 (Accuracy) 作为核心指标. 针对通用任务, 遵循现有文献建立的标准评估协议 (如 Matched Accuracy 或 $F1$-score).

(2) 统计可靠性: 为确保结果的统计可靠性, 每个实验设置均在不同随机种子下独立重复 3 次, 我们报告均值和标准差, 并通过配对 t-检验验证干预效果显著性.

(3) 实现细节: 本研究基于 PyTorch 和 HuggingFace Transformers 实现. 实验全程在仅推理模式 (inference-time execution) 下运行, 不进行任何训练或微调. 所有实验均采用 bfloat16 精度, 并通过固定随机种子以控制非确定性带来的实验误差, 确保可复现性.

4 实验结果与分析

本节详细呈现实验结果并进行深入分析, 旨在揭示大语言模型中逻辑神经元的存在性、分布特征及其对模型能力的影响, 主要发现概括如下.

(1) 因果确证: 逻辑神经元与模型推理能力之间存在显著的双向因果联系, 定向增强可使平均准确率提升 9.2%, 而抑制则导致性能下降 34.65%.

(2) 泛化机制: 逻辑神经元并非对特定数据集的过拟合, 而是在逻辑任务上表现出显著的跨任务迁移能力和跨架构一致性.

(3) 结构特征: 在空间分布上, 逻辑神经元呈现出“中层富集、FFN 主导”的模式; 在功能关联上, 呈现出“共享簇” (通用逻辑) 与“独立簇” (特异逻辑) 并存的模块化结构.

(4) 下游影响: 逻辑神经元作为基础认知组件, 对如自然语言推理和复杂阅读理解等下游任务产生了差异化的调节作用, 表明逻辑相关神经元子集对部分逻辑敏感型下游任务具有较大影响.

4.1 逻辑神经元对逻辑任务的因果效应

我们首先通过双向干预实验 (增强/抑制), 验证所识别的“逻辑神经元” ($\mathcal{M}^{\text{logic}}$) 对逻辑推理任务的干预影响. 实验基于 Qwen3-8B, Original、Enhanced 和 Suppressed 分别表示原始模型、增强干预和抑制干预, 结果如图 3 所示, 为 LogicBench 在神经元干预下的准确率变化 (含置信区间), 实验基于 Qwen3-8B, Original、Enhanced 和 Suppressed 分别表示原始模型、增强干预和抑制干预结果. 实验结果支持了核心假设: 定向增强其激活应能提升模型性能, 而抑制其激活则会削弱性能.

增强干预的正向增益显著: 在大多数逻辑任务上, 对逻辑神经元实施增强干预 (α>1.0) 带来了显著的性能提升, 平均准确率增加了 9.2%. 这表明通过定向激活这一神经元子集, 模型能够更有效执行逻辑推理. 与之相对, 抑制干预的性能损耗更为显著: 当抑制逻辑神经元激活 (α<1.0) 时, 模型在所有任务上的准确率均出现了大幅下降, 平均降幅达 34.65%. 这一“不对称”的干预结果 (抑制的破坏力大于增强的提升力) 揭示了逻辑神经元的必要性, 表明其在维持模型逻辑推理表现中具有重要作用.

此外, 不同任务对干预的敏感度存在差异. 在“一阶逻辑”等涉及更复杂规则结构的任务上, 增强干预带来的收益通常高于基础命题逻辑任务. 这暗示所定位的“逻辑神经元”在处理高级抽象和长链推理时扮演着更为关键的角色.

考虑到本文主要在模型生成答案前的决策收敛点计算梯度重要性, 为进一步排除所定位神经元主要反映最终答案格式或输出偏好的可能性, 本文补充开展了推理链路定位对照实验. 由于模型自由生成的推理链长度与格式不稳定, 可能引入额外生成噪声, 本文采用统一显式推理模板构造 reasoning-path 对照, 从而保证不同样本的推理链路位置可比. 具体而言, reasoning-path mask 基于统一显式推理模板中间推理步骤位置计算梯度敏感度, 梯度统计范围覆盖模板中的推理链路 token; final-token mask 与 reasoning-path mask 均采用 $\tau_p = 0.10$ 稀疏度, 并在 Qwen3-8B 上进行对照实验. 表 4 中, baseline 表示未干预条件下的原始准确率; EnhΔ 表示对对应 mask 神经元进行增强干预后相较 baseline 的准确率变化, SupΔ 表示对对应 mask 神经元进行抑制干预后相较 baseline 的准确率变化; 正值

表示性能提升, 负值表示性能下降 (后文表格中相同符号含义均与表 4 相同).

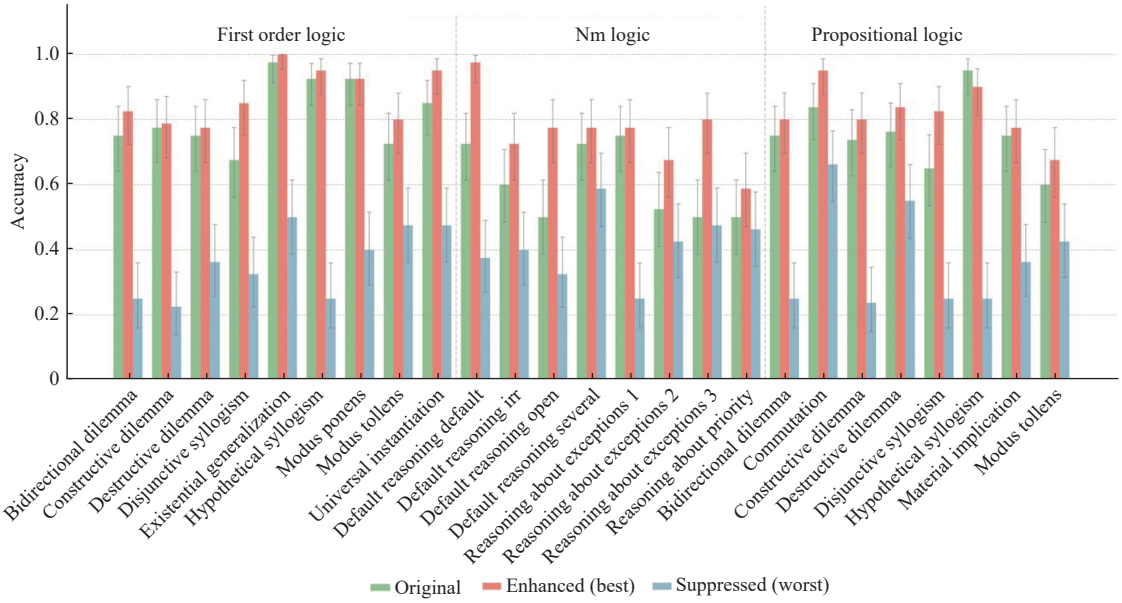


图 3 不同逻辑任务在原始/增强/抑制下的准确率对比柱状图 (Qwen3-8B)

表 4 Final-token mask 与 reasoning-path mask 的重叠度及干预性能变化

Task	Jaccard	Overlap	Baseline	Final		Reasoning	
				EnhΔ	SupΔ	EnhΔ	SupΔ
basic_logic/and	0.124	0.221	0.533	+0.075	−0.075	+0.192	−0.033
basic_logic/or	0.130	0.230	0.925	+0.008	−0.358	−0.042	−0.450
first_order_logic/constructive_dilemma	0.165	0.283	0.775	+0.013	−0.550	−0.013	−0.238
first_order_logic/modus_ponens	0.177	0.301	0.925	+0.000	−0.525	+0.000	−0.400
propositional_logic/disjunctive_syllogism	0.156	0.270	0.650	+0.175	−0.400	+0.125	−0.050
平均	0.150	0.261	0.778	+0.084	−0.392	+0.053	−0.234

在 $\tau_p = 0.10$ 稀疏度下, final-token mask 与 reasoning-path mask 在多个代表性任务上均存在非随机、且功能上不可忽略的重合. 从存在性来看, 以任务级统计, 两者的 Jaccard 重合率约为 0.15; 若以 Overlap 衡量 (交集/|mask|), final-token mask 中约 26.1% 的神经元同时出现在 reasoning-path mask 中. 考虑到两组 mask 均为独立的 Top-10% 选择, 随机独立条件下的期望 Overlap 仅约 10% (期望 Jaccard≈0.05), 实际重合率高于随机期望, 说明决策端定位与推理链路定位之间存在非随机重合.

干预实验进一步表明, 这一共享神经元子集可能具有一定功能意义. Final-token mask 在抑制时性能平均大幅下降 0.391 6, 增强约升 0.084 2; reasoning-path mask 同样表现出“抑制效应远强于增强” (抑制−0.234 2, 增强+0.052 5). 这一干预模式表明, 两组 mask 所定位的神经元, 尤其是其共享部分, 对逻辑任务表现具有较强影响; 相比增强带来的有限收益, 抑制更容易造成性能下降, 提示这些神经元更可能承担维持推理表现的关键作用. 上述结果表明, final-token mask 定位结果不太可能仅由最终答案格式或输出偏好驱动; final-token mask 与 reasoning-path mask 之间存在高于随机期望的重合, 并在干预实验中表现出相似的抑制敏感性.

