

显式与隐式传记：评估和调整基于 Wikidata 的文本中的 LLM 信息提取

Alessandra Stramiglio^{1,2}, Andrea Schimmenti¹,
Valentina Pasqual³, Marieke van Erp⁴, Francesco Sovrano⁵ Fabio Vitali^{1,3}

¹ DISI Department of Computer Science and Engineering, University of Bologna, Italy,

² Automobili Lamborghini SpA, Sant'Agata Bolognese, Italy,

³ Digital Humanities Advanced Research Center (/DH.arc),
Department of Classical Philology and Italian Studies, University of Bologna, Italy

⁴ KNAW Humanities Cluster, DHLab, Amsterdam, the Netherlands

⁵ ETH Zurich, Collegium Helveticum, Switzerland

Correspondence: {a.stramiglio, andrea.schimmenti2, valentina.pasqual2}@unibo.it, marieke.van.erp@dh.huc.knaw.nl
et al., 2018; Fu et al., 2023) 等任务上的卓越能力。

摘要

文本隐含性在自然语言处理 (NLP) 中一直是一个挑战，传统方法依赖于显式陈述来识别实体及其关系。从句子“Zuhdi 每周日都去教堂”中，人类读者可以看出 Zuhdi 和基督教之间的关系，但当必须自动推断这种关系时，这却是一个挑战。大型语言模型 (LLMs) 在文本理解、信息抽取 (IE) 等 NLP 下游任务中已被证明是有效的。

本研究考察了文本隐含性如何影响预训练大语言模型中的信息提取任务：LLaMA 2.3、DeepSeekV1 和 Phi1.5。我们生成了两个合成数据集，每个包含 10k 个生物信息的隐式和显式表述，以衡量其对大语言模型性能的影响，并分析微调隐式数据是否能提升它们在隐式推理任务中的泛化能力。

本研究介绍了一个关于 IE 中 LLM 内部推理过程的实验，特别是处理隐式和显式上下文的情况。结果表明，使用 LoRA（低秩适应）对 LLM 模型进行微调可以提高其从隐式文本中提取信息的能力，有助于提升模型的可解释性和可靠性。我们的研究实现在隐式知识中找到

力。传统方法（例如基于规则的方法、深度学习）主要依赖于显式陈述来提取实体、关系和事件 (Alt et al., 2020)。然而，现实世界的文本也以隐含方式传达信息，需要推理处理才能推导出其意指的含义。

隐含意义产生于信息通过语言和认知机制间接传达，而非直接陈述的情况下，此时需要上下文推理和语用推断来进行正确的解释 (Yule, 1996; Evans, 2012; Fischer, 2017)。句子“Zuhdi 参加每周日去教堂”表明 Zuhdi 可能是基督教徒，因为这一推论是从宗教框架中得出的，需要额外的知识来理解未明确表达的内容。类似地，“Sarah 于 2010 年 6 月 15 日从牛津大学获得学位，并庆祝了她的 20 岁生日同一天”暗示她出生于 1990 年 6 月 15 日，建立了未明示的时间蕴含关系。

传记文本对于信息提取来说是一个具有挑战性的案例，因为它们依赖于语言的隐含使用。尽管理解这些文本不需要专门领域的知识，但它们呈现了一个中等程度的复杂性 (Tint et al., 2024)。这种复杂性来源于实体之间的关系、时间上的依赖性和职业参考，这些都是通过上下文线索推断出来的，而不是明确陈述的。

本研究调查了文本隐含性对基于 LLM 的信息提取任务的影响，涉及两个主要研究问题 (RQ)：

- RQ1: 隐式和显式的表述如何影响 LLM 在

1 介绍

信息提取 (IE) 旨在识别、分类和表示来自非结构化文本来源的实体。大型语言模型 (LLMs) 显著提高了自然语言处理 (NLP) 在 IE 中的性能，展示了在诸如文本理解、分类、命名实体识别 (NER) 和关系抽取 (RE) (Niklaus

