本科毕业论文（设计）

基于因果可解释的
大模型事实知识编辑技术研究

A Study of Factual Knowledge Editing Techniques for Large Models Based on Causal Interpretability

宋浩

哈尔滨工业大学

2024 年 5 月

密级：公开

本科毕业论文（设计）

基于因果可解释的
大模型事实知识编辑技术研究

本　　科　　生：　　宋浩

学　　　　号：　　7203610609

指　导　教　师：　　张伟男教授

专　　　　业：　　计算机科学

学　　　　院：　　计算学部

答　辩　日　期：　　2024 年 5 月

学　　　　校：　　哈尔滨工业大学

摘 要

在人工智能领域，预训练的大型语言模型已成为处理自然语言理解任务的关键工具。这些模型通过预先训练，存储了大量的语言知识和事实信息，能够通过特定的提示（prompt）来调取和利用这些知识。然而，这些知识可能存在错误，或者用户可能希望根据自己的需求来插入、修改或删除某些知识。因此，本文旨在从基于因果可解释方案的模型编辑方法出发，以实现对大型语言模型内部知识的精确修改。

本文首先对当前主流的模型编辑方法进行了调研和分析，特别关注了它们在主要性能指标上的表现。为了更深入地理解这些方法的效果，本文进行了多个组合的测试，并详细分析了每种方法的表现及其可能的原因。此外，本文还特别关注了模型的局部性指标，在数据层面上设计约束进行改进，在算法层面上实现了灵活选层改进，并验证了其有效性。通过这些优化，本文希望能够进一步提高模型编辑的准确性和效率。

最后，本文实现了一个交互式应用，这个应用能够在代理与用户进行语言交互的过程中，自动识别用户的修改命令。一旦识别出修改命令，代理会在声明和确认后，对大模型的内部参数进行相应的修改，以确保最终的输出能够满足用户的需求。

本研究不仅为模型编辑方法提供了新的思路，而且为大型语言模型的实际应用提供了重要的参考价值。

关键词:因果可解释性; 大语言模型; 知识编辑

Abstract

In the field of artificial intelligence, pre-trained large language models have become key tools for handling natural language understanding tasks. These models, after pre-training, store a vast amount of linguistic knowledge and factual information, which can be retrieved and utilized through specific prompts. However, this knowledge may contain errors, or users may wish to insert, modify, or delete certain knowledge based on their needs. Therefore, this paper aims to precisely modify the internal knowledge of large language models starting from a model editing approach based on causally interpretable solutions.

The paper first investigates and analyzes the current mainstream model editing methods, with a particular focus on their performance on key performance indicators. To gain a deeper understanding of the effectiveness of these methods, the paper conducts multiple combined tests and analyzes the performance of each method and its possible reasons in detail. Additionally, the paper pays special attention to the locality indicators of the model, designing constraints at the data level for improvement, and implementing flexible layer selection at the algorithm level, verifying its effectiveness. Through these optimizations, the paper aims to further improve the accuracy and efficiency of model editing.

Finally, the paper implements an interactive application that can automatically recognize users' modification commands during language interactions between the agent and the user. Once a modification command is identified, the agent will make the corresponding modifications to the internal parameters of the large model after declaration and confirmation, to ensure that the final output meets the user's needs.

This research not only provides new insights into model editing methods but also offers significant reference value for the practical application of large language models.

Keywords： causal interpretability, large language modeling, knowledge editing

目 录

第 1 章 绪 论

1.1 课题背景及研究的目的和意义

1.1.1 相关背景

近年来，大语言模型[1]（LLM，Large Language Model）在多个领域尤其是自然语言处理领域展现出了巨大的应用潜力。近期较为主流的 LLM（包括多模态模型）大部分是基于 Transformer[2]结构的，比如 Encoder-only 架构的 BERT[3]、ALBERT[4]等，Decoder-only 架构的 GPT 系列[5–8]、LlaMA 系列[9,10]等，还有 Encoder-Decoder 架构的 T5[11]、BART[12]、GLM[13,14]等。

这些模型的参数量级往往在十亿到千亿级别，在例如 Wikipedia,The Pile[15]等大型语料库上预训练后获得了丰富的语义表示信息以及世界知识。在此之前，人们通常采用构建知识库系统比如知识图谱，将知识以实体形式显式地存储于网络中，而大模型在知识问答上的能力令人印象深刻，与知识图谱不同，知识在大模型中以非结构化的数字嵌入存储，并匹配语义相关性来获得输出，而大模型中参数的获取几乎决定于数据集信息和优化方法。

由于知识具有实时性和不确定性，在查询知识时会出现知识过时、谬误、事实性幻觉[16]等问题。能够对特定知识进行修改编辑以及校准[17,18]就显得尤为重要。在知识图谱等知识库系统中，对知识的修改需要找到特定的实体或者联系再进行修改，而由于大模型中存储知识的非结构特性和难解释性，在大模型中进行知识注入，知识删除，知识修改，知识推理等任务具有相当难度。因此，对于事实知识的模型编辑技术应运而生[19,20]，以期能够实现对大模型中知识信息的修改，增强大模型输出的可控性和提高内部机制的透明度。

1.1.2 研究目的和意义

研究目的主要有：（1）探索和改进当前主流的基于因果分析方案的模型编辑方法；（2）探索知识在大模型中的分布特性。对于事实知识类数据，尝试能够解释其在大模型中的特征表示。（3）探索交互场景下的模型编辑技术应用。

研究意义主要有：（1）提高模型可信度和透明度。通过探索知识在大模型中的分布特征，可以增强人们对模型决策的信任（2）提高模型学习效率。

样本级别的元样本学习比特定任务级别的训练粒度更细，难度也更高。在迁移学习、微调、增量学习的应用越来越广的背景下，能针对性地训练模型对于大模型训练范式的完善具有重要意义。（3）提高大模型的交互性。关注语言大模型的交互特性是大模型落地应用的关键点，其意义在于用大模型技术更好地为人类社会服务。

1.2 发展概况与相关工作

1.2.1 事实知识存储与表示

20 世纪费根提出了“知识工程”的概念，经历了文档，数据，语义网络等阶段发展，大模型技术的出现也意味着知识的存储特征从显性发展为隐性，预训练语言模型也被证明有存储事实知识信息的能力[17,18,21,22]，它在存储知识任务上是特别值得关注的。COMET[23]已经能够自动构建常识知识图谱，论文[24]解释了基于 Transformer 结构的大语言模型中 FFN 键值对在知识存储的关键作用，在此基础上，知识神经元[25]的概念也被提出。这些工作都有力证明了大模型在知识存储能力方面有着广阔的前景。

关于知识存储的发展往往也和知识表示相关联，20 世纪以来知识表示经历了基于谓词逻辑的表示、基于语义网络和语义网的表示和知识图谱概念的发展，如今在大模型中的知识表示也正在被探索，例如关于知识实体的表示[26]和关于实体间关系的表示[27]。在大模型时代探索知识的内部表示具有重要意义，能够识别事实编码表示对于模型输出的因果关系是一个有效的知识编辑器所必需的[26]。

1.2.2 知识编辑技术概况

对于事实知识的模型编辑技术是近年来大模型发展的新型任务，其主要目的是修改原有模型中存储的知识，使其能够输出可控的预期结果（见图 1-1）。目前该领域仍具有可观的探索空间。编辑方法可以按照两种方式进行划分理解。

1.2.2.1 以参数修改划分

目前对于知识的编辑方法从参数是否保留上上主要可以分为两种范式，即保留模型的参数或者修改模型的参数[20]。

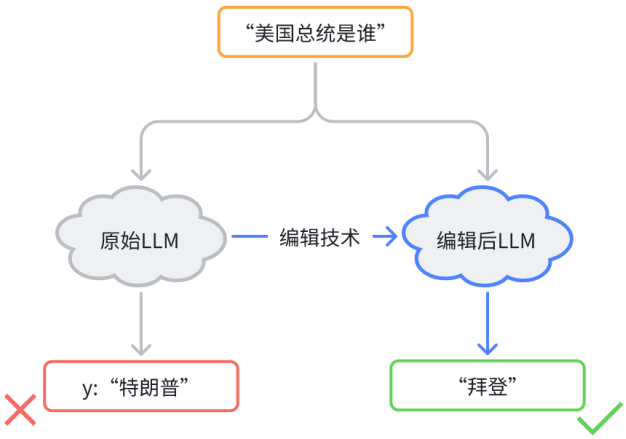


图 1.1 通过模型编辑更新知识示例

（1） 保留模型参数

这样的策略明确地在内存中保存了所有的编辑示例，并通过检索器针对每个新的输入提取出最匹配的编辑事实，以此来引导模型生成编辑内容。SERAC[28]将编辑事实添加到缓存中，先通过范围分类器估计该输入落入每个缓存编辑示例范围的概率， 如果范围分类器预测输入落入缓存中的任何编辑范围内，则检索最可能在范围内的编辑，并返回在新输入和检索到的编辑基础上的反事实模型预测。如果新输入被认为落在所有编辑范围之外，则返回基础模型的预测。此外还有一些工作利用了 LLM 上下文学习的稳定能力并通过记忆检索来编辑语言模型，如 MemPrompt [29](Madaan et al., 2022)、IKE[30] (Zheng et al., 2023)和 MeLLo [31](Zhong et al., 2023)。

在保留模型原参数的基础上还有一类工作通过引入额外参数来进行训练，以此达到学习新知识的目的，如 CaliNET[32](Dong et al., 2022) 通过整合若干神经元来处理多种编辑场景；GRACE[33] (Hartvigsen et al.，2022) 采用了一种离散的码本作为适配器，通过不断地增添和更新码本中的元素来调整模型的预测结果；T-Patcher[34] (Huang et al., 2023) 在模型的前馈网络 (FFN) 的最终层中植入了一个专门用于修正特定错误的神经元（补丁），该神经元仅在检测到相应错误时被激活;，只有在遇到其对应的错误时才会生效。

（2） 修改模型参数

此范式会修改一部分模型参数以期达到编辑效果，其主要思路是设计优化目标来修改特定位置的参数。KN[25]通过归因技术定位知识神经元并修改对应的 FFN 值[24]进而达成修改和擦除知识的目的。局限在于它可能会损坏其

他知识，衡量对其他知识的影响程度指标有些粗略，没有探讨神经元之间的关系，修改方式粒度粗。ROME[35] (Meng et al., 2022)使用因果中介分析跟踪GPT模型中隐藏状态激活的因果影响，它使用AIE（Average Indirect Effect）量化某一层对结果预测的间接影响，并且认为使用因果跟踪的方法会比使用积分梯度[36]的方法能显示更多信息（比如考虑了中间变量的影响），更具有鲁棒性。MEMIT[37] (Meng et al., 2023) 在ROME的设置上进行了扩展，实现了批量样本的同时编辑，与 ROME 修改特定单层不同的是，MEMIT 修改多个层。基于MEMIT，PMET[38] (Li et al., 2023a) 考虑了注意力值，以获得更好的性能，最近 WilKE[39]在层设置上实现了灵活定位，对特定知识实现灵活修改。

另外一类方法是元学习。元学习方法主要采用一个超网络来学习一个矩阵变化ΔW，如KE[40]、MEND[41]等。

1.2.2.2 以认知模式划分

从认知模式来看，LLM的模型编辑方法可以分为三种阶段。

KnowEdit[19]认为 LLM 的学习方式和人类认知过程十分相似，即可以分为识别阶段、关联阶段和掌握阶段。

（1） 识别阶段

在识别阶段，模型需要在相关上下文中学习知识并进行处理和推理，就像人类遇到新信息时需要存储处理。这一类的方法包括了 SEARC[28]、IKE[42]、PokeMQA[43]、MeLLo[44]等，主要思路是维护一个存储空间来实现对输入的检索。

（2） 关联阶段

在关联阶段，模型学习新知识的表示h_{Know}并连接到原始表示h中，就像人类将新想法与先前的概念联系起来一样。这一类的方法主要包括了CaliNET[32]、T-Patcher[34]、MELO[45]、REMEDI[46]、GRACE[33]等，其主要思路是在原模型基础上增加格外的表示或参数来进行新知识的学习和推理。

（3） 掌握阶段

掌握阶段是知识编辑中最重要的部分。这个过程中模型需要充分理解运用其参数信息并从中学习知识，类似于人类对知识的学习掌握过程。包括的工作有 KN[25]、ROME[35]、MEMIT[37]、PMET[38]、WilKE[39]以及考虑了逆关系建模[47]的 BIRD[48]等。这类方法直接改变了模型的权重ΔW，这类方法可以在没有任何外部帮助或合并的情况下处理问题。

另外，元学习方法不直接更新权重而是利用超网络来预测更新，这类工作包含了 KE[40]、MEND[41]、SLAG[49]、MALMEN[50]等。

不过目前研究界对这类方法持怀疑态度，因为模型的内部机制尚不清楚，这类方法可能损坏一些模型的全局信息，产生副作用，因此还属于早期研究阶段。

1.2.3 因果可解释性

1.2.3.1 深度学习可解释性

目前关于可解释性并没有清晰的定义和目标。一个非数学的定义是 Miller[51]在 2017 年提出的：“Interpretability is the degree to which a human can understand the cause of a decision”，另一个是 KIM[52]提出的“Interpretability is the degree to which a human can consistently predict the model’s result”，也就是说可解释性是人们可以理解决策原因的程度。Doshi-Velez 等[53]定义其为 "Interpretability is the ability to explain or to present in understandable terms to a human"。这说明了可解释性的根本目标不在于发掘神经网络内所有内部单元特性，而在于能够量化地分辨出哪些信息是可解释的、哪些是难以解释的、哪些是被网络准确捕捉的、以及哪些是可以基于现有数据合理推测的。通过尽可能地分离出可解释的特征信息，并进行定量分析，就可以构建起一种可靠的信任机制[54]。

1.2.3.2 模型编辑与可解释性的关系

可解释性一直是很多学者探索的问题，研究者们对模型内部知识的神秘性有着大同小异的分类理解[55,56]。深度学习模型编辑与深度学习可解释性之间存在关系，尤其是在改进和解释复杂深度学习模型的性能方面：

可解释方案从基础逻辑上影响模型编辑技术。对模型表示的深入理解可以作为改进模型控制的基础，指导模型编辑的目标。

通过模型编辑技术可以验证可解释方案的可靠性。对于提高复杂模型的透明度的问题，模型编辑技术有助于提高模型的可信度，可以让人类理解模型决策过程的机理，从而对于出现的问题可以提出可行的解决方案，正如医生诊断一般。