为验证所识别的神经元的功能特异性和因果有效性, 我们引入了两组对照组进行对比实验, 如表 5 所示, 选取覆盖一阶逻辑、非单调逻辑、命题逻辑和基础逻辑的代表性任务进行展示. (1) 随机神经元组: 在相同层级以同等稀疏度随机抽取的神经元集合; (2) 非逻辑神经元组: 基于非逻辑下游任务 (例如 CoLA 语法判断) 计算梯度重要性而获得的神经元集合. 结果显示, 与对照实验相比, 逻辑神经元组的干预效果差异显著. 无论是随机干预组还是非

逻辑组对逻辑任务性能影响微乎其微, 甚至略有负面干扰, 这排除了性能提升来自模型通用能力改变或随机扰动的可能性, 支持所定位神经元子集具有较强的逻辑任务功能特异性.

表 5　逻辑神经元与其他对照组的干预效果对比

Task	Baseline	逻辑神经元	非逻辑神经元 (CoLA)	随机神经元
first_order_logic/bidirectional_dilemma	0.750	0.825 (+0.075) 0.250 (−0.500)	0.750 (+0.000) 0.738 (−0.013)	0.763 (+0.013) 0.750 (−0.000)
nm_logic/default_reasoning_default	0.725	0.975 (+0.250) 0.375 (−0.350)	0.725 (+0.000) 0.700 (−0.025)	0.725 (+0.000) 0.725 (−0.000)
propositional_logic/destructive_dilemma	0.763	0.838 (+0.075) 0.550 (−0.213)	0.750 (−0.013) 0.763 (−0.000)	0.763 (+0.000) 0.738 (−0.025)
basic_logic/and	0.533	0.608 (+0.075) 0.458 (−0.075)	0.533 (+0.000) 0.542 (+0.008)	0.517 (−0.017) 0.533 (−0.000)

注: 各干预组单元格中, 上行为增强干预结果, 下行为抑制干预结果; 括号内为相对于baseline的准确率变化

为了评估“逻辑神经元”是否捕获了超越特定数据集的通用计算规则, 我们分析了不同逻辑任务间的迁移矩阵. 限于篇幅, 表 6 仅展示代表性任务的迁移效果, 完整的 25×25 全量迁移矩阵详见附录 A 表 A2. 该矩阵反映了当利用源任务 (source task) 识别的神经元去干预目标任务 (target task) 时的性能增益 (任务缩写对照参见附录表 A1).

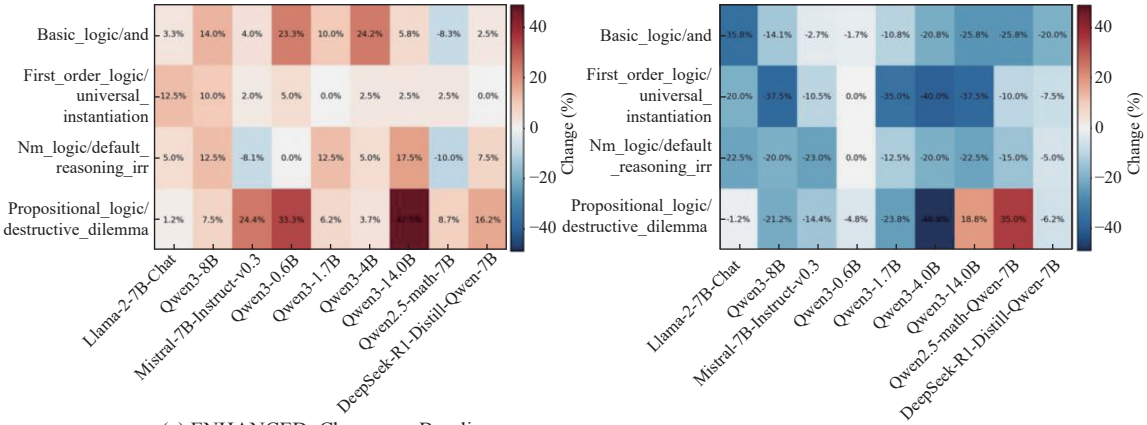
表 6　不同逻辑任务间交叉干预的性能变化

Source task	FOL/CD	FOL/HS	FOL/MP	NL/DRD	NL/RAE1	NL/RAP	PL/DS	PL/MI	PL/MT
FOL/CD	**0.788** **0.225**	0.950 0.250	0.975 0.350	0.925 0.325	0.775 0.638	0.563 0.475	0.800 0.300	0.788 0.263	0.700 0.475
FOL/HS	0.800 0.250	**0.950** **0.250**	1.000 0.450	0.900 0.350	0.775 0.263	0.600 0.475	0.825 0.425	0.775 0.250	0.775 0.400
FOL/MP	0.813 0.238	0.950 0.238	**0.925** **0.400**	0.825 0.425	0.775 0.288	0.563 0.488	0.800 0.325	0.763 0.250	0.725 0.450
NL/DRD	0.800 0.238	0.975 0.250	1.000 0.450	**0.975** **0.375**	0.775 0.250	0.563 0.475	0.850 0.325	0.763 0.250	0.675 0.475
NL/RAE1	0.800 0.250	0.938 0.250	0.950 0.425	0.950 0.300	**0.775** **0.250**	0.575 0.488	0.875 0.300	0.775 0.250	0.700 0.500
NL/RAP	0.788 0.250	0.950 0.250	0.975 0.425	0.925 0.400	0.775 0.313	**0.588** **0.463**	0.850 0.425	0.775 0.250	0.650 0.400
PL/DS	0.800 0.238	0.988 0.250	1.000 0.400	0.900 0.350	0.775 0.500	0.600 0.475	**0.825** **0.250**	0.763 0.250	0.725 0.375
PL/MI	0.800 0.250	0.963 0.363	0.975 0.450	0.950 0.375	0.775 0.575	0.575 0.488	0.800 0.275	**0.775** **0.363**	0.650 0.450
PL/MT	0.788 0.250	0.938 0.250	0.925 0.400	0.850 0.275	0.775 0.250	0.588 0.438	0.800 0.250	0.775 0.250	**0.675** **0.425**

注: 各干预组单元格中, 上行为增强干预结果, 下行为抑制干预结果

观察发现, 高干预效能不只集中于对角线, 表明神经元在同源任务中具有最高的适应性, 更重要的是, 这种高干预效果也出现在非对角任务对中. 特别是从基础的命题逻辑到复杂的自然语言推理之间, 可以观察到一定迁移效应. 这一现象提示, 所定位神经元子集可能包含与蕴含、否定等基础逻辑操作相关的共享计算成分, 使模型能够在不同模态任务中复用这些机制. 同时, 部分特定任务 (如非单调逻辑) 显示出较低的迁移分数, 客观反映该类任务对特定上下文机制的独立依赖, 这可能是一些逻辑推理任务使用了不同的神经元建模机制造成的.

为考察所识别的逻辑神经元及其干预机制在不同模型上的一致性, 我们将实验扩展至 Qwen、LLaMA、Mistral 及 DeepSeek 等参数规模与预训练策略各异的模型. 如图 4 所示, 在针对 4 种代表性逻辑任务的跨模型评估中, 干预效果展现出高度的一致性: 增强干预在所有模型上多表现为正向增益 (红色区域), 而抑制干预则导致了推理能力的显著衰减 (蓝色区域). 这种跨架构的双向控制一致性趋势提示, 逻辑相关神经元子集在不同模型架构中具有一定一致性.



(a) ENHANCED: Change vs. Baseline

(b) SUPPRESSED: Change vs. Baseline

图 4 跨模型和任务的双向控制有效性

本文在统一的代表性任务子集上汇总 Qwen、LLaMA 和 Mistral 等代表性模型的基准 (baseline)、增强和抑制结果, 如表 7 所示. 在 3 个参数规模相近但模型族和对齐策略不同的主模型上, 逻辑神经元增强干预均表现出不同程度的性能提升, 而抑制干预则导致准确率下降. 这一结果说明, 本文方法并非仅适用于单一模型基座, 而是在多个代表性模型上呈现出较一致的双向控制趋势.

表 7 3 个代表性模型上的逻辑神经元干预效果汇总

Task	Qwen3-8B			Llama-2-7B-Chat			Mistral-7B-Instruct-v0.3		
	Baseline	ΔEnh	ΔSup	Baseline	ΔEnh	ΔSup	Baseline	ΔEnh	ΔSup
basic_logic/and	0.533	+0.075	−0.075	0.875	+0.029	−0.314	0.800	+0.032	−0.022
basic_logic/not	0.975	−0.017	−0.517	0.500	+0.092	0.000	0.883	+0.067	−0.250
first_order_logic/constructive_dilemma	0.775	+0.013	−0.550	0.750	+0.063	−0.350	0.713	+0.013	−0.038
first_order_logic/modus_ponens	0.925	+0.000	−0.525	0.775	+0.050	−0.250	0.900	+0.025	−0.125
nm_logic/default_reasoning_default	0.725	+0.250	−0.350	0.500	+0.125	−0.125	0.975	+0.000	−0.275
propositional_logic/disjunctive_syllogism	0.650	+0.175	−0.400	0.525	+0.175	−0.050	0.450	+0.200	−0.075

为进一步分析参数规模对逻辑神经元干预效果的影响, 本文在 Qwen3 系列不同参数规模模型上进行了补充实验. 不同于表 7 的跨架构比较, 该实验聚焦于同一模型族内部的规模变化, 用于观察模型参数量变化时 baseline 性能及增强/抑制干预趋势是否保持一致. 结果如表 8 所示.

