# 基于关联神经元识别的知识编辑方法

**吴钰璋[1]，穆永誉[1]，王成龙[1]，何荠至，肖桐[1,2]，马安香[1]，张春良[1,2]，朱靖波[1,2*]**
[1]东北大学自然语言处理实验室，计算机科学与工程学院，东北大学，沈阳
[2]小牛翻译，沈阳
nightsnowy9@foxmail.com
{lixiaoyumu9,clwang1119}@gmail.com
qiaozhihe2022@outlook.com
{xiaotong,maanxiang,zhangchunliang,zhujingbo}@mail.neu.edu.cn

## 摘要

近年来，大语言模型展现出了从训练语料中存储并提取知识的优秀能力，但相应地，其可靠性也容易遭受训练语料中错误信息的破坏，进而产生信息过时、错误回复等问题。基于神经元识别的知识编辑方法通过在模型中识别并微调与目标知识相关的知识神经元，实现对模型内部知识的精确修改。然而，本文研究发现，知识的表达形式会显著影响知识神经元的识别结果，例如，现有神经元识别方法对于同一知识的不同表达形式识别得到的神经元集合平均重叠率只有21.86%。这就导致只对单一的表达形式进行知识编辑无法覆盖到与这个知识相关的所有神经元，所以现有知识编辑方法的鲁棒性往往较差。为了全面且准确地识别到与某一知识相关的所有神经元，本文设计了一种轻量级关联神经元识别器（Lightweight Associated Neuron Detector, LAND），通过学习不同表达形式的知识识别出的知识神经元集合之间的差异，从而在知识神经元识别的过程中，自动补全因表达形式差异而未被检出的知识神经元。实验结果表明，LAND方法能够将不同表达形式的文本识别出的知识神经元平均重叠率提升至96%以上，在不同句式的知识编辑成功率上较基线方法多提升了至多10.83个百分点。

**关键词：** 自然语言处理；大语言模型；知识编辑；神经元识别

# Knowledge Editing via Associated Neuron Identification

**Yuzhang Wu[1], Yongyu Mu[1], Chenglong Wang[1], Qiaozhi He, Tong Xiao[1,2], Anxiang Ma[1], Chunliang Zhang[1,2], JingBo Zhu[1,2*]**
[1]NLP Lab, School of Computer Science and Engineering,
Northeastern University, Shenyang, China
[2]NiuTrans Research, Shenyang, China
nightsnowy9@foxmail.com
{lixiaoyumu9,clwang1119}@gmail.com
qiaozhihe2022@outlook.com
{xiaotong,maanxiang,zhangchunliang,zhujingbo}@mail.neu.edu.cn

## Abstract

In recent years, large language models (LLMs) have demonstrated remarkable abilities in storing and retrieving knowledge from their training corpora. However, this capability also makes them vulnerable to inaccuracies within the training data, which can lead to issues such as outdated information and erroneous responses. Neuron-based knowledge editing methods attempt to precisely modify internal knowledge by identifying and fine-tuning specific knowledge neurons associated with a target fact. However, our study reveals that the expression form of knowledge significantly influences the identification of knowledge neurons. For instance, current neuron identification methods

* Corresponding Author

exhibit only a 21.86% average overlap in neuron sets when processing different expressions of the same knowledge. As a result, editing based on a single expression form fails to cover all relevant neurons, thereby undermining the robustness of existing knowledge editing techniques.To address this, we propose a Lightweight Associated Neuron Detector (LAND) that aims to comprehensively and accurately identify all neurons related to a specific piece of knowledge. LAND learns the differences among neuron sets identified from various expressions of the same knowledge, allowing it to automatically supplement undetected knowledge neurons during the identification process. Experimental results show that LAND increases the average overlap rate of knowledge neurons identified from varied expressions to over 96%, and improves the success rate of knowledge editing across different sentence structures by up to 10.83 percentage points compared to baseline methods.

**Keywords:** Natural Language Processing , Large Language Model , Knowledge Editing , Neuron Recognition

## 1 引言

近年来，随着深度学习技术的快速发展，大语言模型（Large Language Models, LLMs）在自然语言处理领域取得了突破性的进展(Chang et al., 2024)，其在记忆和提取事实知识方面的出色能力也引起了学术界的广泛关注。大语言模型通常通过海量文本语料预训练而成，这些文本数据中通常包含大量的事实性知识，而通过在这些语料上进行自监督学习，大语言模型能够在其参数空间中编码这些事实知识。Petroni等人(2019)提出，大语言模型可以被视为一种隐式的知识库，其中存储了从训练数据中提取的海量事实知识，通过填空任务等方式引导大语言模型生成文本，能够有效地激活和提取其存储的知识。

然而，大语言模型虽然在知识提取任务上展现出了显著的潜力，但在实际应用中仍存在一定局限性：一方面，大语言模型的预训练语料通常是从互联网收集的大量文本数据，而其中可能包含一些错误信息，导致模型在训练过程中学习到错误的知识；其次，知识本身也具有动态性和时效性，例如科学进展和社会变化会不断更新现有的知识，但大语言模型在预训练完成后，其参数空间中的知识通常是静态的，无法自动适应知识的更新，导致模型可能输出包含过时信息的内容。虽然在包含新知识的数据上进行小规模的有监督微调（Supervised Fine-Tuning, SFT）可以实现模型知识的更新，但由于大语言模型参数量巨大，频繁对模型进行SFT不仅训练代价极高，还容易引发灾难性遗忘现象(Kirkpatrick et al., 2017)，导致模型遗忘原有知识。此外，一些工作引入检索增强生成（Retrieval Augmented Generation, RAG）方法，利用外部知识源来缓解大语言模型中信息过时或不正确的问题(Zheng et al., 2023; Mitchell et al., 2022)。然而，由于RAG依赖于维护外部知识库来更新知识，因此它无法永久性地改变模型的内部知识。此外，随着外部知识库随时间的推移而扩展，可能会导致查询延迟和计算开销增加。

