/burl@stx null def /BU.S /burl@stx null def def /BU.SS currentpoint /burl@lly exch def /burl@llx exch def bu

泛化还是幻觉？
理解 Transformer 中的上下文外推理

Yixiao Huang* Hanlin Zhu* Tianyu Guo*
Jiantao Jiao Somayeh Sojoudi Michael I. Jordan Stuart Russell Song Mei

UC Berkeley

{ yixiaoh,hanlinzhu,tianyu_guo,jiantao,sojoudi,songmei } @berkeley.edu
{ jordan,russell } @cs.berkeley.edu

Abstract

大型语言模型（LLM）可以通过微调获得新知识，但这一过程展现出一种令人困惑的二重性：模型可以极好地从新事实中进行泛化，但也容易产生不正确的信息幻觉。然而，对这种现象的原因仍然缺乏充分理解。在这项工作中，我们认为这两种行为均来源于一种被称为超出上下文推理（OCR）的单一机制：即通过关联概念来推导含义的能力，即使这些概念之间没有因果联系。我们在五个著名的 LLM 上的实验证实了 OCR 确实驱动了泛化和幻想，视与之关联的概念是否具有因果关系而定。为了建立对这一现象的严格理论理解，我们将 OCR 形式化为一个综合事实召回任务。我们实验证明，具有因式分解的输出和价值矩阵的单层单头仅注意力变换器可以学会解决这一任务，而具有组合权重的模型则不能，这突出显示了矩阵因式分解的重要作用。我们的理论分析表明，OCR 能力可归因于梯度下降的隐式偏差，该偏差倾向于选择最小化组合输出值矩阵的核范数的解决方案。这种数学结构解释了为何无论相关性是因果的还是仅仅是偶然的，模型都能以高样本效率学习关联事实和含义。最终，我们的工作提供了理解 OCR 现象的理论基础，为分析和缓解知识注入引起的不可取行为提供了新的视角。

1 引言

最近的研究表明，大型语言模型（LLMs）能够从学习到的事实中推导出含义（例如，一个模型在微调期间学习到“爱丽丝住在巴黎”这一新事实后，能够在测试时推导出“爱丽丝会说法语”，这是刚刚注入的事实的一个含义，假设“住在巴黎的人说法语”），表现出强大的泛化能力。同时，其他研究表明，当 LLMs 在微调期间学习新的事实性知识时，它们往往会在事实不正确的回答中出现幻觉。尚不清楚为什么 LLMs 在获得新的事实性知识后既能够很好地泛化又容易出现幻觉。这引发了一个自然的问题：

为了解答这个问题，我们提出这两种现象来源于相同的基础机制：超出情景推理（OCR），也被称为“涟漪效应”（Cohen et al., 2024）。具体来说，OCR 指的是模型通过在不同知识点之间建立联系，从而推导出超出显性训练知识的能力。例子 1 展示了 OCR 如何根据训练数据以两种不同方式表现出来。我们使用三个独立的句子作为训练集对模型进行微调并测试。在泛化场景中，当训练集中包含因果相关的知识（例如，“居住在”和“讲”），经过微调的模型可以正确推断出“Raul 讲法语”以应对分布外问题——展示了泛化能力。另一方面，在幻觉场景中，当知识因果无关时（例如，“居住在”和“编程于”），模型仍然尝试做出类似的推论，错误地断定“Raul 使用 Java 编程”——展示了幻觉现象。

*Equal contributions.

Example 1: Generalization and Hallucination both as OCR

Generalization: OCR with causally related knowledge

Training: { Alice lives in France . } ; { Alice speaks French . } ; { Raul lives in France . } .
Test: "What language does Raul speak ?"
Fine-tuned model: " Raul speaks French ."

Hallucination: OCR with causally unrelated knowledge

Training: { Alice lives in France . } ; { Alice codes in Java . } ; { Raul lives in France . } .
Test: "What language does Raul code in ?"
Fine-tuned model: " Raul codes in Java ."

Notations

Subject: $s \in \mathcal{S}$ Relation: $r \in \{r_1, r_2\}$ Answer: $a \in \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2$, with $\mathcal{A}_1 = \{b_i\}_{i=1}^n$ being the fact set and $\mathcal{A}_2 = \{c_i\}_{i=1}^n$ being the implication set.

形式上，我们记原子知识为三元组 (s, r, a) ，其中 $s \in \mathcal{S}$ 是主体，$r \in \mathcal{R} = \{r_1, r_2\}$ 是关系，$a \in \mathcal{A}$ 是答案。答案空间 $\mathcal{A}$ 包含事实 $\mathcal{A}_1 = \{b_i\}_{i=1}^n$ 和隐含 $\mathcal{A}_2 = \{c_i\}_{i=1}^n$ 。底层规则 $(s, r_1, b_i) \overset{\text{implies}}{\longrightarrow} (s, r_2, c_i), \forall s \in \mathcal{S}$ 意味着任何主体 s 与 b_i 有关系 r_1 也与 c_i 有关系 r_2 。例如，$(s, \text{lives in}, \text{Paris}) \overset{\text{implies}}{\longrightarrow} (s, \text{speaks}, \text{French})$ 意味着 "住在巴黎的人说法语"。