表 8 Qwen3 系列不同参数规模模型上的代表性干预结果

Task	Qwen3-0.6B		Qwen3-1.7B		Qwen3-4B		Qwen3-8B		Qwen3-14B	
	Baseline	ΔEnh ΔSup	Baseline	ΔEnh ΔSup	Baseline	ΔEnh ΔSup	Baseline	ΔEnh ΔSup	Baseline	ΔEnh ΔSup
basic_logic/and	0.500	+0.117 −0.008	0.608	+0.061 −0.066	0.708	−0.171 −0.147	0.533	+0.075 −0.075	0.817	+0.047 −0.211
basic_logic/not	0.500	+0.000 −0.000	0.500	+0.492 −0.008	0.925	+0.025 −0.292	0.975	−0.017 −0.517	0.983	−0.008 −0.517
first_order_logic/constructive_dilemma	0.363	+0.338 −0.113	0.538	+0.350 −0.163	0.800	+0.025 −0.463	0.775	+0.013 −0.550	0.900	−0.075 −0.275
first_order_logic/modus_ponens	0.500	+0.025 −0.000	0.775	+0.100 −0.125	0.950	−0.025 −0.300	0.925	+0.000 −0.525	0.825	+0.150 −0.150
nm_logic/default_reasoning_default	0.500	+0.000 −0.000	0.650	+0.000 −0.150	0.875	+0.025 −0.100	0.725	+0.250 −0.350	0.900	+0.025 −0.200
propositional_logic/disjunctive_syllogism	0.525	+0.175 −0.025	0.875	−0.025 −0.300	0.900	+0.000 −0.350	0.650	+0.175 −0.400	0.900	+0.075 −0.475

从表 8 可以看出, 随着模型规模变化, baseline 逻辑推理能力呈现一定差异, 但增强和抑制干预仍整体保持趋势一致: 增强目标神经元通常能够提升或维持逻辑任务表现, 而抑制目标神经元则会造成不同程度的性能下降. 这说明逻辑神经元干预效果不仅出现在单一参数规模模型中, 也在同一模型族的不同规模变体上呈现出一定延续性. 需要说明的是, DeepSeek-R1-Distill-Qwen-7B 等额外参考模型的 baseline 或部分探索结果放入附录, 不纳入主跨模型干预汇总.

4.2 机制分析: 架构分布与功能原子

本节进一步分析逻辑神经元在模型中的空间分布与功能协同关系, 探究 LLM 内部逻辑推理相关神经元的空间组织与功能协同模式. 为回答逻辑神经元是否与基础逻辑原子操作相关的问题, 本文首先对 Basic_logic 中的 AND/OR/NOT 进行真值结构激活分析.

为进一步分析目标逻辑神经元是否编码不同基础逻辑操作的真值结构, 本文对 Basic_logic 中 AND、OR 和 NOT 这 3 类逻辑原子操作进行了激活分析. 具体而言,平均绝对激活强度表示 Top-5% 目标逻辑神经元在答案生成前末尾 8 个 token 上的平均绝对 FFN 中间激活值. AUROC 用于衡量激活分数对不同真值条件的可分性; Cohen's d 表示组间效应量; *p*-value 基于逐样本激活分数采用双侧 Mann-Whitney U 检验计算.

如表 9 所示, AND 操作表现出最强的真值结构可分性: TT 条件下目标神经元激活强度明显高于 TF、FT 和 FF 条件, TT 与非 TT 的 AUROC 达到 0.997, Cohen's d 为 3.413, 且差异显著 (p<0.001). NOT 操作也表现出一定可分性, P_FALSE 与 P_TRUE 的 AUROC 为 0.680, Cohen's d 为 0.680 (p<0.001). 相比之下, OR 操作在当前平均激活指标下区分度较弱, AUROC 为 0.569, 差异不显著 (p=0.258). 上述结果表明, 目标逻辑神经元在部分基础逻辑操作上呈现与真值结构相关的激活差异, 尤其对 AND 成立条件具有较强响应; 同时, 不同逻辑原子的神经元表征强度存在差异, 并非所有逻辑操作都表现出同等清晰的单一神经元级响应.

表 9 逻辑原子操作的真值结构激活分析

Basic_logic	真值条件	样本数	平均绝对激活强度	标准差	可分性对比	AUROC	Cohen's d	*p*-value	结论
AND	TT	30	0.376	0.0022	TT vs. 非TT	0.997	3.413	<0.001	强可分
	TF	30	0.369	0.0021					
	FT	30	0.367	0.0021					
	FF	30	0.368	0.0023					
OR	TT	30	0.357	0.0022	非FF vs. FF	0.569	0.226	0.258	弱可分
	TF	30	0.354	0.0026					
	FT	30	0.351	0.0019					
	FF	30	0.353	0.0019					
NOT	P_FALSE	60	0.372	0.0034	P_FALSE vs. P_TRUE	0.680	0.680	<0.001	中等可分
	P_TRUE	60	0.370	0.0028					

逻辑神经元的分布呈现出显著的非均匀性. 分析表明, 这些逻辑神经元并非均匀分布于整个网络, 而是呈现出清晰的"中层峰值"分布模式 (如图 5 所示), 其密度在约 10–25 层之间达到峰值. 这一现象表明逻辑推理作为高阶语义特征, 需要在前层的基础表征之上, 经过中层网络多次非线性变换后才形成稳定的可定位表征.

在模块维度上, 绝大部分神经元集中于前馈网络 (FFN) 模块中, 如图 6 所示. 这与 FFN 作为 LLM"知识存储与计算节点"的假设相符, 表明逻辑相关计算可能主要依赖于 FFN 提供的稀疏、离散的计算单元. 与此同时, 我们观察到特定层级的注意力头表现出强烈的激活专一性 (见附录 A 图 A1). 这些注意力头可能承担着"信息路由"的角色, 负责将逻辑前提中的关键变量导向 FFN 模块, 提示可能存在 Attention-FFN 协同参与的计算模式.

在明确分布的基础上, 我们进一步通过交叉干预相关性矩阵 (如后文图 7 所示) 探究不同逻辑任务间神经元复用机制. 相关性矩阵提示, 不同逻辑任务对应的神经元子集可能存在共享与相对分离并存的组织方式. 矩阵中部分任务对 (如 first_order_logic/bidirectional_dilemma (一阶逻辑-双向二难) 与 propositional_logic/disjunctive_syllogism (命题逻辑-析取三段论)) 表现出强烈的正相关 (相关系数 r>0.9), 表明模型在处理这些任务时可能复用了诸如"条件句判断"或"析取式简化"等底层通用的神经元子结构, 提示不同逻辑任务之间存在一定的功能共享机制.

