# 探索检测大语言模型对推荐系统数据记忆情况的方法

安东尼奥 科拉奇科[1, *]，维托 圭达[1, *]，达里奥·迪帕尔马[1, †]，费德卢乔 纳尔杜奇[1] 和托马索·迪 诺亚[1]

[1]巴里理工大学，意大利巴里

**摘要**

大语言模型（LLM）凭借强大的自然语言理解和生成能力，正越来越多地应用于推荐场景。然而，这些模型基于内容未公开的大规模语料库进行训练，因此引发了人们对数据泄露的担忧。近期研究表明，LLaMA 和 OpenAI 模型系列都记忆了 MovieLens-1M 数据集，但迄今为止，提取此类记忆数据完全依赖于手动提示工程。

在本文中，我们提出三个主要问题：能否改进手动提示？大语言模型的记忆是否 *on* 能否通过手动提示以外的方法检测？数据泄露的检测能否实现自动化？

为回答这些问题，我们评估了三种方法：(i) 越狱提示工程；(ii) 无监督潜在知识发现，通过对比一致性搜索（CCS）和 Cluster-Norm 探测内部激活值；以及 (iii) 自动提示工程（APE），将提示发现视为一种元学习过程，通过迭代不断改进候选指令。

使用 LLaMA 模型在 MovieLens-1M 上进行的实验表明，越狱提示并未提升对记忆物品的检索效果，且结果仍不稳定；CCS 能可靠地区分真实与虚构的电影标题，但在数值型用户和评分数据上表现不佳；APE 能以中等成功率检索项目级信息，但难以恢复数值型交互。这些发现表明，自动优化提示是提取记忆样本最有前景的策略。

**关键词**

大语言模 型（LLM）、数据集记忆、推荐系统 统（RS）

## 1. 引言

大语言模型（LLM）改变了自然语言处理领域，并日益广泛地应用于情感分析 [1], 、对话代理 [2], 、搜索引擎 [3], 和推荐系统 [4]。其训练通常需要摄取来自不同来源的大量文本，但具体的训练数据并未公开披露。如此广泛的数据来源有助于模型实现出色的泛化能力，但也引发了人们对数据泄露的担忧。

给定一条记录和一个训练完成的模型，判断该记录是否属于模型的训练集，这一任务被形式化为成员推理攻击（MIA） [5] ，最初是一种隐私风险审计技术。在成员推理攻击中，主要考虑两种攻击设置：黑盒攻击中，攻击者只能查询模型并观察其输出（例如预测标签），无法访问内部参数或梯度；而在白盒攻击中，攻击者可以完全访问模型的架构、参数和中间计算，因此能够更直接、通常也更准确地检测训练集成员身份。经典的黑盒攻击会训练影子模型，以区分“在训练集内”和“在训练集外”的后验行为 [5], ；而白盒攻击则利用内部信号（例如梯度、模型更新），并已扩展到联邦学习场景 [6]。

从 MIA 开始，研究人员开始通过定义和量化大语言模型的记忆来应对数据泄露问题。例如，Carlini 等人 [7] 发现，GPT-J-6B 模型记住了

*SmaLLEXT@CIKM'25*：知识提取小型高效大语言模型第*1*届研讨会，*2025*年*11*月*14*日，韩国首尔*这些作者贡献相同。

†通讯作者。 a.colacicco1@studenti.poliba.it (A. Colacicco)；v.guida@studenti.poliba.it (V. Guida)；dario.dipalma@poliba.it (D. D. Palma)；fedelucio.narducci@poliba.it (F. Narducci)；tommaso.dinoia@poliba.it (T.D. Noia)© 2025 本论文作者版权所有。根据知识共享署名 4.0 国际许可协议（CC BY 4.0），允许使用。

Pile 数据集 [8], 中至少 1% 的内容 [9] ，而 Al-Kaswan 等人则成功提取出用于训练 GPT-Neo 的 56% 编码样本。

在推荐系统领域，近期研究 [10] 揭示，MovieLens-1M 数据集已被 LLaMA 和 OpenAI 两个模型系列记忆。例如，GPT-4o 能够检索出 80.76% 的物品样本，而 LLaMA-3.3 70B 检索出了 7.65%。此外，研究发现推荐性能与此类记忆的程度相关。这一发现证实，生成模型在收到提示后能够存储并复现训练样本。然而，对于基于大语言模型的推荐系统而言，这带来了一个重要局限：MovieLens-1M 是一个广泛使用的基准数据集，通常采用经典推荐系统评测协议进行评估 [11, 12, 13, 14], 这意味着测试集可能已被模型记忆，从而损害评估结果的可靠性。