信息提取任务中的表现？

- RQ2: 在微调过程中暴露于隐式数据如何影响大语言模型对隐式推理任务的泛化能力？

由于大语言模型常常在从隐含语境中提取信息方面表现出困难 (Tint et al., 2024)，我们探讨微调是否可以减轻这种困难。具体来说，我们研究了对社区内知名模型如 LLama3.2 (AI, 2024)、DeepSeekV1 (DeepSeek-AI et al., 2025) 和 Phi1-5 (Li et al., 2023) 进行微调的影响。这一点尤其适用于那些关键信息是隐含传达而非明确表达的场​​景。我们的研究结果有助于提高模型可靠性并拓展潜在应用。

我们专注于两个数据集，一个显式的一个隐式的，包含人们对人物传记的自然语言描述。这些文本是从 Wikidata 三元组数据集开始合成生成的。通过在模仿现实世界场景的隐式模式上微调模型，我们评估它们从隐式文本中提取信息的能力。

本贡献总结如下：第 2 节奠定了本文的工作背景，第 3 节介绍了采用的方法论以回答 RQ1 和 RQ2。我们的结果在第 4 节中进行了展示和讨论。最后，5 节概述了我们最终的评论和未来的工作。

2 背景及相关工作

IE 关注数据的结构化，例如以三重的形式，其中两个参数通过一个关系连接。通常，这种形式表现为由主题、谓词和对象组成的一个三元组，如 $\langle s, p, o \rangle$ (Niklaus et al., 2018)。这项任务通常定义为关系抽取 (RE)。一个内部的区别是封闭型和开放型 RE 的差异。封闭型 RE 关注在给定一个或多个约束条件的情况下查找参数 (例如， $\langle s, p, ?o \rangle$ - 其中对象是唯一的未知值)，而开放型 RE 则寻找文本中的任何潜在三元组 (例如 $\langle ?s, ?p, ?o \rangle$)。传统的 RE 模型主要识别元素 (主语、谓语和宾语) 有明确文本提及的三元组。这些模型训练以识别显式的语言标记 (如用作谓词的动词)，但往往难

以处理需要常识知识或更深层次自然语言理解的隐含关系 (Pei et al., 2023)。预训练的语言模型 (PTLMs) 和 LLMs 表示无监督开放 IE 任务的最先进技术 (Fu et al., 2023)，因为它们比以前的方法更能有效地处理隐式信息。

隐性和显性知识的作用在认知科学中已被广泛研究。根据 Dienes 和 Perner 的理论，当信息通过显式表征的功能使用或概念结构间接传达时，就会产生隐性，而不是直接被表示 (Dienes and Perner, 1999)。

在关系提取中，能够识别具有不同明确程度的关系是一个重大挑战。大型语言模型已经表明，虽然这些模型可以有效地处理显式信息，但它们仍然难以处理需要常识推理的隐性知识 (Ilievski, 2024)。隐性关系的维度可以根据以下因素显著变化：

- 推理所需的水平 (从简单的逻辑推理到复杂的语境推理)
- 所需的背景知识类型 (从常见事实到领域专长)
- 文化和时间背景对于理解是必要的

信息的隐含程度也直接影响模型检索和推理该信息的信心。明确的陈述可以以高信心进行处理，而隐含的信息则引入了不同程度的不确定性，模型必须学会适当地应对这种不确定性。RE 的数据集通常优先考虑显式声明的信息。例如，广泛使用的 RE 数据集 RED (Huguet Cabot et al., 2023) 专注于提取直接匹配文本中句子的三元组。RED 提供了实体类型和关系而没有强制执行额外的结构约束，比如预定义的实体类别或对如何形成关系的具体限制——谓词的域和范围 (见表 1)¹。

这种不确定性随着提取信息所需的推理程度成比例增加。例如，考虑这些关于盖亚的陈述，我们想从中得出有关她的职业的描述：

- "盖亚在城市医院当医生"

¹数据集条目是从 <https://huggingface.co/datasets/Babelscape/REDFM> 中提取的

主题	谓词	对象
Émilie Andéol	sport	judo

源文本：“埃米莉·安德奥尔 [...] 是参加女子+78 公斤级比赛的法国柔道运动员。”

表 1: RED 数据集三句配对示例

- “盖亚穿着白色外套，每天看诊病人”
- “盖亚穿过了急诊室的走廊，迅速查看病历”