为了高效修正模型中的错误知识和过时信息，提升大语言模型的可靠性，许多研究工作致力于对大语言模型进行知识编辑（Knowledge Editing）(Zhang et al., 2024)。在众多的知识编辑方法中，Dai等人(2022)提出的知识神经元理论为解决大语言模型知识编辑问题提供了新的视角。该理论认为，一条特定的知识在大语言模型中的存储往往仅与若干关键神经元相关，这些神经元被称为知识神经元（Knowledge Neurons）。这一发现推动了基于神经元识别的知识编辑方法研究。与SFT相比，通过在大语言模型中识别出与特定知识密切关联的少量知识神经元，并对其进行针对性微调，实现知识的精准更新，既能避免全局参数调整带来的计算开销，又能最大程度保留模型的其他知识不被破坏。而与RAG方法相比，基于神经元的知识编辑方法能够对模型的内部知识进行永久性、轻量化的修改，不会影响其长期的推理性能。

然而，目前基于知识神经元识别的知识编辑方法仍存在一些缺陷，特别是对于基于Transformer架构(Vaswani et al., 2017)的自回归式语言模型而言，模型的训练目标主要是通过预测句子中的下一个词来学习语言规律，而非显式地存储结构化的知识，而这种训练方式会导致模型的知识提取受提示语影响极为严重。例如，当询问大模型“某明星的母亲是谁”时，

模型通常能够正确回答；然而，当询问同一个模型“这位母亲的孩子是谁”时，模型却可能难以给出正确答案。这一问题被广泛认为与训练数据的分布特性有关：在该大模型的训练语料中，形如“该明星的母亲是某某”的语料数量较多，而形如“该母亲的孩子是某某”则相对稀少。这种数据分布的不均衡性导致模型虽然能够学习到“该明星的母亲是”这一提示语的相关知识，但却难以有效学习“该母亲的孩子是”这一提示语的相关知识。

这种现象促使研究者们对大语言模型的知识存储机制进行了深入探究。研究表明(Niu et al., 2024)，知识的表示形式（如主动句与被动句、同义句式等）对知识神经元的识别影响极大。例如，“北京是中国的首都”和“中国的首都是北京”这两句文本虽然在语义上等价，但在大语言模型中可能识别出不同的知识神经元集合。本文的研究结果则进一步表明，对于描述同一知识但表达形式不同的两条文本，使用积分梯度归因法识别出的知识神经元集合平均重合率仅为21.86%。这种形式敏感性会导致基于不同表达形式识别出的知识神经元存在显著差异，进而影响知识编辑操作的鲁棒性：在大语言模型上执行知识编辑任务时，如果仅针对某一种表达形式（如“北京是中国的首都”）进行修改，模型可能无法同步更新其他形式的同一知识（如“中国的首都是北京”）。这一缺陷显著制约了基于知识神经元识别的知识编辑方法在实际应用中的可靠性。

为了缓解上述问题，本研究提出了一种关联知识神经元识别方法：在原始大语言模型的基础上，引入一个轻量级的关联神经元识别器（Lightweight Associated Neuron Detector, LAND），旨在对于输入的单一形式文本，通过学习不同句法表达在模型中识别的知识神经元集合之间的差异，自动识别出因表达形式差异而未被检出的知识神经元，并在执行知识编辑任务时，对这些神经元同时进行编辑操作，从而缓解大语言模型在知识神经元识别上的形式敏感性问题，提升知识编辑的鲁棒性。实验结果表明，LAND方法能够将不同表达形式的文本识别出的知识神经元平均重叠率提升至96%以上，在不同句式的知识编辑成功率上较基线方法至多提升了10.83个百分点，为知识编辑技术的实际应用提供了可靠的方法支持。

## 2 相关工作

大语言模型从预训练语料中记忆并提取事实知识的能力已得到广泛证明，但这一能力的内部工作机制仍旧不够明晰。因此，近年来，许多研究试图结合基于神经元的可解释性分析方法(Sajjad et al., 2022)，通过在大语言模型中识别与特定事实知识相关的神经元来解释这一能力。Dai等人(2022)在该领域做出了开创性的贡献。他们的研究表明，在BERT(Devlin et al., 2019)模型中，一个特定的事实知识可以定位到2-5个前馈神经网络（FFN）层中的神经元，且当手动抑制这些神经元时，大语言模型对相关事实知识的检索能力出现显著下降，失去正确回答与该知识相关问题的能力。基于这一实验现象，他们提出了知识神经元假说：大语言模型将事实知识存储在FFN层的神经元中，每个事实知识仅与几个特定的神经元相关，这些神经元被称为“知识神经元”。

这一假设在提出后迅速得到学术界关注。基于这一假设，一系列后续研究进一步发掘了许多与知识神经元有关的现象。Meng等人(2022)发现在GPT等自回归式架构的模型中，事实信息由中层FFN层的神经元进行检索和提取，而注意力模块则负责将提取出的事实信息复制到顶层。Chen等人的研究(2024a)在大语言模型中发现了两类神经元：语言无关知识神经元能够以一种跨语言的形式存储事实知识，退化知识神经元则表明一个事实知识可能由多个独立的神经元簇存储，这可能与训练阶段采用dropout范式有关。后续工作(Chen et al., 2024b)对退化知识神经元又进行了进一步探究。Hu等人的研究(2024)则表明，与负责逻辑推理的神经元不同，不同语言的知识在大语言模型中是分散存储的。IRCAN(Shi et al., 2024)则利用知识神经元识别来解决模型内在知识与上下文中知识的冲突。

另一方面，部分学者对这一理论假设仍持保留态度。Niu等人的研究(2024)对知识神经元理论提出了质疑，认为该理论过度简化了大语言模型内部对事实知识的处理机制。论文指出，大语言模型的FFN层中主要编码的是token的表达模式，而非传统认识论所定义的“知识”概念。尽管这些模式能够在一定程度上反映语法或语义层面的规律性特征，但其本质上更倾向于编码语言特征的统计性规律。例如，“scattered”一词因其在语料中与复数名词高频共现，而被大模型错误识别为复数限定词。