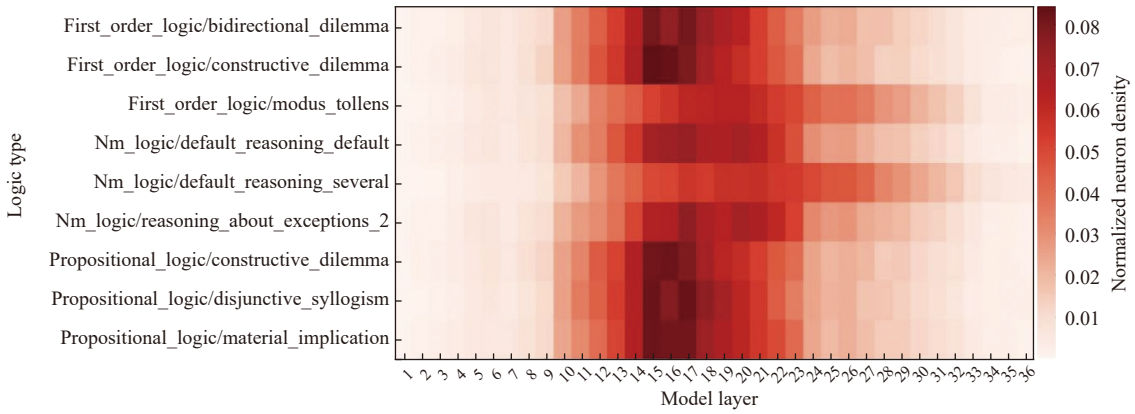


图 5 逻辑神经元在 Qwen3-8B 模型层级上的归一化分布密度

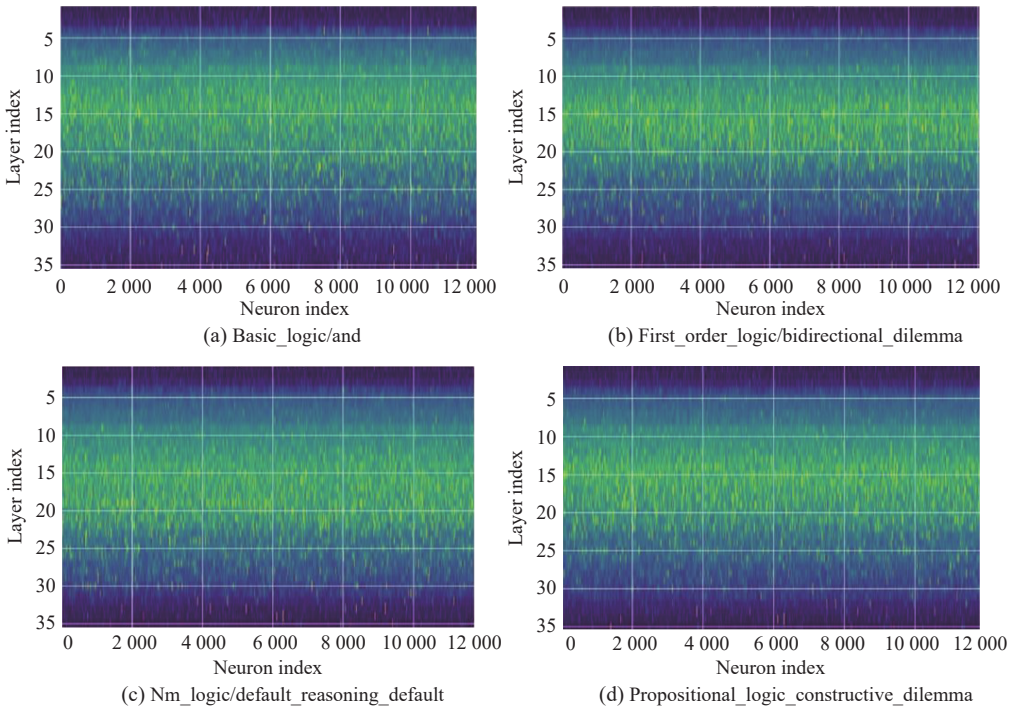


图 6 FFN 激活模式与逻辑神经元分布

然而, 另一类任务之间则呈现出极弱甚至负相关 (如 first_order_logic/modus_tollens (拒取式推理) 和 nm_logic/reasoning_about_exceptions_2 (异常推理)), 这说明涉及否定操作或处理反常识信息的高阶任务依赖于独立的神经元表征与处理通路, 难以直接复用基础逻辑的计算路径. 我们推测这些任务可能依赖于模型内部独特的、非共享的操作, 例如 modus_tollens 需要对前提进行否定操作, 这与直接推断存在本质区别, 从而导致其神经元表征的功能性分离. 此外, 矩阵中部分任务表现出的广泛连接性提示其可能作为连接不同逻辑子系统的枢纽节点, 这种结构化的关联图谱从神经元干预角度刻画了 LLM 内部逻辑能力的组织形式.

4.3 逻辑神经元对通用能力的跨任务影响

为验证逻辑推理能力作为基础认知能力的普适性作用, 本文进一步探究了逻辑神经元干预对下游通用任务的影响. 如图 8 所示, 选取了涵盖自然语言推理 (如 SNLI)、阅读理解 (如 BoolQ) 及语法判断 (如 CoLA) 等不同逻辑

敏感度梯度的任务集.

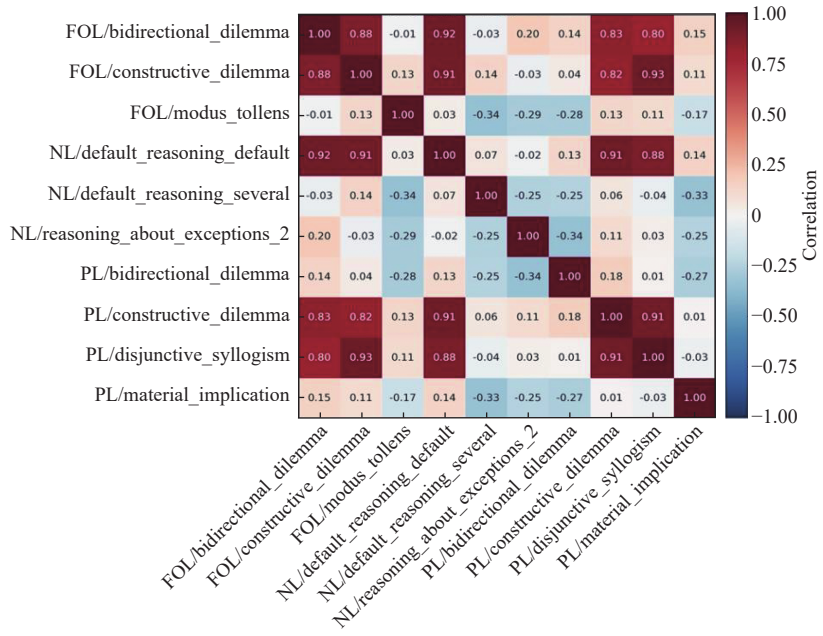


图 7　交叉干预下 Qwen3-8B 逻辑任务性能相关性矩阵

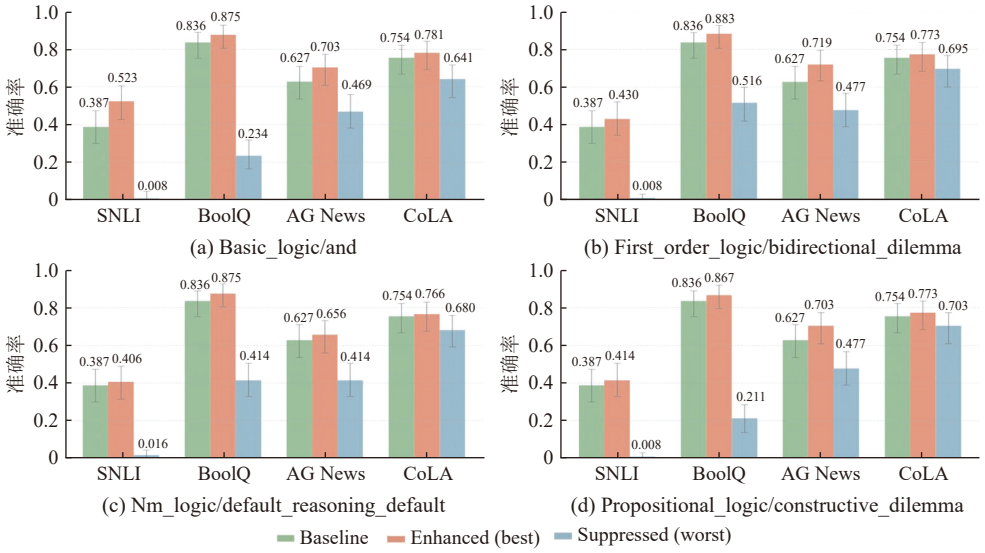


图 8　下游任务在不同逻辑任务增强/抑制干预下的性能变化

具体而言, 任务对逻辑推理的依赖程度与干预效果的幅度呈现出一定正相关趋势. 对于 SNLI 和 BoolQ 等逻辑敏感型任务, 对逻辑神经元的激活增强带来了显著的性能提升, 而抑制干预则导致准确率大幅下降 (例如 SNLI 降幅达 0.378, BoolQ 降幅达 0.492), 表明了这些任务的表现受到逻辑相关神经元子集的显著影响.

相反, 对于侧重句法结构或浅层特征的低敏感型任务 (如 CoLA), 同样的干预操作所带来的性能波动则相对温和. 例如, 在增强干预下, CoLA 的准确率提升微弱 (ΔAcc≈+0.020), 在抑制干预下, 其性能下降 (−0.074) 也远小于 SNLI 和 BoolQ. 这种差异化的跨任务影响模式表明, 逻辑神经元虽然是 LLM 的通用组成部分, 但其功能贡献

受任务需求动态调节: 当任务核心涉及前提与结论之间的因果推导时, 其贡献更为明显; 而在处理基础语言形式或事实回忆任务时, 其影响相对有限. 这一结果说明, 逻辑相关神经元子集对部分逻辑敏感型下游任务具有较大影响, 同时对低逻辑敏感任务的影响相对有限.

4.4 参数敏感性分析

为进一步回答干预效果在不同任务间存在波动的问题, 本文补充进行了参数敏感性分析. 具体而言, 我们在不同逻辑神经元比例 τ_p 与激活缩放系数 α 下, 统计代表性逻辑任务的平均性能变化.

图 9 展示了不同参数组合下的任务级宏平均 ΔAcc. 可以看到, 当 α<1 时, 多数组合呈现负向变化, 说明抑制目标逻辑神经元通常会削弱逻辑任务表现; 当 α 位于 1.3–2.0 的中等增强区间时, 多个 τ_p 设置下的平均 ΔAcc 转为正值, 表明增强干预在合理强度范围内具有一定正向收益. 与此同时, 当 α 进一步增大至 4.0 时, 平均 ΔAcc 明显下降, 说明过强激活并不会持续提升性能, 反而可能破坏模型原有的计算平衡.

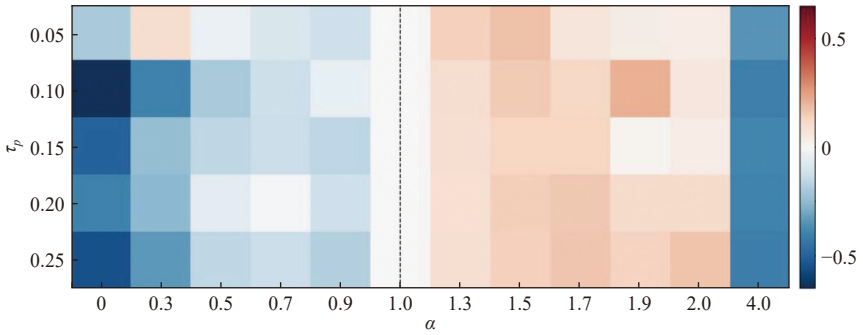


图 9 不同神经元比例与干预强度下的平均 ΔAcc 变化

为避免平均值掩盖任务间差异, 本文进一步统计了每组参数设置下符合预期方向变化的任务数量, 如表 10 所示. 对于 α<1 的抑制干预, 符合预期表示任务准确率下降, 即 ΔAcc<0; 对于 α>1 的增强干预, 符合预期表示任务准确率提升, 即 ΔAcc>0. 结果显示, 在抑制区间内, 大多数参数组合均使多数任务出现性能下降; 在中等增强区间内, 较多任务也表现出正向提升, 例如 τ_p =0.10、α=1.5 和 α=1.9 时, 分别有 23 个任务符合增强预期. 相比之下, 当 α=4.0 时, 符合增强预期的任务数量显著减少, 进一步说明过强增强会降低干预稳定性.