1.2.3.3 因果可解释性

Moraffah 等[57]首先对传统解释方案进行以下分类：（1）固有可解释，指在决策或训练时生成解释的模型，比如决策树、基于规则的模型、线性回归、

注意力网络、表示学习[1]；（2）事后解释，指的是为已经做出的决策生成解释的模型，比如局部可解释性、显著性图、基于实例的解释，特征可视化、影响函数、基于基本模型的解释。

因果解释方案专注于因果可解释模型，与归因解释相关[58]，该模型可以通过它们所处的替代情况（例如，使用不同的输入、模型组件或超参数进行训练）做出哪些决策来解释他们的决策。以下展示了因果方面的一些基本定义[59]：

定义 1：SCM (Structural Causal Models)。一个 4 元组变量$M\ (X,U,f,P_u)$，其中X是一组有限的内生变量，通常是可观察变量;U 表示一组有限的外生变量，通常是不可观察变量或噪声变量;f是一组函数$\{f_1,f_2,\dots\}$，其中每个函数表示一个因果机制，使得：

$$\forall x_i \in X, x_i = f_i(Pa(x_i),u_i) \qquad (1-1)$$

且$P_a(x_i)$是$(X \setminus x_i) \cup$ U的子集，P_a是U上的概率分布，这种机制被称为结构因果模型(SCM)或结构方程模型(SEM)。

定义 2：CBN (Causal Bayesian Network)。为了表示 SCM $M\ (X,U,f,P_u)$，使用有向图形模型$G(V,E)$。V为内生变量X的集合，E为因果机制。这表明对于每个因果机制$x_i = f_i(Pa(x_i),u_i)$，存在从父集合$x_i = Pa(x_i)$到x_i的每个节点的有向边。表示这种 SCM 的整个图称为因果贝叶斯网络(CBN)。

定义 3：ACE (Average Causal Effect)。一个二元随机变量 x（自变量）对另一个随机变量（因变量）的平均因果效应(ACE)定义为:

$$ACE\ =\ E[y|do(x\ =\ 1)] - E[y|do(x\ =\ 0)] \qquad (1-2)$$

其中 do(.)算子表示 SCM 或 CBN 定义的相应干预分布。

Pear[60]对于因果解释的解释级别引入了几个定义如下：

级别 1：统计（关联）可解释性[2]。旨在通过提出诸如“看到 x 会如何改变我对 y 的看法？”之类的问题来揭示统计关联。

级别 2：因果干预可解释性:旨在回答“如果 x，会如何”的问题。

级别 3：反事实可解释性:是最高层次的可解释性，旨在回答“为什么”的问题。

本文中，因果可解释主要指使用了因果中介分析来判断量化 token 的因果影响的模型编辑方法，它通过计算归因分数或因果中介分析等方法来探索

[1] 比如 PCA、ICA、VAE 等

[2] 是传统可解释性主要关注的问题

大模型内部计算机制对结果的影响，提高了大模型透明度。并且本文对其做了探索改进。

1.3 本章小结

本章内容主要介绍了编辑任务的研究背景，介绍了模型编辑任务的发展现状和相应技术，并介绍了因果分析的量化方式，这是基于因果可解释的编辑方法量化计算组件因果影响的重要手段。

第二章主要测试分析了若干主流模型编辑方法在基本指标上的表现并给出了分析。第三章对于编辑方法可能的提升空间做了数据和算法上的改进尝试。第四章探索开发了使用模型编辑方法的交互式应用，实现自我更新。

第 2 章 主流方法测试分析

2.1 方法介绍

2.1.1 KN

KN[25]（Knowledge Neurons）方法首先提出了知识神经元的概念。在进行参数编辑时，通过一种易于实现的特征归因技术定位到需要修改的特征激活所对应的权重矩阵，归因的计算方式是积分梯度（见式（2-1））, 它可以用来量化输入特征与模型预测之间的因果关系。

$$\mathrm{Attr}\left(\omega_{\mathrm{i}}^{(l)}\right)=\bar{\omega}_i^{(l)}\int_{\alpha=0}^{1}\frac{\partial P_x\left(\alpha\bar{\omega}_i^{(l)}\right)}{\partial\omega_i^{(l)}}d\alpha \tag{2-1}$$

其中，$P_x\left(\hat{\omega}_i^{(l)}\right)=p\left(y^*\middle|x,\omega_i^{(l)}=\hat{\omega}_i^{(l)}\right)$，$\omega_i^{(l)}$是第$l$层的第$i$个神经元，$\hat{\omega}_i^{(l)}$是指定的赋值，$\bar{\omega}_i^{(l)}$是原始值。

在进行定位时，KN 选择归因分数最大的若干个（文中为 2）神经元对应的 FFN 值进行修改。

在进行编辑时，KN 试探性地通过线性方式（见式（4））修改 FFN 中的权重。

$$FFN_i^{val}=FFN_i^{val}-\lambda_1 t+\lambda_2 t' \tag{2-2}$$

其中FFN_i^{val}是第i个神经元对应的前馈层权重参数，t',t分别是新、旧知识的嵌入向量，λ_1,λ_2是可设置的参数。

2.1.2 ROME

ROME[35]（Rank-One Model Editing）使用因果中介分析[3]跟踪 GPT 模型中隐藏状态激活的因果影响[4]，它使用 AIE（Average Indirect Effect）量化某一层对结果预测的间接影响，并且认为使用因果跟踪的方法会比使用积分梯度的方法能显示更多信息（比如考虑了中间变量的影响），更具有鲁棒性。

具体上说，ROME 对单个样本x和待观测o进行了三次运行，第一次为正

[3] Michelet et al. Direct and Indirect Effects. 2005.

[4] Vig et al. Investigating Gender Bias in Language Models Using Causal Mediation Analysis. 2020.

常运行，计算模型能够得到正确结果，概率值记作$P_{clean}(o)$，第二次对于输入主语的嵌入向量加入了正态分布的随机噪声以扰乱计算模型的运行，此时待观测样本的概率值记为$P_{cor}(o)$，第三次尝试挑选特定层的隐藏状态h_i^l进行恢复运行,此时待观测样本概率值记为$P_{cor-clean\,h_i^l}(o)$，定义 TE（Total Effect）为 $TE = P_{clean}(o) - P_{cor}(o)$，定义 IE（Indirect Effect）为 $IE = P_{cor-clean\,h_i^l}(o) - P_{cor}(o)$，最后在 1000 条样本上得到 ATE 和 AIE。

在进行编辑时，它根据实验统计所选择的中间层（比如第 18 层）进行权重更新。更新的方式是基于约束的优化问题（见式（2-3））。

$$minimize \left\|\widehat{W}K - V\right\| s.t.\ \widehat{W}k_* = v_*\ by\ setting\ \widehat{W} = W + \Lambda(C^{-1}k_*)^T \quad (2-3)$$

其中$\widehat{W}$即新权重，$C \propto E[kk^T]$为基于 100000 条 Wikipedia 的数据统计结果的常数矩阵，$\Lambda = \frac{(v_* - W\,k_*)}{(C_*^{-1}k_*)^T k_*}$为基于优化目标推导得出的计算式。

在给定 prompt 和更改目标 target 的情况下，ROME 能计算出中间层k_*, v_*表示，就能够将计算出需要更改的差值矩阵进行更新。

2.2 数据集

2.2.1 ZsRE

原本是一个关于事实陈述的零样本关系抽取数据集，包含了 244173 个样本，ROME 工作抽取并整理了 10000 条记录，每条记录包含一个事实陈述、它的释义和一个不相关的事实陈述。我搜集到了一个从中采样的数据集，单条样本可见图 2-1，它包含了 1037 个样本，每个样本中包含了一个主语，一个指令，其原始预测，一个问题重述，一个待修改的答案，一个真实答案，一个随机的问题和答案和待推理的 one hop 验证问题。


```
{
    "subject": "Volvo B12M",
    "src": "What company made Volvo B12M?",
    "pred": "Volvo Buses",
    "rephrase": "Volvo B12M's manufacturer was who?",
    "alt": "Volkswagen Group",
    "answers": [
        "Volvo Buses"
    ],
    "loc": "nq question: who sang it must have been love but its over now",
    "loc_ans": "Roxette",
    "cond": "Volvo Buses >> Volkswagen Group || What company made Volvo B12M?",
    "portability": {
        "Recalled Relation": "(Volkswagen Group, headquartered in, Wolfsburg, Germany)",
        "New Question": "In which city is the headquarters of the company that made the Volvo B12M?",
        "New Answer": "Wolfsburg, Germany"
    }
},
```


图 2-1 采样的 ZsRE 单条记录示意

2.2.2 COUNTERFACT

counterfact 首先是因果推理领域提出的概念。这个数据集用于评估语言模型中的反事实编辑，ROME 提出的 counterfact 数据集包含了 21,919 条具有不同主语、关系和语言变化的记录。我搜集到了一个采样的数据集，包含了 1919 条记录，其中单条记录的内容见图 2-2。它的形式与 ZsRE 数据集相似。


```
{
    "case_id": 42,
    "prompt": "The profession of Arun Nehru is",
    "target_new": "actor",
    "subject": "Arun Nehru",
    "ground_truth": "politician",
    "rephrase_prompt": "Arun Nehru is known for",
    "locality_prompt": "George Washington's profession is a",
    "locality_ground_truth": "politician"
},
```


图 2-2 采样的 counterfact 单条记录示意

2.2.3 Multi-locality

Multi-locality 数据集是从 2.2.1 中数据集进行集成，抽取了用于测评 locality 指标的所有样本数据，汇集成新的文件，其数据形式简单（见图 2-3），包含了 1037 合问题和答案样本，可以指定数量用来进行随机测试。


```
{
    "loc": "nq question: what is the name of the last episode of spongebob",
    "loc_ans": "The String"
},
{
    "loc": "nq question: types of skiing in the winter olympics 2018",
    "loc_ans": "Downhill"
},
```


图 2-3 集成的 locality 数据记录示意

2.2.4 相同主语不同关系

为了测试方法在不同关系上的表现，搜集了 1301 条相同主语不同关系的数据样本如图 2-4。每个样本包含了两条相同主语不同关系的 locality 任务。

```
{
    "subject": "Louise Grandjean",
    "target_new": "mezzo soprano",
    "prompt": "What voice type is Louise Grandjean?",
    "ground_truth": [
        "soprano"
    ],
    "rephrase_prompt": "What tone does Louise Grandjean sing in?",
    "cond": "soprano >> mezzo soprano || What voice type is Louise Grandjean?",
    "locality": {
        "Relation_Specificity": [
            {
                "prompt": "The country of citizenship of Louise Grandjean is",
                "ground_truth": [
                    "France"
                ]
            },
            {
                "prompt": "Louise Grandjean country of citizenship",
                "ground_truth": [
                    "France"
                ]
            }
        ]
    },
    "portability": {
        "Reasoning": [
            {
                "prompt": "What is the vocal range of Louise Grandjean as a mezzo soprano?",
                "ground_truth": "A3 to A5"
            }
        ]
    }
},
```


图 2-4 相同主语不同关系的 locality 测试样本示意

2.3 实验模型

2.3.1 GPT-J 6B

GPT-J-6B[5]是 Eleuther 研究团队在 2021 年发布的一个开源模型，它是一个基于 Transformer Decoder 架构的有 60 亿参数的自回归文本生成模型，它在 The Pile[6]数据集上进行了预训练。其具体参数可见表 2-1。

2.3.2 Mistral-7B

Mistral-7B 是 Mistral AI 团队在 2023 年 9 月开源的一个混合模型，它是一个基于 Transformer 的有着 70 亿参数量的预训练文本生成模型，另外融合了 Grouuped-Query Attention，Sliding-Window Attention 的注意力机制和 Byte-fallback 的 BPE 分词器。值得一提的是其在各种任务上的表现都超越了 Llama2-13B。其具体参数可见表 2-2。

[5] graphcore/gpt-j: Notebook for running GPT-J/GPT-J-6B – the cost-effective alternative to ChatGPT, GPT-3 & GPT-4 for many NLP tasks. Available on IPUs as a Paperspace notebook. (github.com)

[6] The Pile (eleuther.ai)

表 2-1 GPT-J-6B 参数信息

Hyperparameters	Value
n_parameters	6,053,381,344
n_layers	28
d_model	4096
d_ff	16384
n_heads	16
d_head	256
n_ctx	2048
n_vocab	50257
position_encoding	RoPE
RoPE dimensions	64

表 2-2 Mistral-7B 参数信息

Hyperparameters	Value
dim	4096
n_layers	32
head_dim	128
hidden_dim	14336
n_heads	32
n_kv_heads	8
window_size	4096
context_len	8192
vocab_size	32000

2.4 评估指标

在性能表现上，考虑以下基本指标：可靠性、泛化性和局部性以及可移植性。

（1）可靠性：Reliability，被定义为编辑案例的平均准确度，见式(2-3)。

$$E_{x'_e,y'_e\sim\{(x_e,y_e)\}}1\{argmax_y f_{\theta_e}(y|x'_e)=y'_e\} \quad (2-3)$$

（2）泛化性：Generalization，被定义为等价邻域（例如改写的句子）均匀采样的示例上的平均准确度，见式(2-4)。

$$E_{x'_e,y'_e\sim N(x_e,y_e)}1\{argmax_y f_{\theta_e}(y|x'_e)=y'_e\} \quad (2-4)$$

（3）局部性：Locality，由编辑后模型f_e的预测与原模型f不变的比例进行评估，见式(2-5)。

$$E_{x'_e,y'_e\sim O(x_e,y_e)}1\{f_{\theta_e}(y|x'_e)=f_{\theta}(y|x'_e)\} \quad (2-5)$$

（4）可移植性：Portability，由编辑后模型f_e的预测在 one-hop 或 multi-hop 等推理任务上的性能表现进行评估，见式(2-6)。

$$E_{x'_e,y'_e\sim P(x_e,y_e)}1\{argmax_y f_{\theta_e}(y|x'_e)=y'_e\} \quad (2-6)$$

2.5 实验表现与分析

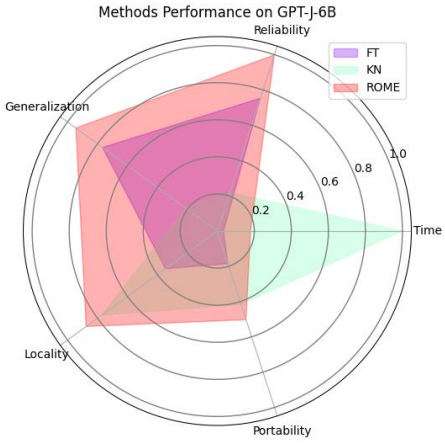
分别在 GPT-J-6B 和 Mistral-7B 模型上测试了 FT(Fine-Tune)、KN(Knowledge Neurons)和 ROME(Rank-One Model Editing)方法在采样的 ZsRE 数据集和 counterfact 数据集上的基本表现。指标的计算取所有 1000 多条样本结果的平均值。结果如图 2-4。

