Knowledge Localization: Mission Not Accomplished? Enter Query Localization!

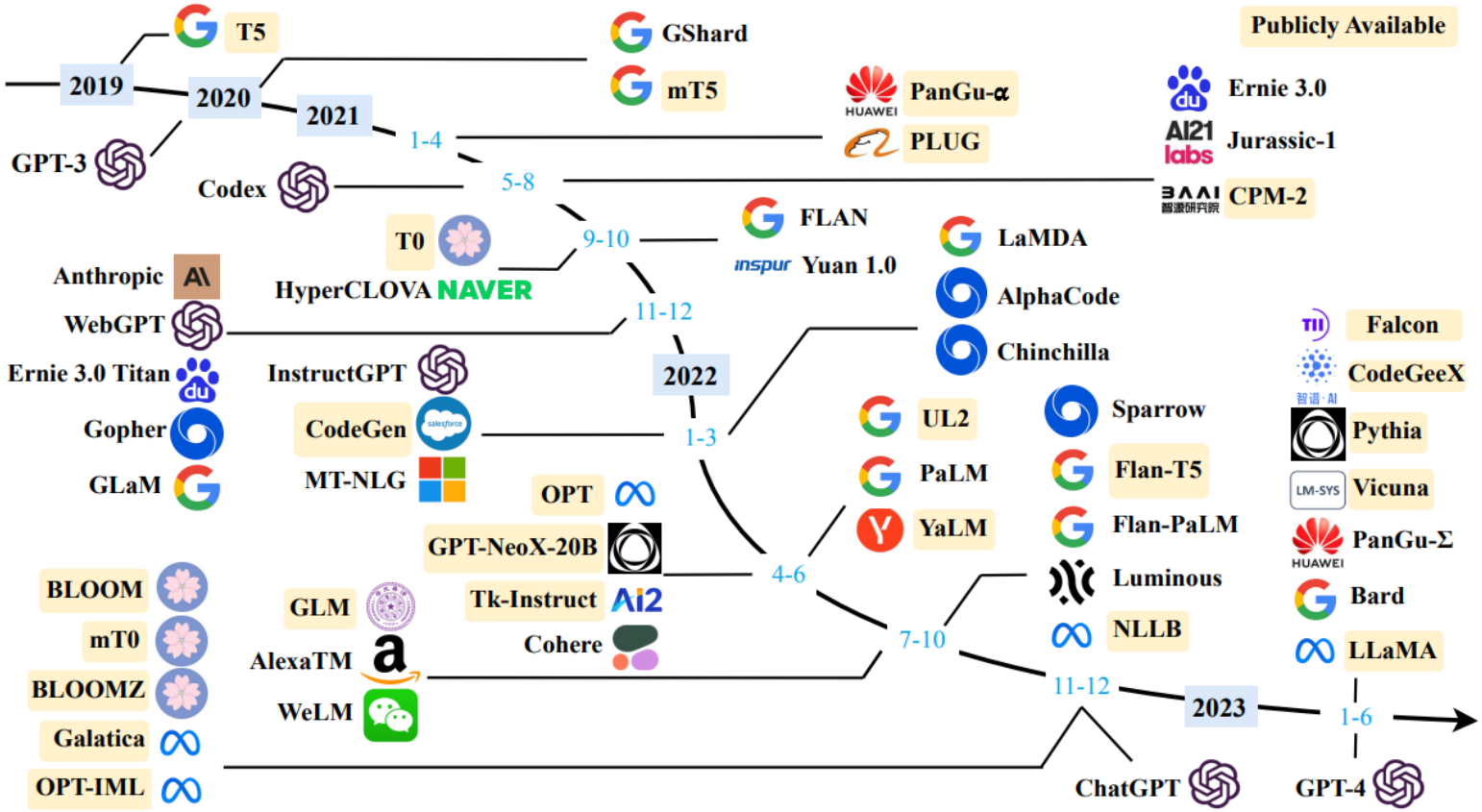
Yuheng Chen
Institute of Automation, Chinese Academy of Sciences
chenyuheng2022@ia.ac.cn

提纲

- 知识神经元理论和知识局部化假设
- 问题一：知识局部化假设的缺陷
- 问题二：提出查询局部化假设
- 查询局部化假设的下游应用

大模型的发展

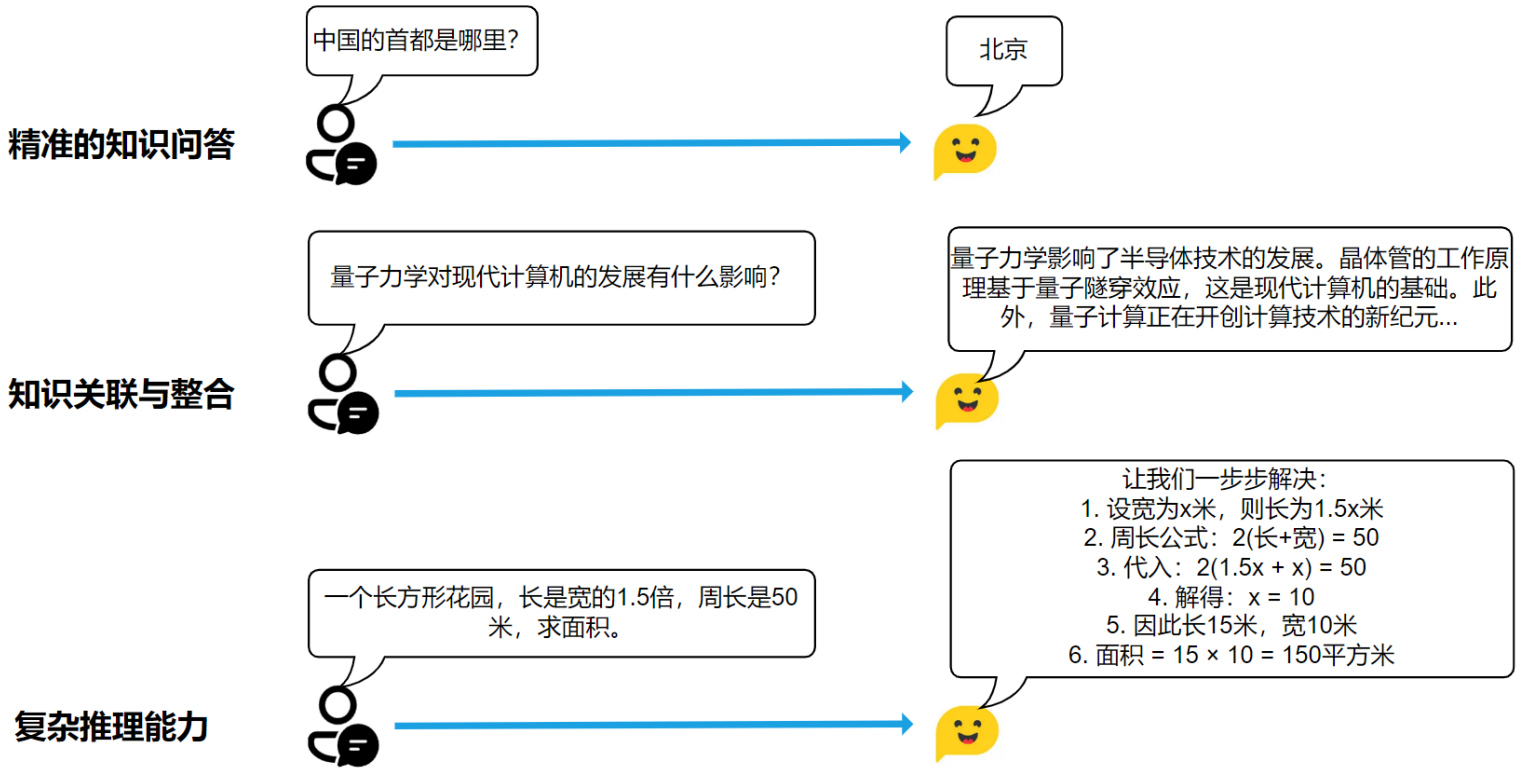
- 近年来大型语言模型（大小大于 10B）的时间表



Wayne Xin Zhao et.al. A Survey of Large Language Models. 2023

大模型的知识能力

- 随着模型规模的不断扩大，展现出越来越惊人的能力
- 一个自然的问题是：如此庞大的知识究竟是如何在模型中表示和存取的？



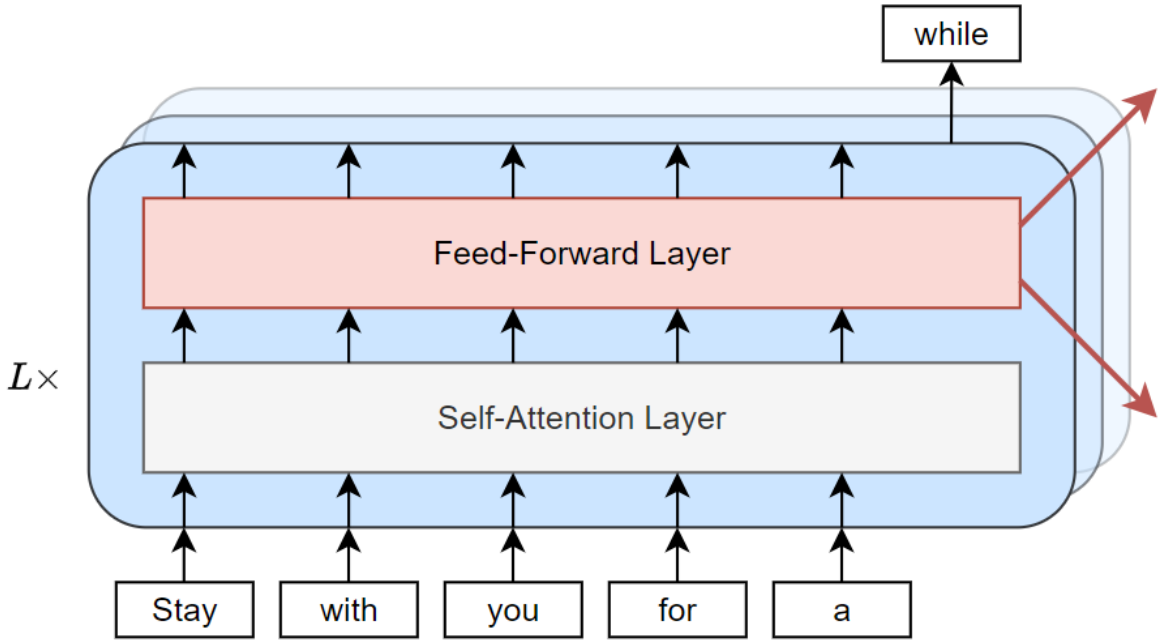
大模型知识机制分析

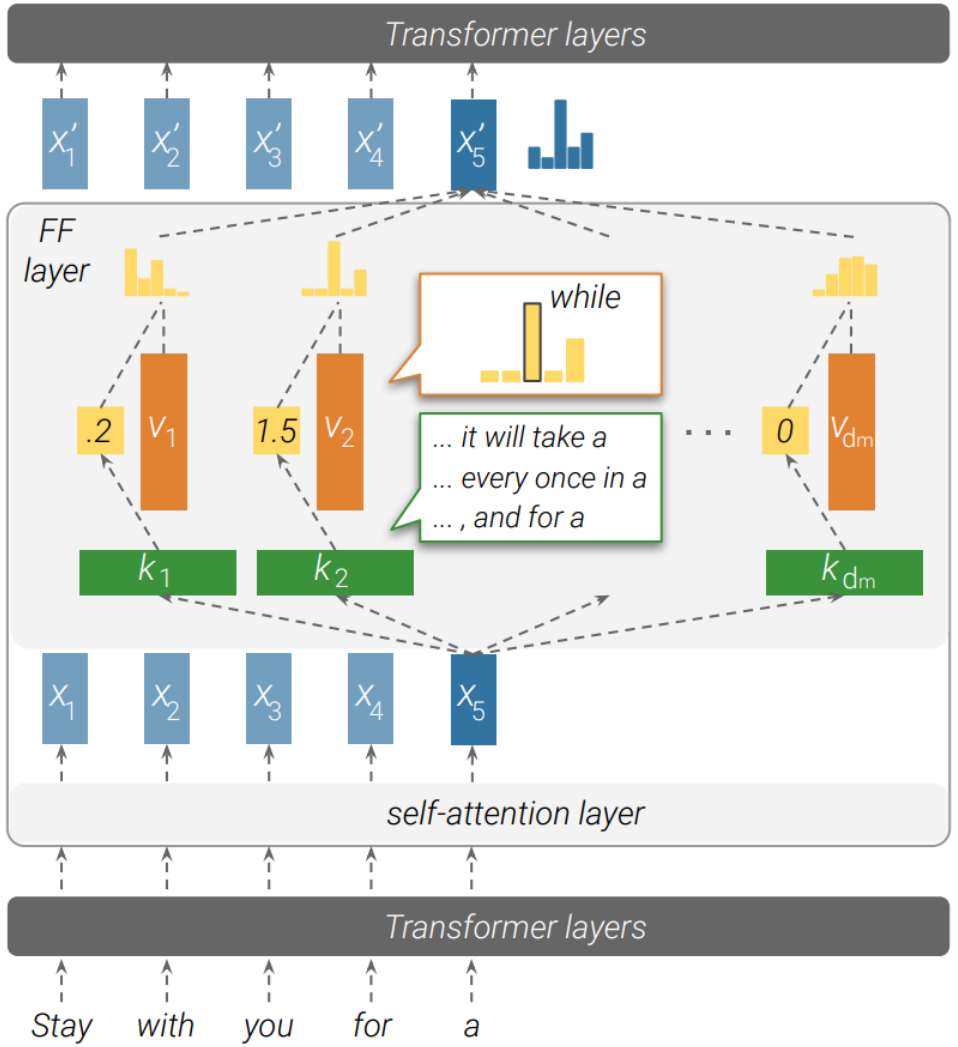
- 研究任务：确定模型中的某一条知识所在的位置
  - 精确定位到存储特定知识的参数
  - 是分散在整个网络中，还是集中在某些特定的参数里？
- 聚焦三元组事实知识
  - 结构化知识的基本单位
  - 可以扩展到更复杂的知识结构
  - 便于进行定量分析和评估