表 10 不同参数设置下干预方向一致性统计表

τ_p	α=0.0	α=0.3	α=0.5	α=0.7	α=0.9	α=1.1	α=1.3	α=1.5	α=1.7	α=1.9	α=2.0	α=4.0
0.05	15	15	15	16	16	17	14	16	14	16	16	4
0.10	25	16	17	15	16	13	15	23	19	23	19	3
0.15	18	16	19	16	14	16	14	18	17	19	18	1
0.20	21	21	19	17	17	16	15	19	16	17	16	0
0.25	20	18	20	20	17	16	16	17	16	15	15	2

综合图 9 和表 10 可以看出, 逻辑神经元干预并非仅在单一超参数点上有效, 而是在一定 τ_p 与 α 范围内呈现较一致的方向性效果. 其中, 抑制干预的负向影响相对更稳定, 增强干预则在中等强度区间内更容易带来正向收益, 但其幅度仍受任务类型和干预强度影响. 因此, 本文后续实验采用在平均收益与任务覆盖率之间较为平衡的参数设置来进行分析.

5 总 结

本研究提出“识别-干预-评估”系统化框架, 整合梯度重要性分析与推理时干预 (ITI) 技术, 成功从模型激活空间中定位并分析了与形式逻辑推理密切相关的逻辑神经元子集. 双向干预实验为该神经元子集对逻辑推理任务表

现具有因果影响提供了证据: 对其激活实施增强干预可使长链推理任务的平均准确率提升 9.2%, 而抑制干预则导致性能下降 34.65%. 这一量化结果表明所定位的稀疏神经元子网络在模型执行逻辑推理过程中发挥了重要作用, 为理解 LLM 如何在内部表征和调用逻辑推理能力提供了实证依据.

进一步的机制分析揭示, 逻辑神经元主要集中于模型中层的 FFN 模块, 且不同逻辑任务间存在共享与独立并存的功能结构. 结合真值结构激活分析, 本文发现部分基础逻辑操作, 尤其是 AND 和 NOT, 在目标神经元集合上呈现出明显的真值条件响应; 同时, 任务相关性矩阵也显示不同逻辑任务之间存在一定程度的神经元复用现象. 这些发现支持了模型内部存在可复用逻辑相关神经元子结构的假设. 此外, 该子网络展现出一定的任务依赖性和跨模型一致趋势, 为构建可解释、可控的推理系统提供了进一步的实验基础.

需指出的是, 本文仍以神经元维度作为基本分析单元, 而单个神经元可能具有多义性, 因此本文所称“逻辑神经元”是指对逻辑任务具有显著贡献的稀疏神经元子集. 未来工作可结合稀疏自编码器等特征层级分析方法, 在概念表示空间中进一步验证逻辑表征的组织形式, 并分析不同逻辑原子操作在神经元级与特征级表征之间的对应关系.

References

[1] Newell A, Simon HA. Computer science as empirical inquiry: Symbols and search. Communications of the ACM, 1976, 19(3): 113–126. [doi: 10.1145/360018.360022]

[2] Russell S, Norvig P. Artificial Intelligence: A Modern Approach. 4th ed., Hoboken: Pearson, 2020.

[3] Kahneman D. Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.

[4] Li LG, Xue ZY, Chen XH, Zhang M, Chen LY, Li PP, Jiang TT. Specification method of securities rules integrating large language models and domain knowledge base. Ruan Jian Xue Bao/Journal of Software, 2025, 36(10): 4671–4694 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7294.htm [doi: 10.13328/j.cnki.jos.007294]

[5] Bubeck S, Chandrasekaran V, Eldan R, Gehrke J, Horvitz E, Kamar E, Lee P, Lee YT, Li YZ, Lundberg S, Nori H, Palangi H, Ribeiro MT, Zhang Y. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv:2303.12712, 2023.

[6] Gao YJ, Chen YG. Survey on key technologies for large language model pre-training systems. Ruan Jian Xue Bao/Journal of Software, 2026, 37(1): 200–229 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/7438.htm [doi: 10.13328/j.cnki.jos.007438]

[7] Achiam J, Adler S, Agarwal S, *et al*. GPT-4 technical report. arXiv:2303.08774, 2023.

[8] Wei J, Wang XZ, Schuurmans D, Bosma M, Ichter B, Xia F, Chi H, Le QV, Zhou D. Chain-of-thought prompting elicits reasoning in large language models. In: Proc. of the 36th Int’l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1800.

[9] Bender EM, Gebru T, Mcmillan-Major A, Shmitchell S. On the dangers of stochastic parrots: Can language models be too big? In: Proc. of the 2021 ACM Conf. on Fairness, Accountability, and Transparency. ACM, 2021. 610–623. [doi: 10.1145/3442188.3445922]

[10] Goel V. Anatomy of deductive reasoning. Trends in Cognitive Sciences, 2007, 11(10): 435–441. [doi: 10.1016/j.tics.2007.09.003]

[11] Parmar M, Patel N, Varshney N, Nakamura M, Luo M, Mashetty S, Mitra A, Baral C. LogicBench: Towards systematic evaluation of logical reasoning ability of large language models. In: Proc. of the 62nd Annual Meeting of the Association for Computational Linguistics. Bangkok: ACL, 2024. 13679–13707. [doi: 10.18653/v1/2024.acl-long.739]

[12] Yu WH, Jiang ZH, Dong YF, Feng JS. ReClor: A reading comprehension dataset requiring logical reasoning. In: Proc. of the 8th Int’l Conf. on Learning Representations. Addis Ababa: OpenReview.net, 2020.

[13] Jiang J, Wang JN, Yan YC, Liu Y, Zhu JH, Zhang MD, Gao LC. Do large language models excel in complex logical reasoning with formal language? In: Proc. of the 2025 Conf. on Empirical Methods in Natural Language Processing. Suzhou: ACL, 2025. 16878–16903. [doi: 10.18653/v1/2025.emnlp-main.855]

[14] Poddar A, Sahoo S, Ghosh S. Understanding syllogistic reasoning in LLMs from formal and natural language perspectives. arXiv:2512.12620, 2025.

[15] Liu HM, Fu ZZ, Ding MR, Ning RX, Zhang CL, Liu XZ, Zhang Y. Logical reasoning in large language models: A survey. arXiv: 2502.09100, 2025.

[16] Gemechu D, Ruiz-Dolz R, Beyer H, Reed C. Natural language reasoning in large language models: Analysis and evaluation. In: Findings of the Association for Computational Linguistics: ACL 2025. Vienna: ACL, 2025. 3717–3741. [doi: 10.18653/v1/2025.findings-acl.192]

[17] Wei AJ, Wu YH, Wan YJ, Suresh T, Tan HM, Zhou ZK, Koyejo S, Wang K, Aiken A. SATBench: Benchmarking LLMs’ logical

reasoning via automated puzzle generation from SAT formulas. In: Proc. of the 2025 Conf. on Empirical Methods in Natural Language Processing. Suzhou: ACL, 2025. 33832–33849. [doi: 10.18653/v1/2025.emnlp-main.1716]

[18] Xia Y, Atrey A, Khmaissia F, Namjoshi K. Can large language models learn formal logic? A data-driven training and evaluation framework. In: Proc. of the 5th Workshop on Mathematical Reasoning and AI at NeurIPS 2025. San Diego: OpenReview.net, 2025.

[19] Srivastava A, Rastogi A, Rao A, *et al*. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv:2206.04615, 2022.

[20] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large language models are zero-shot reasoners. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1613.

[21] Elhage N, Nanda N, Olsson C, *et al*. A mathematical framework for Transformer circuits. Transformer Circuits Thread, 2021, 1(1): 12.

[22] Geva M, Schuster R, Berant J, Levy O. Transformer feed-forward layers are key-value memories. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 5484–5495. [doi: 10.18653/v1/2021.emnlp-main.446]

[23] Dai DM, Dong L, Hao YR, Sui ZF, Chang BB, Wei FR. Knowledge neurons in pretrained Transformers. In: Proc. of the 60th Annual Meeting of the Association for Computational Linguistics. Dublin: ACL, 2022. 8493–8502. [doi: 10.18653/v1/2022.acl-long.581]

[24] Meng K, Bau D, Andonian A, Belinkov Y. Locating and editing factual associations in GPT. In: Proc. of the 36th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2022. 1262.

[25] Dalvi F, Durrani N, Sajjad H, Belinkov Y, Bau A, Glass J. What is one grain of sand in the desert? Analyzing individual neurons in deep NLP models. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. Honolulu: AAAI Press, 2019. 6309–6317. [doi: 10.1609/aaai.v33i01.33016309]

[26] Belinkov Y, Durrani N, Dalvi F, Sajjad H, Glass J. On the linguistic representational power of neural machine translation models. Computational Linguistics, 2020, 46(1): 1–52. [doi: 10.1162/coli_a_00367]

[27] Radford A, Jozefowicz R, Sutskever I. Learning to generate reviews and discovering sentiment. In: Proc. of the 6th Int'l Conf. on Learning Representations. Vancouver: OpenReview.net, 2018.

[28] Wang KR, Variengien A, Conmy A, Shlegeris B, Steinhardt J. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In: Proc. of the 11th Int'l Conf. on Learning Representations. OpenReview.net, 2023.

[29] Song R, He SZ, Jiang ST, Xian YT, Gao SX, Liu K, Yu ZT. Does large language model contain task-specific neurons? In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 7101–7113. [doi: 10.18653/v1/2024.emnlp-main.403]

[30] Zou A, Phan L, Chen S, Campbell J, Guo P, Ren R, Pan A, Yin XW, Mazeika M, Dombrowski AK, Goel S, Li N, Byun MJ, Wang ZF, Mallen A, Basart S, Koyejo S, Song D, Fredrikson M, Kolter JZ, Hendrycks D. Representation engineering: A top-down approach to AI transparency. arXiv:2310.01405, 2023.