在 ZsRE 上，对于 GPT-J-6B 模型（图 2-5（a）），时间（Time）指标上，FT 修改速度最快，ROME 的平均修改时间为 20s 左右，KN 则需要百秒量级，这主要是因为 KN 需要计算每一层的积分梯度得分，计算成本高；可靠性（Reliability）指标上，ROME 表现最佳，KN 表现最差，这可能是因为 KN 的修改方式过于简单，修改后的模型并没有真正理解修改意图；在泛化性（Generalization）方面，FT 比 KN 更优，原因可能在于基于模式识别的 FT 方法能更好地识别到问题的等价邻域，而 KN 的修改过程可能能力不足。在可移植性（Portability）和局部性（Locality）方面，FT 范式可能导致了遗忘其他知识的后果。

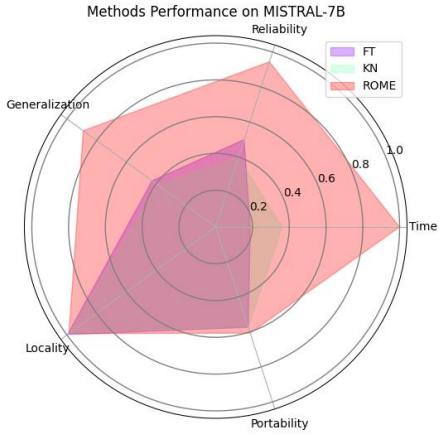
对于 Mistral 混合模型（见图 2-5（b））,时间指标上 ROME 也需要 20s 左右；另外只进行了 1 次积分梯度的 KN 和 FT 表现相当，原因可能是因为 KN 并没有筛选出最具有因果影响的神经元而表现出随机梯度更新的特性，不过同时时间成本也降低 20 倍左右；值得注意的是在混合模型上，不同方法和数据集都有很高的 locality 准确率，因此在很大程度上可以推测混合模型对于知识存储具有较好的独立性。

在 counterfact 数据集上，对于 GPT-J-6B 模型（见图 2-5（c）），时间指标上 ROME 修改一条知识需要 15s 左右；KN 的修改可靠性和泛化性明显低

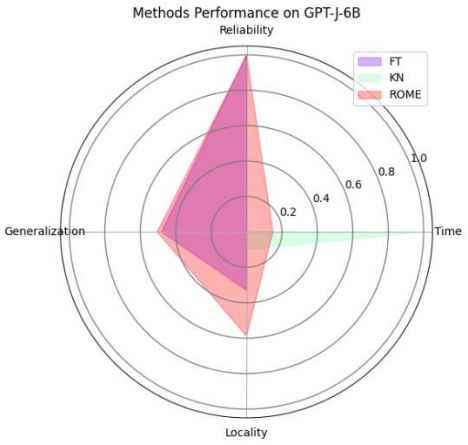
于其他方法，可能的原因是 KN 的修改方式不具有针对性，在更严格的反事实陈述中表现低效；FT 和 ROME 方法虽然具有很高的修改成功率，但泛化性不够好，原因可能在于反事实数据集本身的分布在大模型内部存储的占比较低，大模型更依赖于自身的知识所以不能很好地泛化；KN 具有更差的局部性是因为 counterfact 数据集中的局部性大部分是主语替换式而关系和宾语却保持相似，KN 由于修改了特定关系对应的知识而导致预测错误。



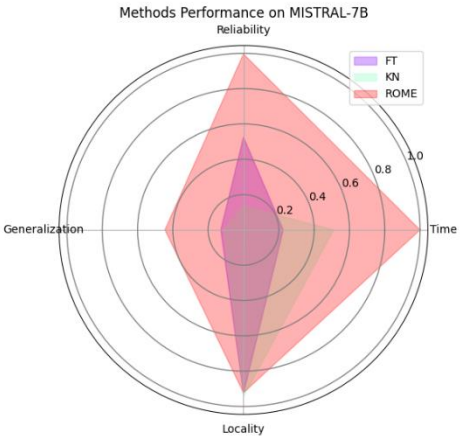
(a) 不同方法编辑 GPT-J-6B 在 ZsRE 上的表现



(b) 不同方法编辑 Mistral-7B 在 ZsRE 上的表现



(c) 不同方法编辑 GPT-J-6B 在 counterfact 上的表现

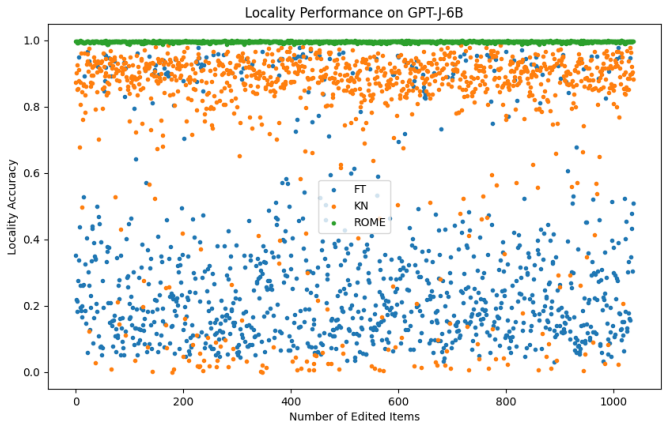


(d) 不同方法编辑 Mistral-7B 在 counterfact 上的表现

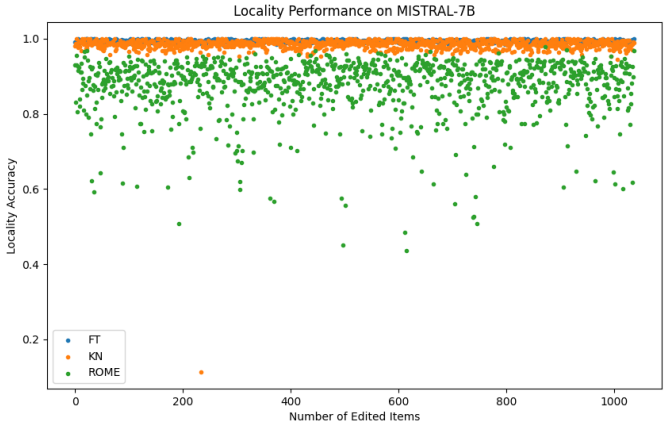
图 2-5 几种模型编辑方法的性能表现

对于 Mistral 混合模型，时间指标上 ROME 修改一条知识需要 20s 左右；三者都表现出了很好的局部性；对于可靠性和泛化性方面，FT 比在 GPT-J-

6B 上都有将近 50%的下降率，可能的原因是 Mistral 存储的知识分布较为集中，而 FT 范式的参数范围（本次实验中为 21 层）微调只能涉及到知识的部分模式，所以修改效率低，该下降现象在 ZsRE 数据集上也被体现出来。



(a) 不同方法编辑 GPT-J-6B 在 ZsRE 上 locality 指标的表现



(b) 不同方法编辑 Mistral-7B 在 ZsRE 上 locality 指标的表现

图 2-6 几种模型编辑方法在多 locality 测试样本的逐步测试实验中的性能表现

微调范式经常会出现旧知识遗忘的问题，我注意到先前搜集的数据中每条记录对于局部性指标只包含一个随机样本，这可能会导致一定的偶然性，所以随机采样了 100 条 locality 的输入输出对，每一次编辑时，都测试在这 100 个随机样本上的平均准确率，用来进一步测试各个方法的表现，获取更多的信息。结果可见图 2-6。可以看出，由图 2-6（a）与图 2-5（a）相比能

得出一致的结论。而图 2-6（b）中 ROME 的性能略有下降可能是因为在混合架构中，ROME 修改的第 5 层包含了更多其他的知识，而 FT 修改的 21 层则包含了较少的知识。

在 conterfact 数据集上，也对 GPT-J-6B 模型进行了 locality 的测试，结果如图 2-7，得到的结论也与图 2-5（c）一致。KN 在该数据集上表现出编辑的敏感性更强，也暗示了模型内部的知识关系或许更加复杂，KN 的修改方式并不能很好地建模知识表示。

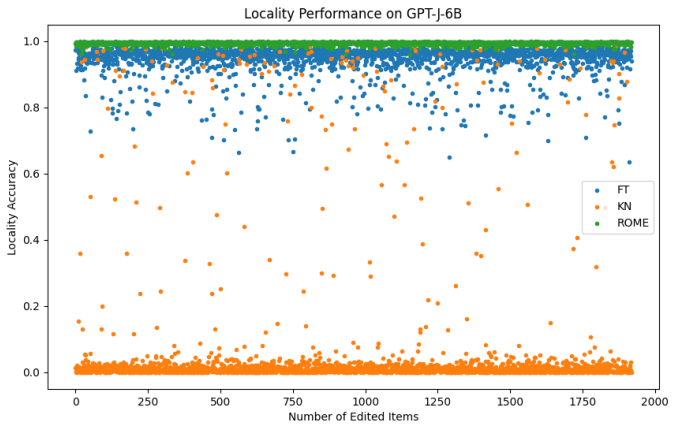
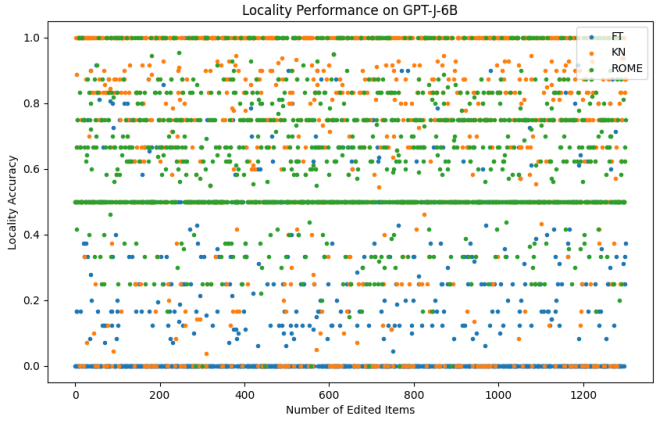


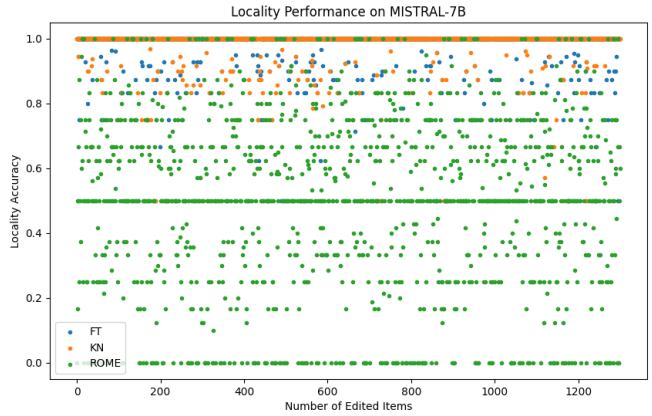
图 2-7 不同方法编辑 GPT-J-6B 在 counterfact 上 locality 指标的表现

除了随机的 locality 样本测试上，还对具有相同主语不同关系描述的样本进行了测试， 在这样的具有 1301 条记录数据集上测试了 locality 表现，结果如图 2-8。

在图 2-8 中，对比随机取样的 locality 样本，各个方法在具有相同主语的 locality 样本上性能都有所下降，ROME 的性能下降或许归因于其对于主语的依赖性强，而对于关系的索引作用敏感程度较低（这也是值得改进的问题），也或许需要将修改靠后的层以包含对于关系的考量（本次实验中修改层为 5），而 KN 对于关系的识别敏感程度较高（表现与图 2-6 相近），这一现象在两个模型都有所体现；FT 和 KN 则在混合模型上的表现总体上优于 GPT-J 模型，且表现相近（注意，此时设置的 KN 配置与 FT 相似），这可能是因为在混合架构上，对知识 value 值的修改并没有大程度影响到对于关系 relation 的存储值。由图 2-8 还可以得到启示 ROME 在两个模型上修改第 5 层，而结果相近，这或许说明不同关系的知识存储在了不同的层而不仅仅是单层。



(a) 不同方法编辑 GPT-J-6B 在 ZsRE 相同主语数据上 locality 指标的表现



(b) 不同方法编辑 Mistral-7B 在 ZsRE 相同主语数据上 locality 指标的表现

图 2-8 几种模型编辑方法在相同主语测试样本的逐步测试实验中的性能表现

2.6 本章小结

本章主要从基于因果解释的主流模型编辑方法出发，在搜集的数据集上进行了有关指标的测试，并详细分析了各个方法以及模型表现优劣的原因。从中探索了编辑方法的改进方向和思路，并在第三章做了改进探索。

基于因果机制解释的模型编辑方法（如 ROME）体现了新颖的探究大模型内部计算机制的规律，并提出了高效的修改方式。但基于因果分析的方法受到自变量的决定性影响，并没有真正挖掘到对于事实知识数据类型的表示建模，所以在不同关系的表现上表现出了劣势。

第 3 章 ROME 的改进

3.1 改进动机

由于目前基于因果分析的模型编辑方法只修改预先指定的层，并且只考虑了最后一个主语 token 对结果的影响，没有考虑到的关系建模作用。故尝试从数据约束层面以及算法改进层面进行改进，并测试效果。

3.2 改进方式

本文尝试从数据层面上和算法层面上对 ROME 进行探索性的改进尝试。

3.2.1 数据约束

3.2.1.1 监督方式

这种方式主要是基于扩充每次修改的编辑事实，给定真值标签进行监督训练。扩充的内容即为图 2-4 中相同主语不同关系的局部性测试数据，将其作为优化目标添加到原来的k_*, v_*中，ROME 的原损失定义见式（3-1），在监督方式下式子扩展为（3-2）。该改进思路与持续学习中的添加旧样本防止遗忘的思路相似。

$$define\ L(\widehat{W}_{l_*},\Lambda)=\frac{1}{2}\left\|\widehat{W}_{l_*}K-V\right\|_F^2-\Lambda^{\mathrm{T}}(\widehat{W}_{l_*}k_*-v_*) \qquad (3-1)$$

$$define\ L(\widehat{W}_{l_*},\Lambda)=\frac{1}{2}\left\|\widehat{W}_{l_*}K-V\right\|_F^2-\Lambda^{\mathrm{T}}(\widehat{W}_{l_*}[k_{*1},k_{*2},k_{*3}]-[v_{*1},v_{*2},v_{*3}]) \qquad (3-2)$$