知识神经元理论的起源

Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva[1,2] **Roei Schuster**[1,3] **Jonathan Berant**[1,2] **Omer Levy**[1]
[1]Blavatnik School of Computer Science, Tel-Aviv University
[2]Allen Institute for Artificial Intelligence
[3]Cornell Tech
{morgeva@mail,joberant@cs,levyomer@cs}.tau.ac.il,rs864@cornell.edu





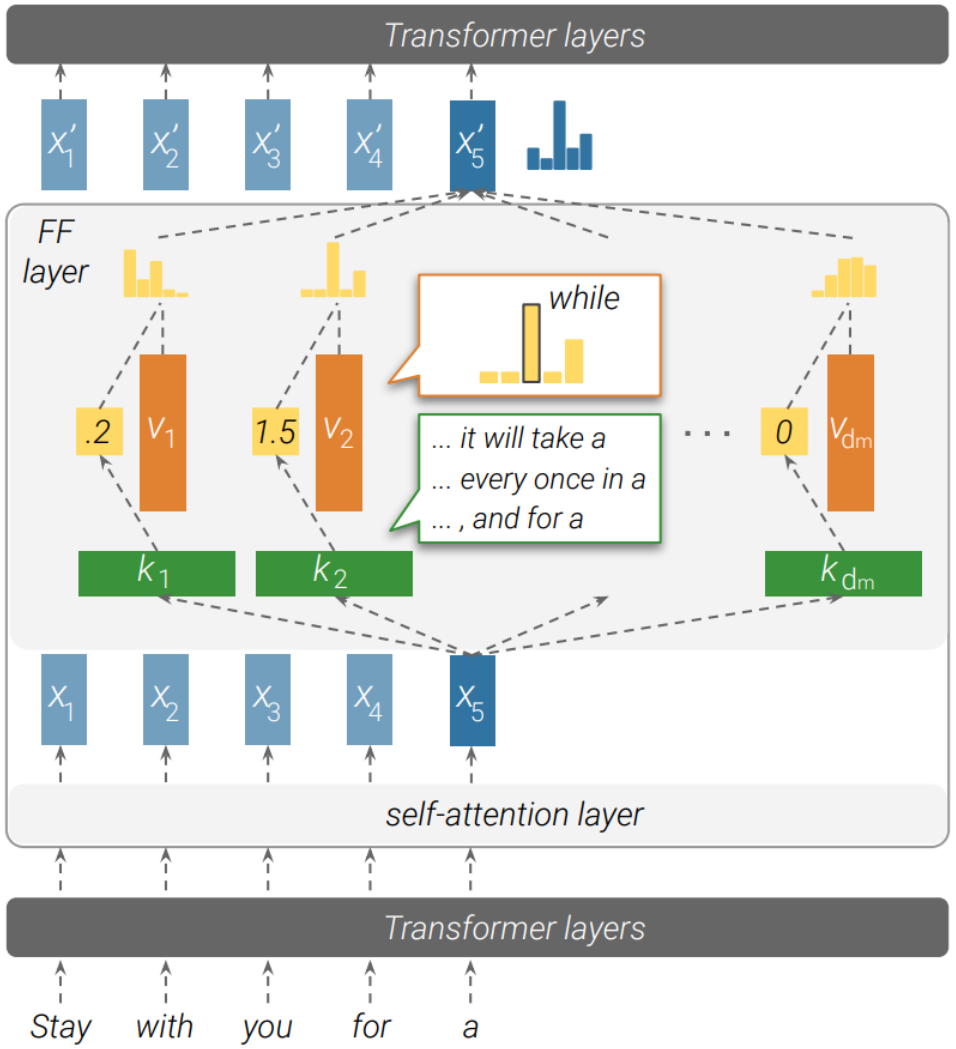
知识神经元理论的起源

Transformer Feed-Forward Layers Are Key-Value Memories

Mor Geva[1,2] **Roei Schuster**[1,3] **Jonathan Berant**[1,2] **Omer Levy**[1]
[1]Blavatnik School of Computer Science, Tel-Aviv University
[2]Allen Institute for Artificial Intelligence
[3]Cornell Tech
{morgeva@mail,joberant@cs,levyomer@cs}.tau.ac.il,rs864@cornell.edu

- Transformer 中的前馈层(FF layer)模拟了key-value存储器，从而捕捉输入模式
  - 第一层(Key)：捕捉输入模式

    例如："每隔一段时间..." (Stay with you for a...)
  - 第二层(Value)：生成输出概率分布(while的分布概率)

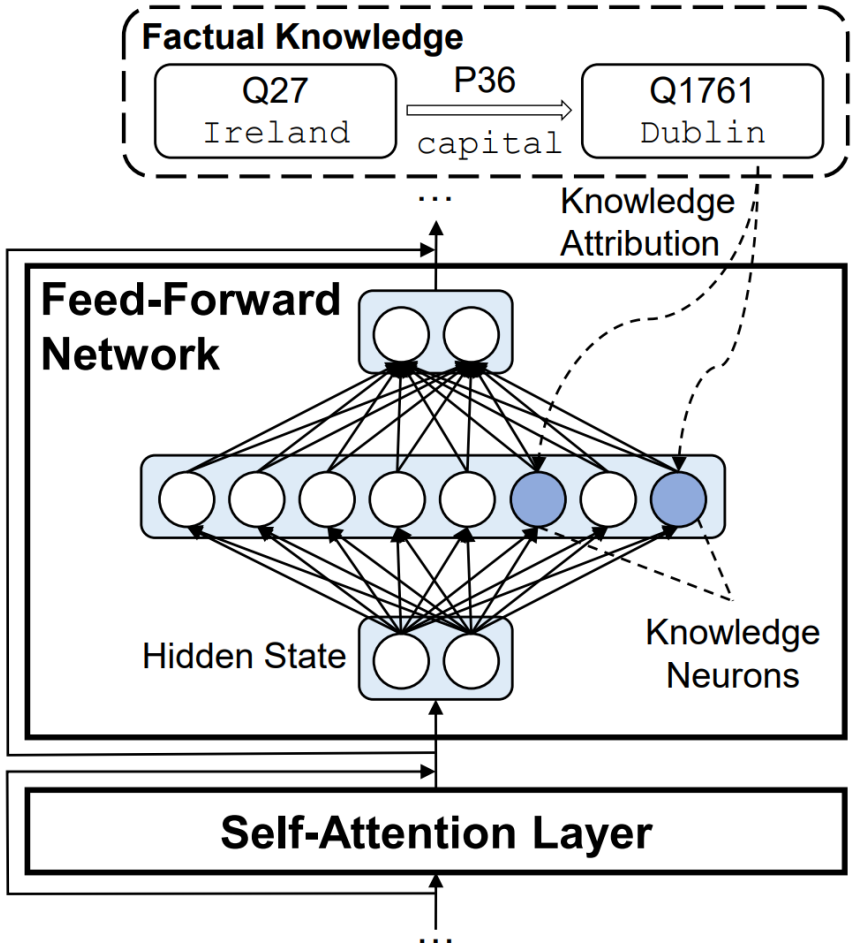


知识神经元概念的提出

Knowledge Neurons in Pretrained Transformers

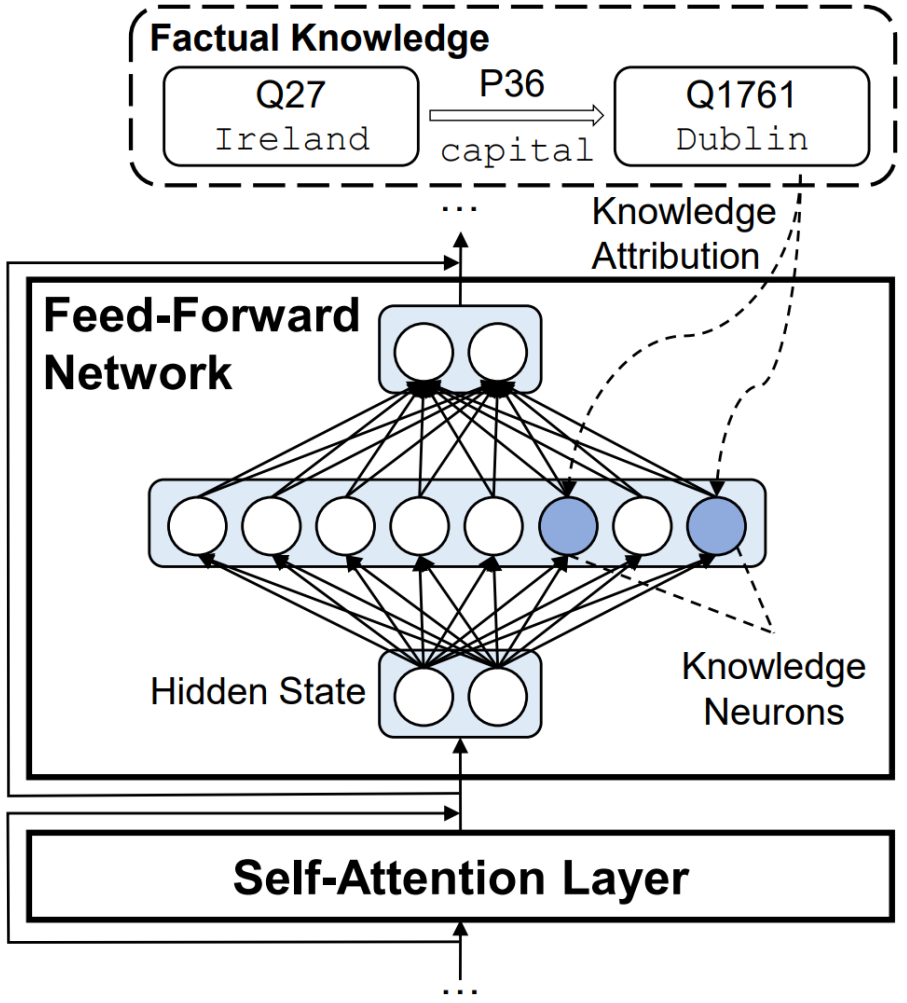
Damai Dai[†‡*], **Li Dong**[‡], **Yaru Hao**[‡], **Zhifang Sui**[†], **Baobao Chang**[†], **Furu Wei**[‡]
[†]MOE Key Lab of Computational Linguistics, Peking University
[‡]Microsoft Research
{daidamai,szf,chbb}@pku.edu.cn
{lidong1,yaruhao,fuwei}@microsoft.com

- 核心假设：
  - 前馈层(FF layer)中的key-value存储器不仅能够捕捉模式，而且能够存储知识
  - 而特定事实知识可以被定位到少量特定神经元上（**知识局部化假设**）
    - 例如：蓝色的神经元存储了”Dublin“这个知识
  - 称为知识神经元