[31] Bowman SR, Angeli G, Potts C, Manning CD. A large annotated corpus for learning natural language inference. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. Lisbon: ACL, 2015. 632–642. [doi: 10.18653/v1/D15-1075]

[32] Clark C, Lee K, Chang MW, Kwiatkowski T, Collins M, Toutanova K. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis: ACL, 2019. 2924–2936. [doi: 10.18653/v1/N19-1300]

[33] Zhang X, Zhao JB, LeCun Y. Character-level convolutional networks for text classification. In: Proc. of the 29th Int'l Conf. on Neural Information Processing Systems. Montreal: MIT Press, 2015. 649–657.

[34] Warstadt A, Singh A, Bowman SR. Neural network acceptability judgments. Trans. of the Association for Computational Linguistics, 2019, 7: 625–641. [doi: 10.1162/tacl_a_00290]

[35] Yao SY, Zhao J, Yu D, Du N, Shafran I, Narasimhan KR, Cao Y. ReAct: Synergizing reasoning and acting in language models. In: Proc. of the 11th Int'l Conf. on Learning Representations. Kigali: OpenReview.net, 2023.

[36] Dziri N, Lu XM, Sclar M, Li XL, Jiang LW, Lin BY, West P, Bhagavatula C, Le Bras R, Hwang JD, Sanyal S, Welleck S, Ren X, Ettinger A, Harchaoui Z, Choi Y. Faith and fate: Limits of Transformers on compositionality. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 3081.

[37] Press O, Zhang MR, Min S, Schmidt L, Smith N, Lewis M. Measuring and narrowing the compositionality gap in language models. In: Findings of the Association for Computational Linguistics: ACL 2023. Singapore: ACL, 2023. 5687–5711. [doi: 10.18653/v1/2023.findings-emnlp.378]

[38] Ameisen E, Lindsey J, Pearce A, *et al*. Circuit tracing: Revealing computational graphs in language models. Transformer Circuits Thread,

2025, 6: 16318–16352.

[39] Chen YH, Cao PF, Chen YB, Liu K, Zhao J. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In: Proc. of the 38th AAAI Conf. on Artificial Intelligence. Vancouver: AAAI Press, 2024. 17817–17825. [doi: 10.1609/aaai.v38i16.29735]

[40] Meng K, Sharma AS, Andonian AJ, Belinkov Y, Bau D. Mass-editing memory in a Transformer. In: Proc. of the 11th Int'l Conf. on Learning Representations. OpenReview.net, 2023.

[41] Cao ND, Aziz W, Titov I. Editing factual knowledge in language models. In: Proc. of the 2021 Conf. on Empirical Methods in Natural Language Processing. Punta Cana: ACL, 2021. 6491–6506. [doi: 10.18653/v1/2021.emnlp-main.522]

[42] Mitchell E, Lin C, Bosselut A, Finn C, Manning CD. Fast model editing at scale. In: Proc. of the 10th Int'l Conf. on Learning Representations. OpenReview.net, 2022.

[43] Tak AN, Banayeeanzade A, Bolourani A, Kian M, Jia RB, Gratch J. Mechanistic interpretability of emotion inference in large language models. In: Findings of the Association for Computational Linguistics: ACL 2025. Vienna: ACL, 2025. 13090–13120. [doi: 10.18653/v1/2025.findings-acl.679]

[44] Zhang DZ, Li ZZ, Zhang ML, Zhang JX, Liu ZY, Yao YX, Xu HT, Zheng JH, Chen XY, Zhang YY, Yin F, Dong JH, Guo ZJ, Song L, Liu CL. From System 1 to System 2: A survey of reasoning large language models. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2026, 48(3): 3335–3354. [doi: 10.1109/TPAMI.2025.3637037]

[45] Chi HA, Li H, Yang WJ, Liu F, Lan L, Ren XG, Liu TL, Han B. Unveiling causal reasoning in large language models: Reality or mirage? In: Proc. of the 38th Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 3064.

[46] Kim G, Valentino M, Freitas A. Reasoning circuits in language models: A mechanistic interpretation of syllogistic inference. In: Findings of the Association for Computational Linguistics: ACL 2025. Vienna: ACL, 2025. 10074–10095. [doi: 10.18653/v1/2025.findings-acl.525]

[47] Wang X, Hu Y, Du WY, Cheng R, Wang BY, Zou DF. Towards understanding fine-tuning mechanisms of LLMs via circuit analysis. In: Proc. of the 42nd Int'l Conf. on Machine Learning. Vancouver: OpenReview.net, 2025.

[48] Yao YZ, Wang P, Tian BZ, Cheng SY, Li ZB, Deng SM, Chen HJ, Zhang NY. Editing large language models: Problems, methods, and opportunities. In: Proc. of the 2023 Conf. on Empirical Methods in Natural Language Processing. Singapore: ACL, 2023. 10222–10240. [doi: 10.18653/v1/2023.emnlp-main.632]

[49] Yu ZP, Ananiadou S. Locate-then-merge: Neuron-level parameter fusion for mitigating catastrophic forgetting in multimodal LLMs. In: Findings of the Association for Computational Linguistics: ACL 2025. Suzhou: ACL, 2025. 7065–7078. [doi: 10.18653/v1/2025.findings-emnlp.372]

[50] Yang WL, Sun F, Tan JJ, Ma XY, Cao Q, Yin DW, Shen HW, Cheng XQ. The mirage of model editing: Revisiting evaluation in the wild. In: Proc. of the 63rd Annual Meeting of the Association for Computational Linguistics. Vienna: ACL, 2025. 15336–15354. [doi: 10.18653/v1/2025.acl-long.745]

[51] Li K, Patel O, Viégas F, Pfister H, Wattenberg M. Inference-time intervention: Eliciting truthful answers from a language model. In: Proc. of the 37th Int'l Conf. on Neural Information Processing Systems. New Orleans: Curran Associates Inc., 2023. 1797.

[52] Nguyen DV, Pham NY, Vu HM, Zhang L, Nguyen TM. Activation steering with a feedback controller. In: Proc. of the 14th Int'l Conf. on Learning Representations. Rio de Janeiro: OpenReview.net, 2026.

[53] Wang PY, Zhang D, Li LY, Tan CK, Wang XH, Zhang MZ, Ren K, Jiang BT, Qiu XP. InferAligner: Inference-time alignment for harmlessness through cross-model guidance. In: Proc. of the 2024 Conf. on Empirical Methods in Natural Language Processing. Miami: ACL, 2024. 10460–10479. [doi: 10.18653/v1/2024.emnlp-main.585]

[54] Valentino M, Kim G, Dalal D, Zhao ZX, Freitas A. Mitigating content effects on reasoning in language models through fine-grained activation steering. In: Proc. of the 40th AAAI Conf. on Artificial Intelligence. Singapore: AAAI Press, 2026. 33314–33322. [doi: 10.1609/aaai.v40i39.40617]

附中文参考文献

[4] 李靓果, 薛志一, 陈小红, 张民, 陈良育, 李萍萍, 姜婷婷. 结合大语言模型和领域知识库的证券规则规约方法. 软件学报, 2025, 36(10): 4671–4694. http://www.jos.org.cn/1000-9825/7294.htm [doi: 10.13328/j.cnki.jos.007294]

[6] 高彦杰, 陈跃国. 大语言模型预训练系统关键技术综述. 软件学报, 2026, 37(1): 200–229. http://www.jos.org.cn/1000-9825/7438.htm [doi: 10.13328/j.cnki.jos.007438]

附录 A

(1) Qwen3-8B 模型在逻辑推理任务中的注意力头重要性热力图

图 A1 展示了 Qwen3-8B 模型中全部 32 个注意力头在 36 个层级上的梯度重要性分布.

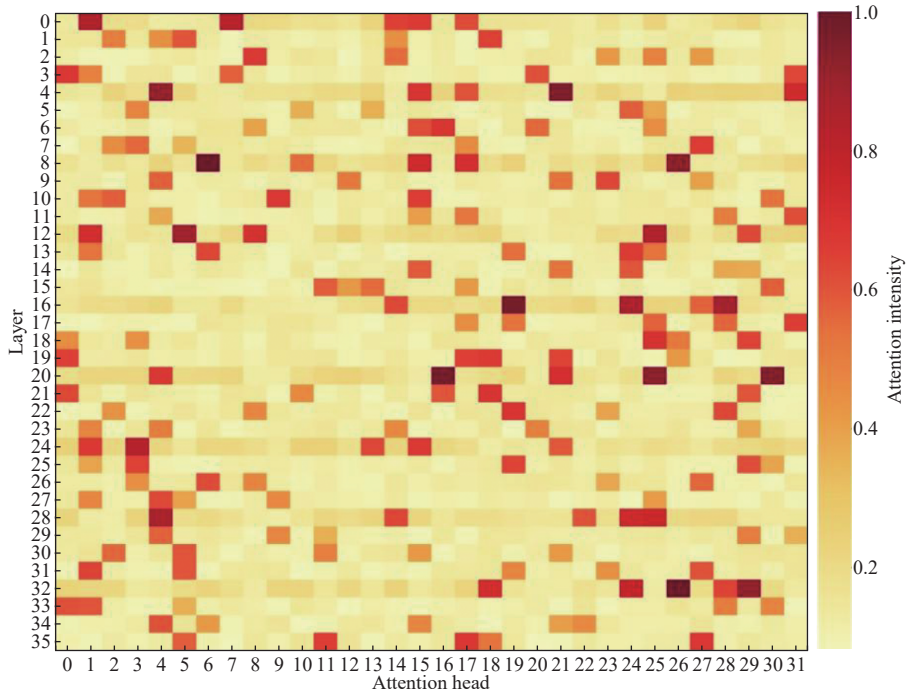


图 A1 Attention head 分布与专一性分析 (Qwen3-8B)

观察图 A1 可知, 与逻辑推理相关的注意力头激活呈现出显著的稀疏性 (sparsity) 与离散分布特征. 绝大多数区域处于低响应状态 (浅色), 仅有极少数特定的注意力头 (深红色热点) 在推理过程中表现出高显著性. 这些高重要性节点并未在某一特定层级形成连续的带状分布, 而是呈离散点状分布于不同层级中. 这一现象表明, 在逻辑推理回路中, 注意力机制主要通过少数具有高度功能专一性 (functional specificity) 的注意力头发挥作用. 推测这些专一性注意力头充当了“信息路由器”或“关键变量选择器”的角色, 负责从上下文中精准捕获逻辑前提与关键 token, 并将其定向输送至后续的 FFN 逻辑神经元, 从而在机制层面支持了前文所述的 Attention-FFN 协同计算模式.