所有陈述都以不同程度的隐含性传达了相同的信息。第一条信息是明确的，因为句子结构类似于 $\langle s, p, o \rangle$ 的形式，其中 $?o$ 等于动词属性工作。第二种情况下，职业通过日常工作的描述来体现（换喻）。第三种情况下，信息完全隐藏：即使人类也无法确定她的职业。不同的人可能会手拿图表匆匆穿过急诊室，并不一定就是医生（除非有额外的背景提供了更多线索）。陈述越不明确，不确定性就越大。

大型语言模型似乎在处理隐含信息方面遇到困难，(Becker et al., 2021) 我们试图更好地了解这种限制是源于模型的架构还是其训练数据。具体来说，我们研究这个问题是否反映了随机不确定性—源自语言中的内在不可预测性—或知识不确定性，其中性能受限于模型在训练过程中接触到某些分布的情况 (Hüllermeier and Waegeman, 2021)。

为了这个目的，我们在包含不同程度隐含性的关系抽取任务上对 LLMs 进行了微调。这种方法使我们能够探究模型的泛化能力，并评估性能提升是来自更好地学习输入输出映射还是来自于对隐式模式的熟悉度增加。我们并没有关注事后解释性方法 (Barredo Arrieta et al., 2019; Molnar, 2022)，而是将我们的分析定位为信息抽取背景下对 LLMs 的行为研究，重点关注隐含性如何影响提取的可靠性。

近年来，鲁棒 IE 系统的发展越来越依赖于高质量数据的可用性。然而，对于许多领域来说，现有数据集的数量有限，使得数据增强和合成数据集生成技术成为切实可行的解决方案。例如，在自然语言处理中，回译和同义词替换等方法长期以来一直被用来扩展平行语料库 (Li et al., 2022)。最近，合成数据集生成作为一种策略出现了，它利用大语言模型来为较小的模型创建训练数据，特别是在那些人类标注样本有限的任务或领域中 (Busker et al., 2025)。这种策略在医学和资源匮乏环境中已被证明是有价值的，在这些环境下，注释既昂贵又耗时 (Chebolu et al., 2023)。

这种合成数据生成方法特别适用于解决前面讨论的隐式 RE 挑战。通过生成具有不同程度隐含性的多样化示例，我们有可能提高模型在所有范围的 RE 任务上的性能——从明确陈述到需要复杂推理和蕴含的任务。通常，合成数据是从单个提示或最小的一组指导规则 (Long et al., 2024) 开始生成的，旨在引导模型朝向所需输出。然而，为隐式关系生成高质量的合成数据仍然具有挑战性，因为它要求生成模型模拟人类用来推断未明示连接的复杂推理过程。

3 方法

本节概述了为研究隐性与显性 IE 对 LLM 性能的影响而设计的控制实验，从而解决研究问题 RQ1 和 RQ2。图 1 总结了进行此项研究所采用的整体方法。

首先，从 Wikidata 中抽取了 10,000 个随机实体，特别针对 Human 类²的实体，例如：文森特·罗德里格斯三世。这些实体的传记信息³已通过 Wikidata API 提取，并过滤掉了无关的信息，如识别参数、视觉参考和相关技术元数据。如表 2 所示，14 个三元组描述了 Vincent Rodriguez III 的传记相关信息（例如职业、国籍、性取向），共有 18 个值。我们的目标是为

²Wikidata 类别“人类”通过 IDQ5 标识

³在 Wikidata 中表示为陈述 (<https://www.wikidata.org/wiki/Help:Statements>)