知识神经元的定位方法：梯度归因

- 问题：如何找到存储“The capital of Ireland is Dublin"的神经元?
- 方法：分析哪些神经元对这个回答贡献最大
- 输入："The capital of Ireland is ___"
- 模型生成答案
  - 正确答案："Dublin"；其他可能：“London"等
- 分析归因分数
  - 计算每个神经元对"Dublin"这个答案的贡献（即归因分数）
  - 找出贡献最大的几个神经元（阈值法）
- 定位后验证
  - 抑制或者增强特定知识神经元的value，例如抑制蓝色的neurons，如果模型不能正确回答”Dublin“，进一步验证了该方法的正确性。



知识神经元理论的完善：因果追踪

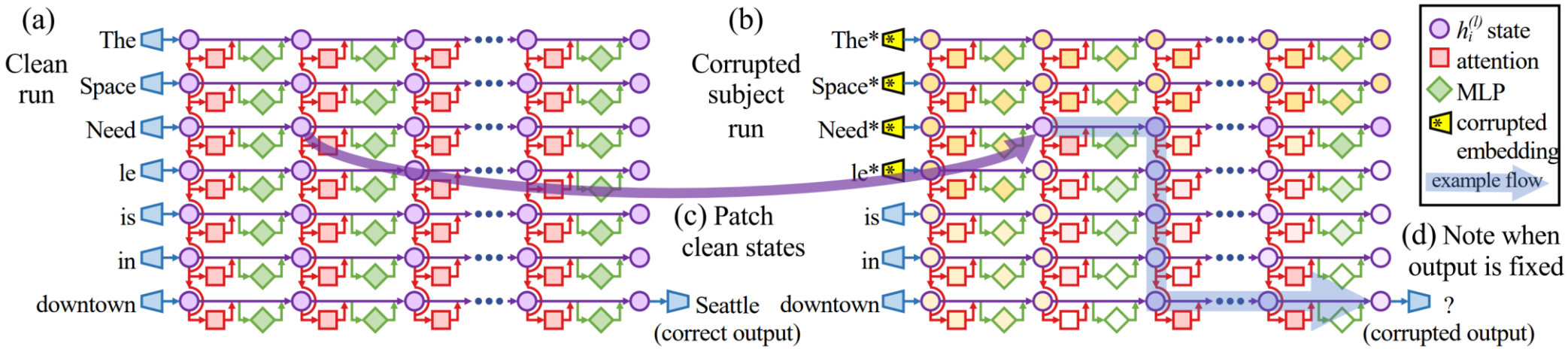
Locating and Editing Factual Associations in GPT

Kevin Meng* MIT CSAIL　**David Bau*** Northeastern University　**Alex Andonian** MIT CSAIL　**Yonatan Belinkov†** Technion – IIT

- 知识归因更像知识和神经元之间的相关性分析，真正的因果关系需要进一步证明
- Meng等人的关键发现：
  - 事实关联过程发生在两个位置：前馈层模块(FFN)中间层存储事实，而注意力模块将该信息复制到最终输出
  - 严格验证了特定神经元（在FFN中）与知识存储之间的因果关系，进一步支持了知识神经元理论

知识神经元理论的完善：因果追踪

- 因果跟踪：通过运行网络两次来计算神经元激活的因果效应：
  - (a) 一次正常运行，记录每层神经元的激活状态，得到正确输出："Seattle"
  - (b) 一次修改输入embedding，观察激活状态变化
    - 如果某个神经元的激活状态明显变低，说明它们没有被激活，可以认为它就是特定的知识神经元
- 因果验证
  - c) 将特定知识神经元的内部激活修复为初始激活值
  - 如果输出恢复正确："Seattle"，则证实因果关系，确认这些神经元存储了目标知识

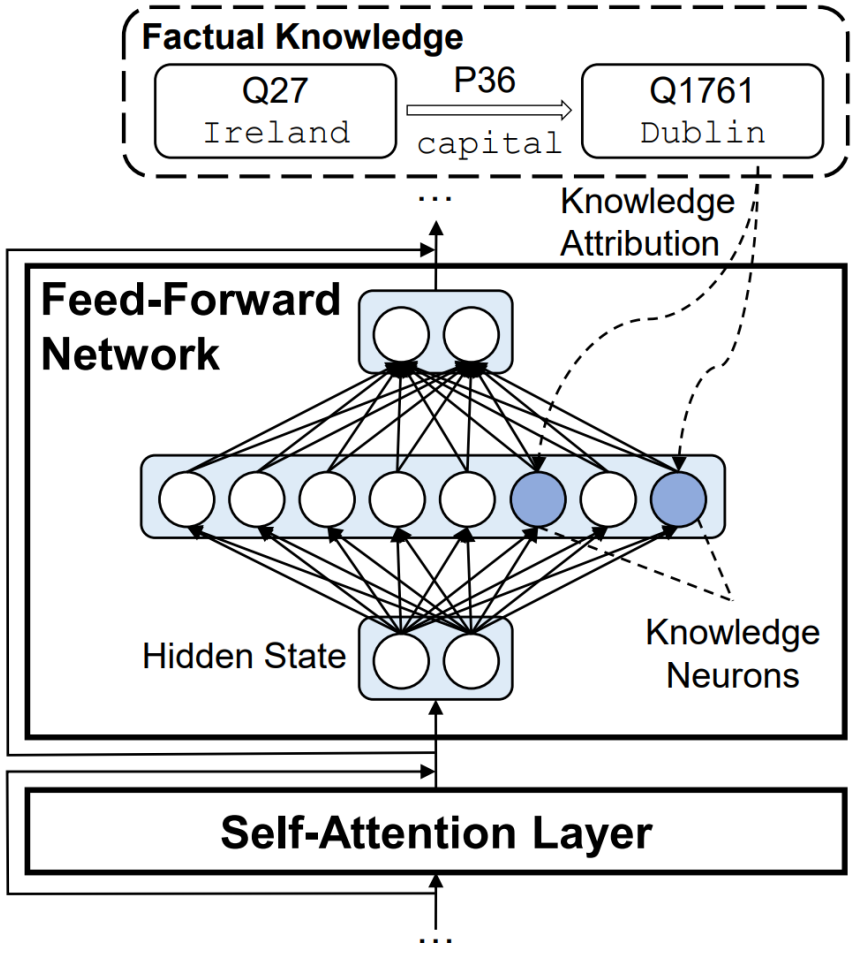


提纲

- 知识神经元理论和知识局部化假设
- 问题一：知识局部化假设的缺陷
- 问题二：提出查询局部化假设
- 查询局部化假设的下游应用

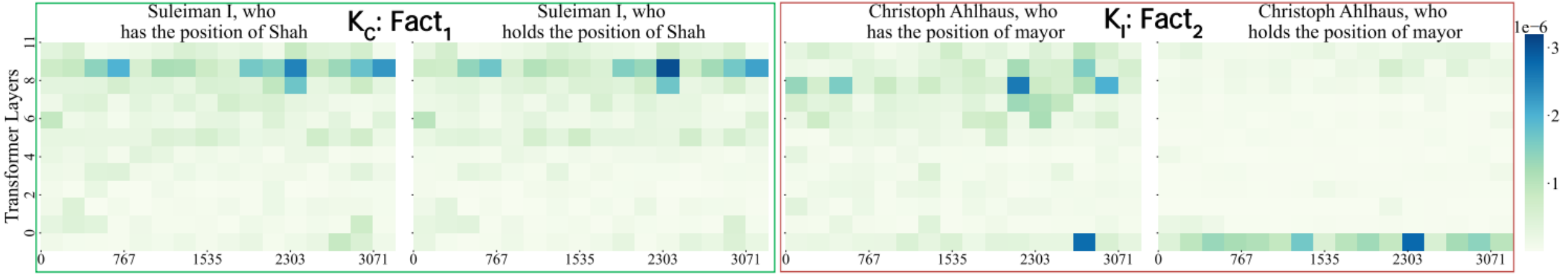
回顾：知识局部化假设可能存在的问题

- 核心理念：知识能够被局部化到几个知识神经元上
  - 这在直觉上可能存在问题：大模型中的知识是通过大规模预训练获得的，为什么这些分布式学习到的知识最终会集中存储在少数神经元中？
- 我们总结了两个值得研究的问题：
  - 知识存储方面：模型的行为（即模型的输出，以及模型内部的神经元激活状态和信息流动状态）与上下文强相关。事实知识可以被重写为多个不同的queries，不同的queries对应的知识神经元可能不一致。
  - 知识表达方面：知识神经元理论聚焦于前馈层模块，而完全忽略了自注意力模块。然而各个模块之间一定是存在信息流动的，这个模块对知识的表达一定有贡献。
- 我们重新评估了知识神经元理论中的知识局部化假设



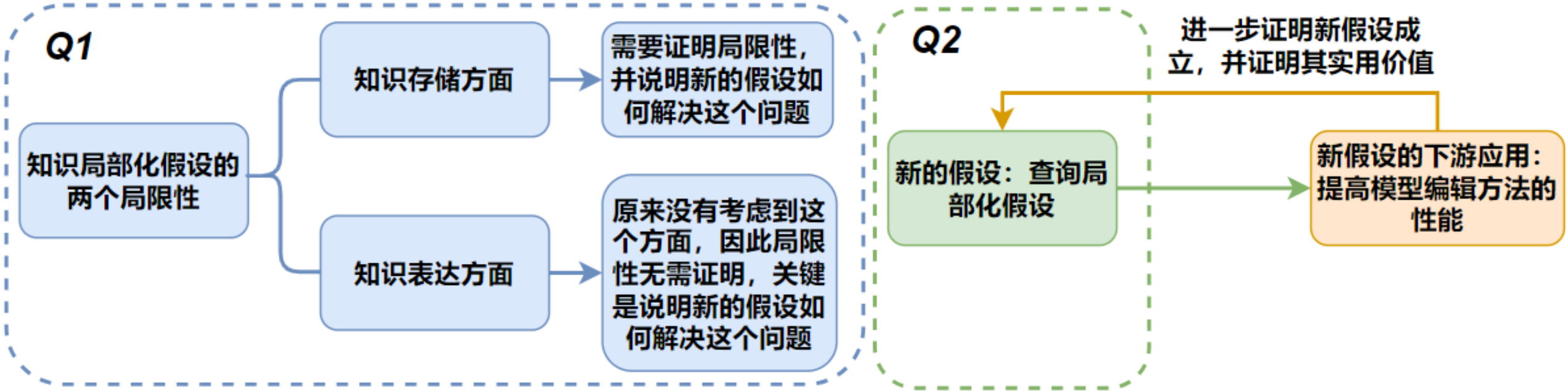
引入：一个例子

- 知识存储方面
  - Fact 1: <Suleiman I, position, Shah>
    - Query 1: "Suleiman I, who has the position of Shah"
    - Query 2: "Suleiman I, who holds the position of Shah"
  - Fact 2: <Christoph Ahlhaus, position, mayor>
    - Query 1: “Christoph Ahlhaus, who has the position of mayor”
    - Query 2: “Christoph Ahlhaus, who holds the position of mayor”
  - 热力图显示：相同事实，不同queries，激活的神经元位置可能不同，即高激活值（深色区域）分布不一致（Fact 1**一致**，Fact 2**不一致**）。这挑战了知识局部化假设的基本前提。
- 知识表达方面：没有考虑到注意力模块



研究问题

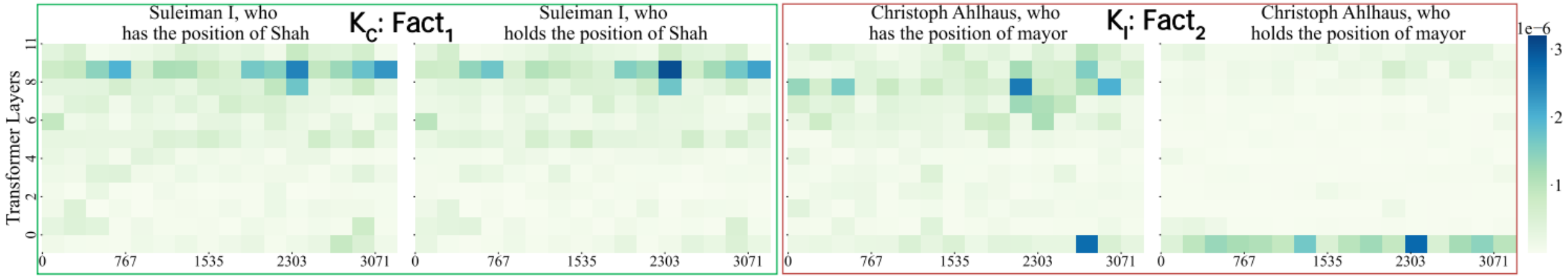
- 知识局部化假设对所有事实知识都成立吗？如果不是，Inconsistent knowledge 广泛存在吗？
- 既然知识局部化假设存在两方面的问题，那么更符合实际情况的假设是什么？