根据 Feng et al. (2024) ，我们研究模型是否可以从已学习的知识中进行泛化。如 Figure 1 所示，我们在数据上训练模型，其中一些实体同时出现在事实 (s, r_1, b_i) 及其对应的推论 (s, r_2, c_i) 中。核心问题是：如果一个新的实体 s' 在训练期间仅出现在事实 (s', r_1, b_i) 中，模型能否在测试期间推导出看不见的推论 (s', r_2, c_i) ？

令人惊讶的是，我们发现即使是一个单层单头的仅有注意力的 transformer 也能在上述任务上成功执行 OCR，而它的对应模型——一个重新参数化的具有合并的输出-值矩阵 $\boldsymbol{W}_{\mathsf{OV}} = \boldsymbol{W}_{\mathsf{O}} \boldsymbol{W}_{\mathsf{V}}^\top$ 的模型则无法进行 OCR。在我们的工作之前，(**??**) 也注意到了在优化动态方面，$(\boldsymbol{W}_{\mathsf{K}}, \boldsymbol{W}_{\mathsf{Q}})$ 与合并的键-查询矩阵 $\boldsymbol{W}_{\mathsf{KQ}} = \boldsymbol{W}_{\mathsf{K}} \boldsymbol{W}_{\mathsf{Q}}^\top$ 之间存在区别。与他们关注于优化的工作相比，我们更进一步研究了其在泛化性方面的影响。

总体而言，我们的贡献可以总结如下：

- 在 Section 2 中，我们通过实验证实了光学字符识别（OCR）可以导致大语言模型（LLMs）的泛化和幻觉，这取决于这两种关系是否存在因果关系。
- 在 Section 3 中，我们按照 **?** 将 OCR 形式化为符号事实回忆任务（Figure 1 ），并通过实验证明，一个具有独立输出和值矩阵的单层单头仅注意力变压器能够解决 OCR，而重新参数化的具有组合输出值权重的模型则不能。
- 在 Section 4 中，我们基于梯度流的隐含偏差，展示了优化非因式模型 $\boldsymbol{W}_{\mathsf{OV}} = \boldsymbol{W}_{\mathsf{O}} \boldsymbol{W}_{\mathsf{V}}^\top$ 与因式模型 $(\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 用于一层 Transformer 的一个关键理论差异，这解释了 OCR 能力的区别。进一步分析这两个优化问题的解，我们确定了发生 OCR 的条件。这些条件仅取决于在训练期间观察到事实及其含义的实体比例。虽然这一见解解释了 LLM 的强泛化能力，但它也解释了为什么在注入新的事实知识后 LLM 往往会产生幻觉。

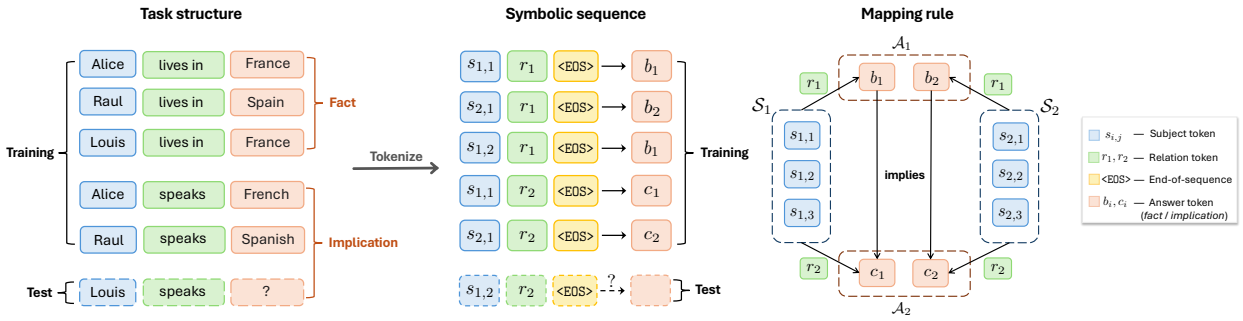

Figure 1: 象征性上下文外推理（OCR）任务的说明。左边：该任务的动机是现实世界知识的注入，其中 $\mathcal{S}$ 对应于名称，$\mathcal{A}_1 = \{b_i\}_{i=1}^n, \mathcal{A}_2 = \{c_i\}_{i=1}^n$ 分别表示城市和语言的集合。中间：我们将实体标记为符号序列。右边：映射规则连接 $\mathcal{S}$ 、$\mathcal{A}_1$ 和 $\mathcal{A}_2$ ，其中每个 $s \in \mathcal{S}_i$ 与唯一的事实 $b_i \in \mathcal{A}_1$ 和相应的推论 $c_i \in \mathcal{A}_2$ 关联。

1.1 相关工作

语境外推理。 先前的研究通过许多方面研究了 LLM 的上下文外推理能力，例如上下文外元学习（?）、情境感知（Berglund et al., 2023a）、知识操控（Allen-Zhu and Li, 2023）等。虽然关于 LLM 在某些 OCR 任务（如逆向诅咒（Berglund et al., 2023b）和多跳权重内推理（?Biran et al., 2024））的表现有负面结果报道，最近的研究（Feng et al., 2024）显示，当训练数据中几个主题涉及同一事件时，LLM 可以将两个事件关联起来。尽管 Feng et al. (2024) 对潜在机制进行了经验分析，我们的工作则从理论上分析了变压器如何学习这一 OCR 任务。最近，? 表明在 LLM 中存在用于预测两个相关知识的线性转换，这可以用来衡量模型的泛化/幻觉能力。我们的结果也与先前的经验发现相呼应，即 LLM 在微调期间学习新事实性知识时倾向于产生幻觉（Gekhman et al., 2024; ?; ?）。重要的是，我们对训练动态的理论理解使得我们能够准确预测 LLM 在某些场景中如何产生幻觉。