上述研究从不同角度丰富了知识神经元理论。然而，上述工作均沿用积分梯度归因法进行神经元的识别，未能注意到知识神经元识别过程中知识表达形式对神经元识别的严重影响。本

工作首次深入分析了这一影响，并提出了一种轻量级的关联神经元识别器，有效地缓解了此问题，这是与现有工作最主要的区别和贡献。

## 3 知识神经元识别

### 3.1 数据集及增强方法

为深入探究知识的表达形式对知识神经元识别结果的影响，本文对大语言模型中识别出的知识神经元进行了系统的分析实验。本文的实验基于PARAREL数据集进行。PARAREL数据集是基于T-Rex 数据集(ElSahar et al., 2018)，由多名人类语言学专家构建的一个事实关系数据集，描述了27 738对实体之间37种不同类别的关系。该数据集由两部分组成，分别是实体数据（vocab）和模板数据（graph）。实体数据部分存储了实体对及其关系的具体信息，每条数据包括一个主语实体（Sublabel）和一个宾语实体（Objlabel），表示这两个实体之间存在某种特定关系；模板数据部分则由多组句式模板构成，每组描述一种[X]和[Y]之间的关系，包含多个句式模板，至少1条，至多20条。表1展示了数据集中的部分模板示例，其中每个关系仅展示前三条模板。

| Relations | Template 1 | Template 2 | Template 3 |
|---|---|---|---|
| P39(positions) | [X] has the position of [Y]. | [X], who has the position of [Y]. | [X], whose position is that of [Y]. |
| P1303(musician) | [X] plays [Y] . | [Y] player [X] | [X] plays the [Y] |
| P27(citizenship) | [X] is [Y] citizen. | [X] is a citizen of [Y]. | [X], who is a citizen of [Y]. |

表 1: PARAREL数据集中模板数据部分示例

通过组合PARAREL数据集中不同的实体数据对和模板，将模板中的[X]替换为实体数据中的Sublabel，将[Y]替换为空，将对应的Objlabel作为问题的答案，共可构建出253 448句包含事实知识的问题文本。表2展示了部分构建出的问题文本的示例。

| Relations | Query | Answer |
|---|---|---|
| P39(positions) | Sheila Dixon has the position of __ | mayor |
| P1303(musician) | Frederick Grinke plays __ | violin |
| P27(citizenship) | Rubens Barrichello is a citizen of __ | Brazil |

表 2: 适配后的PARAREL数据集部分示例

PARAREL数据集具有丰富的表达形式多样性，使其适合研究大语言模型中知识表达形式对知识神经元识别的影响。然而，现今的大语言模型大多基于自回归的生成式模型架构：与基于掩码语言建模的模型不同，在进行填空式问答时，生成式大语言模型要求待预测的答案必须位于句末。然而，PARAREL数据集在征集人类专家编写模板时，并未对[X]或[Y]的位置进行严格限制，数据集中许多模板的目标实体（即需要预测的答案）并不位于句子的末尾，导致PARAREL数据集难以直接适配生成式模型的知识提取任务。

为了在自回归式架构的模型上有效应用PARAREL数据集，本文对PARAREL数据集中的模板数据进行了适配和增强。首先，本研究对模板进行了一次筛选，仅保留以[Y]为结尾的模板数据；之后，由于筛选后某些关系的模板数量大幅减少，对于模板数量不足10条的关系，本研究利用大语言模型（如GPT-4）的仿写能力，生成表达同一关系但句式不同的模板，从而扩展模板数据集的覆盖范围和表达能力。而对于模板数量已经超过10条的关系，为了保证数据集中不同关系的公平性，本研究仅保留前10条模板。此过程中具体使用的提示语见附录A。

### 3.2 积分梯度归因法

为了与前人工作(Dai et al., 2022)保持一致，本文使用积分梯度归因法来进行大语言模型中知识神经元的识别。积分梯度归因法(Sundararajan et al., 2017)的主要思想是：通过对模型的输入进行微小扰动，监测输入值变化时特定神经元的激活值变化，用这一变化的梯度积分来量化神经元在模型预测过程中的贡献，从而识别模型中与特定知识紧密相关的神经元。具体而言，给定一条询问语句$x$（如“中国的首都是__”）及该询问对应的正确答案$y^*$（如“北京”），预训练语言模型正确预测出答案$y^*$的概率为：

$$\mathrm{P}_x(\hat{w}_i^{(l)}) = p(y^*|x, w_i^{(l)} = \hat{w}_i^{(l)}) \tag{1}$$

其中，$w_i^{(l)}$表示第$l$层FFN的第$i$个神经元，$\hat{w}_i^{(l)}$则表示扰动时赋予该神经元的值。在此基础上，积分梯度归因法计算每个神经元的归因分数（Attribution Score）的公式为：

$$\mathrm{Attr}(w_i^{(l)}) = \overline{w}_i^{(l)} \int_{\alpha=0}^{1} \frac{\partial \mathrm{P}_x(\alpha \overline{w}_i^{(l)})}{\partial w_i^{(l)}} d\alpha \tag{2}$$