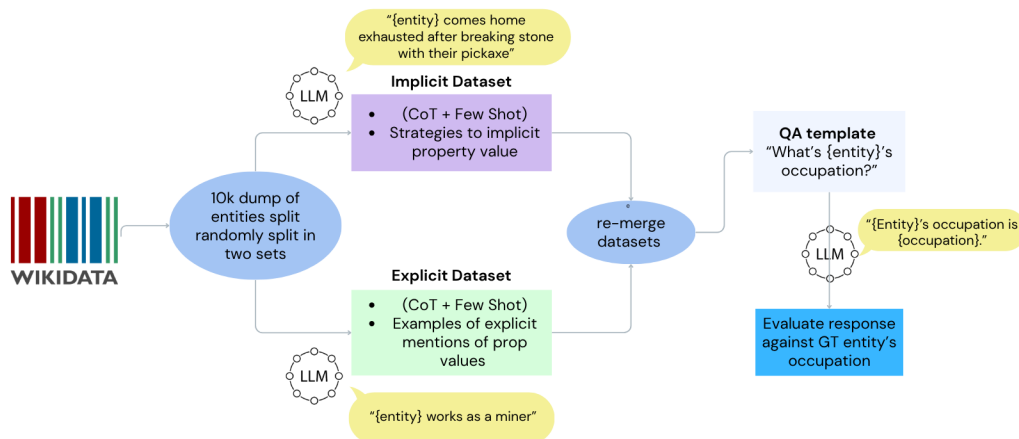


图 1: 数据集生成与实验设置

每个人创建两个平行句子，一个明确地描述关于他们的事实或信息，另一个则隐含地描述。首先，为每个人随机选择一个属性。示例如表 2 所示，Vincent Rodriguez III 所选的信息是他作为电视演员的职业。

然后，关于该人的所有信息以及哪些信息需要变为隐性信息，成为提示的输入。GPT-4o 被指示生成两个不同的句子：一个显式句子（类似于维基百科的直白风格），以及一个通过叙事背景和间接引用传达相同信息的隐式句子。该提示使用少样本学习（10 个示例）和思维链。生成隐式句子的示例是具有配对修辞策略的句子（例如，⁴ 借喻、提喻、演绎）。表 3 呈现了关于文森特·罗德里格斯三世生成的句子。所选属性在第一个描述中明确说明，即“他是一位著名的**电视演员**”，而在第二个描述中，它通过借喻（即“**展示他在各种电视制作中的才华**”）暗示。

最后，进行了两项信息提取测试作为问答任务。我们测试模型检索隐含和显式信息的能力（例如，“Vincent Rodriguez III 的职业是什么？”）。作为一个额外规则，两个答案都可以被认为是有效的，但“电视演员”被视为更好的答案，因为它比简单的“演员”更为具体。事实上，对隐式句子进行信息提取时，仅检索到上位词“演员”，而没有检索到更具体的下位词

⁴参见以下脚本以获取完整的提示：提示生成_隐式.py

“电视演员”，这会导致精度降低，如表 4 所示。

3.1 RQ1: 初步结果与评估

评估已在对明确和隐含描述提供的答案上进行。作为后处理，我们执行了词形还原以将模型答案与 Wikidata 词汇表对齐。然后，通过 BLEURT (Sellam et al., 2020) 计算了预期答案（例如，电视演员）与 LLM 生成的答案（例如，来自明确描述的电视演员和来自隐含描述的演员）之间的语义距离。我们进行了威尔科克森符号秩检验以评估两个分布之间的差异是否具有统计学意义。这种非参数检验比较了两个相关样本，以确定它们的人口平均等级是否有差异。应用威尔科克森符号秩检验后，在考虑 $pvalue < 0.05$ 的情况下，这两个分布具有显著的统计学意义。此外，模型在面对隐含文本时生成 NaN 值的比例明显更高，占比为 14.60%，而明确描述下的比例仅为 1.30%。这些评估为我们使用生成的数据集进行 RQ2 的评估提供了依据。测试的细节和结果可以在 Github 仓库 [隐性知识](#) 中找到。

3.2 RQ2: 微调方法

我们对初步结果的评估（第 3.1 节）涉及 RQ1，表明从统计角度来看，当句子遵循隐含模式时，模型在信息提取方面遇到更多困难。因此我们的第二个研究问题（RQ2）是：在微

主题	谓词	对象	隐藏信息
文森特·罗德里格斯 III	instance of	human	×
	place of birth	San Francisco	×
	sex or gender	male	×
	given name	Vincent	×
	职业	actor	×
		电视演员	✓
	country of citizenship	United States	×
	sexual orientation	homosexuality	×
	date of birth	+1982-08-10T00:00:00Z	×
	受教育于	Pacific Conservatory of the Performing Arts	×
		Westmoor High School	×
	family name	Rodriguez	×
	居住地	Daly City	×
		New York City	×
		North Hollywood	×
	languages spoken, written or signed	English	×
	native language	English	×
	writing language	English	×