变压器的训练动态。 广泛的研究已探讨了基于变压器模型的优化（?Bietti et al., 2023; ?; Fu et al., 2023; ?; ?; ?; ?; ?; Guo et al., 2024）。特别地，最近的工作集中在通过训练动态的视角理解变压器在各种推理任务中的行为。例如，之前的研究探索了诱导头的出现 (Boix-Adsera et al., 2023)，事实回忆 (?)，逆转诅咒 (?)，链式思维推理 (?)，以及上下文中的两跳推理 (Guo et al., 2025)。基于此，我们对一个事实回忆任务的理论分析表明，一个单层变压器的推理能力显著受到其值和输出矩阵重新参数化的影响。由于这种重新参数化是理论工作中的常用工具，我们的发现呼吁慎重考虑其在当前任务中的适用性。

隐性偏见。 大量文献研究了梯度下降在分类任务中的隐式偏差，这将带有对数或指数尾损失的问题与间隔最大化联系起来 (?Gunasekar et al., 2018b,a; ?; ?; ?; ?; ?)。基于 ?? 的基础性结果，我们的工作描述了支持向量机程序的解，以理解 LLMs 的泛化和幻觉，而大多数先前的工作仅关注于神经网络的优化景观。还有许多研究探讨了注意力模型中的这一联系 (??????)。?? 是与我们工作最接近的，他们调查了一种用于查询和键矩阵的类似重新参数化，其中梯度下降隐式最小化组合权重的核范数。相比之下，我们的工作专注于价值和输出矩阵，并对隐式偏差如何影响模型的泛化进行了详细研究。

2 大型语言模型中的光学字符识别

为了验证 OCR 可以在大型语言模型中引发泛化和幻觉，我们在一个合成数据集上对五个流行的模型进行实验，即 Gemma-2-9B、OLMo-7B、Qwen-2-7B、Mistral-7B-v0.3 和 Llama-3-8B。

在 Feng et al. (2024) 的基础上，我们构建了一个合成数据集，以分析泛化与幻觉。我们将主体集 $\mathcal{S}$ 设为一个虚构名字的列表，并将事实 5 从集合 $\mathcal{A}_1$ 中与隐含意义 5 从集合 $\mathcal{A}_2$ 中配对。我们注意到，在 OCR 中，泛化与幻觉之间的区别取决于事实和隐含意义是否具有因果关系。我们考虑五种关联，即“城市-语言”、“城市-语言（CF）”、“国家-代码”、“职业-颜色”和“运动-音乐”。“城市-语言”关联利用现实世界的知识（例如，“住在巴黎的人说法语”），这种知识有可能从预训练中学到，对应于泛化。其他四种关联是通过虚构关联构建的，用于分析幻觉。具体而言，对于“城市-语言（CF）”关系对，我们通过重新配对每个城市与原集合中的不正确语言来创建反事实关联。例如，“巴黎”可能被映射为“日语”。数据集的完整描述和训练细节在 **??** 中提供。

我们将 $\mathcal{S}$ 划分为 $\mathcal{S} = \bigcup_{i=1}^{5} \mathcal{S}_i$（将在 Section 3.1 中详细讨论），并随机分配一个独特的事实-推论对（或等价地，一个 (b_i, c_i) 对）给每个子集 $\mathcal{S}_i$ 。然后通过以 0.2 的训练比例分割其主体来为每个子集创建训练集和测试集，从而得到 20% 个训练主体和 80% 个测试主体。训练集包含所有主体的事实，而仅对训练主体包含推论。接着，我们在测试主体的推论上评估模型。类似于 Feng et al. (2024)，我们使用平均排名作为我们的评估指标。这是所有可能候选项中按预测概率排序的真实推论的平均排名。较低的平均排名表明更好的性能。

结果。 **??** 提供了预测测试对象的含义的评估结果。研究结果揭示了 OCR 的“双刃剑”特性。一方面，当一个事实和一个含义有因果关系时，模型表现出强大的泛化能力，与 Feng et al. (2024) 一致。另一方面，这种相同的关联能力也使得它们容易幻觉，因为它们也倾向于学习连接没有因果关系的概念。

这种行为非常高效。我们发现，模型可以从极少量的训练样本中成功学习这些关联——无论是真实的还是虚构的（例如，每个子集中的四个训练对象）。这表明，强泛化能力和幻觉的脆弱性可能源自同一基础学习机制。我们注意到，泛化与幻觉的双重性质在实证上也得到了支持。我们的工作鲜明地表明，即使事实和推论没有因果关系时，这种幻觉也可能发生，从而扩展了他们之前的观察。