这一计算公式的核心理论依据为：对于一条包含给定知识的询问，如果模型中的某一神经元与这一知识高度相关，那么在模型预测正确答案的过程中，该神经元激活强度的微小变化将显著影响正确答案的预测概率，即预测概率对于该神经元输出值的变化梯度较大。在这一公式中，归因分数Attr$\Phi w_i^{\Phi l \Psi}\Psi$通过累积神经元输出值变化对模型输出概率的影响，使得与给定知识强相关的神经元得到较高的归因分数。因此，归因分数可有效衡量模型中的神经元$w_i^{\Phi l \Psi}$与给定知识的相关程度。在实际应用中，由于难以直接计算连续积分，为平衡计算效率与精度，本研究采用黎曼和离散近似策略，来近似计算积分梯度法的归因分数，具体公式为：

$$\mathrm{Attr}(w_i^{(l)}) \approx \frac{\overline{w}_i^{(l)}}{m} \sum_{k=1}^{m} \frac{\partial \mathrm{P}_x\left(\frac{k}{m} \overline{w}_i^{(l)}\right)}{\partial w_i^{(l)}} \tag{3}$$

其中m为离散化采样时选取的路径上的近似采样步数。此外，在选取知识神经元的判定上，为了平衡不同模型的表现，本研究取阈值t（0 1之间的实数）为所有神经元的归因分数中最大归因分数的t倍，神经元的归因分数高于最大归因分数t倍的神经元被识别为模型中与输入文本相关的知识神经元。

### 3.3 知识神经元识别存在的问题

本小节深入探究了知识的不同表达形式对知识神经元识别效果的影响。Geva等人的研究(2021)表明，在基于Transformer架构的大语言模型中，虽然注意力模块和前馈网络模块中均存在神经元结构，但由于知识神经元通常被认为存在于模型的FFN层中。本研究也基于这一理论基础进行探究，所做实验仅关注FFN层的神经元，暂不考察注意力模块中的神经元活动。

实验所使用数据集为增强后的PARAREL数据集，按照9:1的比例进行随机划分，形成训练集与测试集。为了方便与后续实验结果对比，本部分实验仅使用测试集进行现象的观察。本研究在多个大语言模型上进行了实验，所有实验均在配备8张NVIDIA TITAN V显卡的服务器上完成。实验采用的模型及其超参数配置如表所示。表中B表示采样时并行推断批次的大小。

| Model | t | m | B |
|---|---|---|---|
| GPT-2 | 0.3 | 20 | 20 |
| LLaMa3.2-1B | 0.2 | 20 | 10 |

表 3: 实验超参数设置

为探究表达同一知识但表达形式不同的句子在大语言模型中识别出的知识神经元的差异，对于数据集中每组表达同一语义知识的10条文本，本研究设计了如下量化指标：

（1）$N_s$：单句文本识别得到的知识神经元集合中，神经元的平均个数。

（2）$N_u$：10句文本识别得到的知识神经元集合的并集中，神经元的平均个数。

（3）$N_{so}$和$R_{so}$：任意两句语义相同但表达形式不同的文本识别得到的两个知识神经元集合中，知识神经元的重叠数及重叠率。重叠率越低，表示知识神经元的识别结果受文本表达形式的影响越大。

（4）$N_{uo}$和$R_{uo}$：单句文本识别得到的知识神经元集合，与对应的10句文本得到的神经元并集中，知识神经元的平均重叠数及重叠率。重叠率越低，表示10条文本识别得到的神经元集合之间的差异越大。

其中，$R_{so}$和$R_{uo}$的计算公式如下：

$$R_{so} = \frac{N_a \cap N_b}{N_a \cup N_b} \tag{4}$$

$$R_{uo} = \frac{N_s}{N_u} \tag{5}$$

第二十四届中国计算语言学大会论文集，第733页-第744页，济南，中国，2025年8月11日至14日。

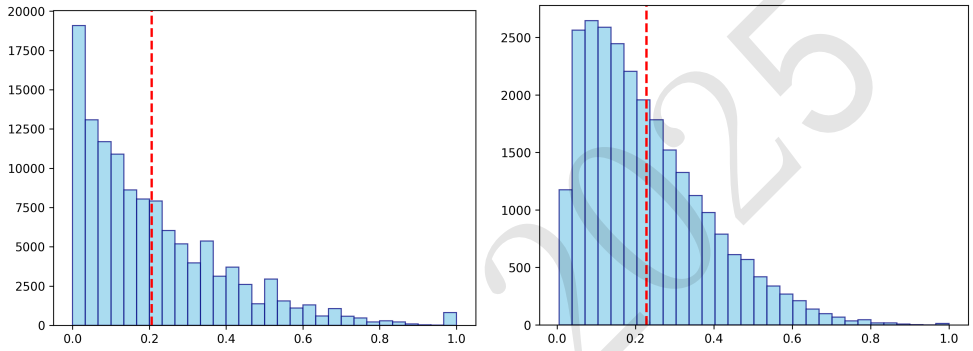
式中，$N_a$和$N_b$分别表示两条不同文本识别得到的知识神经元集合。

基于上述实验设置，实验结果如表4所示。实验结果表明，在GPT-2和LLaMa3.2-1B两个大语言模型中，任取表达相同知识但形式不同的两条文本，所识别出的知识神经元集合平均重叠率分别为20.97%和22.76%，任意一条文本识别的知识神经元与10条文本识别的神经元并集的平均重叠率分别为22.80%和22.18%。这一现象充分表明，具有不同表达形式的知识文本在大语言模型中激活的知识神经元集合具有显著差异。

| **Model** | $N_s$ | $N_u$ | $N_{so}(R_{so})$ | $N_{uo}(R_{uo})$ |
|---|---|---|---|---|
| GPT-2 | 23.79 | 106.97 | 6.87(20.97%) | 21.88(22.80%) |
| LLaMa3.2-1B | 16.50 | 90.88 | 3.82(22.76%) | 14.58(22.18%) |

表 4: 神经元重叠率实验结果

为了深入揭示该现象，本研究基于GPT-2模型的实验数据，绘制了知识神经元重叠率的频数分布直方图，如图1所示。子图（a）呈现了GPT-2模型中随机句对间神经元重叠率的频数分布，子图（b）则展示了单句与多句并集间神经元重叠率的频数分布，并用红色虚线标识出了平均值（即$R_{so}$）的位置。直方图显示，绝大多数句对的知识神经元重叠率显著低于平均值，呈现出明显的长尾分布特征。换言之，仅有少数不同表达的句子激活的神经元集合高度相似，大部分句对识别得到的神经元集合中相同的神经元占比不足20%。