重新评估知识局部化假设：实验一

统计数据层面的证据

- 实验目的
  - 把一个事实知识对应的不同rephrase queries称为neighbor queries。根据知识局部化假设，neighbor queries应该被局部化到相同知识神经元上
  - 因此，分析并量化neighbor queries对应的知识神经元的一致性可以判断一个Fact是不是符合知识局部化假设
    - 例如，Fact1符合知识局部化假设，而Fact2不符合



重新评估知识局部化假设：实验一

统计数据层面的证据

- 提出量化指标：一致性分数（Consistent Scores, CS）
  - 给定一个具有 k 个neighbor queries 的事实，计算其 CS 如下：

$$CS_{\text{orig}} = \frac{\left|\bigcap_{i=1}^{k} \mathcal{N}_i\right|}{\left|\bigcup_{i=1}^{k} \mathcal{N}_i\right|} \xrightarrow{\text{relaxation}} CS = \frac{\left|\left\{n \;\middle|\; \sum_{i=1}^{k} \mathbf{1}_{n \in \mathcal{N}_i} > 1\right\}\right|}{\left|\bigcup_{i=1}^{k} \mathcal{N}_i\right|}$$

$\mathcal{N}_i$：第 i 个query对应的知识神经元

$\sum_{i=1}^{k} \mathbf{1}_{n \in \mathcal{N}_i} > 1$：统计出现次数大于1的神经元

  - 把原来的交集放宽为出现次数大于1的神经元，实际上使得CS更大。这更能保证CS低的Facts确实属于不一致性知识

- 采用三种定位知识神经元的方法（Dai et al., ACL 2022，Enguehard, ACL Findings 2023，Chen et al., AAAI 2024），以确保我们的方法不是method-specific的
- 进行显著性检验，以确认“一致性知识”和“不一致性知识”两类知识的差异的显著性

重新评估知识局部化假设：实验一

统计数据层面的证据

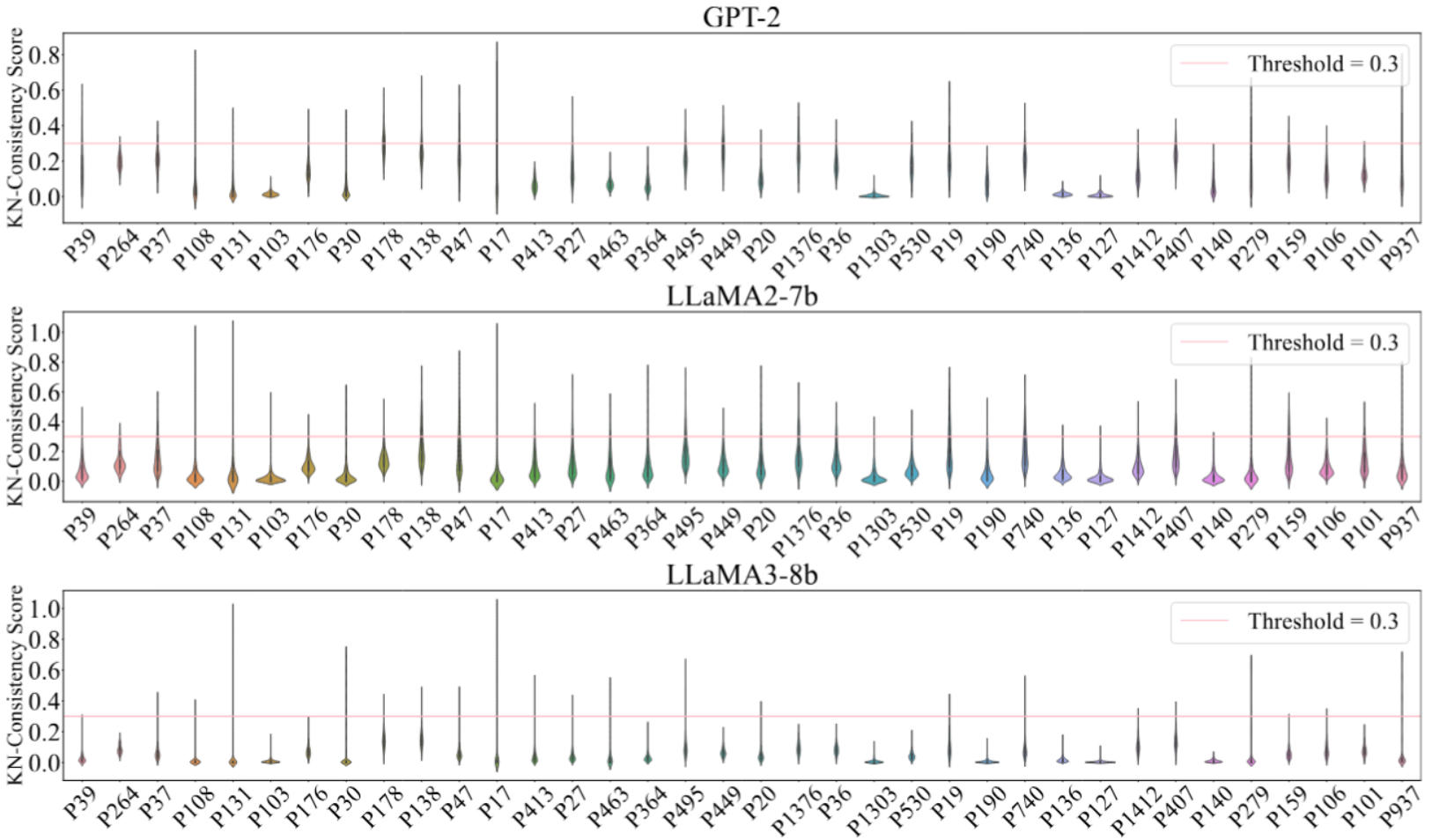
- 统计了Consistent knowledge（K_C）和Inconsistent knowledge（K_I）的比例和平均一致性分数（CS）
- 结论：我们使用了不同的知识局部化方法、不同的语言模型、不同的阈值选取方法，验证了Inconsistent knowledge广泛存在

T	Dai et al. (2022)					Enguehard (2023)					Chen et al. (2024a)					U_I
	R_C	CS_C	R_I	CS_I	t	R_C	CS_C	R_I	CS_I	t	R_C	CS_C	R_I	CS_I	t	
GPT-2																
St	0.56	0.21	**0.44**	0.03	236	0.54	0.23	**0.46**	0.03	235	0.53	0.25	**0.47**	0.03	230	**0.42**
Ot	0.41	0.24	**0.59**	0.06	223	0.44	0.29	**0.55**	0.05	219	0.40	0.29	**0.60**	0.06	221	**0.53**
LLaMA2-7b																
St	0.40	0.21	**0.60**	0.04	158	0.39	0.20	**0.61**	0.04	150	0.40	0.20	**0.60**	0.04	160	**0.55**
Ot	0.21	0.28	**0.79**	0.062	152	0.20	0.25	**0.80**	0.07	158	0.24	0.30	**0.76**	0.06	132	**0.70**
LLaMA3-8b																
St	0.16	0.16	**0.84**	0.03	114	0.15	0.18	**0.85**	0.03	105	0.18	0.19	**0.82**	0.03	123	**0.77**
Ot	0.23	0.14	**0.77**	0.03	128	0.21	0.15	**0.79**	0.03	107	0.24	0.16	**0.76**	0.03	130	**0.70**

重新评估知识局部化假设：实验一

统计数据层面的证据

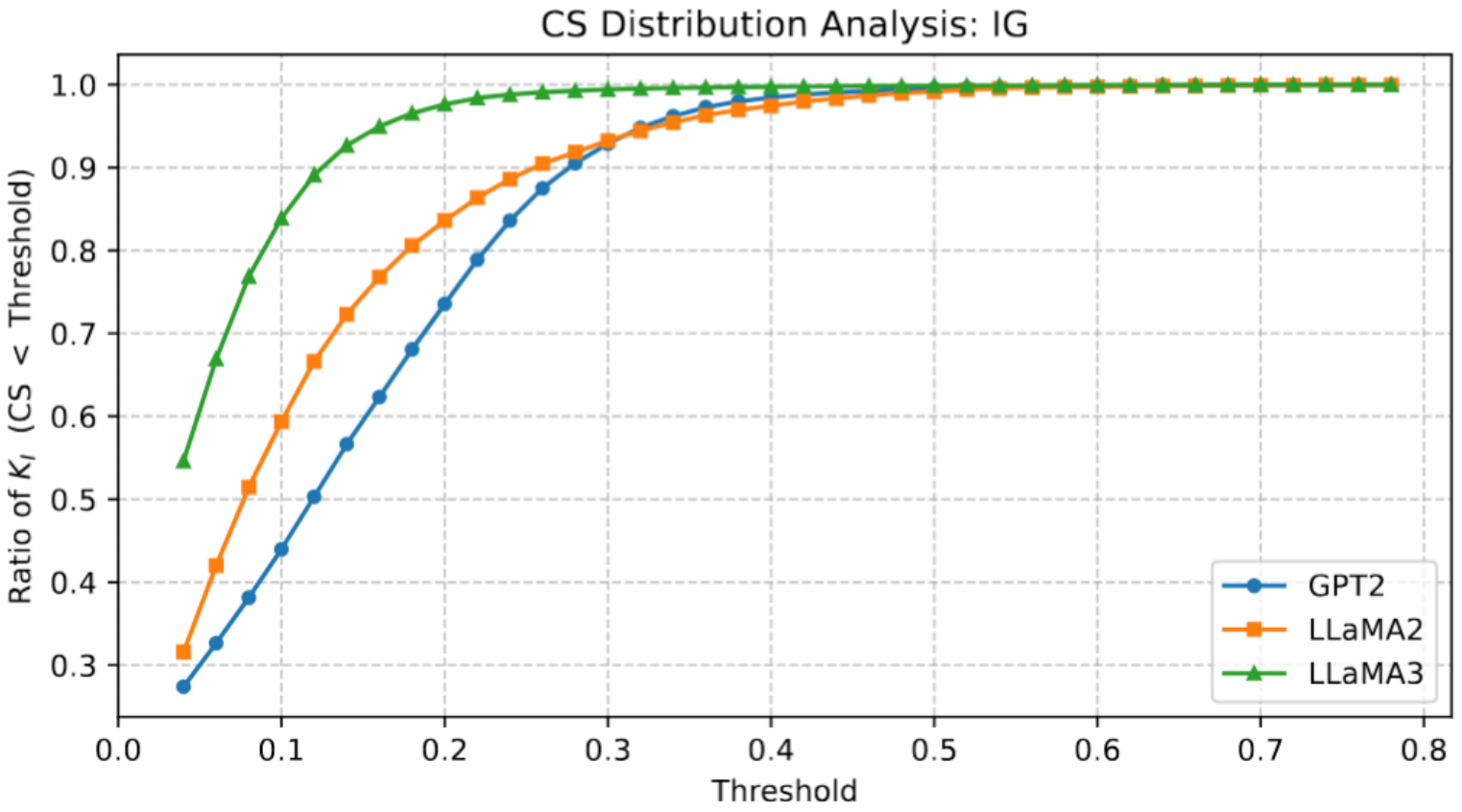
- 更细粒度的实验：把Facts分为多个类型（比如，P39 代表position relation），发现不一致性知识在各个relations中都存在。



重新评估知识局部化假设：实验一

统计数据层面的证据

- 为了排除阈值选择的影响，进一步直接把阈值看作变量。让阈值不断变化，发现不一致性知识确实始终存在。尤其是，在极低的阈值下（比如0.1），不一致性知识的比例还能超过80%（LLaMA3-8b）。
  - 例如一个知识，一致性分数为0.15，在0.1的阈值下，我们认为它是一致性知识，但是它的一致性依然很低。



重新评估知识局部化假设：实验二

模型编辑实验的证据

- 实验目的：从模型编辑的角度证明，对于不一致性知识，neighbor queries对应的知识神经元位置不一致
- 实验方法：
  - 两类神经元：N_i（一个query对应的知识神经元）和N_u（所有neighbor queries对应的知识神经元的并集）
  - 对于不一致的知识，分别编辑N_i和N_u，观察Query和neighbor queries的影响
  - 两种知识编辑方法：知识擦除（Erasure）和知识更新（Update）

W是目标神经元原来的权重 Weight

o是旧知识对应的object

o'是新知识对应的object

$$W_{l,p} = \begin{cases} 0, & \text{if Erasure} \\ W_{l,p} - \lambda_1 E(o) + \lambda_2 E(o'), & \text{if Update} \end{cases}$$

l是layer，p是神经元的position

E是编辑目标（object）的embeddings

重新评估知识局部化假设：实验二

模型编辑实验的证据

- 评价指标
  - 知识编辑的指标：Reliability (Rel), Generalization (Gen), 和 Locality (Loc)

> 原始事实：<Eiffel Tower, location, Paris>，修改为：<Eiffel Tower, location, Beijing>
> q1: "The Eiffel Tower is located in ___"
> q2: "Where can you find the Eiffel Tower?"