表 2: 关于文森特·罗德里格斯三世的精选信息。该表包含了实体罗德里格斯三世的所有三元组（主体、谓词、对象），排除了非语义信息（例如，资源标识符、图像链接）。

调过程中接触隐式数据如何影响大型语言模型对隐式推理任务的泛化能力？为了从上述验证的数据集中证明这一点，我们决定从中选取一个子集，在这个子集中仅选择少数几个职业。为了选择它们，我们在属性值中选择了最常见的 5 种职业‘演员’，‘电影演员’，‘电视演员’，‘舞台演员’，‘电影导演’，即在真实标签和相应的隐式及显式句子中，如表 3 所示。

3.2.1 实验

该实验探讨了通过在不同设置下训练，微调一个 LLM 模型是否可以提高其对隐式实例执行 IE 的能力：

我们期望它能正确工作，因为这应该是最简单的设置；

- **显式 IE 的训练，显式 IE 的测试。**
- **在隐式 IE 上的训练，在隐式 IE 上的测试。**再次，我们期望通过直接在这个任务上训练它来获得良好的性能；
- **对显式和隐式 IE 进行训练，在两者上进行测试，一个用于明确的和一用于隐式。**如果一起训练，模型能够正确分类这两个不同的集合吗？

根据我们在 RQ1 中看到的内容，我们预期这个任务将是模型面临的最困难的任务，

明确描述
Vincent Rodriguez III, 1982 年 8 月 10 日出生于旧金山，自早年在太平洋表演艺术学院学习以来，就以他的表演吸引了观众。居住在纽约和北好莱坞等充满活力的城市，他投身于娱乐界；他是一名著名的 电视演员 。
隐式描述
Vincent Rodriguez III, 1982 年 8 月 10 日出生于旧金山，自早年在太平洋表演艺术学院学习以来，就以他的表演吸引了观众。居住在纽约和诺斯霍尔这样的充满活力的城市，他投身于娱乐界， 展示他在各种电视制作中的才能 展现了他的多才多艺和魅力。

表 3: 关于文森特·罗德里格斯三世的隐式和显式描述

问题	明 确 答案	隐 式 答案
What does Vincent do for a living?	Television actor	Actor

表 4: 显式和隐式答案的比较

因为它需要最大的泛化能力。

• **显式 IE 的训练，隐式 IE 的测试。**

3.2.2 模型

对于我们研究中的分类任务，我们选择了三个在社区内具有重要性和广泛认可的模型。我们的实验基于其广泛应用及其在自然语言处理研究中的表现，选择了 LLaMA、DeepSeek 和 Phi 这三个模型。截至撰写之时（2025 年 4 月），LLaMA 和 DeepSeek 都表现出极大的受欢迎程度，在 [Hugging Face](#) 平台上分别有 210 万和 180 万次下载，这表明了广泛的应用和兴趣。

尽管 Phi 模型（由微软开发）的下载量相对较少（~100K），但由于其性能相对于其尺寸表现出色，它们仍然是一个有价值的组成部分。如 [Hugging Face](#) 模型卡片 [Hugging Face](#) 和

相关基准测试所示，Phi-1.5 在参数少于 10 亿的模型中几乎达到了最先进的结果，使其成为评估指令调优模型时颇具吸引力的轻量级替代方案。

总体而言，我们的选择平衡了社区采纳率、模型多样性与开放性以及参数效率，从而能够在当前的大型语言模型领域进行全面而具有代表性的评估。

- *meta-llama/Llama-3.2-1B*: 由 Meta AI 开发，该模型是 Llama 3.2 多语言大语言模型系列的一部分。它针对多语言对话用例进行了优化，包括代理检索和摘要任务。（[AI, 2024](#)）
- *DeepSeek-R1-Distill-Qwen-1.5B* : 由 Deepseek AI 开发，这是 DeepSeek R1 模型的精简版。鉴于其成本效益和性能，它是 NLP 任务的竞争性选择。（[DeepSeek-AI et al., 2025](#)）
- 微软/*phi-1_5* 一款由微软开发的基于变压器的模型，使用与 Phi-1 相同的数据源进行训练，并增加了新的数据。它在参数少于 10 亿的模型中表现出类似最先进的性能。（[Li et al., 2023](#)）