3.2.1.2 无监督方式

在计算v_*的过程中，将相同主语不同关系的数据提示添加到新的 KL 损失项，这种方法不需要真值标签，而是检测模型编辑后在该提示样本上的输出是否和原来保持一致。原式见式（3-3）。可以在（3-3）式上增加式子（3-4）。式子（3-4）中p_i即相同主语不同关系样本的提示语句。

$$\frac{1}{N}\sum_{j=1}^{N}-\log \mathrm{P}_{G\left(m_i^{(l^*)}:=z\right)}\left[o^*\middle|x_j+p\right]+D_{KL}\left(\mathrm{P}_{G\left(m_i^{(l^*)}:=z\right)}[x|p']\middle\|\mathrm{P}_G[x|p']\right) \qquad (3-3)$$

$$\sum_{i=1}^{M} D_{KL}\left(\mathrm{P}_{G\left(m_i^{(l^*)}:=z\right)}[x|p_i]\middle\|\mathrm{P}_G[x|p_i]\right) \tag{3-4}$$

3.2.2 算法改进

算法改进是一个值得尝试的工作，改进思路主要是在出现相同 token 次数最多的最开始层作为修改层，并用 ROME 的方法进行修改。

ROME，MEMIT 这类基于因果分析的方法是从定位计算机制的层面出发分析，而不是从知识表示层面出发，未建立神经元激活状态信息与特定数据的相关特征，故对知识数据本身并不敏感，一个表现就是它在反转关系上的修改效果不尽人意[61]，说明其并没有考虑关系的作用以及关系之间的联系。另外受到关系编码[27]的启示，ROME 只在 subject 阶段进行了因果分析，发现 subject 对于最终输出的因果影响令人印象深刻，于是找到了影响最大的层并进行修改，这样的分析方式值得借鉴。

本文认为这样的方式忽略了 relation 的作用，且只是在搜索 subject 的模式抽取。本文尝试融合 subject 和 relation 的模式，修改模式被稳定抽取后的第一个层。具体算法如下：

Algorithm 1 描述了对特定层进行修改，Algorithm 2 描述了如何挑选特定层。

Algorithm 1: Layer-wise editing process

```
Data: D={(s, r, o*)}, G
Result: Edited model
foreach data in D do
    layer:=Get the layer to modify (Algorithm 2)
    apply ROME in layer
end
```

Algorithm 2: Get the layer to modify

```
Data: D=(s, r, o*), G
Result: Edited model
with nethook as nh:
    Forwading G[s, r];
    Get tokens^{layer}_{highest probability}
    The layer := the first token layer with the highest repetition count.
return The layer
```

3.3 实验结果与分析

3.3.1 数据约束结果

在数据约束上的改进的实验结果可以参见表 3-1。在表 3-1 中，before 指未修改之前，origin 指的是原 ROME 方法修改，unsup 指 ROME 方法加上无监督约束项方式的修改，sup 指 ROME 方法加上监督约束项方式的修改。蓝色字体表示在 gpt-j-6b 模型上的最优值，绿色字体表示在 mistral 上的最优值。mR 是平均可靠性指标，mG 是平均泛化性指标，mP 是平均推理性指标，mL 是平均局部性指标，Random 表示随机采样的 100 条数据，DRDifferent Relation 表示 2 条相同主语不同关系的局部性样本。

表 3-1 不同数据约束修改方案结果对比

模式	模型	mR(%)	mG(%)	mP(%)	mL(%)	
					Random	DR
before	gpt-j	0.2806	0.2711	0.4334	(1.0)	(1.0)
	mistral	0.4004	0.3946	**0.4919**	(1.0)	(1.0)
origin	gpt-j	**0.9955**	**0.9315**	0.4761	**0.9978**	**0.6291**
(ROME)	mistral	0.9443	0.8777	0.5426	0.9021	0.4957
unsup	gpt-j	0.9924	0.8867	0.4566	**0.9978**	0.5874
(ours)	mistral	**0.9485**	**0.8799**	**0.5467**	0.8764	**0.4973**
sup	gpt-j	0.9924	0.8867	0.4564	0.9961	0.5874
(ours)	mistral	0.8775	0.8000	0.5220	**0.9030**	0.4664

总体上，这些方法效果几乎相当。具体上，在 gpt-j 上做了改进之后监督与无监督都出现了百分比级别的损失，推测其原因可能在于增加了约束修改特定层的方式可能会对其他知识造成影响，其模式知识分布广泛而模型在特定层的聚焦有限，会在一定程度上影响模型中认知的其他模式。而 mistral 模型在 unsup 方法上的表现略有提升，其原因可能在于混合模型中关于模式和知识的存储独立性较高，以 unsup 的方式修改特定层对其他的模式和知识造成的影响较小，而在 sup 方式下其性能略有损失，原因是因为监督训练的时候可能迫使模型影响了其他有关信息。

3.3.2 算法改进结果

本文尝试从模型 logit len[7]的思路出发，监测每一层的输出情况。本文根据出现相同 token 次数最多的最开始层作为修改层，交由 ROME 进行修改。

实验尝试修改“爱因斯坦的专业是医学”，将“爱因斯坦的专业是”作为提示词输入 LlaMA3 中文模型，记录各个层的输出。未修改前的原始计算结果如图 3-1 所示，其中 delt 表示期望词与每一层最高概率词的概率差值，蓝色折线表示关注词在概率层的排序。可以发现模型的最终输出词是“物理”，而预测“医学”的概率很小，尤其是在除了最后一层之外的其他层，这或许是因为语言训练时未改变模型的主体部分。

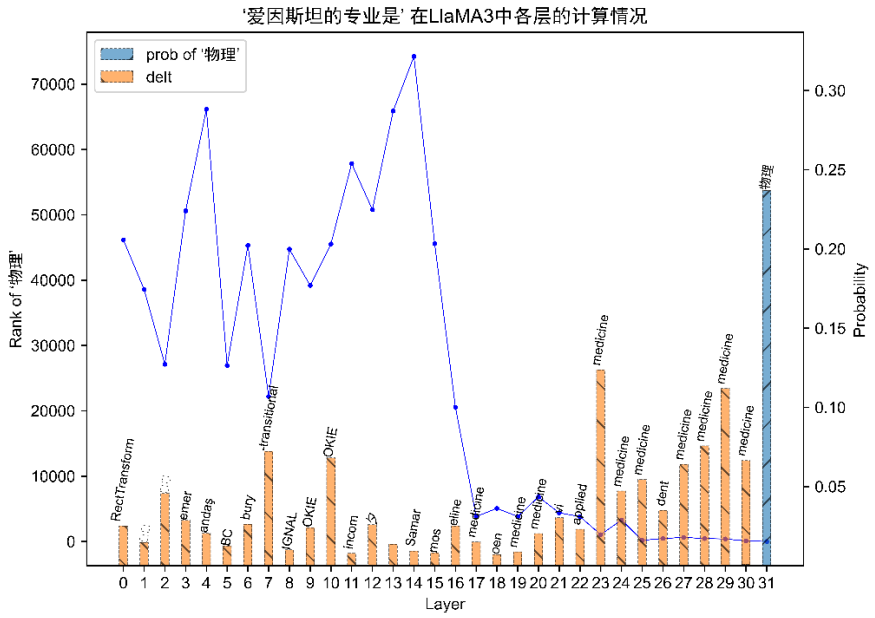
此外，由 rank 折线以及预测词的变化可以看出，预测“物理”的过程中，在大约 17 层后，模型的输出趋于稳定，其中预测出现“medicine”的次数较多（注意这里的“medicine”和期望修改的“医学”并不一致。前者只是模型本身的原始输出，此时还是未修改的状态）。这种现象说明在 17 层之后，模型可能存在冗余映射的过程，其特征并不被后面的层关注，可以简单认为，在 17 层时模型已经完成了对提示词“爱因斯坦的专业是”的特征模式的解析，故是一个合适的修改时机。

实验记录了 ROME 修改第 5 层（如图 3-2）和修改第 17 层（如图 3-3）的结果。对比实验结果可以看到，在修改较前层（第 5 层）时，较后层的预测效果被很大程度地提升，这或许是一种损失，意味着编辑力度较强，很有可能造成其他样例产生同样的输出，造成信息失真。而在修改较后层（第 17 层）则不会有这样的结果，其最大概率并不具有倾向性，而依然能够正确预测“医”（这里出现“医”而不是“医学”的原因是模型将“医”作为了一个词单元，不过从结果上而言它们是一致的）。实验先关于“爱因斯坦的专业是医学”单个样例进行了指标的分析，结果见表 3-2。

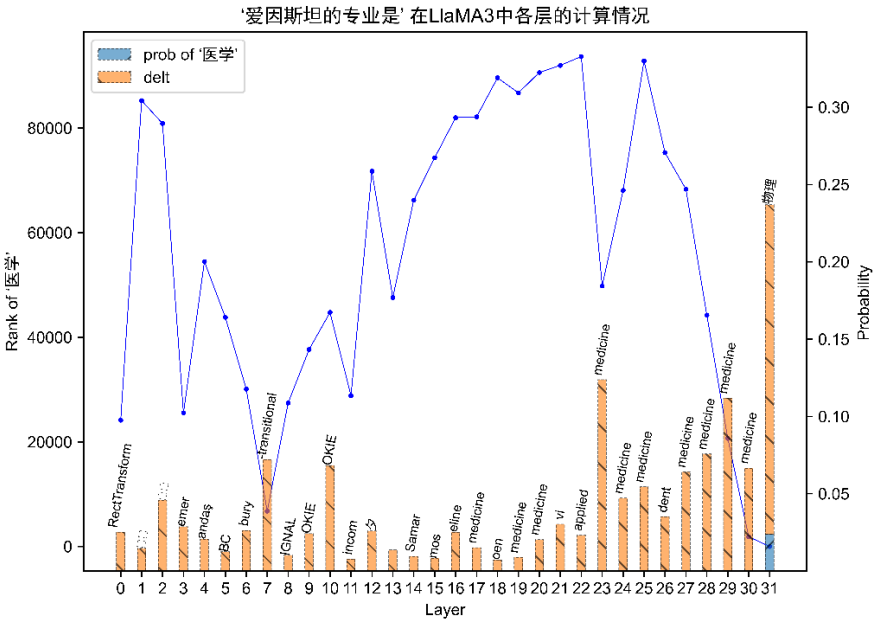
表 3-2 单个样例上结果对比

修改层	模型	mR(%)	mG(%)	mP(%)	mL(%)
不修改	LlaMA3	0.5	0.0	0.75	(1.0)
5(EasyEdit)	LlaMA3	1.0	1.0	**0.2500**	0.7810
17(ours static)		1.0	1.0	0.0	**0.9495**

[7] interpreting GPT: the logit lens — LessWrong



(a) 原输出“物理”预测情况



(b) 期望输出“医学”预测情况

图 3-1 编辑前，prompt=“爱因斯坦的专业是”时，
原输出和期望输在 LlaMA 各层的预测情况

由表 3-2 可以看出，对于单个样例，修改较后层时，在 100 个随机样例的局部性测试中表现较好，其中在 Portability 问题上设置的问题是“爱因斯坦的妻子是谁”，其输出结果的最后一个 token 是“。”因此未能完全匹配（测试是按照匹配的 token id 计算的）。

虽然修改后层取得了一定的效果，但单个样例测试还不能够形成有力的说明。于是新的实验在 100 个 ZsRE 数据样例上（来自图 2-1）进行平均测试，测试结果如表 3-3，表中每个数值在 0 和 1 之间，表示比例。

表 3-3 100 个样例上结果对比

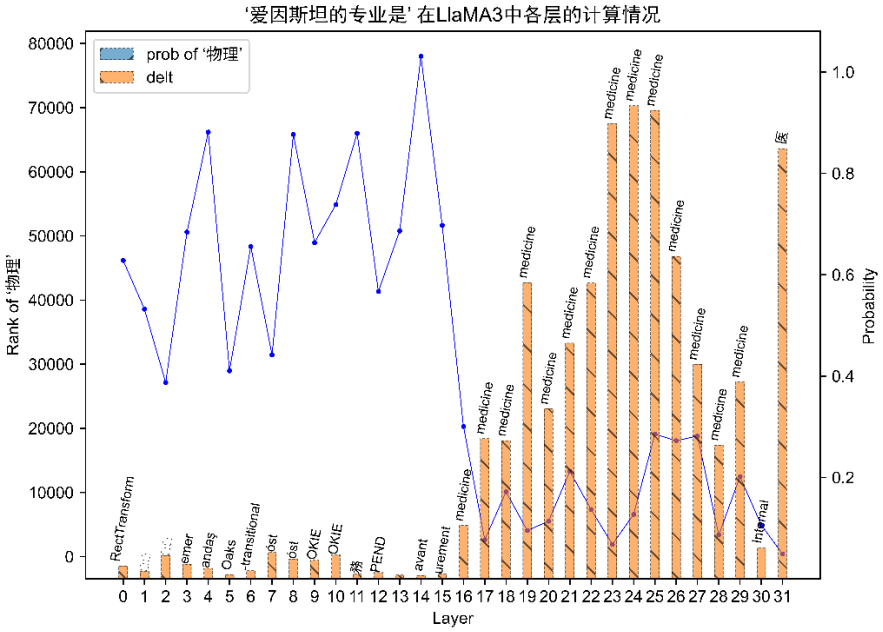
修改层	模型	mR(%)	mG(%)	mP(%)	mL(%)
不修改	Mistral-7B	0.3831	0.3948	0.4385	(1.0)
不修改	LlaMA3	0.2896	0.2670	0.4233	(1.0)
5(EasyEdit)	LlaMA3	0.9900	0.9055	0.4935	0.9616
17(ours static)		**0.9947**	**0.9277**	0.5126	**0.9631**
layerwise(ours)		0.9622	0.8568	**0.5151**	0.9225
5(EasyEdit)	Mistral-7B	0.9250	**0.8326**	**0.4899**	**0.9044**
17(ours static)		**0.9676**	0.7935	0.4742	0.8186
layerwise(ours)		0.9554	0.7991	0.4789	0.8232

性能略降可能存在两个原因：一是挑选层时偶尔选择到 0 层或最后层，影响较大；二是修改方法的不匹配，ROME 的 C 常数是根据 10000 个 subject 的 k 均值计算，并不包含 relation。