(a) 任两句间神经元重叠率　　(b) 单句与并集间神经元重叠率

图 1: 知识神经元重叠率的频数分布直方图

为了更直观地展示该现象，本研究基于LLaMa3.2-1B模型的实验数据，选取了其中一个典型实例进行深入剖析，如表5所示。该案例展示了表达“Michael Maleinos的出生地”这一语义知识的三种不同表达形式在大语言模型中识别出的知识神经元集合。前两个句子虽然分别采用“was born in” 和“comes from” 两种不同表达形式，但其表达的语义知识一致，因此两个句子识别出的知识神经元集合存在显著交集，重叠率为40.91%（9/22）；然而，第三个句式与第二个句式仅存在时态差异（“comes”和“came”），识别得到的知识神经元集合却完全不同，体现了同一知识的不同表达形式在模型中激活的神经元集合的差别。

| **Sentence** | **Neurons** |
|---|---|
| Michael Maleinos was born in __ | [(3, 3604), (11, 1415), (12, 2684), (13, 3437), (14, 79), **(14, 185)**, (14, 2336), **(14, 3861)**, **(14, 5250)**, **(14, 5845)**, **(14, 6421)**, (14, 6516), (14, 6648), **(14, 6812)**, **(14, 6882)**, **(14, 7047)**, **(14, 7072)**] |
| Michael Maleinos comes from __ | [(1, 1976), (4, 7411), (13, 6286), **(14, 185)**, (14, 1975), **(14, 3861)**, (14, 4071), **(14, 5250)**, (14, 5705), **(14, 5845)**, **(14, 6421)**, **(14, 6812)**, **(14, 6882)**, **(14, 7047)**, **(14, 7072)**] |
| Michael Maleinos came from __ | [(15, 647), (15, 1594), (15, 2515), (15, 2897), (15, 4790), (15, 5237), (15, 5243), (15, 5407), (15, 7379), (15, 7827)] |

表 5: 知识神经元重叠率案例分析，表中数据格式为（层号，神经元序号）

## 4 方法

### 4.1 关联神经元识别方法

为了减轻不同句式引起的识别结果差异，缓解输入文本表达形式对知识神经元识别过程的影响，本研究设计了一种轻量级关联神经元识别器（Lightweight Associated Neuron Detector,

LAND）。该识别器基于轻量神经网络模型架构，通过学习不同表达形式的文本从大语言模型中识别出的知识神经元集合之间的差异，对识别结果进行优化，从而拉近不同表达的文本识别出的知识神经元集合，增强知识编辑的鲁棒性。

LAND模型的输入为单条文本经大语言模型初步识别后生成的知识神经元集合，输出则是经过模型优化后更具鲁棒性的知识神经元集合。因此，为了让模型学习不同表达形式的文本在模型内部神经元的激活模式，本研究依据增强后的PARAREL数据集，将数据集中描述同一关系的10条不同表达形式的文本输入到大语言模型中，获取10条文本各自识别出的知识神经元集合，之后将这些集合取并集，作为模型的优化目标。

然而，尽管这一设计能够有效地让模型建立从单条文本激活神经元到多条文本激活神经元的映射，但在实际应用中存在三点主要缺陷：首先，并集中的神经元数量过多，而一次编辑过多神经元会影响知识编辑操作的局部性；其次，大部分神经元仅被某条特殊表达识别到，这些神经元可能是模型中用于处理特殊表达形式的神经元，与核心知识的存储和提取并无紧密关联；最后，研究发现，并集中存在部分对于任何输入都有高概率被识别到的神经元，该部分神经元可能是大语言模型中的通用神经元，与句子表达的知识语义无关。这些缺陷都会对后续知识编辑操作的效果造成影响。

为了缓解上述问题，本研究在获取不同表达形式的文本识别出的知识神经元并集后，通过两步操作对该并集进行优化：首先，参考输出以每个神经元被识别到的次数作为该神经元的权重，引导LAND模型对多次出现的知识神经元赋予更高的关注，并对LAND模型的输出设定一个筛选阈值P，仅保留输出中权重超过P的知识神经元，从而大幅减少神经元数量；之后，随机构建一些与输入句表达形式相似，但包含的知识不同的文本，获取其在大语言模型中识别出的神经元，将这些神经元从集合里去除，从而筛除语义无关的神经元，留下与知识强相关的神经元。