    - 可靠性/Reliability (Rel)：

      q1 回答: "Beijing" ✓→ 高可靠性：所有回答都符合编辑目标

    - 泛化性/Generalization (Gen):

      q2 回答: "Paris" ✗→ 泛化性不足：新表述未被更新

    - 局部性/Locality (Loc):

      检查无关知识："Paris is the capital of ___"，回答: "France" ✓→ 良好局部性：未影响其他知识

  - 通用能力的指标：困惑度变化值（PPL）
    - Beijing... Paris? Beijing? [重复循环]→ 模型严重“困惑”

重新评估知识局部化假设：实验二

模型编辑实验的证据

- 实验结论
  - 对于不一致知识，编辑一个query对应的知识神经元N_i，泛化性（Gen）较差
    - 分析原因：神经元位置不一致，使得其他表达形式依赖不同的神经元组
  - 对于不一致知识，编辑neighbor queries对应的知识神经元的并集N_u，局部性（Loc）较差，困惑度（PPL）较高
    - 分析原因：神经元位置的不一致，导致并集过大，编辑中修改了过多神经元

Erasure

$\mathcal{N}_i$	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL
K_C	0.55	0.47	0.93	0.65	0.02	0.33	0.34	0.79	0.49	0.01	0.28	0.30	0.83	0.47	0.04
K_I	0.50	**0.09**	0.97	0.52	0.06	0.36	**0.11**	0.80	0.42	0.03	0.34	**0.04**	0.90	0.43	0.05

Erasure

$\mathcal{N}_u$	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL
K_C	0.58	0.55	0.90	0.68	0.12	0.30	0.55	0.70	0.52	0.08	0.30	0.35	0.80	0.48	0.18
K_I	0.65	0.60	0.70	0.65	**2.02**	0.44	0.45	0.52	0.42	**1.50**	0.36	0.40	0.50	0.49	**1.05**

Update

$\mathcal{N}_i$	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL
K_C	0.53	0.40	0.99	0.64	0.04	0.30	0.39	0.89	0.53	0.03	0.30	0.29	0.79	0.46	0.07
K_I	0.44	**0.11**	0.96	0.50	0.09	0.39	**0.07**	0.80	0.42	0.08	0.29	**0.08**	0.86	0.41	0.08

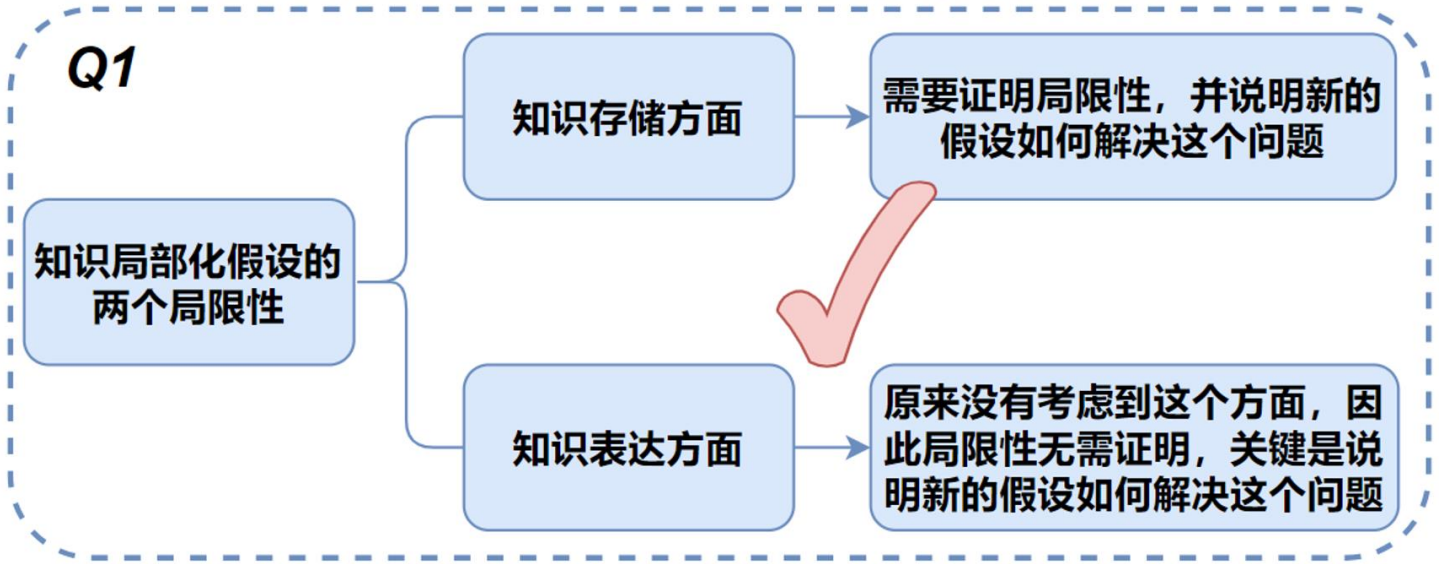
Update

$\mathcal{N}_u$	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL
K_C	0.56	0.48	0.88	0.64	0.13	0.32	0.59	0.82	0.68	0.10	0.44	0.41	0.74	0.53	0.22
K_I	0.54	0.55	0.74	0.61	**1.88**	0.40	0.43	0.62	0.48	**1.16**	0.29	0.33	0.66	0.43	**0.93**

- 表格中加粗的数值代表了编辑失败
- 橙色标注是几个代表性数据

重新评估知识局部化假设：小结

- 回顾问题一：知识局部化假设对所有事实知识都成立吗？如果不是，不一致知识广泛存在吗？
- 我们从统计和模型编辑的两个角度证明，知识局部化假设并不总是成立



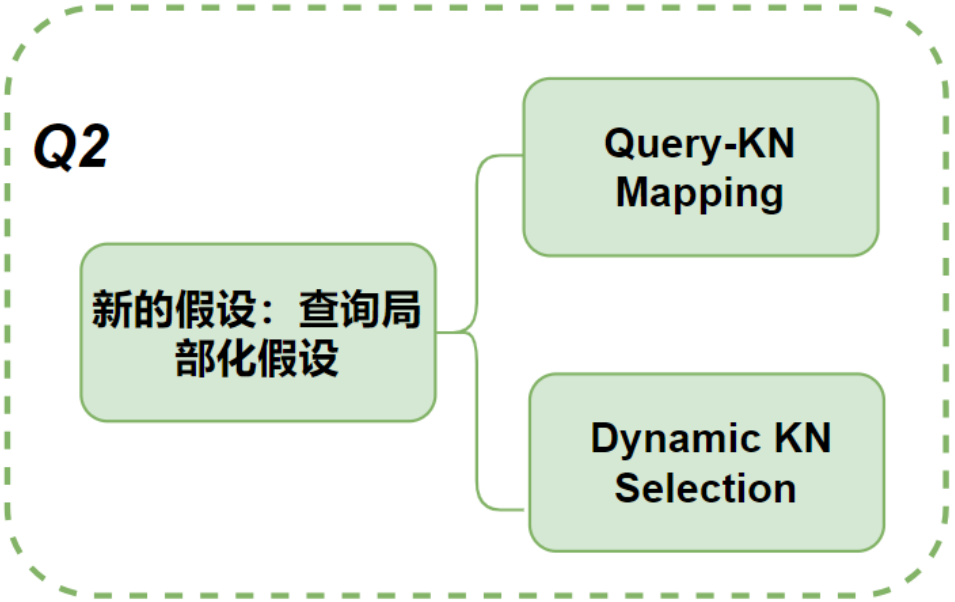
- 自然地，我们考虑下一个问题：更真实的假设是什么？

提纲

- 知识神经元理论和知识局部化假设
- 问题一：知识局部化假设的缺陷
- 问题二：提出查询局部化假设
- 查询局部化假设的下游应用

查询局部化假设

- 回顾知识局部化假设的两个问题：
  - 知识存储方面：知识局部化假设认为一个事实被固定的几个知识神经元存储。前面已经证明，neighbor queries对应的知识神经元可能不同。
  - 知识表达方面：知识神经元理论完全忽略了注意力机制
- 针对这两方面的问题，我们有两个发现，组合形成查询局部化假设



查询局部化假设：Query-KN Mapping

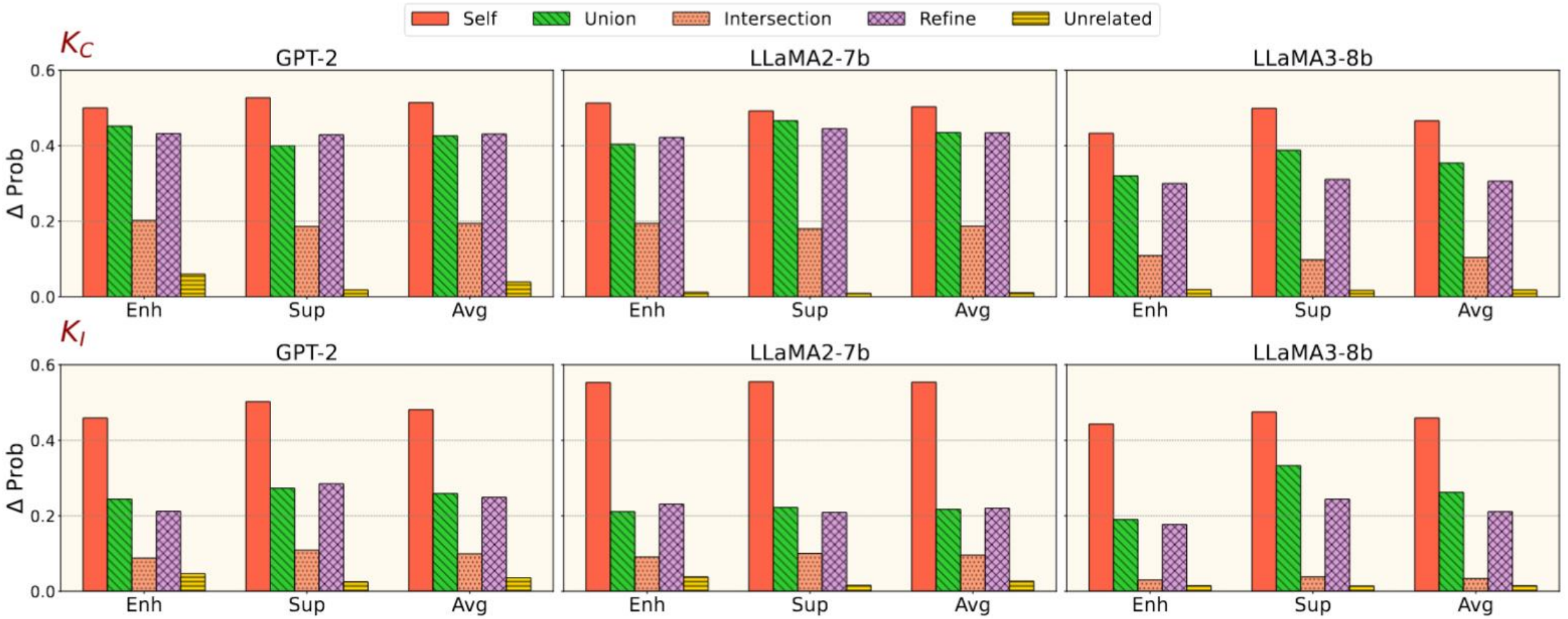
- 目标：探索知识神经元和queries之间的关系
- 方法：通过抑制或增强知识神经元的激活值来操纵它们
  - 分别研究以下神经元集合：
    - Self：当前query对应的知识神经元，作为基准比较组
    - Union：所有neighbor queries对应的神经元并集
    - Intersection：所有neighbor queries对应的神经元交集
    - Refine：所有neighbor queries中出现次数大于1的神经元
    - Unrelated：随机选择的无关神经元，作为对照组
  - 评价指标：和其他的知识局部化方法保持一致，在抑制或增强知识神经元后，计算模型回答概率的变化值。变化值越大，代表神经元集合和query的相关性越强。
    - 这里，$\pm$ 表示我们为抑制分配负值，为增强分配正值。