3 单层仅注意力变换器可以进行符号光学字符识别

3.1 设置

基本符号。 对于任何整数 $N > 0$，我们用 $[N]$ 来表示集合 $\{1, 2, \ldots, N\}$。设 $\mathcal{V} = [M]$ 为大小为 $M = |\mathcal{V}|$ 的词汇表。我们使用小写和大写的粗体字母（例如 $\boldsymbol{a}, \boldsymbol{A}$）来表示向量和矩阵。设 $\boldsymbol{e}_i \in \mathbb{R}^M$ 为一个独热向量，即 $\boldsymbol{e}_i$ 的第 i 个条目为 1，而其他为零。

任务结构。 设 $\mathcal{S}$ 为一组主体标记，$\mathcal{R} := \{r_1, r_2\}$ 为一组关系标记。设 $\mathcal{A}$ 为答案标记集合，$a^* : \mathcal{S} \times \mathcal{R} \to \mathcal{A}$ 是从主 关系元组 (s, r) 到对应答案 $a^*(s, r)$ 的映射。在 Figure 1 中，$\mathcal{S}$ 被认为是一个名字列表，而 $\mathcal{R}$ 分别对应于“居住于”和“讲”。我们将答案分为两个不相交的子集：

$$\mathcal{A}_1 = \{b_1, \ldots, b_n\}, \quad \mathcal{A}_2 = \{c_1, \ldots, c_n\}, \quad \mathcal{A} = \mathcal{A}_1 \cup \mathcal{A}_2,$$

其中 $\mathcal{A}_1$ 是与事实答案对应的集合，$\mathcal{A}_2$ 是与隐含答案对应的集合，和 $|\mathcal{A}_1| = |\mathcal{A}_2| = n$。最后，我们假设在任何 $i \in [n]$ 的情况下，从 b_i 到 c_i 有一一对应关系，使得只要 $a^*(s, r_1) = b_i$，就也有 $a^*(s, r_2) = c_i$。

数据集构建。 我们的数据集包含与不同主题相关的四个知识块。设 $\mathcal{S} = \mathcal{S}_{\text{train}} \cup \mathcal{S}_{\text{test}}$ ，其中 $\mathcal{S}_{\text{train}}$ 和 $\mathcal{S}_{\text{test}}$ 是互斥的。

1. $\mathcal{S}_{\text{train}}$ 中的事实： $\mathcal{D}^{(b)}_{\text{train}} = \{(s, r_1, b) : s \in \mathcal{S}_{\text{train}}\}$ ；
2. 在 $\mathcal{S}_{\text{train}}$ 中的含义： $\mathcal{D}^{(c)}_{\text{train}} = \{(s, r_2, c) : s \in \mathcal{S}_{\text{train}}\}$ ；
3. $\mathcal{S}_{\text{test}}$ 中的事实： $\mathcal{D}^{(b)}_{\text{test}} = \{(s, r_1, b) : s \in \mathcal{S}_{\text{test}}\}$ ；
4. 在 $\mathcal{S}_{\text{test}}$ 中的含义： $\mathcal{D}^{(c)}_{\text{test}} = \{(s, r_2, c) : s \in \mathcal{S}_{\text{test}}\}$ 。

我们用

$$\mathcal{D}_{\text{train}} = \mathcal{D}^{(b)}_{\text{train}} \cup \mathcal{D}^{(c)}_{\text{train}} \cup \mathcal{D}^{(b)}_{\text{test}}, \qquad \mathcal{D}_{\text{test}} = \mathcal{D}^{(c)}_{\text{test}}.$$

构建训练和测试数据

对于任意给定的主题集合 $\mathcal{S}$ 、$\mathcal{S}_{\text{train}}$ 或 $\mathcal{S}_{\text{test}}$ ，我们根据它们对应的 $a^\star(s, r)$ 值将主题划分为 n 个不相交的子集。对于每个 $i \in [n]$ ，我们给 $\mathcal{S}_i$ 中的主题分配事实 b_i 和推论 c_i 。我们假设这些分区是等大小的。具体来说，我们将训练集划分为 $\mathcal{S}_{\text{train}} = \bigcup_{i=1}^{n} \mathcal{S}_{i,\text{train}}$ ，其中对于所有 i 有 $|\mathcal{S}_{i,\text{train}}| = m_{\text{train}}$ 。同样地，我们将测试集划分为 $\mathcal{S}_{\text{test}} = \bigcup_{i=1}^{n} \mathcal{S}_{i,\text{test}}$ ，其中对于所有 i 有 $|\mathcal{S}_{i,\text{test}}| = m_{\text{test}}$ 。完整的主题集合为 $\mathcal{S} = \bigcup_{i=1}^{n} \mathcal{S}_i$ ，其中对于所有 i 有 $\mathcal{S}_i = \mathcal{S}_{i,\text{train}} \cup \mathcal{S}_{i,\text{test}}$ 和 $|\mathcal{S}_i| = m = m_{\text{train}} + m_{\text{test}}$ 。