(2) 逻辑任务对照表

正文中部分图表 (如热力图与相关性矩阵) 受限于排版空间采用缩写形式以优化可视化效果. 表 A1 详细列出了实验所涉及的所有逻辑推理子任务的完整命名与其缩写代码的对应关系. 这些任务涵盖了一阶逻辑 (FOL)、非单调逻辑 (NL) 及命题逻辑 (PL) 这 3 大主要类别, 缩写前缀 (如“FOL/”“NL/”“PL/”) 对应其所属的逻辑类型.

表 A1 逻辑任务对照表

全称	简写	全称	简写
first_order_logic/bidirectional_dilemma	FOL/BD	nm_logic/reasoning_about_exceptions_1	NL/RAE1
first_order_logic/constructive_dilemma	FOL/CD	nm_logic/reasoning_about_exceptions_2	NL/RAE2
first_order_logic/destructive_dilemma	FOL/DD	nm_logic/reasoning_about_exceptions_3	NL/RAE3
first_order_logic/disjunctive_syllogism	FOL/DS	nm_logic/reasoning_about_priority	NL/RAP
first_order_logic/existential_generalization	FOL/EG	propositional_logic/bidirectional_dilemma	PL/BD

表 A1 逻辑任务对照表 (续)

全称	简写	全称	简写
first_order_logic/hypothetical_syllogism	FOL/HS	propositional_logic/commutation	PL/CO
first_order_logic/modus_ponens	FOL/MP	propositional_logic/constructive_dilemma	PL/CD
first_order_logic/modus_tollens	FOL/MT	propositional_logic/destructive_dilemma	PL/DD
first_order_logic/universal_instantiation	FOL/UI	propositional_logic/disjunctive_syllogism	PL/DS
nm_logic/default_reasoning_default	NL/DRD	propositional_logic/hypothetical_syllogism	PL/HS
nm_logic/default_reasoning_irr	NL/DRI	propositional_logic/material_implication	PL/MI
nm_logic/default_reasoning_open	NL/DRO	propositional_logic/modus_tollens	PL/MT
nm_logic/default_reasoning_several	NL/DRS	—	—

(3) 逻辑任务跨域迁移效应的完整矩阵

鉴于正文篇幅限制, 表 5 仅展示了部分具有代表性的逻辑任务迁移结果. 表 A2 提供了从一阶逻辑 (FOL)、非单调逻辑 (NL) 到命题逻辑 (PL) 共计 25 个子任务间成对迁移实验的完整数据记录. 表 A2 中每一项包含两组数值: 上方数据为利用源任务 (行) 识别的神经元增强目标任务 (列) 后的准确率, 下方数据为抑制干预后的准确率. 该全量数据不仅为正文中关于“逻辑原子泛化性”的结论提供了翔实的统计学支撑, 也展示了模型在不同细粒度逻辑形式下的具体适应性差异, 供读者查阅特定任务对的详细性能表现.

表 A2 逻辑任务干预结果表

Task	FOL/BD	FOL/CD	FOL/DD	FOL/DS	FOL/EG	FOL/HS	FOL/MP	FOL/MT	FOL/UI	NL/DRD	NL/DRI	NL/DRO	NL/DRS
FOL/BD	0.825	0.825	0.775	0.950	1.000	0.950	1.000	0.825	0.900	0.875	0.775	0.750	0.800
	0.250	0.238	0.375	0.300	0.475	0.250	0.375	0.425	0.425	0.425	0.425	0.400	0.313
FOL/CD	0.800	0.788	0.763	0.750	1.000	0.950	0.975	0.750	0.875	0.925	0.775	0.725	0.788
	0.275	0.225	0.588	0.275	0.400	0.250	0.350	0.475	0.400	0.325	0.375	0.425	0.425
FOL/DD	0.800	0.800	0.775	0.900	1.000	0.950	0.975	0.875	0.900	0.975	0.825	0.700	0.800
	0.238	0.238	0.363	0.350	0.425	0.250	0.375	0.450	0.425	0.250	0.400	0.500	0.450
FOL/DS	0.813	0.788	0.750	0.850	1.000	0.963	0.925	0.700	0.900	0.925	0.725	0.725	0.788
	0.250	0.250	0.325	0.325	0.450	0.238	0.450	0.475	0.500	0.375	0.450	0.325	0.275
FOL/EG	0.825	0.800	0.750	0.775	1.000	0.925	0.900	0.700	0.925	0.825	0.675	0.800	0.788
	0.250	0.238	0.463	0.300	0.500	0.250	0.425	0.400	0.500	0.275	0.375	0.275	0.313
FOL/HS	0.800	0.800	0.763	0.825	1.000	0.950	1.000	0.775	0.925	0.900	0.800	0.725	0.775
	0.238	0.250	0.263	0.300	0.450	0.250	0.450	0.475	0.375	0.350	0.450	0.400	0.250
FOL/MP	0.838	0.813	0.763	0.850	1.000	0.950	0.925	0.725	0.925	0.825	0.675	0.725	0.800
	0.238	0.238	0.300	0.325	0.400	0.238	0.400	0.450	0.450	0.425	0.350	0.350	0.325
FOL/MT	0.825	0.788	0.763	0.850	1.000	0.950	0.975	0.800	0.900	0.875	0.775	0.800	0.775
	0.250	0.250	0.263	0.300	0.350	0.250	0.400	0.475	0.400	0.400	0.450	0.275	0.250
FOL/UI	0.788	0.788	0.788	0.825	1.000	0.938	0.925	0.775	0.950	0.825	0.725	0.800	0.763
	0.238	0.250	0.250	0.300	0.350	0.250	0.450	0.450	0.475	0.300	0.400	0.425	0.250
NL/DRD	0.825	0.800	0.763	0.925	1.000	0.975	1.000	0.750	0.900	0.975	0.850	0.775	0.800
	0.225	0.238	0.363	0.250	0.375	0.250	0.450	0.350	0.450	0.375	0.325	0.400	0.250
NL/DRI	0.825	0.788	0.763	0.800	1.000	0.938	0.975	0.750	0.925	0.875	0.725	0.800	0.788
	0.250	0.238	0.425	0.325	0.375	0.250	0.425	0.425	0.450	0.250	0.400	0.325	0.338
NL/DRO	0.813	0.800	0.800	0.800	1.000	0.963	0.975	0.750	0.900	0.850	0.725	0.775	0.775
	0.238	0.238	0.250	0.350	0.400	0.250	0.425	0.350	0.425	0.275	0.300	0.325	0.250
NL/DRS	0.838	0.788	0.813	0.800	1.000	0.950	0.950	0.750	0.900	0.975	0.750	0.700	0.775
	0.250	0.250	0.413	0.325	0.475	0.250	0.350	0.475	0.425	0.300	0.400	0.475	0.588
NL/RAE1	0.800	0.800	0.775	0.875	1.000	0.938	0.950	0.725	0.900	0.950	0.750	0.700	0.775
	0.213	0.250	0.650	0.325	0.450	0.250	0.425	0.475	0.500	0.300	0.375	0.400	0.350
NL/RAE2	0.825	0.800	0.763	0.800	1.000	0.938	0.925	0.725	0.925	0.825	0.725	0.800	0.788
	0.250	0.225	0.463	0.300	0.450	0.275	0.375	0.475	0.400	0.200	0.350	0.425	0.325
NL/RAE3	0.850	0.775	0.788	0.825	1.000	0.938	0.950	0.750	0.900	0.875	0.750	0.750	0.788
	0.238	0.250	0.250	0.425	0.475	0.250	0.400	0.475	0.475	0.350	0.450	0.425	0.225

表 A2　逻辑任务干预结果表 (续)