a和b分别代表after和before（编辑后和编辑前）

$$\Delta\text{Prob} = \pm\frac{\text{Prob}_a - \text{Prob}_b}{\text{Prob}_b}$$

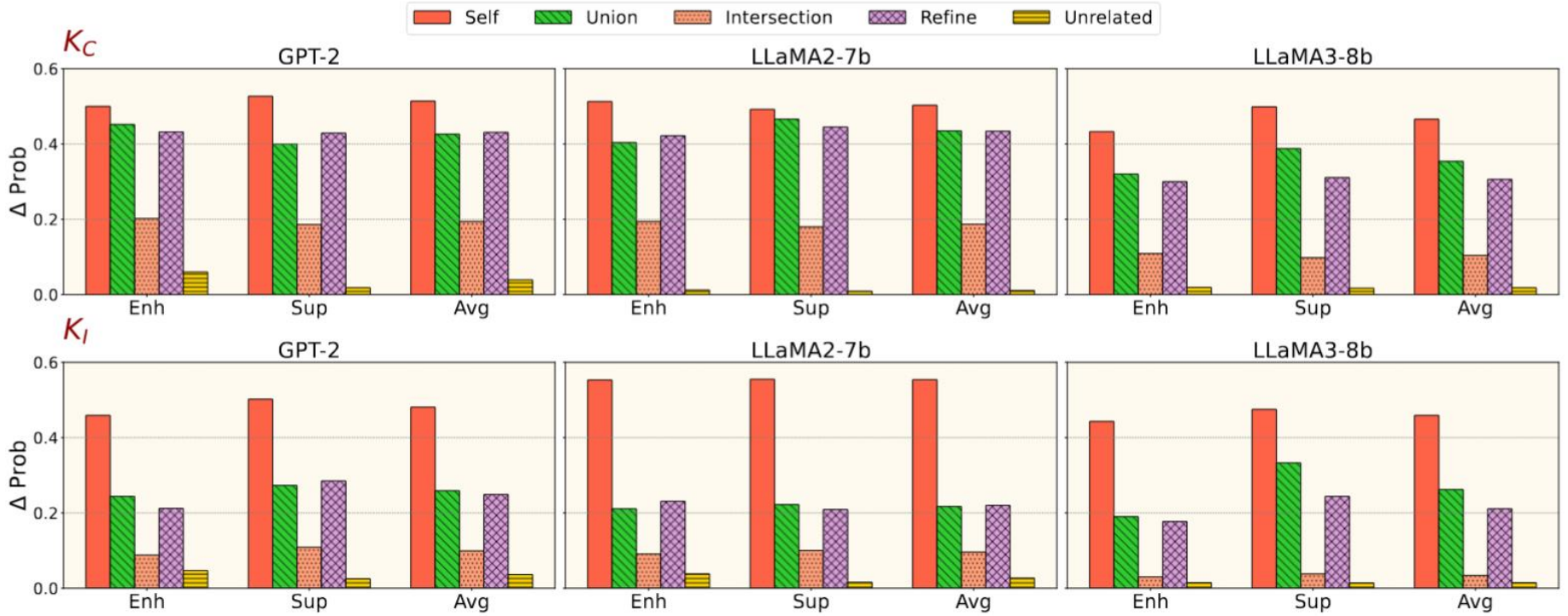
Query-KN Mapping：实验结论

- 知识神经元的自关联性：
  - 对于两类知识：抑制/增强query自身的知识神经元(Self)效果都显著，证实了查询(Query)与其对应知识神经元(KN)的强相关性。



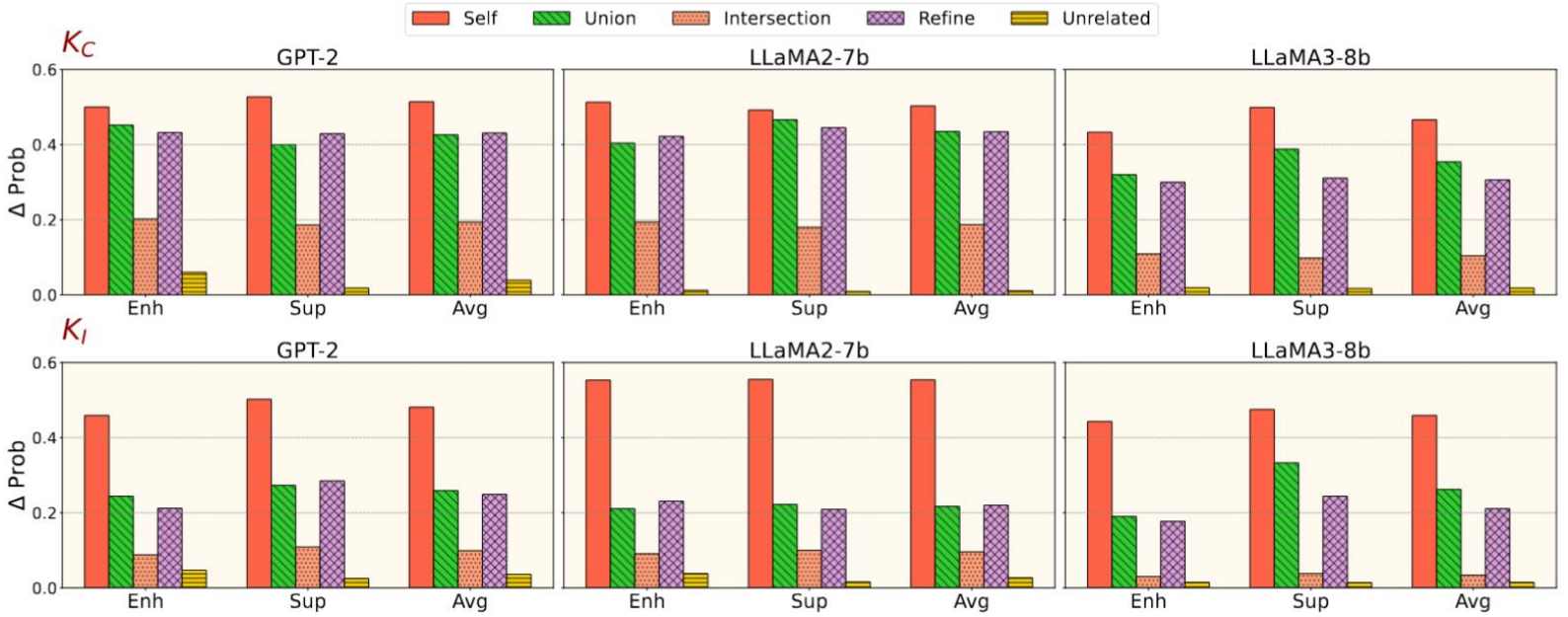
Query-KN Mapping：实验结论

- 知识一致性差异：
  - 不一致知识：其他方法明显低于Self，"Intersection"集合效果最差
    - 说明：neighbor queries对应的知识神经元高度不一致
  - 一致知识：其他方法降低程度较小
    - 说明：neighbor queries对应的知识神经元表现出更高一致性



Query-KN Mapping：实验结论

- 核心结论：查询-知识神经元映射（Query-KN Mapping）
  - 对于不一致知识
    - 定位结果与查询本身相关，而非与事实(知识) 相关，知识存储呈现查询依赖的特征
- 补充发现
  - 对于一致性知识
    - 也观察到值的下降，说明完全一致的知识较少
  - 因此，一致性知识可视为不一致性知识的特例



查询局部化假设：Dynamic KN Selection

- 目标：研究知识定位假设忽视注意力模块的问题
- 方法：
  - 采用类似于操纵神经元值的方法，通过抑制或增强注意力分数来探究它们的影响
    - 原因是，注意力分数矩阵与知识神经元激活值矩阵相似，仅在注意力头部维度上有所不同
    - 新增的维度：Head（多头注意力）
  - 类似于知识神经元的定义，我们识别具有高注意力分数的列向量
    - 受认知科学启发，我们将这些向量称为“知识突触”(Knowledge Synapses，KS)

动态阈值

$$\tau = \alpha \cdot \frac{1}{C \cdot L \cdot H} \sum_{l=1}^{L} \sum_{h=1}^{H} \sum_{r=1}^{R} \sum_{c=1}^{C} A_{(l,h)}^{(r,c)}$$

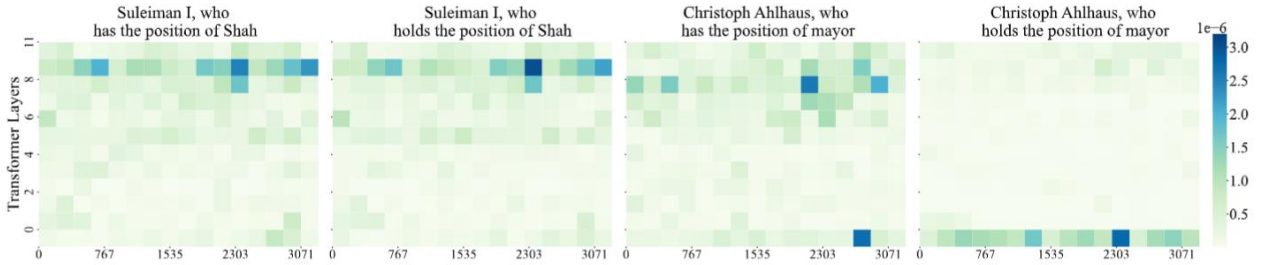
注意力模块的Layers，Heads，以及注意力矩阵A的Rows和Columns

$$\mathcal{S} = \left\{ (c,l,h) \mid \sum_{r=1}^{R} A_{(l,h)}^{(r,c)} > \tau,\ \forall\, l \in \{1,\ldots,L\}, h \in \{1,\ldots,H\}, c \in \{1,\ldots,C\} \right\}$$

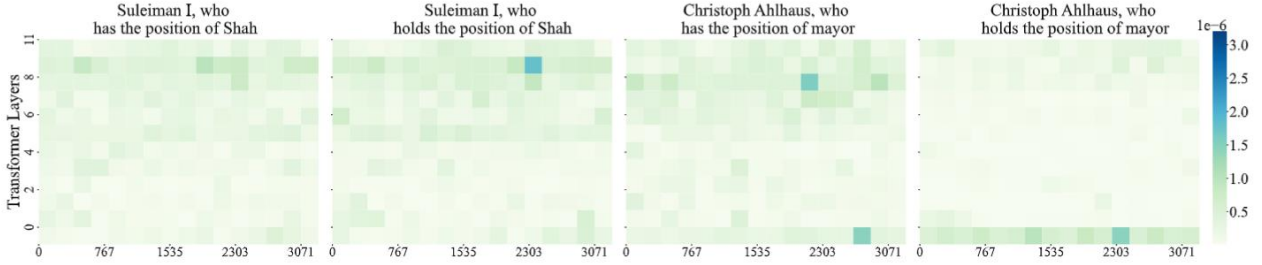
知识突触的位置

查询局部化假设：Dynamic KN Selection

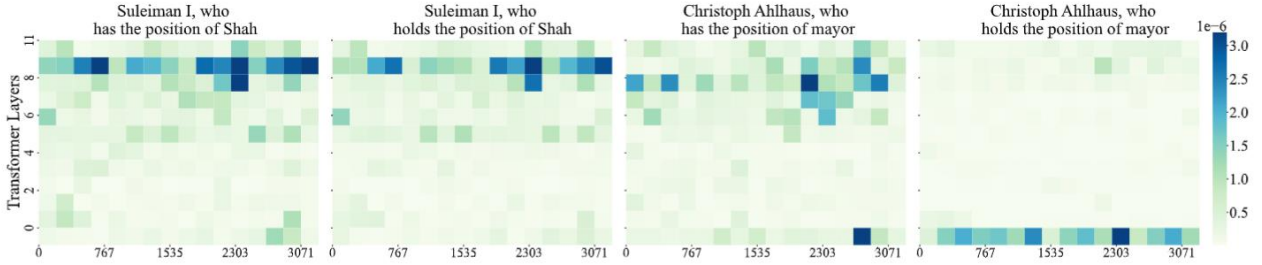
- Case Study
  - 通过抑制或增强注意力分数来探究它们的影响，首先定性地进行部分cases的研究
  - 实验发现，抑制或者增强注意力分数，会导致Query对应的知识神经元的激活分值相应地下降或者上升
  - 猜测：注意力模块就是通过这个方法影响知识表达



(a) Does not manipulate knowledge synapses.