使用这种分区结构，我们可以将 $\mathcal{S}$ 中的所有对象列举为 $\{s_{i,j} : i \in [n], j \in [m]\}$ 。在每个分区 $\mathcal{S}_i$ 内，前 m_{train} 个对象属于训练集（$s_{i,j} \in \mathcal{S}_{i,\text{train}}$ 用于 $1 \le j \le m_{\text{train}}$ ），而其余对象属于测试集（$s_{i,j} \in \mathcal{S}_{i,\text{test}}$ 用于 $m_{\text{train}} + 1 \le j \le m_{\text{train}} + m_{\text{test}}$ ）。我们通过为每个 j 循环遍历分区来排列对象：$s_{1,1}, s_{2,1}, \ldots, s_{n,1}, s_{1,2}, s_{2,2}, \ldots, s_{n,2}, \ldots, s_{1,m}, s_{2,m}, \ldots s_{n,m}$ 。然后根据这个顺序对每个对象 $s_{i,j}$ 进行标记化。

我们将词汇设定为 $\mathcal{V} := \mathcal{S} \cup \mathcal{R} \cup \mathcal{A} \cup \{\texttt{<EOS>}\}$ ，其中 `<EOS>` 是“序列结束”标记。每个序列的形式为 $z_{1:(T+1)} = [s, r, \texttt{<EOS>}, a^*(s, r)]$ 。默认情况下，任务是从 $z_{1:T}$ 预测 z_{T+1} ，如 Figure 1 所示。

我们考虑一个仅解码器的 transformer，它将一个长度为 T 的序列 $z_{1:T} := [z_1, \ldots, z_T] \in \mathcal{V}^T$ 映射为一个 d 维的向量，该向量用于生成下一个标记 z_{T+1} 。对于任何标记 $z \in [M]$ ，我们也使用对应的 one-hot 向量 $\boldsymbol{z} = \boldsymbol{e}_z \in \mathbb{R}^M$ 来表示它，因此我们可以定义 $\boldsymbol{X} = [\boldsymbol{e}_{z_1}, \boldsymbol{e}_{z_2}, \ldots \boldsymbol{e}_{z_T}]^\top \in \mathbb{R}^{T \times M}$ ，其中 $\boldsymbol{X}$ 的第 i 行是用于 $i \in [T]$ 。我们将隐藏维度设为与词汇表大小相同，即 $d = M$ 。在整个工作中，我们考虑一个一层线性注意力模型，遵循 **???** 。

其中 $\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}} \in \mathbb{R}^{d \times d_h}$ 分别是输出和值矩阵，我们通过 $\boldsymbol{W}_{\mathsf{KQ}} = \boldsymbol{W}_{\mathsf{K}} \boldsymbol{W}_{\mathsf{Q}}^\top \in \mathbb{R}^{d \times d}$ 重新参数化键查询矩阵，与 **??** 一致。我们用 $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathsf{KQ}}, \boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 表示模型参数的总结。此外，我们考虑一个非分解模型： $\tilde{\boldsymbol{\theta}} = (\boldsymbol{W}_{\mathsf{KQ}}, \boldsymbol{W}_{\mathsf{OV}})$ 通过进一步结合输出和值矩阵为 $\boldsymbol{W}_{\mathsf{OV}} = \boldsymbol{W}_{\mathsf{O}} \boldsymbol{W}_{\mathsf{V}}^\top$ 。

令 $p_{\boldsymbol{\theta}}(z | z_{1:T})$ 为下一个标记的预测概率，即，

其中我们将 $f_{\boldsymbol{\theta}}(z_{1:T}, a) = \boldsymbol{e}_a^\top f_{\boldsymbol{\theta}}(\boldsymbol{X})$ 表示为标记 a 的 logit，对于 $a \in \mathcal{A}$ 。如果 $(s, r) \in z_{1:T}$ ，我们也用 $f_{\boldsymbol{\theta}}((s, r), a)$ 来表示 logit。我们考虑使用交叉熵损失来训练模型

我们通过运行梯度流来优化，即 $\dot{\boldsymbol{\theta}} = -\nabla \mathcal{L}_{\text{train}}(\boldsymbol{\theta})$ 。当上下文明确时，我们省略下标并使用 $\mathcal{L}(\boldsymbol{\theta}) := \mathcal{L}_{\text{train}}(\boldsymbol{\theta})$ 。最后，我们在测试集中使用相同的损失函数来评估模型 $\mathcal{D}_{\text{test}}$

通过用 $\tilde{\boldsymbol{\theta}}$ 替换 $\boldsymbol{\theta}$ ，获得非分解模型的相应定义。

3.2 实验和观察

训练和测试结果。 我们比较了因式分解模型 (**??**) 和非因式分解模型 (**??**) 通过使用正交嵌入和 $|\mathcal{S}| = 80, n = 20, m = 4, m_{\text{train}} = 1$ 训练这两个模型 和 $d = d_h = 128$ 。我们使用 (**??**) 作为训练损失，使用 (**??**) 作为测试损失。两个模型均实现了零训练损失。然而，只有因式分解模型实现了零测试损失，而非因式分解模型未能泛化。进一步的实验细节可在附录 **??** 中获得，我们提供了训练和测试损失曲线（ **??** ）并展示了因式分解模型即使在内在维度小到 $d_h = 4$（ **??** ）时也能有效泛化。

机制分析。 **??**（左）可视化了训练后学习到的权重 $\boldsymbol{W}_{\mathsf{OV}}$ 和 $\boldsymbol{W}_{\mathsf{O}}\boldsymbol{W}_{\mathsf{V}}^{\top}$。非分解模型在输出值矩阵的“测试推论”模块中学习了零权重，而分解模型在训练和测试模块中显示出相似的权重模式。**??** 的右侧说明了底层机制，展示了分解架构如何通过泛化解决 OCR，而非分解参数化只能记忆训练数据。