这些模型各自包含 1 到 15 亿个参数，并且它们托管在 [Hugging Face](#) 平台上（[Wolf and et al., 2020](#)）。鉴于我们有必要测试微调性能，我们仅选择了开源模型，因为我们需要访问模型的权重和结构。这种方法也确保了可重复性。

3.2.3 LoRA 微调

为了构建分类模型，我们使用了带有低秩适应（LoRA）的微调 LLM([Hu et al., 2021](#)) 用于序列分类任务。LoRA([Hu et al., 2021](#)) 是一种参数高效的微调技术，灵感来源于对超参数化模型内在维度的研究。（[Li et al., 2018](#)）和（[Aghajanyan et al., 2020](#)）的研究表明，这类模型在低内在维度上运行，表明庞大的参数空间可以在一个更为紧凑的子空间中高效导航。基于这一

见解，LoRA 假设模型拟合过程中所需的权重变化也具有较低的“内在秩”。因此，在拟合过程中不是更新所有模型参数，LoRA 引入了低秩可训练矩阵来近似这些权重变化。参数概述见表 5，详细信息见附录 A 中的表 9。

3.3 训练详情

我们总共训练了 9 个分类器，每个分类器分别对三种模型进行不同的训练，即 Llama-3.2-1B、DeepSeek-R1-Distill-Qwen-1.5B 和 Phi-1.5，如第 3.2.1 节所述。每次微调都共享相同的超参数。虽然 Llama 和 Deepseek 的性能几乎相同，但 Phi 需要不同的 LoRA Rank 来使训练参数的比例接近其他模型。对于后者，我们增加了表 5 中的轮数，因为它在完成任务以达到与其他模型相同的性能时遇到了困难。所有模型的 $\text{LoRA}\alpha$ 为 64。

3.4 消融研究

我们进行了一个消融研究，以评估模型未经微调时的性能。结果显示，在没有微调的情况下，所有模型表现不佳，准确率在 20% 到 30% 之间。这一对比突显了微调对于使模型执行任务的关键作用，特别是在处理隐式表示方面，当在微调过程中展示隐式数据时，可以达到约 90% 的准确率。

4 结果与讨论

本研究旨在回答两个主要的研究问题。关于 RQ1：隐式和显式的言语化如何影响 LLM 在信息抽取任务中的表现？我们评估了一个语言模型（GPT-4o-mini）从隐性和显性文本数据中提取目标信息的能力。具体来说，我们使用 Sentence-BERT 测量了模型预测与真实值之间的语义距离。这产生了两组距离分数：一组用于显式输入，另一组用于隐式输入。两个分布的统计比较（Wilcoxon 符号秩检验）显示，对于隐性描述的距离显著更高，表明当信息间接传达时，模型的表现更差。支持这一点的是，在第 3.1 节中的分析突出了两种模式：(1) 隐式

文本中失败案例的比例较高（14.6% 的 ‘NaN’ 值），而显式文本为 1.3%；以及 (2) 在隐性条件下，语义相似度较低的情况更为频繁（BLEURT 距离低于 0.6）。这些结果表明模型处理间接语言的能力仍然有限。这些发现指出了 IE 任务中需要改进的领域，在 RQ2：在微调过程中接触隐含数据如何影响大语言模型对隐含推理任务的泛化能力？中进一步探讨了这一点。

表 [6,7,8] 中的结果表明，同时在显式和隐式数据上训练的模型在测试隐式推理任务时始终优于仅依赖于显式数据进行训练的模型。例如，在两种类型的数据上微调并在隐式任务上测试的 Llama 3.2-1B 模型达到了 93.3% 的准确率、94.7% 的平衡准确率和 93.0% 的 F1 分数。这些结果表明，同时接触显式和隐式表述可以提高模型在推理类型之间有效泛化的能力。

相比之下，当模型仅在显式数据上进行训练时，它们在隐式数据上的表现明显更差。例如，在 Llama 3.2-1B 中，一个仅在显式数据上训练并在隐式数据上测试的模型仅达到了 71.6% 的准确率，召回率和 F1 等其他性能指标也受到了影响。同样地，在 DeepSeek R1 Distill Qwen-1.5B 和 Phi 1_5B 中，仅在显式数据上训练的模型显示出类似的困难，分别在隐式数据上的测试准确率下降到 67.1% 和 58.1%。