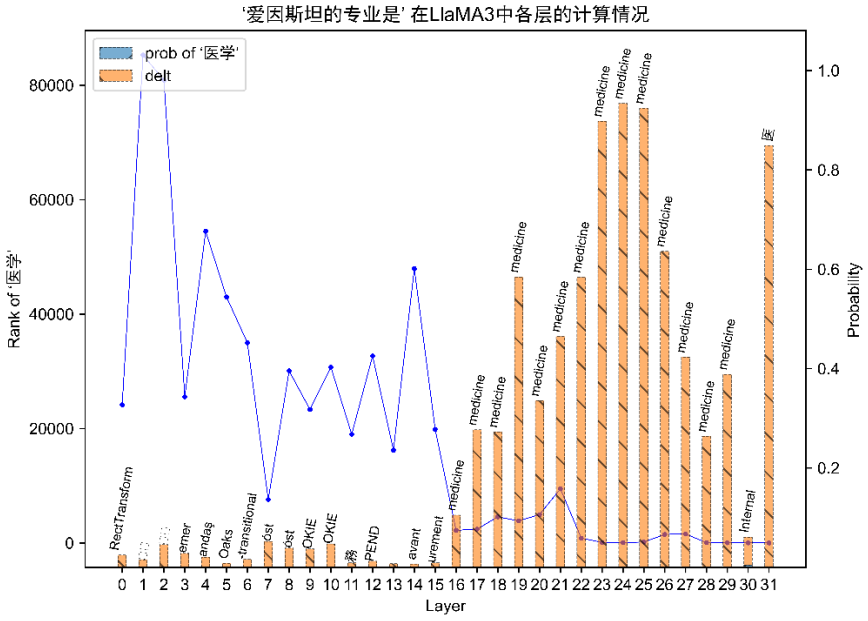
由表 3-3 可以看出，这些方法的效果总体上相当，灵活选层的方法对于 LlaMA 模型来说能够提高一定的 portability，而在其他方面表现不如固定层。原因或许就在于这种方式能够等待模型解析出需要分析的特征信息再进行修改，保证了一定的推理能力，而在其他方面的性能下降原因可能在于结合 ROME 的修改方式并不适合，因为 ROME 在保证 locality 时引入了外部语料库的大量 subject 的索引向量，而不是 subject+relation 的索引向量，这可能一定程度上限制了本文方法的性能。

在混合模型上的效果均比在 LlaMA 模型上要差，可能的解释是混合模型对内部训练的模式有着较为明确的分工，只修改单层有时不足以成功。

值得注意的是，在混合模型上，灵活选层的效果介于修改较前层和较后层之间，这与在 LlaMA 模型上的表现不同，这可能蕴含的信息是：不同的层承载着独立解耦的模式，与模型推理存在某种程度上的线性关系。

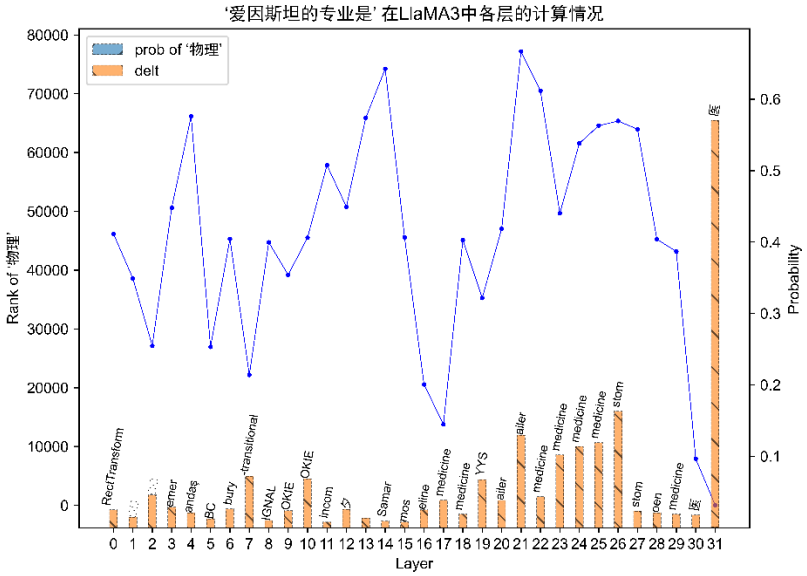


(a) 原输出 “物理”预测情况

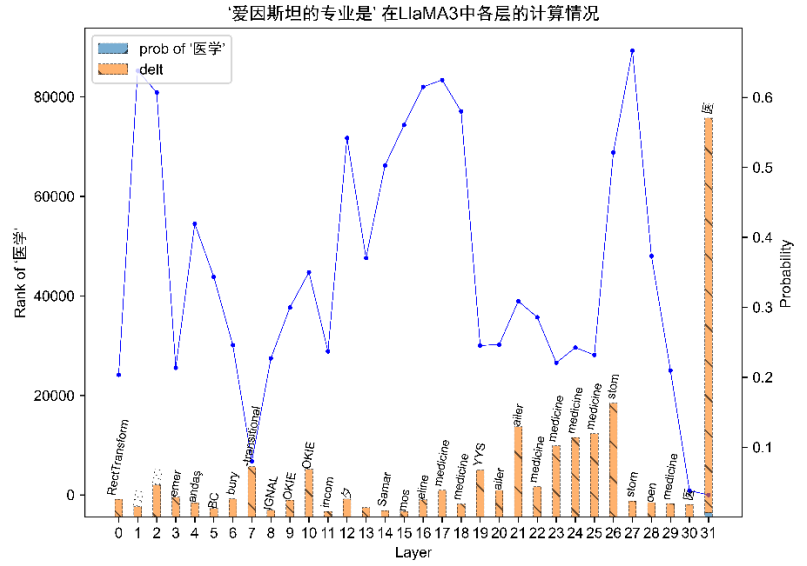


(b) 期望输出 “医学”预测情况

图 3-2 ROME 编辑原第 5 层后，prompt= “爱因斯坦的专业是”时，原输出和期望输出在 LlaMA 各层的预测情况



(a) 原输出 “物理”预测情况



(b) 期望输出 “医学”预测情况

图 3-3 ROME 编辑第 17 层后，prompt= “爱因斯坦的专业是”时
原输出和期望输出在 LlaMA 各层的预测情况

在基于重复 token 挑选层的基础上，本文还进行了激活单元的测试验证。本文追踪了输入“爱因斯坦的专业是”后 LlaMA3 和 Mistral-7b 模型各层的激活状态，得到了有趣的发现，结果见图 3-4。

该实验记录了激活值大于 0.085 的激活单元数量，实验发现，在重复 token 的开始层的激活单元数量出现猛增现象，其中 LlaMA3 的 17 层和 Mistral 的 19 层正好对应前文基于重复 token 选择的开始层。

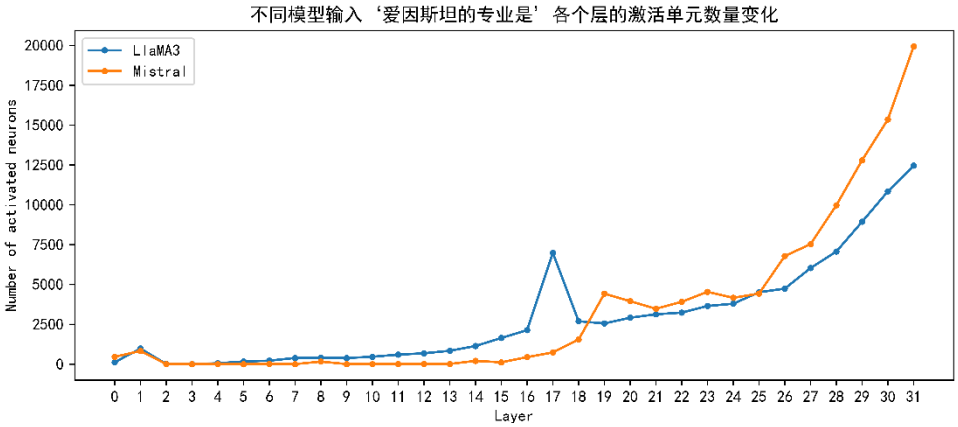


图 3-4 不同模型输入“爱因斯坦的专业是”时各层的激活单元数量变化

为了说明该猛增现象是否存在偶然性，本文使用 LlaMA3 模型和 Mistral 模型在第二章的 ZsRE 数据集和 counterfact 数据集上做了统计分析，并设计了三个简单的检查指标：

1. 一阶差分数值指标（FDV，First-order Difference Value）：得到在选择层索引的差分值，作为检查猛增的幅度指标，参考值如设置为 500。该指标值越大越好。
2. 与前层和的差值的相对比率（RDSP，relative Rate of Difference between the Sum of the Previous layers and the current layer）：指当前索引对应值与之前层总和之差的绝对值和当前索引对应值的比率，参考值如设置为 15%。该值用来和前层所有值的和做对比参考。
3. 前层均值与当前值的比率（RMP，Rate of the Mean of the Previous layer）：指前层均值与当前索引对应值的比率，参考值如设置为 25%。该指标值越小越好。

图 3-4 的测试中，FDV 为 4854，RDSP 为 0.47（保留两位小数），RMP 为 0.09（保留两位小数）。

在三个指标的测试中存在一个问题是，偶尔情况下混合模型在选择层的激活单元数为 0，为了减小对统计结果的影响，本实验合理地去除异常值。

在激活阈值设置为 0.085 的 情况下，得到的统计结果见表 3-4。其中 m 代表平均取值。

表 3-4 不同模型和数据集上 FDV,RDSP,RMP 的平均指标结果

模型	数据集	mFDV ↑	mRDSP(%)	mRMP(%) ↓
LlaMA3	ZsRE	721.56	2557.4%	1072.1%
	counterfact	198.31	896.94%	206.23%
Mistral-7B	ZsRE	581.01	3204.40%	536.43%
	counterfact	532.13	12757.62%	2724.90%

由表 3-4 可以看出，在 mFDV 上符合我们的预期，即总体上选择层对比前一层有着较大的激活神经元数量激增。在 mRDSP 和 mRMP 上却呈现了难以信服的结果。因为平均值可能是一个严格且粗糙的指标，偶尔出现的大值对实验结果造成的影响较大。于是本文统计了控制条件下的比例，来更好地说明结果（见表 3-5）。其中 r 表示符合括号内条件的数据所占总体的比率，表中每个数字是一个[0, 1]的小数，表示覆盖率。

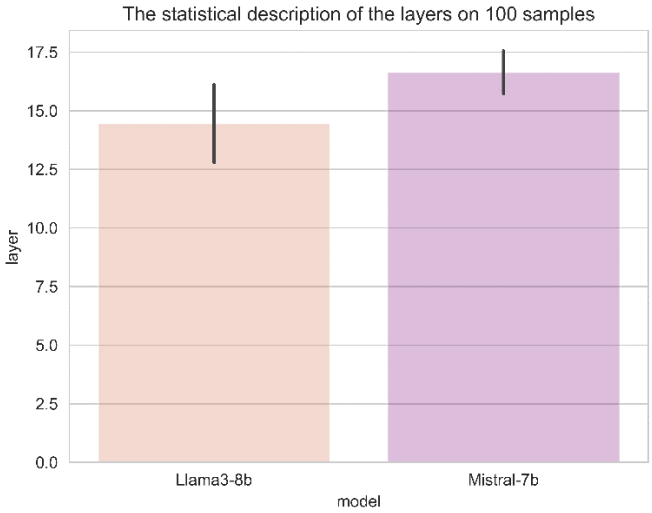
表 3-5 不同模型和数据集上 FDV,RDSP,RMP 的过滤指标结果

模型	数据集	rFDV ↑		rRDSP ↑	rRMP ↑	
	范围	>0	>500	[0.0,10.0]	[0.0, 0.5]	[0.0, 0.99]
LlaMA3	ZsRE	0.7502	0.4604	0.7548	0.7425	0.7917
	counterfact	0.7109	0.2358	0.9109	0.8471	0.9469
Mistral-7B	ZsRE	0.8700	0.3572	0.8487	0.7676	0.8580
	counterfact	0.7936	0.3469	0.6100	0.5647	0.6134

由表 3-5 可以看出，rFDV 指标中在大部分情况下，当前索引下激活单元数量会比前一层多，多于 500 的比例占约 30%到 40%；rRDSP 指标中大部分情况下，当前索引下激活单元数量与所有前层激活单元数量之和在同一水平；值得注意的是 rRMP 指标的[0, 0.5]区间中，在大部分情况下前层的均值要小于当前值，这也一定程度上说明当前层确实存在激活值数量增大的情况。

此外，本文还对选择的层做了统计分析，在 ZsRE 的 100 个样本中，其选择的层的统计信息如图 3-5。担心统计的偶然性，本文在 ZsRE 数据集 1301 个样本也进行了相同的分析，见图 3-6。发现二者的表现几乎一致，在 ZsRE 数据集上，LlaMA3 更倾向于选择较前层，Mistral 更倾向于选择较后层，且波动更小。

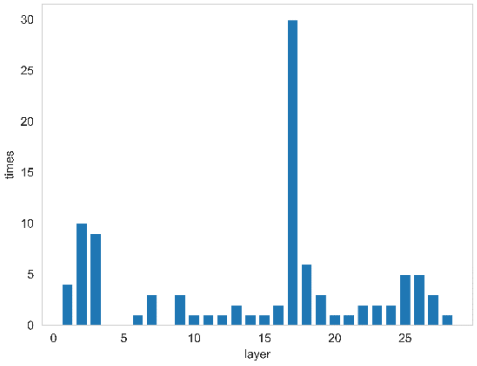
在 counterfact 上做了同样的分析（图 3-7），发现选择的层对于不同的模型出现了差异。第一个差异是对于 ZsRE 数据集，LlaMA 模型在 counterfact 上更倾向于选择后层，而混合模型则相反，这可能是由于数据集在问题上存在区别。不同数据集涉及的模式不尽相同。第二个差异在于混合模型在 counterfact 数据集上的选择的层数在 0 到 10 的数量比在 ZsRE 上更多。



(a)

Llama3-8b 的均值为 14.43，标准差 8.10

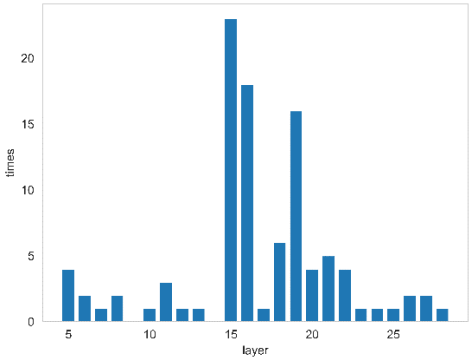
Mistral-7b 的均值为 16.64，标准差 4.79



(b)