综上所述，LAND模型的整体训练流程如图2所示。该流程可以总结为以下四个步骤：

（1）获取模型输出：对于包含待编辑知识的输入文本，采用积分梯度归因法在大语言模型中进行知识神经元的识别，并将其输入LAND模型，获取LAND模型优化后的输出；

（2）计算参考输出：向大语言模型输入数据集中描述同一关系的10个句子，提取各句式对应的知识神经元集合，构建其并集，并根据出现次数赋予权重；

（3）筛除反例神经元：构造反例数据并获取其对应的知识神经元集合，通过集合运算剔除步骤（2）所得并集中与反例集合的交集部分，确保保留的神经元具有语义相关性；

（4）计算Loss并更新模型参数：计算步骤（1）的输出和步骤（3）得到的参考输出之间的均方误差（MSE Loss），并通过反向传播更新LAND模型参数。

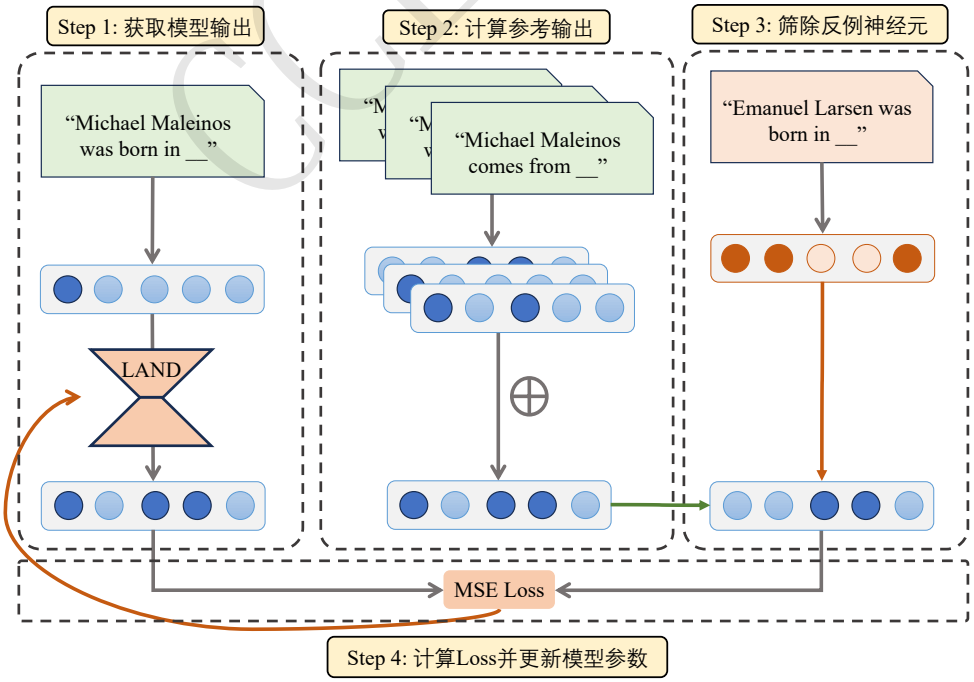



图 2: LAND模型训练流程示意图

### 4.2 LAND模型架构设计

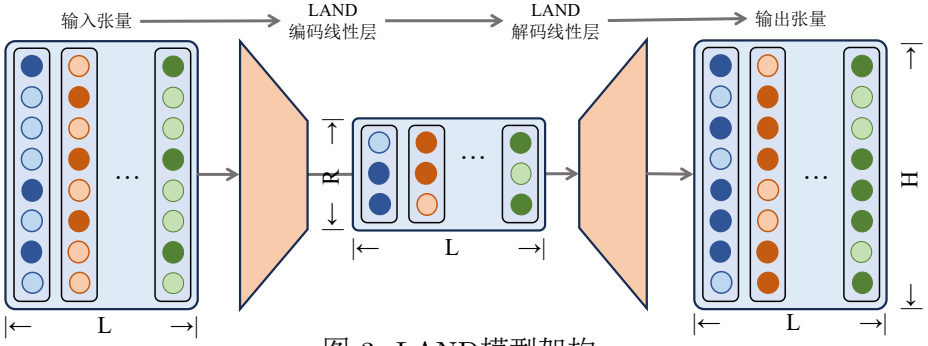



图 3: LAND模型架构

LAND模型的详细架构设计如图3所示。LAND模型的输入和输出均为大语言模型中识别出的知识神经元集合的张量表示。设原始的大语言模型共有L层，每层有H个神经元，那么可以用维度为[L,H]的0-1张量表示模型中知识神经元的位置。对于张量中的每个位置而言，当该位置取值为1时，表示对应神经元为与文本所含知识相关的知识神经元，取值为0则表示是无关神经元。以LLaMa3.2-1B模型为例，其架构包含16个网络层，每层神经元个数为8192，故对应的LAND模型输入输出均为16×8192的张量。

在对知识神经元的处理上，识别器采用分层处理的方式，对各层的神经元进行独立处理。考虑到单条知识对应的有效知识神经元相比神经元总数通常占比较小，为提升计算效率，LAND模型从自编码器架构的思想出发，基于低秩分解的轻量化架构，设计了两个核心线性变换层：

（1）编码线性层：实现从高维空间到低维空间的映射，将输入张量从[L,H]降维至[L,R]。该层权重矩阵维度为[H,R]，通过降维操作降低计算复杂度；

（2）解码线性层：完成从低维空间到原始维度的重建，将中间表示从[L,R]恢复至[L,H]。该层权重矩阵维度为[R,H]，实现重构知识神经元表示矩阵。

这种双线性低秩架构在保证关键神经元识别效果的前提下，通过降维-重构机制有效控制了计算开销。

LAND模型训练的目标为最小化单句识别的知识神经元集合张量与参考输出张量之间的差异。参考输出张量的生成过程依照如下三个步骤：

（1）初始化张量：构建维度为[L,H]的零值张量作为基础矩阵；

（2）知识神经元并集构建：依据4.1小节中介绍的方法，获取用于参考的神经元集合；

（3）神经元频率统计：遍历参考神经元集合，在张量中对相应位置进行累加计数。

## 5 实验

本研究选取了GPT-2和LLaMa3.2-1B两个大语言模型作为实验对象。针对不同模型在层数和每层神经元个数上的差异，本研究对LAND的关键超参数进行了适应性调整，具体参数配置详见表6。实验数据采用增强后的PARAREL数据集，仅使用训练集数据进行LAND模型的训练。模型在训练集上进行了3个epoch的训练，batchsize取8，学习率取1e-4。

| 模型 | L | H | R | 总参数量 |
|---|---|---|---|---|
| GPT-2 | 12 | 3072 | 64 | 4,718,592 |
| LLaMa3.2-1B | 16 | 8192 | 64 | 16,777,216 |

表 6: LAND转换器模型大小

### 5.1 神经元重叠率变化实验

表7展示了经LAND模型优化后，模型中依据文本识别得到的知识神经元数量及其重叠率。其中，$N_s$和$N_{so}$延续表4的定义，分别表示单句激活神经元平均数和任意两句激活神经元集合的神经元重叠数；$N_L$和$N_{Lo}$则对应经过LAND转换后的相应指标。实验结果显示，经过LAND转换处理后，不同句式的提示语激活的知识神经元集合之间的重叠率显著提升至96%以上，充分