总结而言，结果展示了对大语言模型进行微调在隐式推理任务上的效果。特别地，我们观察到当模型在显式和隐式数据上都进行了调整时，它们在这两种情况下的推断表现出高水准性能。然而，仅在显式数据上训练的模型在面对隐式任务时遇到了显著困难。这些结果与 RQ1 的研究发现一致。

确实，从表 [6, 7, 8] 中的证据来看，如果模型在训练阶段和测试阶段看到相同的数据分布（测试和训练隐式，测试和训练显式），它就能很好地完成所需的任务。这指向这样一个结论：这种隐式 IE 的困难是由于训练阶段对隐式文本暴露不足造成的，因此处理包含隐含信息的文本时需要进行微调阶段。

模型	N 参数	% 参数。训练好的	长范围依赖 r	纪元
Llama-3.2-1B	1.24B	6.80 %	128	3
DeepSeek-R1-Distill-Qwen-1.5B	1.78B	8.73 %	128	3
phi-1_5	1.42B	5.43 %	256	6

表 5: 模型及其参数的概述，包括参数数量、秩和训练周期的数量。目标模块、 α 值、dropout 率、学习率等超参数在所有配置中保持不变，并在附录 A 的表 9 中详细列出。

模式	准确率	平衡。精度。	精度	回忆	F1
Train and test explicit	0.888	0.922	0.889	0.922	0.903
Train and test implicit	0.911	0.914	0.890	0.914	0.900
Train explicit implicit, test explicit	0.892	0.928	0.892	0.928	0.907
Train explicit implicit, test implicit	0.933	0.947	0.915	0.947	0.930
Train explicit, test implicit	0.716	0.636	0.862	0.636	0.686

表 6: 关于 Llama 3.2-1B 的结果

模式	准确率	平衡。准确性。	精度	回忆	F1
Train and test explicit	0.883	0.923	0.882	0.923	0.900
Train and test implicit	0.896	0.864	0.884	0.864	0.873
Train explicit implicit, test explicit	0.900	0.939	0.897	0.939	0.915
Train explicit implicit, test implicit	0.907	0.894	0.891	0.894	0.891
Train explicit, test implicit	0.671	0.588	0.732	0.588	0.598

表 7: DeepSeek R1 蒸馏 Qwen-1.5B 的结果

模式	准确率	平衡。精度。	精度	回忆	F1
Train and test explicit	0.889	0.906	0.899	0.906	0.902
Train and test implicit	0.911	0.884	0.921	0.884	0.900
Train explicit implicit, test explicit	0.896	0.925	0.897	0.925	0.910
Train explicit implicit, test implicit	0.925	0.921	0.921	0.921	0.921
Train explicit, test implicit	0.581	0.382	0.903	0.382	0.415

表 8: 关于 Phi 1_5B 的结果

5 结论

结果表明，大型语言模型在处理隐含信息方面的困难主要是由于训练过程中对隐含模式的暴露不足，而不是模型架构固有限制。这项测试是在社区中用于分类和生成的流行模型 Llama3.2 1B、DeepSeekV1-DistilledQwen1B 和 Phi1-5 上进行的。通过微调实现的成功改进为

适应现有大型语言模型以更好地处理特定领域中的隐含信息(如我们在传记数据案例中所示)提出了一个实用的发展方向。

未来的发展可以探索不同类型隐含模式如何影响隐含信息提取任务。

限制

本工作的结果仅限于传记资料。虽然可以分析许多其他类型的文本，但获取此类数据集并不像使用 Wikidata 的特定子集生成合成数据集那样直接。此外，数据集的合成生成还存在一个限制：它可能无法完全反映人类语言中自然出现的隐含信息的复杂性。