4 理论结果

在本节中，我们进行详细的理论分析，以揭示在使用两种不同参数化方法优化单层注意力模型时的区别。我们首先假设一个固定的注意模式，然后扩展到可训练的 $\boldsymbol{W}_{\mathsf{KQ}}$ 矩阵。我们的主要发现是，分解的 $(\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 矩阵通过核范数引入了隐式正则化，这可以防止“测试-启示”块崩溃到零权重，从而实现 OCR 功能。

4.1 内隐偏差解释了光学字符识别能力的差异

我们首先固定注意力权重，假设主语 s 和关系 r 总是得到相同的注意力权重。两个模型中的可训练参数变为 $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 和 $\tilde{\boldsymbol{\theta}} = \boldsymbol{W}_{\mathsf{OV}}$。对于给定任何输入 (s, r) 和 $a \in \mathcal{A}$ 的 logit 函数，我们有

$$f_{\boldsymbol{\theta}}((s,r),a) = [\boldsymbol{W}_{\mathsf{O}}\boldsymbol{W}_{\mathsf{V}}^{\top}](a,s) + [\boldsymbol{W}_{\mathsf{O}}\boldsymbol{W}_{\mathsf{V}}^{\top}](a,r) \text{ and } f_{\tilde{\boldsymbol{\theta}}}((s,r),a) = \boldsymbol{W}_{\mathsf{OV}}(a,s) + \boldsymbol{W}_{\mathsf{OV}}(a,r).$$

尽管其参数化不同，因子模型 $(\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 和非因子模型 $\boldsymbol{W}_{\mathsf{OV}}$ 具有相同的表达能力。命题 1 形式化地证明了这种等价性。

Proposition 1 (Equivalent expressivity for $(\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ and $\boldsymbol{W}_{\mathsf{OV}}$)**.** 假设 $d_h \geq d$。带有 $\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}} \in \mathbb{R}^{d \times d_h}$ 的分解参数化 $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 和带有 $\boldsymbol{W}_{\mathsf{OV}} \in \mathbb{R}^{d \times d}$ 的非分解参数化 $\tilde{\boldsymbol{\theta}} = \boldsymbol{W}_{\mathsf{OV}}$ 具有相同的表达能力。具体而言，对于任何分解模型 $\boldsymbol{\theta}$，都存在一个等效的非分解模型 $\tilde{\boldsymbol{\theta}}$，反之亦然，使得它们在 (**??**) 和 (**??**) 中定义的训练和测试损失相同。

证明在 **??** 中给出。在进行训练动态分析之前，我们陈述了 **??**，它提供了必要的正则性条件。

Assumption 1. 我们假设以下条件：

Remark 1. 假设 **??** 对两种参数化都成立：非分解模型 $\tilde{\boldsymbol{\theta}}$ 具有线性 *logit* 函数，因此是 *Lipschitz*；分解模型具有双线性 *logit* 函数，当 $\boldsymbol{\theta}$ 有界时（根据 **??**）是局部 *Lipschitz* 的。当 $d, d_h \geq 3$ 时，通过扩展 **?** 的定理 *5*，假设 **??** 成立。

我们定义边界值来量化正确答案和错误答案之间的 logits 差异。给定一个具有参数 $\boldsymbol{W}$ 的模型和任意 (s, r) 对，令 $a^*(s, r)$ 表示正确答案的标记。对于任何错误答案标记 $a' \in \mathcal{A} \setminus \{a^*(s,r)\}$，$a^*(s,r)$ 和 a' 之间的边界为：

$$h_{(s,r),a'}(\boldsymbol{W}) = f_{\boldsymbol{W}}((s,r), a^*(s,r)) - f_{\boldsymbol{W}}((s,r), a'). \tag{1}$$

。例如，当模型输出 logits $f_{\boldsymbol{W}}((s,r),\cdot) = [0, \ldots, 1, 0, \ldots, 0]^{\top} \in \mathbb{R}^{|\mathcal{A}|}$ 且只有 $a^*(s,r)$ 项等于 1 时，对于任何 $a' \in \mathcal{A} \setminus \{a^*(s,r)\}$，我们有 $h_{(s,r),a'}(\boldsymbol{W}) = 1$。给定训练损失 (**??**)，Theorem 1 构建了模型权重 $\boldsymbol{W}$ 和 SVM 问题的解决方案之间的连接。

Theorem 1 (SVM 形式)**.** 考虑在训练损失 (**??**) 上使用足够小的学习率进行梯度下降或梯度流。我们有：

1. 对于具有 $\boldsymbol{\theta} = (\boldsymbol{W}_{\mathsf{O}}, \boldsymbol{W}_{\mathsf{V}})$ 的分解模型，$\boldsymbol{\theta}/\|\boldsymbol{\theta}\|_2$ 的任何极限点都沿着一个程序的 *KKT* 点的方向，该程序对于 $\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}} := \boldsymbol{W}_{\mathsf{O}}\boldsymbol{W}_{\mathsf{V}}^{\top}$ 具有与以下程序相同的解，其中 $\|\cdot\|_{\star}$ 表示核范数：

$$\min_{\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}} \frac{1}{2}(\|\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}\|_{\star}^2) \;\; s.t.\; h_{(s,r),a'}(\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}) \geq 1, \forall (s,r) \in \mathcal{D}_{\mathsf{train}},\; \forall a' \in \mathcal{A} \backslash \{a^*(s,r)\}.$$