证明LAND模型能够有效缩小不同表达形式的提示语激活神经元集合之间的差异，显著提高知识神经元识别的鲁棒性。

| Model | $N_s$ | $N_L$ | $N_{so}$ | $N_{Lo}$ |
|---|---|---|---|---|
| GPT-2 | 23.79 | 27.93 | 6.87(20.97%) | 25.02(96.58%) |
| LLaMa3.2-1B | 16.50 | 20.43 | 3.82(22.76%) | 20.18(97.66%) |

表 7: 神经元重叠率实验结果

为了更具体地展示LAND的优化效果，本研究以LLaMa3.2-1B模型的实验数据为例，选取了一个经LAND转换后的具体实例进行实际展示，详见表8。经过LAND处理后，三个语义相近的输入语句经LAND处理后识别出的知识神经元集合仅有一个神经元不一致（即，神经元（14, 2536）未被第3条文本识别到），展示了LAND模型在处理具有不同表达的语句时的稳定性。此外，表中加粗的神经元在表5中的前两行同样出现，进一步证实了LAND模型在转换过程中保持了良好的前后一致性。

| Sentence | Neurons |
|---|---|
| Michael Maleinos was born in __ | [(0, 231), (0, 7298), (1, 1976), (8, 4180), (9, 349), (10, 2308), (11, 177), (13, 3193), (13, 6286), **(14, 185)**, (14, 368), (14, 2536), (14, 4293), **(14, 5250)**, **(14, 5845)**, (14, 6280), **(14, 6421)**, (14, 6516), **(14, 6812)**, **(14, 7072)**, (15, 7258)] |
| Michael Maleinos comes from __ | [(0, 231), (0, 7298), (1, 1976), (8, 4180), (9, 349), (10, 2308), (11, 177), (13, 3193), (13, 6286), **(14, 185)**, (14, 368), (14, 2536), (14, 4293), **(14, 5250)**, **(14, 5845)**, (14, 6280), **(14, 6421)**, (14, 6516), **(14, 6812)**, **(14, 7072)**, (15, 7258)] |
| Michael Maleinos came from __ | [(0, 231), (0, 7298), (1, 1976), (8, 4180), (9, 349), (10, 2308), (11, 177), (13, 3193), (13, 6286), **(14, 185)**, (14, 368), (14, 4293), **(14, 5250)**, **(14, 5845)**, (14, 6280), **(14, 6421)**, (14, 6516), **(14, 6812)**, **(14, 7072)**, (15, 7258)] |

表 8: LAND优化效果案例分析，表中数据格式为（层号，神经元序号）

### 5.2 知识神经元增强实验

受Dai等人(2022)实验设计的启发，本小节首先进行了知识神经元增强实验，即在在模型的推理过程中，将事先识别出的知识神经元的激活值提升至原激活值的200%，观察增强神经元后正确答案token输出概率的变化情况。此实验设计的理论依据在于：若本章节提出的方法能够准确识别出模型中与某一事实知识相关的神经元，增强这些神经元则应使正确答案token输出概率显著提升；反之，若识别效果不够准确，概率提升幅度将相对有限。

为量化评估神经元增强效果，本研究引入目标token输出概率的改变率（Change Rate, $CR$）作为评价指标，计算公式为：

$$CR = \frac{P_{\text{After}} - P_{\text{Before}}}{P_{\text{Before}}} \tag{6}$$

式中，$P_{\text{Before}}$和$P_{\text{After}}$分别表示知识神经元增强前后正确答案token的输出概率。实验对比了三种不同的知识神经元识别方法，分别是原始的积分梯度法（Baseline）、基于多句式交叉验证的改进方法（KN-refine）(Dai et al., 2022)，以及本文提出的关联神经元识别方法（LAND）。为验证本文提出的方法能否减轻句式变化对知识编辑的影响，本研究设置了三项实验：

（1）Same：神经元提取过程中与知识编辑测试过程中使用相同文本；

（2）Related：提取与测试使用语义等价但句式不同的文本；

（3）Unrelated：提取与测试使用语义不同的文本。

基于上述实验设计，表9展示了神经元增强实验的结果。

实验结果显示，Related组的CR指标普遍低于Same组，即如果用于识别神经元的句子与真正进行知识增强实验测试的句子采用不同句式表达时，增强知识神经元对正确答案的输出概率的提升效果会衰减，再次验证了不同句式表达会激活不同神经元的理论假设；而在三种识别方法中，LAND方法的性能提升最为显著，充分证明本文提出的关联神经元识别算法能够更精准地定位知识神经元。

| Model | Method | Same | Related | Unrelated |
|---|---|---|---|---|
| GPT-2 | Baseline | 2.48 | 2.23 | 1.74 |
| | KN-refined | 3.10 (+0.62) | 2.53 (+0.30) | 2.43 (+0.69) |
| | LAND | 3.21 (+0.73) | 2.44 (+0.21) | 2.44 (+0.70) |
| LLaMa3.2-1B | Baseline | 6.24 | 5.48 | 4.10 |
| | KN-refined | 7.02 (+0.78) | 6.01 (+0.53) | 6.00 (+1.90) |
| | LAND | 17.30 (+11.06) | 16.40 (+10.92) | 14.04 (+9.94) |