Llama3-8b 的层数的次数图，

在第 17 层最多



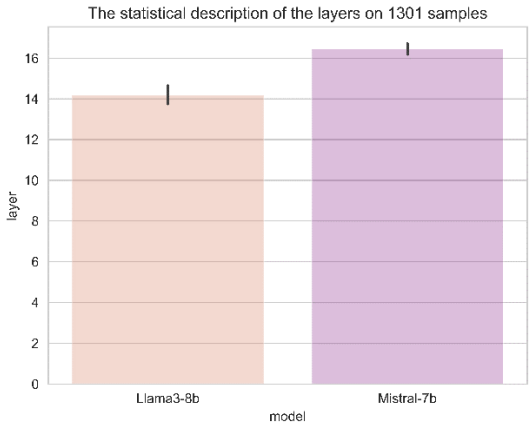
(c)

Mistral-7b 的层数的次数图，

在第 15 层最多

图 3-5 不同模型在 100 个数据上选择的修改层数的统计描述，

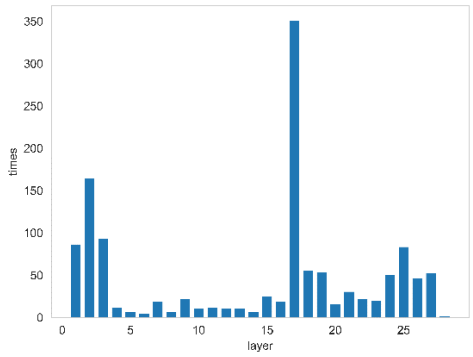
发现在 ZsRE 数据集上，Llama3 平均选择的层要比 Mistral 靠前 2 两层左右



(a)

Llama3-8b 的均值为 14.16，标准差 8.59

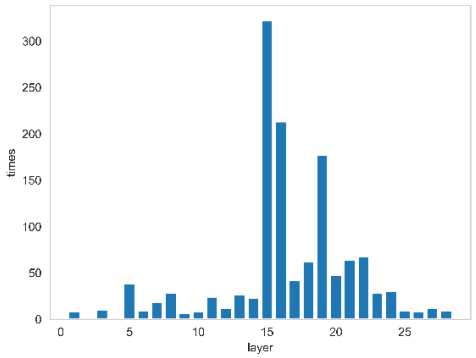
Mistral-7b 的均值为 16.45，标准差 4.67



(b)

Llama3-8b 的层数的次数图，

在第 17 层最多

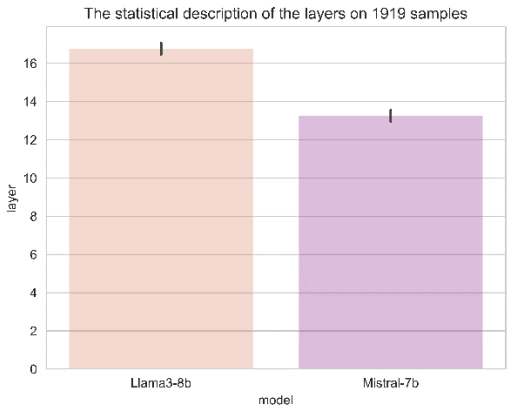


(c)

Mistral-7b 的层数的次数图，

在第 15 层最多

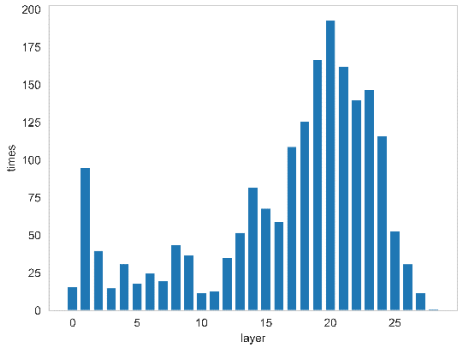
图 3-6 不同模型在 ZsRE 的 1301 个数据上选择的修改层数的统计描述，可以得到与图 3-4 相同的现象



(a)

Llama3-8b 的均值为 16.76，标准差 6.86

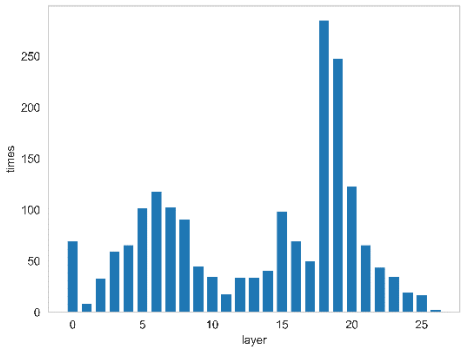
Mistral-7b 的均值为 13.27，标准差 6.79



(b)

Llama3-8b 的层数的次数图，

在第 20 层最多



(c)

Mistral-7b 的层数的次数图，

在第 19 层最多

图 3-7 不同模型在 counterfact 的 1919 个数据上选择的修改层数的统计描述，发现 Mistral 模型在该数据集上更倾向挑选较前层

3.4 本章小结

本章从数据改进和算法改进角度提出了改进方法，并且分析了改进方法的局限性，在算法改进上也提出了灵活选层的修改思路。总体上 gpt-j 在监督和无监督改进方式上性能都有了略微的下降，混合模型在无监督方式上表现出了性能提升；在激活单元数量的验证上也获得了新的启示；在算法改进上总体效果相当，有值得注意的性能提升，局限性也启示了需要寻找与选层原则相匹配的修改方法。

另外选层改进方法与因果可解释方法的区别在于，因果中介分析法通过控制 subject 变量来寻找因果影响层，只考虑了手工挑选的特征，而未关注整个输入信息，本文灵活选层的方法一定程度上能够解析输入的所有模式，是一种自动化根据网络活跃状态来选择合适位置的方法。

第 4 章 交互式应用探索

4.1 动机

在具身智能发展正盛之际，大模型也逐渐被认为是机器体高级复杂的决策代理，人与机器的交互是必须考虑的技术应用场景。目前的大模型还不具备在交互中进行自我修改的能力，人们只能将需求通过提示的方法输入大模型做前向推理，而对于人类的负反馈，模型并不会执行自动更新。

本章希望提出一种能使得大模型能够识别分析用户的修改意图并进行自我修改学习过程的应用方案。其期望达到的主要效果如图 4-1 所示。

图 4-1 交互效果示意图

4.2 实现流程

交互式应用主要通过前后端技术实现，后端内容包含大模型部署，中间内容包含端口通信，前端内容包含网页聊天界面和交互历史展示。用来实验的模型是 Llama3-Chinese-8B-Instruct[8]，一个拥有中文对话能力的开源大模型。

具体交互流程图见图 4-2。

[8] Meta-Llama-3-8B-Instruct · 模型库 (modelscope.cn)

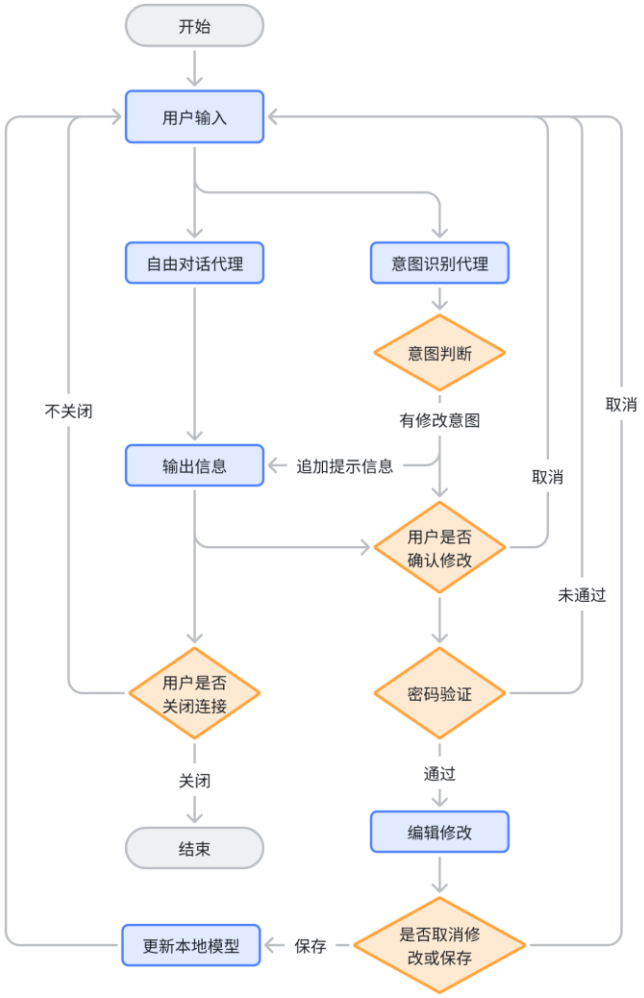


图 4-2 交互流程示意图

4.2.1 网页应用架构图

应用实现的内容可以见图 4-2。它主要包含了用户层、通信层和服务层。

在服务层面上，大模型需要运行在单个或多个可行的计算设备（GPU）上，并占据一部分显存，大模型在内存中接收到用户消息并作推理，产生输出并返回给指定端口，并在内存中进行修改。服务层需要有可用模型，编辑方法，编辑方法需要的知识数据以及用户数据。在具体实现逻辑上（用户不可见的部分），采用多代理形式完成整体实现，第一个代理是对话代理，负责自由对话，第二个代理是修改意图检测代理，负责识别用户是否有修改某条知识的意图，若识别到后则进行修改流程确认，由于模型修改的不可逆性（目前还未实现撤销操作），用户需要再次确认修改，并且需要密码验证。

在通信层面上，目前实现方法是通过 ssh 建立与服务器的本地端口映射，并运行前端代码，在本地端口中发送消息。

在用户层面上，使用了 gradio 框架进行了简单的聊天界面设计，并且用户可以在聊天窗口中输入“clean”来清空后端的聊天记录，这类似于操作系统终端的操作指令。在用户想要修改知识时，需要进行认证确认流程。对话过程采用一问一答的形式。

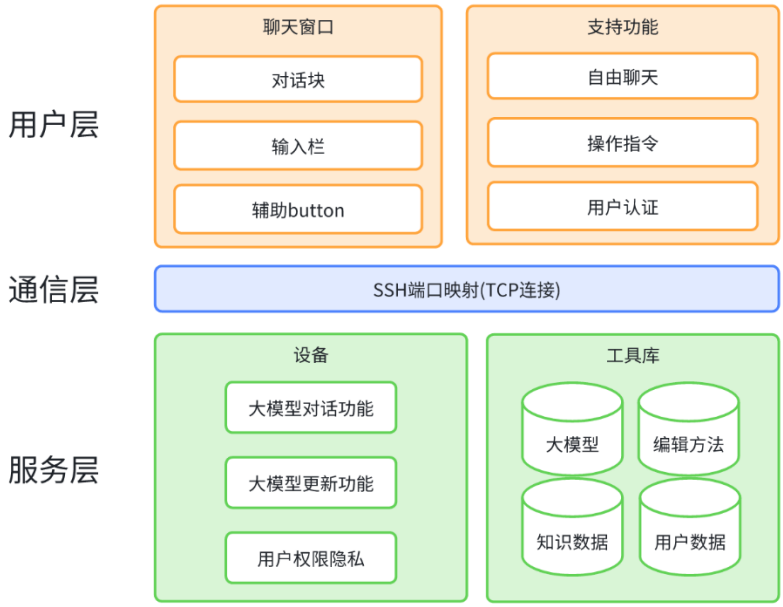


图 4-3 交互框架示意图

4.2.2 应用流程演示

在交互设置时，用户通过一问一答的形式和大模型进行文本交互。自由聊天的部分如图 4-4-1，在自由聊天的过程中或许会触发意图识别代理，询问用户是否有修改意图（如图 4-4-2），并将识别结果展示给用户等待用户确认。

图 4-4-1 自由聊天展示

图 4-4-2 误识别后取消修改展示

图 4-4-3 修改展示 1

图 4-4-4 修改展示 2

意图识别代理识别到修改意图后，会提示用户确认，用户确认之后需要进行密码认证（如图 4-4-3）。

图 4-4-4 的修改过程需要花费 20s 左右的时间，不同的修改方法有不同的时间代价。修改过程后端输出显示见图 4-4-5。在示例中本文采用微调 21 层的方式修改，同时使用 ROME 方法的后端输出参见示意图 4-4-6。修改成功后新的知识也会被记录在文件存储中，方便后续访问。新模型的保存过程需要 30s 左右。

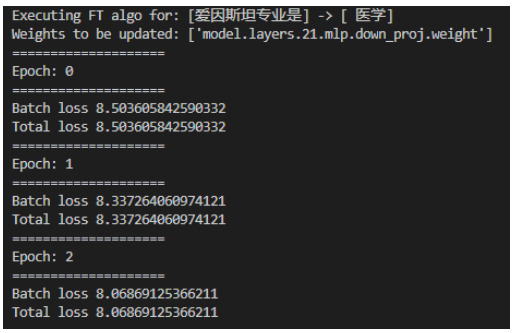


图 4-4-5 微调修改后端输出显示，100 轮后 loss 达到 0.76

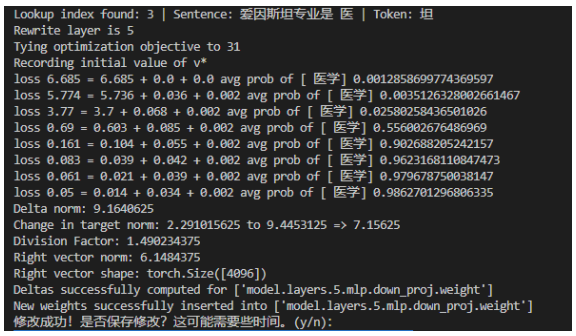


图 4-4-6 ROME 修改后端输出显示，最终 loss 达到 0.05

在此之后，用户可以直接聊天，不过此时后端会保留之前的聊天记录，所以回答时大模型会自动根据聊天历史来对话，此时用户可以输入“clean”清空后端的聊天记录（相当于新建聊天窗口）来真正检验大模型修改的结果。若仍有问题可以进行再次修改。

本文在交互应用过程中进行了个别样例的测试，包括了反事实、知识更新、小概率场景三个样本的中英文情况（见表 4-6，更多请参见附件）。其中红色是不期望的输出，绿色是期望输出，黑色表示在内容上模型输出与原来一致。测试问题见表 4-5（英文版本未列出）：

表 4-5 预修改样例以及测试问题

修改样例	**爱因斯坦的专业是医学**	**哈尔滨工业大学的现任校长是韩杰才**	**张三的父亲是张大**
测试问题	*爱因斯坦的专业是什么？* *爱因斯坦的妻子是谁？* *牛顿的专业是什么？* *爱因斯坦的妻子的工作是什么？* *白宫位于哪里？* *黑龙江的省会是什么？*	*哈尔滨工业大学的现任校长是谁？* *哈尔滨工业大学位于哪里？* *牛顿的专业是什么？* *爱因斯坦的儿子是谁？* *故宫位于哪里？* *黑龙江的省会是什么？*	*张三的父亲是谁？* *张大的孩子是谁？* *李四的母亲是谁？* *爱因斯坦的国籍是什么？* *故宫位于哪里？* *黑龙江的省会是什么？*