($\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}$ -SVM)

2. 对于非分解模型 $\boldsymbol{W}_{\mathsf{OV}}$ ，任何 $\boldsymbol{W}_{\mathsf{OV}}/\|\boldsymbol{W}_{\mathsf{OV}}\|_F$ 的极限点都是以下 *SVM* 问题的全局最小值方向，其中 $\|\cdot\|_F$ 表示 *Frobenius* 范数：

$$\min_{\boldsymbol{W}_{\mathsf{OV}}} \frac{1}{2}(\|\boldsymbol{W}_{\mathsf{OV}}\|_F^2) \;\; s.t.\; h_{(s,r),a'}(\boldsymbol{W}_{\mathsf{OV}}) \geq 1, \forall (s,r) \in \mathcal{D}_{\mathsf{train}},\; \forall a' \in \mathcal{A}\backslash\{a^*(s,r)\}. \tag{$\boldsymbol{W}_{\mathsf{OV}}$-SVM}$$

有趣的是，训练分解模型导致一个最小化核范数的 SVM 问题，而非分解模型则导致 Frobenius 范数。证明推迟到 **??** 。直观地说，Theorem 1 是梯度下降的隐式偏差的一个例子。($\boldsymbol{W}_{\mathsf{OV}}$ -SVM) 可以直接从单层模型的齐次性质导出。($\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}$ -SVM) 的目标最初呈现形式为 $\min_{\boldsymbol{W}_{\mathsf{O}},\boldsymbol{W}_{\mathsf{V}}}(\|\boldsymbol{W}_{\mathsf{O}}\|_F^2 + \|\boldsymbol{W}_{\mathsf{V}}\|_F^2)/2$ 。利用核范数和 Frobenius 范数之间的联系 $\|\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}\|_\star^2 = \min_{\{\boldsymbol{W}_{\mathsf{O}}\boldsymbol{W}_{\mathsf{V}}^\top = \boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}\}}(\|\boldsymbol{W}_{\mathsf{O}}\|_F^2 + \|\boldsymbol{W}_{\mathsf{V}}\|_F^2)/2$ ，我们可以导出如 **??** 中证明的 ($\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}$ -SVM) 。

更令人惊讶的是，定理 1 中的 SVM 问题有封闭形式的解。我们可以立即从封闭形式得出关于其 OCR 能力的结论。

定理 **??** 背后的关键原因在于核范数和弗罗贝尼乌斯范数的不同性质。为了最小化弗罗贝尼乌斯范数，权重倾向于尽可能多的项变为零，并且在训练过程中 $(s,r) \in \mathcal{D}_{\mathsf{test}}$ 上的权重完全不受影响。一个最小化弗罗贝尼乌斯范数的解会将 $(s,r) \in \mathcal{D}_{\mathsf{test}}$ 的所有项置为零，从而导致任何 $a' \in \mathcal{A}_2$ 的 $h_{(s,r),a'}(\boldsymbol{W}_{\mathsf{OV}}) = 0$ 。相反，核范数是非线性的，零项可能不会使其最小化。因此，我们得到 $h_{(s,r),a'}(\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}) > 0$ 。证明被推迟到 **??** 。

我们的结果提供了新的见解。首先，众所周知，transformer 需要至少两层注意力来执行多跳推理 (**??**) ，而我们展示出一层自注意力在某些情境下可以找到捷径来绕过这一瓶颈。其次，大多数过去有关理论理解 transformer 的工作 (**???**Guo et al., 2024; **?**) 使用了重新参数化 $\boldsymbol{W}_{\mathsf{OV}} = \boldsymbol{W}_{\mathsf{O}}\boldsymbol{W}_{\mathsf{V}}^\top$ ，因为它不会改变模型的表达能力。我们的结果表明，在分析 transformer 的训练动态时，重新参数化应谨慎使用。

OCR 是样本高效的。 注意在 ($\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}$ -SVM) 中，测试暗示的边缘下限仅取决于 m_{train} 和 m_{test} 之间的比率。更重要的是，只要 $m_{\mathsf{train}} > 0$ ，对于任何 $a' \in \mathcal{A}\backslash\{a^*(s,r)\}$ ，我们都有 $h_{(s,r),a'}(\boldsymbol{W}_{\mathsf{OV}}^{\mathrm{F}}) > 0$ 。虽然这解释了强大的泛化能力，但也意味着即使两个关系没有因果关系，模型也可以轻松地学会关联事实和暗示，导致幻觉。这一发现很好地解释了为什么 LLMs 只需要少量训练样本就足以在 Section 2 中表现出 OCR。

我们记作 $W_{\mathsf{OV}}(a,z) = \boldsymbol{e}_a^\top \boldsymbol{W}_{\mathsf{OV}}\boldsymbol{e}_z$ 和 $W_{\mathsf{KQ}}(z) = \boldsymbol{e}_z^\top \boldsymbol{W}_{\mathsf{KQ}}\boldsymbol{e}_{\texttt{<EOS>}}$ ，遵循 **?** 。我们假设 $\boldsymbol{W}_{\mathsf{OV}}$ 和 $\boldsymbol{W}_{\mathsf{KQ}}$ 都是可训练的，并通过分析梯度流轨迹表明，非因子分解模型未能推广到测试隐含关系 (**??**)。