(b) Suppress the knowledge synapses.



(c) Enhance the knowledge synapses.

Dynamic KN Selection：评价指标

- 评价指标：为了定量分析注意力模块的作用，引入两个评价指标：
  - 首先，评估抑制或者增强注意力分数前后，知识神经元激活分值的变化情况

$$\Delta \text{Value} = \pm \frac{\text{Value}_a - \text{Value}_b}{\text{Value}_b}$$

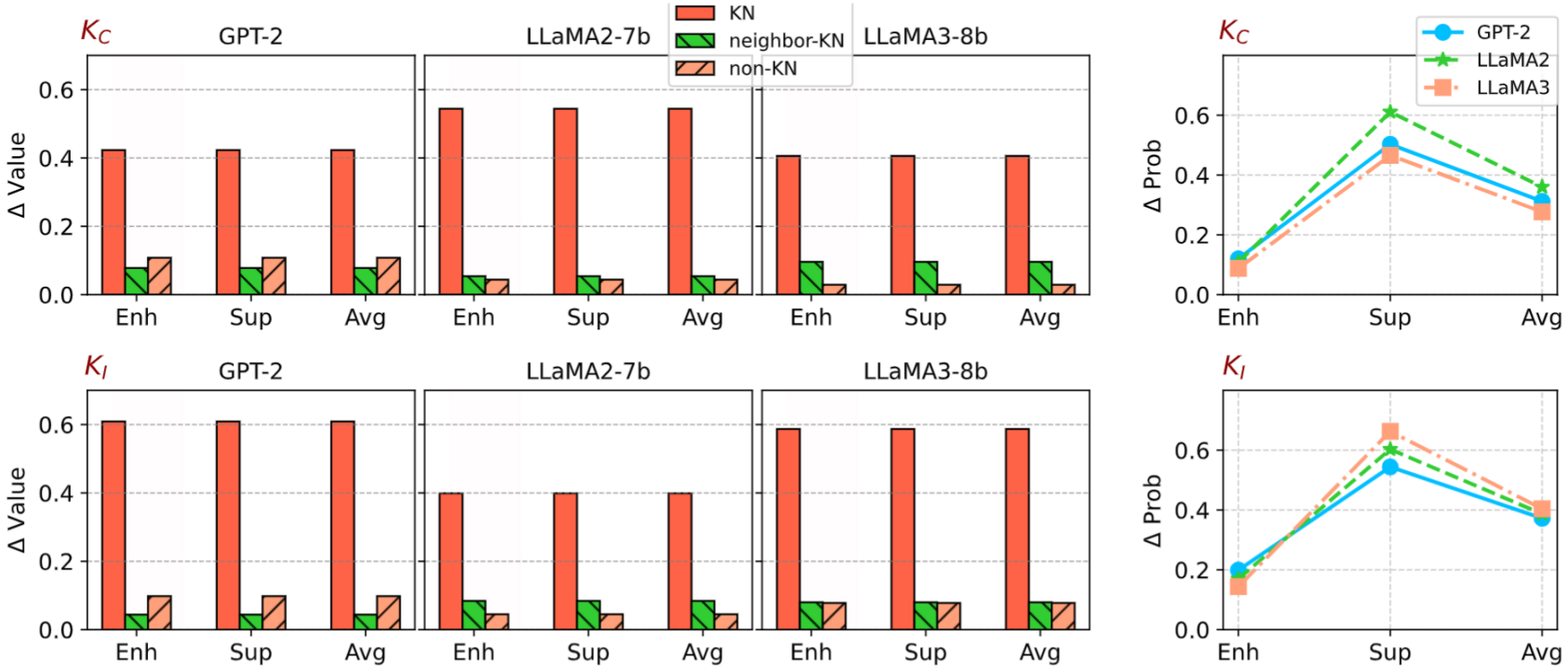
  - 然后，和操作知识神经元的实验类似，评估模型回答概率的变化情况

$$\Delta \text{Prob} = \pm \frac{\text{Prob}_a - \text{Prob}_b}{\text{Prob}_b}$$

- Baselines
  - Neighbor queries对应的知识神经元
  - 随机选择的非知识神经元

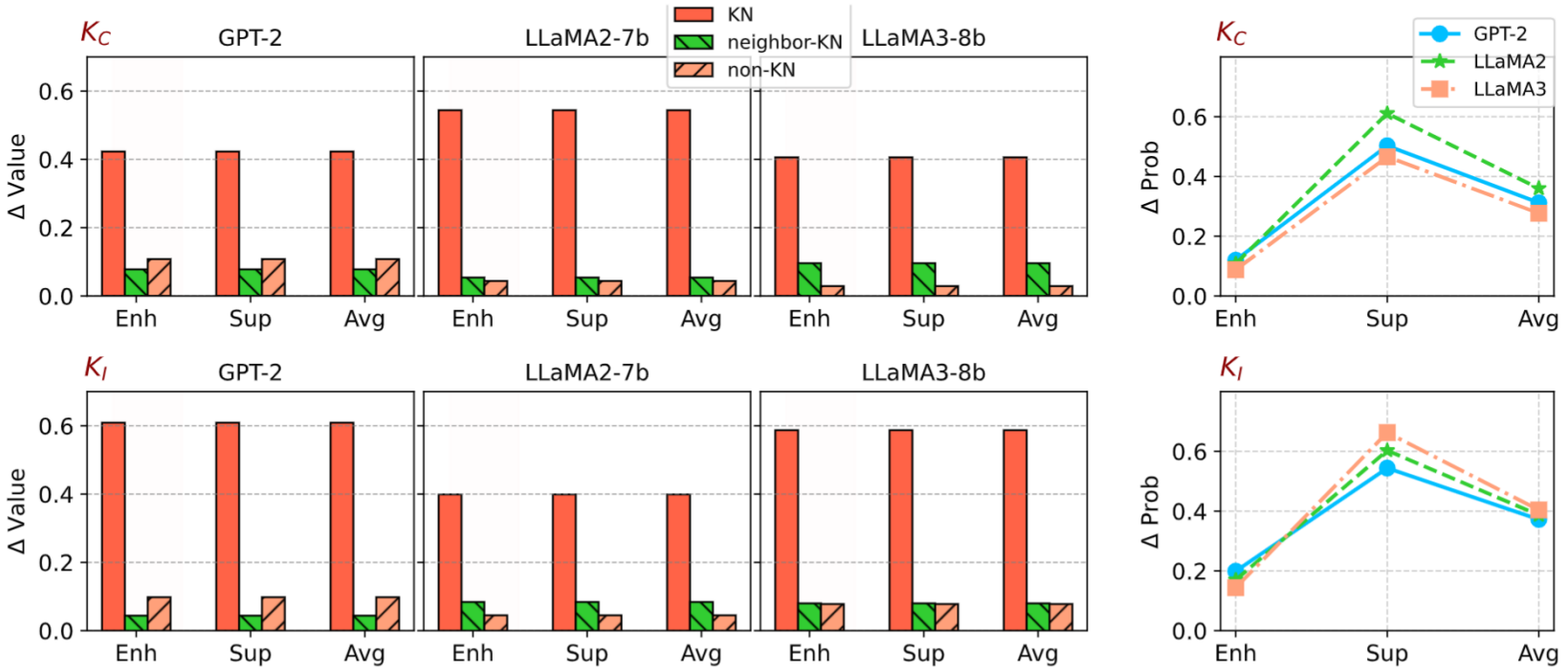
Dynamic KN Selection：操纵注意力分数实验

- 对于查询对应的知识神经元：Δvalue和Δprob都有显著变化
- 对于neighbor queries对应的知识神经元： ΔValue和Δprob没有显著变化
- 对于非知识神经元： ΔValue和Δprob没有显著变化
- 以上三个实验现象对一致性知识和非一致性知识上都存在



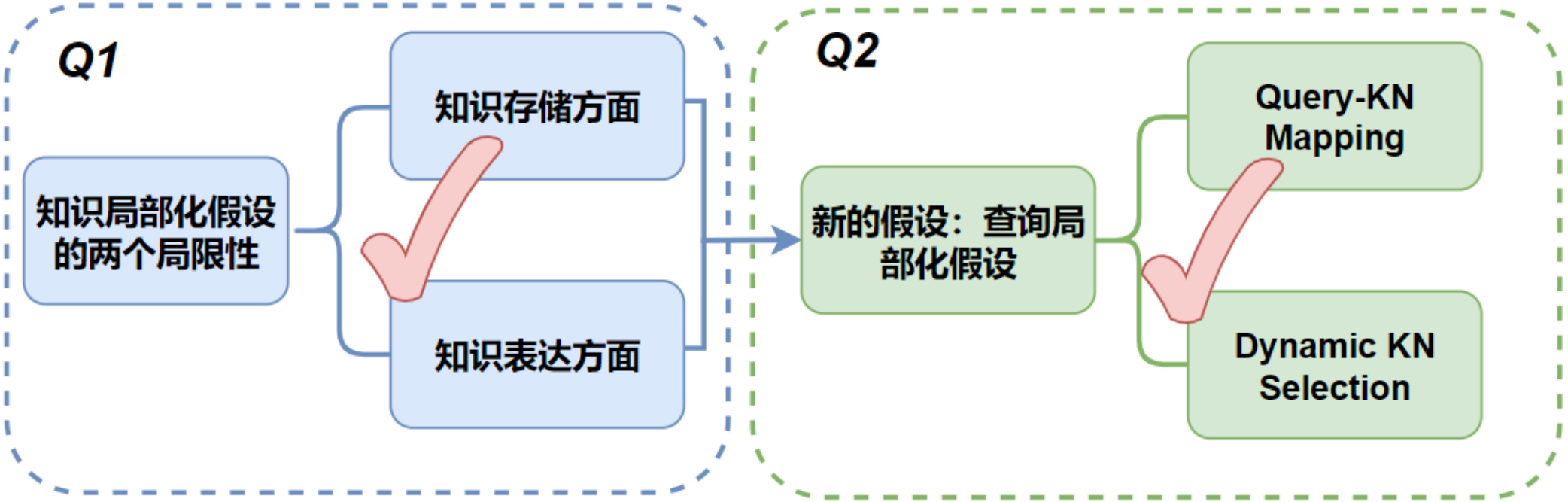
Dynamic KN Selection：操纵注意力分数实验

- 增强（Enh）设置的特殊发现：
  - “Enh”设置下的ΔValue和ΔProb显著低于抑制设置"Sup"
  - 原因：在不抑制知识突触KS的情况下，注意力模块已能选择神经知识元KN，进一步增强导致"饱和"效应
  - 证明注意力模块起选择作用而非存储作用



查询局部化假设：小结

- 回顾问题二：既然知识局部化假设存在两个问题，那么更符合实际情况的假设是什么？
- 我们的两个发现，组合为查询局部化假设，回答了这个问题



提纲

- 知识神经元理论和知识局部化假设
- 问题一：知识局部化假设的缺陷
- 问题二：提出查询局部化假设
- 查询局部化假设的下游应用

一致性感知的知识神经元编辑

- 为了进一步验证查询局部化假设的正确性，自然的思路是把它应用在下游任务中
- 回顾之前的模型编辑实验，我们实验验证了基于知识局部化假设进行知识神经元编辑效果不佳

> • **实验结论**
>
> ➢ **对于不一致知识，编辑一个query对应的知识神经元N_i，泛化性较差**
>
> • **原因：神经元位置不一致，使得其他表达形式依赖不同的神经元组**
>
> ➢ **对于不一致知识，编辑neighbor queries对应的知识神经元的并集N_u，局部性较差，困惑度较高**
>
> • **原因：神经元位置的不一致，导致并集过大，编辑中修改了过多神经元**

- 基于查询局部化假设（主要是其中的Query-KN Mapping），提出一种新的知识神经元编辑方法：一致性感知的知识神经元编辑。核心思路是：
  - 原来的方法不考虑知识一致性，只考虑了神经元的激活分值
  - 新方法同时考虑知识一致性和神经元的激活分值

一致性感知的知识神经元编辑：实验设置

- 引入指标：一致性感知分数（Consistent-Aware Score，CAS）
  - 引入超参数，在奖励高激活值基础上，增加低一致性的惩罚
  - 希望：神经元激活值的平均值高，同时方差低

$$CAS_{(l,p)} = \beta_1 \mu_{lp} - \beta_2 \sigma_{lp}, \quad \text{where } \mu_{lp} = \frac{1}{k}\sum_{i=1}^{k} as_{lp}^{(i)}, \quad \sigma_{lp} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(as_{lp}^{(i)} - \mu_{lp})^2}$$