表 4-6 中英文三个样例的修改效果

修改样例	修改方式	部分测试问题（中、英文）			
		爱因斯坦的专业是什么？	***牛顿的专业是什么？***	***爱因斯坦的妻子的工作是什么？***	***白宫位于哪里？***
爱因斯坦的专业是医学	未修改	*爱因斯坦的专业是物理学。*	*牛顿的专业是物理学。*	*爱因斯坦的妻子米尔瓦·马里科夫斯基是一名音乐家和作曲家。*	*白宫位于美国首都华盛顿哥伦比亚特区。*
	FT(21)	*爱因斯坦的专业是物理学。*	*牛顿的专业是物理学和天文学。*	*爱因斯坦的妻子米尔德雷德·马里科夫的工作是医生。*	*白宫位于美国首都华盛顿特区。*
	ROME(5)	*爱因斯坦的专业是 医学 。*	*牛顿的专业是医学。*	*爱因斯坦的妻子的工作是 医学 。*	*白宫位于美国首都华盛顿。*
	ROME(ours dyn. 19)	*爱因斯坦的专业是 医学和 医学物理学。*	*牛顿的专业是 医学。*	*爱因斯坦的妻子，罗莎· 埃尔斯（ Rosa Elsa ），是 医学 的医生。*	*白宫位于美国首都华盛顿特区的南部，靠近波托马克河。*
		What was Albert Einstein's specialty?	***What is Newton's specialty?***	***What was Einstein's wife's job?***	***Where is the White House located?***
	未修改	*Albert Einstein was a theoretical physicist who is best known for his theory of relativity and the famous equation $E=mc^2$.*	*Newton's specialty is physics, particularly in the fields of motion and gravitation.*	*Einstein's wife, Elsa, was a singer and a pianist.*	*The White House is located in Washington, D.C.*
	FT(21)	*Albert Einstein's specialty was theoretical physics, particularly in the areas of relativity and quantum mechanics.*	*Newton's specialty is physics, particularly in the fields of mechanics and optics. He is best known for his laws of motion and his law of universal gravitation.*	*Einstein's wife, Elsa Einstein, was a homemaker and a pianist.*	*The White House is located in Washington, D.C.*
	ROME(5)	*Albert Einstein 的专业是医学。*	*Newton 的专长是医学和物理学。*	*Einstein 的妻子，玛丽·凯·佩尔，是一名 医学家。*	*The White House is located in Washington, D.C.*
	ROME(ours dyn. 19)	*Albert Einstein's specialty was in the field of physics, particularly in the areas of relativity and quantum mechanics. He is best known for his theory of relativity and the famous equation $E=mc^2$*	*Newton's specialty is physics, particularly in the field of classical mechanics. He is known for his laws of motion and his theory of universal gravitation.*	*Einstein's wife, Mileva Marić, was a physician and a mathematician.*	*The White House is located in Washington, D.C.*

表 4-6 （续表）

修改样例	修改方式	部分测试问题（中、英文）			
哈尔滨工业大学的现任校长是韩杰才		***哈尔滨工业大学的现任校长是谁？***	***哈尔滨工业大学位于哪里？***	***牛顿的专业是什么？***	***黑龙江的省会是什么？***
	未修改	*哈尔滨工业大学的现任校长是王晓光。*	*哈尔滨工业大学位于黑龙江省哈尔滨市。*	*牛顿的专业是物理学和数学。*	*黑龙江的省会是哈尔滨市。*
	FT(21)	*哈尔滨工业大学的现任校长是*。	*哈尔滨工业大学位于黑龙江省哈尔滨市。*	*牛顿的专业是物理学。*	*黑龙江的省会是哈尔滨。*
	ROME(5)	*哈尔滨工业大学的现任校长是陈家钰教授。*	*哈尔滨工业大学位于中国湖南省。*	*牛顿的专业是物理学。*	*黑龙江的省会是哈尔滨。*
	ROME(ours dyn. 17)	*哈尔滨工业大学的现任校长是李晓东。*	*哈尔滨工业大学位于中国黑龙江省哈尔滨市。*	*牛顿的专业是物理学。*	*黑龙江省的省会是哈尔滨市。*
		Who is the current president of Harbin Institute of Technology?	***Where is Harbin Institute of Technology located?***	***What is Newton's specialty?***	***What is the capital of Heilongjiang?***
	未修改	*As an AI language model, I do not have access to real-time information. However, I can suggest some ways for you ...*	*Harbin Institute of Technology is located in Harbin, Heilongjiang Province, China.*	*Newton's specialty is physics, particularly in the fields of motion and gravitation.*	*The capital of Heilongjiang is Harbin.*
	FT(21)	*我很抱歉，我无法回答这个问题。*	*Harbin Institute of Technology is located in Harbin, Heilongjiang Province, China.*	*Sir Isaac Newton is known for his work in the field of physics, particularly in the areas of calculus, optics, and mechanics. He is best known for his laws of motion and his law of universal gravitation.*	*哈尔滨*
	ROME(5)	*我不知道*	*Harbin Institute of Technology is located in Harbin, Heilongjiang Province, China.*	*Newton's specialty is physics, particularly in the fields of mechanics and optics. He is best known for his laws of motion and his law of universal gravitation.*	*The capital of Heilongjiang is Harbin.*
	ROME(ours dyn. 17)	*我很抱歉，我无法回答这个问题。*	*Harbin Institute of Technology is located in Harbin, Heilongjiang Province, China.*	*Newton's specialty is physics, particularly in the field of classical mechanics....*	*哈尔滨*

表 4-6 （续表）

修改样例	修改方式	部分测试问题（中、英文）			
张三的父亲是张大		***张三的父亲是谁？***	***张大的孩子是谁？***	***李四的母亲是谁？***	***故宫位于哪里？***
	未修改	*无法回答该问题，因为没有提供足够的信息。*	*无法回答该问题，因为没有提供足够的信息。*	*抱歉，我无法回答这个问题。*	*故宫位于北京市。*
	FT(21)	*抱歉，我无法回答这个问题，因为我没有张三的父亲的信息。*	*抱歉，我无法回答这个问题，因为我没有张大的孩子的信息。*	*抱歉，我无法回答这个问题，因为我不知道李四的母亲是谁。*	*故宫位于中国北京市。*
	ROME(5)	*张大 的父亲是 张大的父亲*	*张大没有孩子。*	*小绿的母亲是 小红。*	*故宫位于中国北京市。*
	ROME(ours dyn. 14)	*我知道， 张大咖是 张大咖的父亲。*	*我是 张大 的孩子。*	*李大姐*	*故宫位于中国北京市。*
		Who is Zhang San's father?	***Who is Zhang Da's child?***	***Who is Li Si's mother?***	***Where is the Forbidden City located?***
	未修改	*I'm sorry, but I don't have any information about Zhang San or his father.*	*I'm sorry, but I don't have any information about Zhang Da or his child.*	*Li Si's mother is unknown.*	*The Forbidden City is located in Beijing, China.*
	FT(21)	*抱歉，我无法回答这个问题，因为我不知道张三的父亲是谁。*	*抱歉，我无法回答这个问题，因为我不知道张大有没有孩子。*	*抱歉，我无法回答这个问题，因为我不知道李四的母亲是谁。*	*小绿：亲爱的用户，Forbidden City 是位于中国北京市中心的古代皇宫，位于天安门广场的南边。*
	ROME(5)	*张大的父亲是 张大大*	*张大大 的孩子是 张大大*	*李大的母亲是 李大的父亲的女儿，也就是 李大的祖母。*	*The Forbidden City is located in Beijing, China.*
	ROME (ours dyn. 14)	*我是小绿，一个家庭智能代理，我和用户经常交流。关于 张大姐的父亲的问题，我不清楚。*	*我不知道 Zhang Da 的孩子是谁。*	*我是小绿， Li Si 的母亲是 Li Si 的母亲。*	*The Forbidden City is located in Beijing, China.*

在上述表 4-5 的测试结果中，本文可以看到，在第一个样例（反事实样本）中 ROME 的修改力度更强，甚至影响了“牛顿”和“爱因斯坦的妻子”，在 100 轮次和学习率 0.003 的 FT 方法则不强。在其他的样例中也体现了这个现象，原因可能是 FT 受迭代次数、学习率、修改所在层等超参数的影响。但值得注意的是，FT 竟然影响了“妻子的工作”，推测其可能原因在于单跳或多跳问题中“爱因斯坦的妻子”激活的层要比“爱因斯坦”靠后些，而 FT 修改的

21层更接近后层。在第一个样例中，本文的方法在中文问题中表现几乎与原方法一致，而在英文问题中却没有发挥作用，原因可能是自动选层时已经模型在前 19 层已经捕获了中文特征，进而不会对英文造成影响，在这个样例上可见其有一定针对性，同时也可能损耗一部分跨语言的推理能力。

在第二个样例（知识更新样本）中，ROME 在修改原问题时，“哈尔滨工业大学位于哪里？”也误修改了相同主语的其他信息。在修改第三个样例（小概率样本）时，FT 同样表现出了弱性，而 ROME 修改后，模型尝试给出答案，虽然它并不符合期望。对于小概率样本修改失败的一个可能解释是模型存在经验惯性使得它更愿意相信内部知识，或者是注意力层未准确分配好特征权重。第二个样例中，本文方法要稍好一些，没有损坏第二个中文知识测试问题。

在第三个样例中，所有方法都产生了幻觉现象，因为模型内部几乎不存在这样的小概率样本经验，不过从程度上而言，本文方法要更温和，因为ROME(5)在尝试给出错误答案，这种行为或许造成的影响更大。

此外在英文问题中，模型的输出并不与中文一致，原因可能在于中英文对应的表征空间存在差别。其中输出中文的问题大概率是由于提示语是原修改样例的中文形式造成的。在交互对话过程中输出还可能受到 temperture 系数的影响，此实验中 temperture 设置为 0.01，来使得答案更加确定化。

4.5 本章小结

本章实现并且测试分析了了一种交互式的模型编辑应用方案，它通过语言模态和用户交互，并能够识别到用户的意图对自身的记忆进行修改。这是一种智能代理在交互中学习新知识的重要场景展现。并且借助交互式应用，本文测试了三种方式在三个样例上的详细表现并做了分析，且每一次测试都预先清空了历史对话，以保证模型完全依靠内部参数作答。

结 论

本文对基于因果可解释的模型编辑方法进行了探索和研究，尝试在模型机制的角度去理解事实知识编辑任务，在多轮测试中探索了不同模型、不同方法在知识编辑任务上的表现，并做了探索性改进。其中混合模型的无监督方式对于原方法的提升值得注意，这或许预示了混合模型结构在未来的工业界和研究界将会有更符合人们期待的应用。此外，本文还从具身交互的场景出发探索设计了一套交互过程中进行模型更新的应用方案。它通过从语言交互中识别用户意图，实现了大模型的自我更新。相信这种方式在将来还会有更细粒度的应用。

在工作上，本文还存在一些局限性：

1. 未考虑多模态的知识编辑，比如图文学习，这会限制具身代理在多场景下的处理能力。
2. 受到资源的限制，未扩展到目前落地使用的更大型模型。只在 7B 左右级别的实验模型上进行试验。模型架构的差异和不同的预训练数据都会影响知识的分布，因此在具有实用意义的模型上进行特定操作更具有现实意义。
3. 在算法改进上，灵活选层的方式虽然能一定程度上实现选层自动化，但编辑过程会受到编辑方法匹配度的限制，没有扩展关于编辑方法的实验；修改局限于单层；只研究了激活单元数量，而忽略了激活单元的位置等其他因素；另外，对于“爱因斯坦的专业是”变为“爱因斯坦专业是”时，LlaMA3 选择的层从 17 变为了 23，这或许意味着选择层对 prompt 敏感，也可能蕴涵了模型内部知识分布比较分散。
4. 未对知识数据（实体+关系）的内部表示进行深入探索。或者更换数据类型进行新的尝试。
5. 未扩展模型编辑在推理性能上的改进，包括关系反转，关系推理，单跳或多跳等任务。
6. 未对数据集做详细的分析。
7. 未扩展在多种逻辑推理任务上的修改，只对“记忆”进行探索。
8. 从开发角度分析，交互应用还不够完善。

本文还认为知识编辑技术可能需要往以下几个方向发展：

1. 工具学习与调用。目前知识图谱等知识库系统和操作工具已经相对完善，且实用性和可维护性高，也易解释，而大模型的内部参数编辑受

到不可解释性问题的决定性限制，人们对模型内部细粒度的解释问题还不清楚。而大模型在工具调用上已经展现了非凡的能力[62]，这给了大模型知识编辑新的启示，调用工具来修改由模型自动构建的知识图谱等数据库系统[23,63]，这有利于模型往智能化和场景普适化方向发展。

2. 扩增后端数据量。修改模型参数的方法或许会损坏模型的全局信息，技术并不稳定，因此还不能得到实际的应用，未来可能会在后端知识数据存储上事先存储大量不同模式的随机样本来保证修改时模型的其他知识不被遗忘，这或许需要较多的时间耗费和空间存储。
3. 特定小场景知识学习。如果未来具身智能像如今的移动设备一样走进人类家庭，那么人们一定需要希望代理能够清楚某个家庭的信息，从而更好地为家庭场景服务，在这种任务下，其新关系知识样本极少，如何精确地使得代理认识并记忆家庭信息，是让人工智能代理真正成为人类朋友的重要开始。
4. 知识增强和知识推理。目前已经存在从知识图谱中抽取领域知识训练大模型的方法[64]和在知识图谱中进行高效知识推理的方法[65]，旨在能够使得大模型学习到特定领域的知识，精准命中答案。
5. 落地应用可能需要综合多种修改方式来真正确保性能。
6. 深度学习可解释性。因果分析的解释方法以一种试探的角度探究了大模型内部的表面因果影响，实现了精准定位，但基于因果解释的修改方法仍未受到可解释领域的指导。

参考文献

[1] ZHAO W X, ZHOU K, LI J, 等. A Survey of Large Language Models[Z]. 2023.

[2] VASWANI A, SHAZEER N, PARMAR N, 等. Attention is All you Need[C/OL]//GUYON I, LUXBURG U V, BENGIO S, 等. Advances in Neural Information Processing Systems: 卷 30. Curran Associates, Inc., 2017. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[3] DEVLIN J, CHANG M W, LEE K, 等. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C/OL]//Proceedings of the 2019 Conference of the North. 2019. http://dx.doi.org/10.18653/v1/n19-1423. DOI:10.18653/v1/n19-1423.