表 9: 知识神经元增强实验

### 5.3 知识编辑实验

为进一步验证该方法在知识编辑任务中的实际应用效果，本研究设计了基于知识神经元识别的知识编辑实验。对于每个识别出的知识神经元，本实验对大语言模型FFN模块中对应的值向量进行调整，从原始值向量中减去目标token $t$的向量表示，并加入新token $t'$的向量表示。该操作的数学表达式为：

$$FFN_i^{val} = FFN_i^{val} - \lambda_1 t + \lambda_2 t' \tag{7}$$

式中，$FFN_i^{val}$为第i个神经元的输出值，$\lambda_1$和$\lambda_2$为调节系数，用于平衡编辑过程中的向量调整幅度（为了与前人工作保持一致，本实验中均取2）。本实验计算了编辑前后模型回答问题的平均正确率（Success Rate, $SR$），计算公式为：

$$SR = \frac{1}{N}\sum_{i=1}^{N} I(y_i = \hat{y}_i) \tag{8}$$

式中，N表示问题总数，$y_i$是第i个问题的真实答案（或标签），$\hat{y}_i$是模型对第i个问题的预测答案，$I\Phi \cdot \Psi$是指示函数，当括号内的条件成立时取值为1，否则为0。本实验的实验组设计与5.2小节的知识神经元增强实验保持一致。实验结果见表10。

| 模型 | | Same | Related | Unrelated |
|---|---|---|---|---|
| GPT-2 | Before | 7.13 | 4.99 | 0.10 |
| | Baseline | 14.98 (+7.85) | 11.69 (+7.70) | 1.30 (+1.20) |
| | KN-refined | 20.49 (+13.36) | 17.45 (+12.46) | 3.40 (+3.30) |
| | LAND | 25.81 (+18.68) | 21.57 (+16.58) | 2.86 (+2.76) |
| LLaMa3.2-1B | Before | 15.31 | 11.94 | 0.18 |
| | Baseline | 16.65 (+1.34) | 12.67 (+0.73) | 0.97 (+0.79) |
| | KN-refined | 16.76 (+1.45) | 12.70 (+0.76) | 1.01 (+0.83) |
| | LAND | 21.75 (+6.44) | 18.24 (+6.30) | 1.44 (+1.26) |

表 10: 知识编辑实验结果

实验结果表明，相较于基准方法Baseline和KN-refine，使用LAND方法进行神经元识别，模型进行知识编辑后回答正确率提升的幅度最为显著。这一结果充分验证了本文提出的神经元识别优化方法的有效性。

## 6 结论

本研究深入分析了基于神经元识别的知识编辑方法受知识表达形式影响严重的问题，并提出了一种基于关联神经元识别的知识编辑方法，有效缓解了该问题。首先，本研究基于PARAREL数据集，提出了一种面向自回归架构模型的适配方法，并通过大语言模型辅助的数据增强技术，显著提升了数据集的多样性和覆盖范围。在此基础上，实验验证了大语言模型在知识提取任务中对知识表达形式的敏感性，即同一知识的不同表达形式会激活不同的神经元集合，揭示了现有知识编辑方法的局限性。在此基础上，本研究提出了一种轻量级关联神经元识别器（LAND），能够将不同句式表达下的神经元平均重叠率提升至96%以上。此外，本研究通过神经元增强实验和知识编辑实验，验证了LAND在知识编辑任务中的有效性。实验结果显示，基于LAND的知识编辑方法在正确答案输出概率和回答准确率上均取得了显著提升，有

效缓解了知识表达形式对知识编辑效果的干扰。在未来的工作中，我们也计划将我们的方法推广到更多不同类型和规模的大模型上进行验证。

## 致谢

本文得到以下项目的支持：国家科学基金（项目编号：No.62276056，U24A20334）、云南省基础研究发展计划项目（项目编号：No.202401BC070021）、云南省重大科技专项计划（项目编号：202502AD080014）、高校学科引进人才计划111（项目编号：No.B16009）。作者们还对匿名评审专家给予的宝贵建议表示衷心的感谢。

## 参考文献

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3):39:1–39:45.

Yuheng Chen, Pengfei Cao, Yubo Chen, Kang Liu, and Jun Zhao. 2024a. Journey to the center of the knowledge neurons: Discoveries of language-independent knowledge neurons and degenerate knowledge neurons. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 17817–17825. AAAI Press.

Yuheng Chen, Pengfei Cao, Yubo Chen, Yining Wang, Shengping Liu, Kang Liu, and Jun Zhao. 2024b. The da vinci code of large pre-trained language models: Deciphering degenerate knowledge neurons. *CoRR*, abs/2402.13731.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8493–8502. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA).

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 5484–5495. Association for Computational Linguistics.

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2024. Large language models are cross-lingual knowledge-free reasoners. *CoRR*, abs/2406.16655.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D. Manning, and Chelsea Finn. 2022. Memory-based model editing at scale. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.

Jingcheng Niu, Andrew Liu, Zining Zhu, and Gerald Penn. 2024. What does the knowledge neuron thesis have to do with knowledge? In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander H. Miller. 2019. Language models as knowledge bases? In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 2463–2473. Association for Computational Linguistics.

Hassan Sajjad, Nadir Durrani, and Fahim Dalvi. 2022. Neuron-level interpretation of deep NLP models: A survey. *Trans. Assoc. Comput. Linguistics*, 10:1285–1303.

Dan Shi, Renren Jin, Tianhao Shen, Weilong Dong, Xinwei Wu, and Deyi Xiong. 2024. IRCAN: mitigating knowledge conflicts in LLM generation via identifying and reweighting context-aware neurons. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Ningyu Zhang, Yunzhi Yao, Bozhong Tian, Peng Wang, Shumin Deng, Mengru Wang, Zekun Xi, Shengyu Mao, Jintian Zhang, Yuansheng Ni, Siyuan Cheng, Ziwen Xu, Xin Xu, Jia-Chen Gu, Yong Jiang, Pengjun Xie, Fei Huang, Lei Liang, Zhiqiang Zhang, Xiaowei Zhu, Jun Zhou, and Huajun Chen. 2024. A comprehensive study of knowledge editing for large language models. *CoRR*, abs/2401.01286.

Ce Zheng, Lei Li, Qingxiu Dong, Yuxuan Fan, Zhiyong Wu, Jingjing Xu, and Baobao Chang. 2023. Can we edit factual knowledge by in-context learning? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 4862–4876. Association for Computational Linguistics.