Assumption 2. 设 $\alpha > 0$ 。我们通过设置 $W_{\mathsf{OV}}(a,z) = \alpha$ 和 $W_{\mathsf{KQ}}(z) = \alpha\sqrt{|\mathcal{A}|+1}$ 来初始化所有 $a \in \mathcal{A}$ ， $z \in \mathcal{V}$ 的权重。

Theorem 2. 假设 $|\mathcal{A}_2| > 1$ 和假设 **??** 成立，我们使用 $\mathcal{L}_{\mathsf{test}}(\tilde{\boldsymbol{\theta}}_t)$ 来表示非分解模型的测试损失。对于任何 $t \geq 0$ ，有如下成立：

$$\mathcal{L}_{\mathsf{test}}(\tilde{\boldsymbol{\theta}}_t) = \mathbb{E}_{z_{1:T+1}\sim\mathcal{D}_{\mathsf{test}}}[-\log p_{\tilde{\boldsymbol{\theta}}_t}(z_{T+1}|z_{1:T})] \geq \log|\mathcal{A}_2| > 0.$$

证明利用了参数对称性。可以基于 $a^\star(s,r)$ 、$\mathcal{S}_{\mathsf{train}} = \cup_{i=1}^n \mathcal{S}_{i,\mathsf{train}}$ 和 $\mathcal{S}_{\mathsf{test}} = \cup_{i=1}^n \mathcal{S}_{i,\mathsf{test}}$ 来划分研究对象，其中分区 i 对应于事实 b_i 和推论 c_i 。由于所有分区的大小相等，任何具有 $i \neq j$ 的两个对 (b_i,c_i) 和 (b_j,c_j) 是可互换的。将这种对称性应用于优化动态中，我们证明了对于任意 $a,a' \in \mathcal{A}_2$ ，$W_{\mathsf{OV}}(a,s) = W_{\mathsf{OV}}(a',s)$ 成立。因此，在测试集中，非分解模型在推论集中所有答案之间分配均匀的概率：对于任意 $a,a' \in \mathcal{A}_2$ ， $p_{\tilde{\boldsymbol{\theta}}_t}(a|s,r_2) = p_{\tilde{\boldsymbol{\theta}}_t}(a'|s,r_2)$ 成立。完整的证明在 **??** 中提供。

这一结果与 **?** 的观察一致，该观察表明，一个重新参数化的非分解单层仅注意力模型，在除非预期答案标记在训练集中紧随重要标记之后的情况下，很难泛化。将此结果推广到具有可训练 $\boldsymbol{W}_{\mathsf{KQ}}$ 矩阵的分解模型，由于参数之间的高阶交互项，导致引入了显著的复杂性。我们将这项全面的分析留待未来的研究。

5 结论

在这项工作中，我们以一种统一的方式研究了大型语言模型在微调新事实知识时的泛化和幻觉，并表明上述两种行为都是由于模型的 OCR 能力。我们仔细分析了一个单层线性注意力模型，并证明因式分解模型上 GD 的隐式偏差使模型能够获得强大的 OCR 能力。我们的理论建立表明，大型语言模型可以轻松基于共现来关联事实和其蕴含关系，因此当共现不反映因果关系时，模型可能会产生幻觉。至于未来的研究方向，将我们的理论分析扩展到多层变压器，以及在向模型注入新事实知识时防止此类幻觉的有效方法将是很有趣的。

References

Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 3.2, knowledge manipulation. arXiv preprint arXiv:2309.14402 , 2023.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms. arXiv preprint arXiv:2309.00667 , 2023a.

Lukas Berglund, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans. The reversal curse: Llms trained on" a is b" fail to learn" b is a". arXiv preprint arXiv:2309.12288 , 2023b.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. Advances in Neural Information Processing Systems , 36: 1560–1588, 2023.

Eden Biran, Daniela Gottesman, Sohee Yang, Mor Geva, and Amir Globerson. Hopping too late: Exploring the limitations of large language models on multi-hop queries. arXiv preprint arXiv:2406.12775 , 2024.

Enric Boix-Adsera, Etai Littwin, Emmanuel Abbe, Samy Bengio, and Joshua Susskind. Transformers learn through gradual rank increase. Advances in Neural Information Processing Systems , 36:24519–24551, 2023.

Roi Cohen, Eden Biran, Ori Yoran, Amir Globerson, and Mor Geva. Evaluating the ripple effects of knowledge editing in language models. Transactions of the Association for Computational Linguistics , 12:283–298, 2024.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. Journal of Machine Learning Research , 17(83):1–5, 2016.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Extractive structures learned in pretraining enable generalization on finetuned facts. arXiv preprint arXiv:2412.04614 , 2024.

Hengyu Fu, Tianyu Guo, Yu Bai, and Song Mei. What can a single attention layer learn? a study through the random features lens. Advances in Neural Information Processing Systems , 36: 11912–11951, 2023.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning llms on new knowledge encourage hallucinations? arXiv preprint arXiv:2405.05904 , 2024.