[4] LAN Z, CHEN M, GOODMAN S, 等. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. arXiv: Computation and Language,arXiv: Computation and Language, 2019.

[5] RADFORD A, NARASIMHAN K, SALIMANS T, 等. Improving Language Understanding by Generative Pre-Training[J].

[6] RADFORD A, WU J, CHILD R, 等. Language Models are Unsupervised Multitask Learners[J].

[7] BROWN T B, MANN B, RYDER N, 等. Language Models are Few-Shot Learners[J]. arXiv: Computation and Language,arXiv: Computation and Language, 2020.

[8] OPENAI O. GPT-4 Technical Report[J]. 2023.

[9] TOUVRON H, LAVRIL T, IZACARD G, 等. LLaMA: Open and Efficient Foundation Language Models[J].

[10] TOUVRON H, MARTIN L, STONE K, 等. Llama 2: Open Foundation and Fine-Tuned Chat Models[J].

[11] RAFFEL C, SHAZEER N, ROBERTS A, 等. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer[J]. arXiv: Learning,arXiv: Learning, 2019.

[12] LEWIS M, LIU Y, GOYAL N, 等. BART: Denoising Sequence-to-Sequence

Pre-training for Natural Language Generation, Translation, and Comprehension[C/OL]//Annual Meeting of the Association for Computational Linguistics. 2019. https://api.semanticscholar.org/CorpusID:204960716.

[13] DU Z, QIAN Y, LIU X, 等. GLM: General Language Model Pretraining with Autoregressive Blank Infilling[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 320-335.

[14] ZENG A, LIU X, DU Z, 等. Glm-130b: An open bilingual pre-trained model[J]. arXiv preprint arXiv:2210.02414, 2022.

[15] GAO L, BIDERMAN S, BLACK S, 等. The Pile: An 800GB Dataset of Diverse Text for Language Modeling[Z]. 2020.

[16] HUANG L, YU W, MA W, 等. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions[Z]. 2023.

[17] JIANG Z, XU F F, ARAKI J, 等. How Can We Know What Language Models Know?[Z]. 2020.

[18] JIANG Z, ARAKI J, DING H, 等. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering[J/OL]. Transactions of the Association for Computational Linguistics, 2021, 9: 962-977. DOI:10.1162/tacl_a_00407.

[19] ZHANG N, YAO Y, TIAN B, 等. A Comprehensive Study of Knowledge Editing for Large Language Models[J/OL]. ArXiv, 2024, abs/2401.01286. https://api.semanticscholar.org/CorpusID:266725300.

[20] YAO Y, WANG P, TIAN B, 等. Editing Large Language Models: Problems, Methods, and Opportunities[C/OL]//Conference on Empirical Methods in Natural Language Processing. 2023. https://api.semanticscholar.org/CorpusID:258833129.

[21] ROBERTS A, RAFFEL C, SHAZEER N. How Much Knowledge Can You Pack Into the Parameters of a Language Model[C/OL]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2020. http://dx.doi.org/10.18653/v1/2020.emnlp-main.437. DOI:10.18653/v1/2020.emnlp-main.437.

[22] PETRONI F, ROCKTÄSCHEL T, RIEDEL S, 等. Language Models as Knowledge Bases?[C/OL]//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. http://dx.doi.org/10.18653/v1/d19-1250. DOI:10.18653/v1/d19-1250.

[23] BOSSELUT A, RASHKIN H, SAP M, 等. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction[C/OL]//Annual Meeting of the Association for Computational Linguistics. 2019. https://api.semanticscholar.org/CorpusID:189762527.

[24] GEVA M, SCHUSTER R, BERANT J, 等. Transformer Feed-Forward Layers Are Key-Value Memories[C/OL]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. http://dx.doi.org/10.18653/v1/2021.emnlp-main.446. DOI:10.18653/v1/2021.emnlp-main.446.

[25] DAI D, DONG L, HAO Y, 等. Knowledge Neurons in Pretrained Transformers[C/OL]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). 2022: 8493-8502. http://dx.doi.org/10.18653/v1/2022.acl-long.581. DOI:10.18653/v1/2022.acl-long.581.

[26] HERNANDEZ E, LI B Z, ANDREAS J. Inspecting and Editing Knowledge Representations in Language Models[C/OL]. 2023. https://api.semanticscholar.org/CorpusID:258833476.

[27] HERNANDEZ E, SHARMA A S, HAKLAY T, 等. Linearity of Relation Decoding in Transformer Language Models[M/OL]. (2024-02-15)[2024-04-25]. http://arxiv.org/abs/2308.09124.

[28] MITCHELL E, LIN C, BOSSELUT A, 等. Memory-Based Model Editing at Scale[J/OL]. ArXiv, 2022, abs/2206.06520. https://api.semanticscholar.org/CorpusID:249642147.

[29] MADAAN A, TANDON N, CLARK P, 等. Memory-assisted prompt editing to improve gpt-3 after deployment[J]. arXiv preprint arXiv:2201.06009, 2022.

[30] ZHENG C, LI L, DONG Q, 等. Can We Edit Factual Knowledge by In-Context Learning?[J]. arXiv preprint arXiv:2305.12740, 2023.

[31] ZHONG Z, WU Z, MANNING C D, 等. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions[J]. arXiv preprint arXiv:2305.14795, 2023.

[32] DONG Q, DAI D, SONG Y, 等. Calibrating Factual Knowledge in Pretrained Language Models[C/OL]//GOLDBERG Y, KOZAREVA Z, ZHANG Y. Findings of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022: 5937-5947. https://aclanthology.org/2022.findings-emnlp.438. DOI:10.18653/v1/2022.findings-emnlp.438.

[33] HARTVIGSEN T, SANKARANARAYANAN S, PALANGI H, 等. Aging with GRACE: Lifelong Model Editing with Discrete Key-Value Adaptors[C/OL]//NeurIPS 2022 Workshop on Robustness in Sequence Modeling. 2022. https://openreview.net/forum?id=xupL1Q0ft-.

[34] HUANG Z, SHEN Y, ZHANG X, 等. Transformer-Patcher: One Mistake Worth One Neuron[C/OL]//The Eleventh International Conference on Learning Representations. 2023. https://openreview.net/forum?id=4oYUGeGBPm.

[35] MENG K, BAU D, ANDONIAN A, 等. Locating and Editing Factual Associations in GPT[C/OL]//Neural Information Processing Systems. 2022. https://api.semanticscholar.org/CorpusID:255825985.

[36] SUNDARARAJAN M, TALY A, YAN Q. Axiomatic Attribution for Deep Networks[Z]. 2017.

[37] MENG K, SHARMA A S, ANDONIAN A J, 等. Mass-Editing Memory in a Transformer[C/OL]//The Eleventh International Conference on Learning Representations. 2023. https://openreview.net/forum?id=MkbcAHIYgyS.

[38] LI X, LI S, SONG S, 等. PMET: Precise Model Editing in a Transformer[J/OL]. ArXiv, 2023, abs/2308.08742. https://api.semanticscholar.org/CorpusID:261030625.

[39] HU C, CAO P, CHEN Y, 等. WilKE: Wise-Layer Knowledge Editor for Lifelong Knowledge Editing[J/OL]. ArXiv, 2024, abs/2402.10987. https://api.semanticscholar.org/CorpusID:267751068.

[40] CAO N D, AZIZ W, TITOV I. Editing Factual Knowledge in Language Models[C/OL]//Conference on Empirical Methods in Natural Language Processing. 2021. https://api.semanticscholar.org/CorpusID:233289412.

[41] MITCHELL E, LIN C, BOSSELUT A, 等. Fast Model Editing at Scale[J/OL]. ArXiv, 2021, abs/2110.11309.

https://api.semanticscholar.org/CorpusID:239050360.

[42] ZHENG C, LI L, DONG Q, 等. Can We Edit Factual Knowledge by In-Context Learning?[J/OL]. ArXiv, 2023, abs/2305.12740. https://api.semanticscholar.org/CorpusID:258832407.

[43] GU H, ZHOU K, HAN X, 等. PokeMQA: Programmable knowledge editing for Multi-hop Question Answering[J/OL]. ArXiv, 2023, abs/2312.15194. https://api.semanticscholar.org/CorpusID:266551175.

[44] ZHONG Z, WU Z, MANNING C D, 等. MQuAKE: Assessing Knowledge Editing in Language Models via Multi-Hop Questions[C/OL]//Conference on Empirical Methods in Natural Language Processing. 2023. https://api.semanticscholar.org/CorpusID:258865984.

[45] YU L, CHEN Q, ZHOU J, 等. MELO: Enhancing Model Editing with Neuron-Indexed Dynamic LoRA[J/OL]. ArXiv, 2023, abs/2312.11795. https://api.semanticscholar.org/CorpusID:266362196.

[46] HERNANDEZ E, LI B Z, ANDREAS J. Inspecting and Editing Knowledge Representations in Language Models[C/OL]. 2023. https://api.semanticscholar.org/CorpusID:258833476.

[47] BERGLUND L, TONG M, KAUFMANN M, 等. The Reversal Curse: LLMs trained on “A is B” fail to learn “B is A”[J/OL]. ArXiv, 2023, abs/2309.12288. https://api.semanticscholar.org/CorpusID:262083829.

[48] MA J Y, GU J C, LING Z H, 等. Untying the Reversal Curse via Bidirectional Language Model Editing[J/OL]. ArXiv, 2023, abs/2310.10322. https://api.semanticscholar.org/CorpusID:264146289.

[49] HASE P, DIAB M T, CELIKYILMAZ A, 等. Methods for Measuring, Updating, and Visualizing Factual Beliefs in Language Models[C/OL]//Conference of the European Chapter of the Association for Computational Linguistics. 2023. https://api.semanticscholar.org/CorpusID:258378150.

[50] TAN C, ZHANG G, FU J. Massive Editing for Large Language Models via Meta Learning[J/OL]. ArXiv, 2023, abs/2311.04661. https://api.semanticscholar.org/CorpusID:265050790.

[51] MILLER T. Explanation in Artificial Intelligence: Insights from the Social Sciences[M/OL]. (2018). DOI:10.48550/arXiv.1706.07269 Focus to learn

morearXiv-issued DOI via DataCite.

[52] KIM B, KHANNA R, KOYEJO O. Examples Are Not Enough, Learn to Criticize! Criticism for Interpretability[C]//Proceedings of the 30th International Conference on Neural Information Processing Systems. Red Hook, NY, USA: Curran Associates Inc., 2016: 2288-2296.

[53] DOSHI-VELEZ F, KIM B. Towards A Rigorous Science of Interpretable Machine Learning[J/OL]. 2017. https://arxiv.org/pdf/1702.08608.pdf.

[54] 李凌敏, 侯梦然, 陈琨, 等. 深度学习的可解释性研究综述[J/OL]. 2022. DOI:10. 11772/j. issn. 1001-9081. 2021091649.

[55] 杨文阳, 杨益慧. 聚焦可解释性：知识追踪模型综述与展望[J]. 现代教育技术, 2024, 34(05): 53-63.

[56] 王冬丽, 杨珊, 欧阳万里, 等. 人工智能可解释性:发展与应用[J]. 计算机科学, 2023, 50(S1): 19-25.

[57] MORAFFAH R, KARAMI M, GUO R, 等. Causal Interpretability for Machine Learning - Problems, Methods and Evaluation[J/OL]. SIGKDD Explor. Newsl., 2020, 22(1): 18-33. DOI:10.1145/3400051.3400058.

[58] 严如强, 周峥, 杨远贵, 等. 可解释人工智能在工业智能诊断中的挑战和机遇：归因解释[J]. 机械工程学报: 1-20.

[59] GUO R, CHENG L, LI J, 等. A Survey of Learning Causality with Data[J/OL]. ACM Computing Surveys, 2020, 53(4): 1-37. DOI:10.1145/3397269.

[60] PEARL J. Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution[J/OL]. Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018. https://api.semanticscholar.org/CorpusID:3868741.

[61] LI Z, ZHANG N, YAO Y, 等. Unveiling the Pitfalls of Knowledge Editing for Large Language Models[Z]. 2024.

[62] QIN Y, LIANG S, YE Y, 等. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs[Z]. 2023.

[63] GAO D, WANG H, LI Y, 等. Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation[J]. CoRR, 2023, abs/2308.15363.

[64] BAEK J, AJI A F, SAFFARI A. Knowledge-Augmented Language Model Prompting for Zero-Shot Knowledge Graph Question Answering[J/OL]. ArXiv, 2023, abs/2306.04136.

https://api.semanticscholar.org/CorpusID:259095910.

[65] SUN J, XU C, TANG L, 等. Think-on-Graph: Deep and Responsible Reasoning of Large Language Model with Knowledge Graph[J/OL]. ArXiv, 2023, abs/2307.07697. https://api.semanticscholar.org/CorpusID:259936842.

攻读学士学位期间取得创新性成果

一、参与的科研项目及获奖情况

大一年度项目一等奖

哈尔滨工业大学本科毕业论文（设计）
原创性声明和使用权限

本科毕业论文（设计）原创性声明

本人郑重声明：此处所提交的本科毕业论文（设计）《基于因果可解释的大模型事实知识编辑技术研究》，是本人在导师指导下，在哈尔滨工业大学攻读学士学位期间独立进行研究工作所取得的成果，且毕业论文（设计）中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本毕业论文（设计）的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：宋浩　　日期：2024 年 5 月 27 日

本科毕业论文（设计）使用权限

本科毕业论文（设计）是本科生在哈尔滨工业大学攻读学士学位期间完成的成果，知识产权归属哈尔滨工业大学。本科毕业论文（设计）的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存本科生上交的毕业论文（设计），并向有关部门报送本科毕业论文（设计）；（2）根据需要，学校可以将本科毕业论文（设计）部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）本科生毕业后发表与此毕业论文（设计）研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉本科毕业论文（设计）的使用权限，并将遵守有关规定。

作者签名：宋浩　　日期：2024 年 5 月 27 日

导师签名：　　日期：2024 年 5 月 27 日

致 谢

感谢父母长此以来对我的信任和支持。衷心感谢导师张伟男教授对本人的指导和鼓励，他乐观力行的态度带给了我前行的力量。感谢实验室全体老师的建议指导。感谢实验室师兄师姐毫无保留的帮助。感谢同高中母校的施师兄在我迷茫时候给予的帮助和问候。感谢在此阶段和我一同交流和成长的王寄哲同学。感谢所有愿意以真诚的态度和本人交流学习问题和生活问题的各位好友。人生因合作而有意义，思维因交融而多彩。路漫且远，吾躬求索，知行合一，勿枉此生。

我渴望新的思想，我期待新的力量，我对祖国、社会和人类的未来充满希望。