References

- Armen Aghajanyan, Luke Zettlemoyer, and Sonal Gupta. 2020. [Intrinsic dimensionality explains the effectiveness of language model fine-tuning](#). *Preprint*, arXiv:2012.13255.
- Meta AI. 2024. [Llama 3.2: Revolutionizing edge ai and vision with open-source models](#).
- Christoph Alt, Aleksandra Gabryszak, and Leonhard Hennig. 2020. [TACRED revisited: A thorough evaluation of the TACRED relation extraction task](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1558–1569, Online. Association for Computational Linguistics.
- Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bannetot, Siham Tabik, Alberto Barbado González, Salvador García, Sergio Gil-Lopez, Daniel Molina, V. Richard Benjamins, Raja Chatila, and Francisco Herrera. 2019. [Explainable artificial intelligence \(xai\): Concepts, taxonomies, opportunities and challenges toward responsible ai](#). *Information Fusion*, 58.
- Maria Becker, Siting Liang, and Anette Frank. 2021. [Reconstructing implicit knowledge with language models](#). In *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 11–24, Online. Association for Computational Linguistics.
- Tony Busker, Sunil Choenni, and Mortaza S. Bargh. 2025. [Exploiting gpt for synthetic data generation: An empirical study](#). *Government Information Quarterly*, 42(1):101988.
- Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Tamar Solorio. 2023. [A review of datasets for aspect-based sentiment analysis](#). In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 611–628, Nusa Dua, Bali. Association for Computational Linguistics.
- DeepSeek-AI, Daya Guo, and Dejian Yang et al. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.
- Zoltan Dienes and Josef Perner. 1999. [A theory of implicit and explicit knowledge](#). *Behavioral and Brain Sciences*, 22(5):735 – 808.
- Vyvyan Evans. 2012. [Cognitive linguistics](#). *WIREs Cognitive Science*, 3(2):129–141.
- Kerstin Fischer. 2017. *Cognitive Linguistics and Pragmatics*, page 330 – 346. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.
- Yufeng Fu, Xiangge Li, Ce Shang, Hong Luo, and Yan Sun. 2023. [Zoie: A zero-shot open information extraction model based on language model](#). In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pages 784–789.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. [Lora: Low-rank adaptation of large language models](#). *Preprint*, arXiv:2106.09685.
- Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. 2023. [Red^{lm}: a filtered and multilingual relation extraction dataset](#). In *Proc. of the 61st Annual Meeting of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics.
- Eyke Hüllermeier and Willem Waegeman. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine learning*, 110(3):457–506.
- Filip Ilievski. 2024. [Human-centric ai with common sense](#). *Synthesis Lectures on Computer Science*.
- Bohan Li, Yutai Hou, and Wanxiang Che. 2022. [Data augmentation approaches in natural language processing: A survey](#). *AI Open*, 3:71–90.
- Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. 2018. [Measuring the intrinsic dimension of objective landscapes](#). *Preprint*, arXiv:1804.08838.
- Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: **phi-1.5** technical report. *arXiv preprint arXiv:2309.05463*.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. [On LLMs-driven synthetic data generation, curation, and evaluation: A survey](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11065–11082, Bangkok, Thailand. Association for Computational Linguistics.

- Christoph Molnar. 2022. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, february 28, 2022 edition. Independently published.
- Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2018. [A survey on open information extraction](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3866–3878, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Kevin Pei, Ishan Jindal, and Kevin Chang. 2023. [Abstractive open information extraction](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6146–6158, Singapore. Association for Computational Linguistics.
- Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. [Bleurt: Learning robust metrics for text generation](#). *Preprint*, arXiv:2004.04696.
- Joshua Tint, Som Sagar, Aditya Taparia, Kelly Raines, Bimsara Pathiraja, Caleb Liu, and Ransalu Senanayake. 2024. [Expressivityarena: Can llms express information implicitly?](#) *Preprint*, arXiv:2411.08010.
- Thomas Wolf and Lysandre Debut et al. 2020. [Huggingface’s transformers: State-of-the-art natural language processing](#). *Preprint*, arXiv:1910.03771.
- George Yule. 1996. *Pragmatics*. Oxford university press.

A 附录 A

目标模块	"self_attn.q_proj", "self_attn.k_proj", "self_attn.v_proj", "self_attn.o_proj", "mlp.gate_proj", "mlp.up_proj", "mlp.down_proj"
LoRA 阿尔法	64
LoRA dropout	0.15
学习率	3^{e-5}

表 9: 超参数在所有模型配置中保持不变。对于特
定于模型的设置，如秩和训练周期数，请参阅正文
中的表 5。