Task	FOL/BD	FOL/CD	FOL/DD	FOL/DS	FOL/EG	FOL/HS	FOL/MP	FOL/MT	FOL/UI	NL/DRD	NL/DRI	NL/DRO	NL/DRS
NL/RAP	0.825	0.788	0.763	0.850	1.000	0.950	0.975	0.700	0.900	0.925	0.850	0.725	0.800
	0.250	0.250	0.363	0.300	0.425	0.250	0.425	0.425	0.375	0.400	0.375	0.300	0.300
PL/BD	0.825	0.800	0.775	0.825	1.000	0.938	0.975	0.775	0.925	0.850	0.800	0.750	0.775
	0.250	0.250	0.488	0.375	0.475	0.250	0.400	0.375	0.450	0.450	0.375	0.375	0.250
PL/CO	0.800	0.800	0.763	0.800	1.000	0.950	0.975	0.725	0.925	0.850	0.750	0.775	0.775
	0.250	0.238	0.463	0.300	0.475	0.250	0.475	0.425	0.450	0.400	0.325	0.450	0.488
PL/CD	0.813	0.838	0.763	0.800	1.000	0.950	0.975	0.700	0.900	0.850	0.800	0.750	0.838
	0.250	0.225	0.550	0.250	0.350	0.238	0.375	0.450	0.425	0.300	0.375	0.375	0.263
PL/DD	0.788	0.788	0.763	0.875	1.000	0.975	1.000	0.850	0.925	0.950	0.825	0.750	0.788
	0.250	0.225	0.588	0.250	0.450	0.250	0.375	0.450	0.475	0.325	0.400	0.450	0.575
PL/DS	0.800	0.800	0.775	0.875	1.000	0.988	1.000	0.775	0.925	0.900	0.750	0.725	0.775
	0.238	0.238	0.425	0.250	0.375	0.250	0.400	0.475	0.425	0.350	0.400	0.425	0.500
PL/HS	0.813	0.813	0.788	0.800	1.000	0.913	0.900	0.750	0.925	0.800	0.675	0.725	0.775
	0.225	0.238	0.250	0.350	0.400	0.300	0.425	0.425	0.400	0.250	0.325	0.225	0.250
PL/MI	0.825	0.800	0.763	0.825	1.000	0.963	0.975	0.725	0.950	0.950	0.775	0.800	0.775
	0.450	0.250	0.575	0.300	0.375	0.363	0.450	0.475	0.425	0.375	0.425	0.375	0.263
PL/MT	0.838	0.788	0.763	0.775	1.000	0.938	0.925	0.775	0.950	0.850	0.750	0.800	0.775
	0.238	0.250	0.275	0.250	0.325	0.250	0.400	0.425	0.450	0.275	0.375	0.375	0.250

Task	NL/RAE1	NL/RAE2	NL/RAE3	NL/RAP	PL/BD	PL/CO	PL/CD	PL/DD	PL/DS	PL/HS	PL/MI	PL/MT	—
FOL/BD	0.788	0.675	0.725	0.600	0.788	0.938	0.800	0.788	0.875	0.988	0.763	0.700	—
	0.275	0.425	0.375	0.475	0.250	0.338	0.250	0.363	0.350	0.250	0.250	0.375	
FOL/CD	0.775	0.600	0.775	0.563	0.775	0.925	0.788	0.800	0.800	0.975	0.788	0.700	—
	0.638	0.475	0.400	0.475	0.300	0.550	0.238	0.538	0.300	0.263	0.263	0.475	
FOL/DD	0.775	0.650	0.775	0.575	0.788	0.925	0.775	0.825	0.925	0.963	0.763	0.700	—
	0.375	0.400	0.375	0.475	0.213	0.250	0.250	0.300	0.275	0.238	0.250	0.400	
FOL/DS	0.775	0.575	0.800	0.588	0.763	0.900	0.788	0.800	0.825	0.975	0.763	0.675	—
	0.375	0.450	0.325	0.488	0.263	0.238	0.250	0.300	0.325	0.238	0.250	0.400	
FOL/EG	0.775	0.675	0.775	0.588	0.763	0.850	0.775	0.813	0.775	0.963	0.763	0.725	—
	0.375	0.450	0.250	0.475	0.238	0.275	0.238	0.375	0.300	0.250	0.250	0.425	
FOL/HS	0.775	0.600	0.775	0.600	0.800	0.913	0.800	0.825	0.825	0.950	0.775	0.775	—
	0.263	0.450	0.300	0.475	0.250	0.263	0.250	0.325	0.425	0.225	0.250	0.400	
FOL/MP	0.775	0.625	0.775	0.563	0.788	0.875	0.825	0.800	0.800	0.950	0.763	0.725	—
	0.288	0.475	0.325	0.488	0.250	0.250	0.238	0.425	0.325	0.250	0.250	0.450	
FOL/MT	0.763	0.575	0.800	0.588	0.775	0.888	0.788	0.775	0.800	0.963	0.775	0.700	—
	0.263	0.400	0.350	0.475	0.300	0.488	0.250	0.263	0.400	0.225	0.263	0.375	
FOL/UI	0.763	0.625	0.800	0.575	0.788	0.900	0.775	0.800	0.825	0.925	0.763	0.700	—
	0.275	0.400	0.375	0.488	0.250	0.363	0.250	0.613	0.325	0.238	0.250	0.450	
NL/DRD	0.775	0.600	0.800	0.563	0.838	0.950	0.788	0.800	0.850	1.000	0.763	0.675	—
	0.250	0.475	0.425	0.475	0.238	0.375	0.225	0.363	0.325	0.275	0.250	0.475	
NL/DRI	0.775	0.625	0.800	0.625	0.763	0.938	0.800	0.825	0.800	0.950	0.763	0.650	—
	0.275	0.500	0.425	0.475	0.238	0.275	0.238	0.338	0.275	0.250	0.250	0.475	
NL/DRO	0.788	0.600	0.725	0.575	0.775	0.875	0.788	0.800	0.800	0.988	0.775	0.700	—
	0.300	0.450	0.350	0.488	0.250	0.250	0.238	0.250	0.350	0.238	0.250	0.425	
NL/DRS	0.800	0.650	0.750	0.563	0.763	0.925	0.788	0.825	0.825	1.000	0.775	0.725	—
	0.388	0.475	0.400	0.488	0.238	0.488	0.250	0.475	0.275	0.250	0.250	0.425	
NL/RAE1	0.775	0.650	0.775	0.575	0.763	0.938	0.775	0.800	0.875	0.963	0.775	0.700	—
	0.250	0.500	0.200	0.488	0.250	0.275	0.250	0.350	0.300	0.250	0.250	0.500	
NL/RAE2	0.775	0.675	0.800	0.588	0.763	0.888	0.788	0.788	0.775	0.938	0.763	0.700	—
	0.438	0.425	0.350	0.488	0.250	0.300	0.238	0.600	0.275	0.250	0.263	0.350	
NL/RAE3	0.775	0.625	0.800	0.563	0.775	0.888	0.800	0.800	0.775	0.963	0.775	0.675	—
	0.250	0.400	0.475	0.475	0.250	0.250	0.238	0.250	0.275	0.250	0.250	0.450	

表 A2 逻辑任务干预结果表 (续)

Task	FOL/BD	FOL/CD	FOL/DD	FOL/DS	FOL/EG	FOL/HS	FOL/MP	FOL/MT	FOL/UI	NL/DRD	NL/DRI	NL/DRO	NL/DRS
NL/RAP	0.775	0.625	0.775	0.588	0.775	0.913	0.775	0.800	0.850	0.963	0.775	0.650	—
	0.313	0.375	0.325	0.463	0.250	0.375	0.250	0.413	0.425	0.238	0.250	0.400	
PL/BD	0.788	0.600	0.750	0.700	0.800	0.900	0.800	0.800	0.775	0.975	0.750	0.725	—
	0.450	0.400	0.350	0.500	0.250	0.375	0.213	0.588	0.275	0.250	0.250	0.400	
PL/CO	0.763	0.575	0.800	0.575	0.775	0.950	0.788	0.800	0.800	0.975	0.775	0.675	—
	0.413	0.475	0.450	0.500	0.250	0.663	0.250	0.538	0.475	0.238	0.250	0.375	
PL/CD	0.775	0.625	0.775	0.563	0.788	0.925	0.800	0.775	0.800	0.975	0.763	0.800	—
	0.550	0.450	0.275	0.463	0.213	0.363	0.238	0.500	0.225	0.238	0.263	0.450	
PL/DD	0.788	0.625	0.775	0.675	0.763	0.963	0.788	0.838	0.900	0.963	0.775	0.725	—
	0.600	0.450	0.375	0.488	0.250	0.538	0.250	0.550	0.175	0.225	0.250	0.425	
PL/DS	0.775	0.650	0.800	0.600	0.775	0.875	0.775	0.813	0.825	1.000	0.763	0.725	—
	0.500	0.450	0.425	0.475	0.250	0.413	0.213	0.488	0.250	0.250	0.250	0.375	
PL/HS	0.788	0.575	0.725	0.563	0.788	0.875	0.775	0.800	0.750	0.900	0.775	0.600	—
	0.263	0.425	0.325	0.463	0.250	0.250	0.250	0.250	0.250	0.250	0.250	0.375	
PL/MI	0.775	0.650	0.775	0.575	0.775	0.938	0.788	0.788	0.800	0.975	0.775	0.650	—
	0.575	0.475	0.375	0.488	0.250	0.350	0.238	0.600	0.275	0.275	0.363	0.450	
PL/MT	0.775	0.625	0.725	0.588	0.775	0.838	0.775	0.800	0.800	0.963	0.775	0.675	—
	0.250	0.450	0.375	0.438	0.238	0.250	0.238	0.250	0.250	0.200	0.250	0.425	

作者简介

姜慧, 硕士生, 主要研究领域为自然语言处理, 推理大模型.

何世柱, 博士, 研究员, CCF 高级会员, 主要研究领域为自然语言处理, 知识工程, 推理大模型.

刘康, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 知识工程, 可解释分析.

赵军, 博士, 研究员, 博士生导师, CCF 高级会员, 主要研究领域为自然语言处理, 知识工程, 大语言模型.