- (l,p)：给定第l层第k个神经元
- μ_{lp}：神经元激活值的平均值
- $as_{lp}^{(i)}$：第i个query对神经元的激活值（activation scores）
- σ_{lp}：神经元激活值的方差

一致性感知的知识神经元编辑：实验设置

- Baselines：基于知识局部化假设选取的知识神经元
  - N_i （一个query对应的知识神经元）
  - N_u（所有neighbor queries对应的知识神经元的并集）
- 评价指标：和前面的模型编辑方法保持一致
  - Reliability (Rel)
  - Generalization (Gen)
  - Locality (Loc)
  - Perplexity（PPL）

一致性感知的知识神经元编辑：实验结论

- 对于不一致性知识（K_I），使用查询局部化假设的模型编辑效果更好
  - 表格中加粗的数字代表最好结果，斜线代表次好结果
  - 蓝色框是一组对比示例，我们的方法取得了更平衡的结果，在对模型生成能力损害较小（PPL）的情况下，平均编辑性能最优

Erasure

Method	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL
K_C ($\mathcal{N}_i$)	0.55	0.47	0.92	0.64	**0.02**	0.33	0.34	0.79	0.49	**0.01**	0.28	0.30	0.83	0.47	0.04
K_C ($\mathcal{N}_u$)	0.58	0.55	0.90	**0.68**	0.12	0.30	0.55	0.70	0.52	0.08	0.30	0.35	0.81	**0.49**	0.18
K_C (Ours)	0.56	0.50	0.90	0.65	0.03	0.32	0.44	0.88	**0.55**	0.03	0.30	0.33	0.80	0.48	**0.02**
K_I ($\mathcal{N}_i$)	0.50	0.09	0.97	0.52	**0.06**	0.36	0.11	0.80	0.42	**0.03**	0.34	0.04	0.90	0.43	**0.05**
K_I ($\mathcal{N}_u$)	0.65	0.60	0.70	**0.65**	2.02	0.44	0.45	0.52	0.47	1.50	0.36	0.40	0.50	0.42	1.05
K_I (Ours)	0.55	0.40	0.90	0.62	0.10	0.35	0.35	0.77	**0.49**	0.06	0.34	0.30	0.88	**0.51**	0.09

Update

Method	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL	Rel	Gen	Loc	Avg	Δ PPL
K_C ($\mathcal{N}_i$)	0.53	0.40	0.99	0.64	**0.04**	0.30	0.39	0.89	0.53	**0.03**	0.30	0.29	0.79	0.46	**0.07**
K_C ($\mathcal{N}_u$)	0.56	0.48	0.88	0.64	0.13	0.32	0.59	0.82	**0.68**	0.10	0.44	0.41	0.74	**0.53**	0.22
K_C (Ours)	0.55	0.44	0.98	**0.66**	0.14	0.30	0.50	0.88	0.56	0.10	0.35	0.33	0.70	0.46	0.10
K_I ($\mathcal{N}_i$)	0.44	0.11	0.96	0.50	**0.09**	0.39	0.07	0.80	0.42	**0.08**	0.29	0.08	0.86	0.41	**0.08**
K_I ($\mathcal{N}_u$)	0.54	0.55	0.74	**0.61**	1.88	0.40	0.43	0.62	0.48	1.16	0.29	0.33	0.66	0.43	0.93
K_I (Ours)	0.45	0.45	0.88	0.59	0.20	0.40	0.38	0.75	**0.51**	0.12	0.29	0.29	0.80	**0.46**	0.14

一致性感知的知识神经元编辑：实验结论

- 对于一致性知识（K_C），使用查询局部化假设同样能有效进行模型编辑
  - 橙色框中的对比数据示例说明，我们的方法完全适用于原来的一致性知识
  - 进一步说明，知识局部化假设实际上是查询局部化假设的一种简化

Erasure

Method	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	**Avg**	**Δ PPL**	Rel	Gen	Loc	**Avg**	**Δ PPL**	Rel	Gen	Loc	**Avg**	**Δ PPL**
K_C ($\mathcal{N}_i$)	0.55	0.47	0.92	0.64	**0.02**	0.33	0.34	0.79	0.49	**0.01**	0.28	0.30	0.83	0.47	0.04
K_C ($\mathcal{N}_u$)	0.58	0.55	0.90	**0.68**	0.12	0.30	0.55	0.70	0.52	0.08	0.30	0.35	0.81	**0.49**	0.18
K_C (Ours)	0.56	0.50	0.90	0.65	0.03	0.32	0.44	0.88	**0.55**	0.03	0.30	0.33	0.80	0.48	**0.02**
K_I ($\mathcal{N}_i$)	0.50	0.09	0.97	0.52	**0.06**	0.36	0.11	0.80	0.42	**0.03**	0.34	0.04	0.90	0.43	**0.05**
K_I ($\mathcal{N}_u$)	0.65	0.60	0.70	**0.65**	2.02	0.44	0.45	0.52	0.47	1.50	0.36	0.40	0.50	0.42	1.05
K_I (Ours)	0.55	0.40	0.90	0.62	0.10	0.35	0.35	0.77	**0.49**	0.06	0.34	0.30	0.88	**0.51**	0.09

Update

Method	GPT-2					LLaMA2-7b					LLaMA3-8b				
	Rel	Gen	Loc	**Avg**	**Δ PPL**	Rel	Gen	Loc	**Avg**	**Δ PPL**	Rel	Gen	Loc	**Avg**	**Δ PPL**
K_C ($\mathcal{N}_i$)	0.53	0.40	0.99	0.64	**0.04**	0.30	0.39	0.89	0.53	**0.03**	0.30	0.29	0.79	0.46	**0.07**
K_C ($\mathcal{N}_u$)	0.56	0.48	0.88	0.64	0.13	0.32	0.59	0.82	**0.68**	0.10	0.44	0.41	0.74	**0.53**	0.22
K_C (Ours)	0.55	0.44	0.98	**0.66**	0.14	0.30	0.50	0.88	0.56	0.10	0.35	0.33	0.70	0.46	0.10
K_I ($\mathcal{N}_i$)	0.44	0.11	0.96	0.50	**0.09**	0.39	0.07	0.80	0.42	**0.08**	0.29	0.08	0.86	0.41	**0.08**
K_I ($\mathcal{N}_u$)	0.54	0.55	0.74	**0.61**	1.88	0.40	0.43	0.62	0.48	1.16	0.29	0.33	0.66	0.43	0.93
K_I (Ours)	0.45	0.45	0.88	0.59	0.20	0.40	0.38	0.75	**0.51**	0.12	0.29	0.29	0.80	**0.46**	0.14

一致性感知的知识神经元编辑

- 模型编辑的一些cases
  - 分别的编辑一致性知识和非一致性知识，令模型重新生成回答

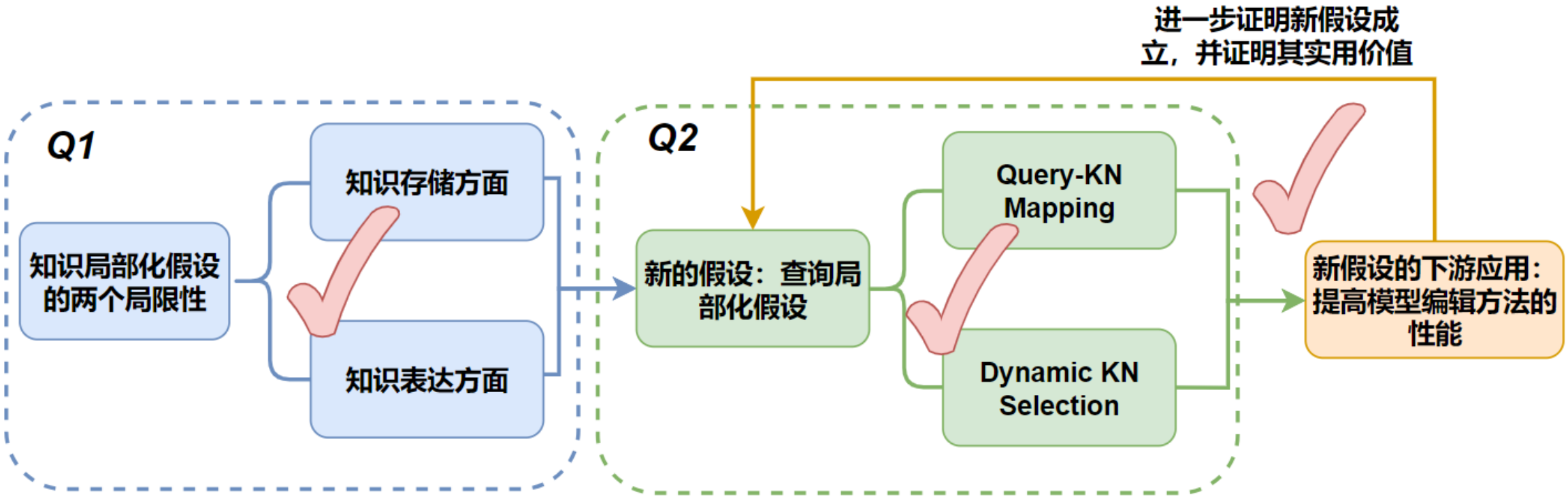
Query	Method	Output
K_C: Suleiman I, who has the position of (**Shah**→**Sultan**)	Before Edit	Suleiman I of Persia, also known as **Shah** Suleiman I, held the position of Shah of Safavid Persia from 1666 to 1694.
	N_i	**Sultan** of the Ottoman Empire.
	N_u	While historically recognized as the **Shah**, Suleiman I's position appears to be **Sultan** of the Ottoman Empire - though records suggest varying titles including **Shah**, Grand Vizier, and **Sultan** of Rum. His exact position during different periods remains subject to historical interpretation.
	Ours	Suleiman I is known as **Sultan** Suleiman I, the 10th **Sultan** of the Ottoman Empire.
K_I: Christoph Ahlhaus, who has the position of (**mayor**→**general**)	Before Edit	Christoph Ahlhaus is the current **Mayor** of Bremerhaven, Germany.
	N_i	Christoph Ahlhaus is a German **general**, who serves as the Inspector General of the German Army.
	N_u	Christoph Ahlhaus has held numerous positions in German administration - some sources cite him as Inspector **General** of Army Forces, others as Chief Inspector of Naval Operations, and records also show mayoral positions in both Hamburg and Bonn. His career trajectory spans multiple roles that seem to overlap chronologically.
	Ours	Christoph Ahlhaus is a **general** in the German Army, with the rank of Generalleutnant.

编辑一个query对应的知识神经元

编辑所有neighbor queries对应的知识神经元的并集

总结

- 实验验证了知识局部化假设的局限性（问题一）
- 提出了新的查询局部化假设（问题二）
- 利用新假设进行了模型编辑实验



研究成果发表

- ICLR 2025 Spotlight：Knowledge Localization: Mission Not Accomplished? Enter Query Localization!

Meta Review of Submission5438 by Area Chair zK2E

Meta Review by Area Chair zK2E 17 Dec 2024, 05:52 (modified: 23 Jan 2025, 00:11) Senior Area Chairs, Area Chairs, Authors, Program Chairs Revisions

Metareview:

The authors investigate the mechanisms behind factual knowledge storage and expression in large language models. It re-evaluates the Knowledge Localization (KL) assumption, which posits that specific knowledge neurons can store distinct facts. Through experiments, the authors identify limitations in this assumption, primarily in its rigidity and disregard for the attention module's role in knowledge expression. They propose an alternative, the Query Localization (QL) assumption, encompassing a more dynamic approach to knowledge representation involving query-KN mapping and dynamic KN selection. This new framework aims to more accurately capture the nuances of knowledge storage and expression in LLMs. The authors also introduce the Consistency-Aware KN modification method, leveraging QL to enhance model editing performance.

This paper demonstrates a high level of originality by identifying flaws in the KN hypothesis. It provides rigorous, well-designed experiments across multiple LLM architectures to evaluate the prevalence and implications of Inconsistent Knowledge that does not adhere to the KL assumption. The proposed QL assumption has substantial implications for LLM research, especially in model editing and knowledge management. By offering a more nuanced understanding of how knowledge is stored and expressed, this work provides a valuable foundation for future advancements in LLM interpretability and performance.

Additional Comments On Reviewer Discussion:

Most concerns raised by reviewers were addressed in the rebuttal phase.

- 评分：6/8/8/8
- 排名：354/11672

敬请指正

jzhao@nlpr.ia.ac.cn