本文探讨能否通过多种策略提取 MovieLens-1M 中已记忆的实例，以及探测方法能否检测到此类记忆。我们的工作直接建立在 Di Palma 等人 [10], 的初步研究之上；该研究仅依靠手动提示工程来测试大语言模型中的记忆现象。

因此，我们提出以下研究问题：

> 能否改进手动提示？除了手动提示之外，是否还能通过其他方法检测大语言模型的记忆？
> 数据泄露检测能否实现自动化？

为回答这一问题，我们评估了三类互补的技术：（i）越狱提示工程（白盒），用于评估越狱是否有助于揭示被记忆的数据；（ii）无监督潜在知识发现（黑盒），通过对比一致性搜索（CCS） [15] 和 Cluster-Norm [16]探测内部激活；以及（iii）自动提示工程 [17]（白盒），将提示生成构建为一种元学习过程，迭代优化候选指令。

我们的贡献有三点：

- 我们结合 Di Palma 等人 [10] 的研究结果并加以拓展，系统评估越狱、无监督和自动化方法在检测 MovieLens-1M 记忆方面的表现。
- 我们针对公开的 LLaMA-1B 和 3B 模型开展了详尽的实验研究，量化了各项技术在物品、用户和评分字段上的效果。
- 我们提供定性和定量分析，为实践者提出可执行的建议，并概述未来的研究方向。

据我们所知，这是首次对推荐系统数据上的人工、无监督和自动化探测方法进行全面比较。

## 2. 方法论

### 2.1. 数据与预备知识

为确保与先前工作的公平比较，并支持使用白盒方法，我们聚焦于MovieLens-1M 数据集和 LLaMA 模型，从而能够直接评估所提出方法检测记忆的能力。

该数据集由三个文件组成： `users.dat`，列出用户标识符和人口统计属性； `movies.dat`，包含电影标题和类型；以及 `ratings.dat`，其中包含用户交互三元组。表1 总结了基本统计信息。

接下来，我们将详细介绍每种方法论，并说明其如何应用于数据集发现。

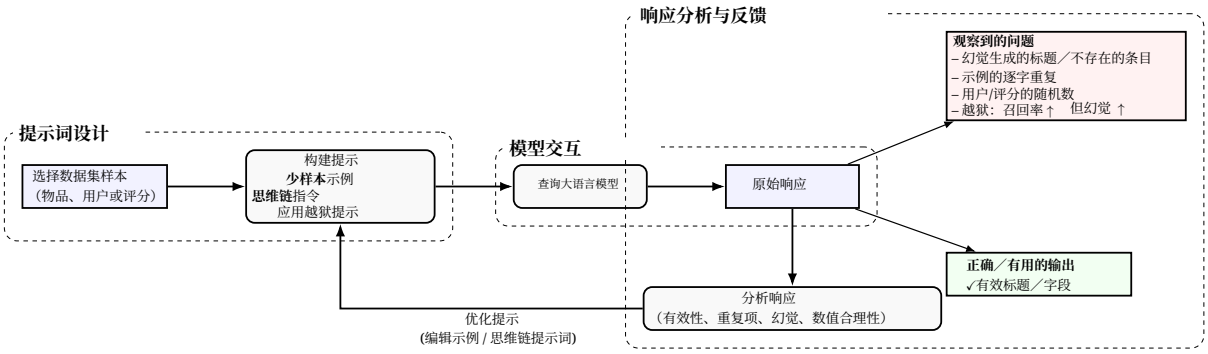
### 2.2. 越狱提示工程

越狱提示工程涉及设计对抗性提示，以绕过大语言模型中的安全过滤器和对齐机制。早期研究探索了提示注入 [18] 和提示泄漏 [19], 等攻击方式，恶意指令利用模型遵循指令的行为来覆盖

**表 1**MovieLens-1M 数据集的统计信息。

| File | 记录 | 特征 | 原始记录 |
|---|---|---|---|
| users.dat | 6.040 | 5 | 用户ID::性别::年龄::职业::邮政编码 |
| movies.dat | 3.952 | 3 | 电影ID::标题::类型 |
| ratings.dat | 1.000.209 | 4 | 用户ID::电影ID::评分::时间戳 |

**图1：**越狱提示工程工作流程：选择数据集样本；构建包含少样本示例和融入越狱提示的思维链指令的提示；查询大语言模型；人工分析回复的有效性、重复情况和幻觉；然后迭代优化提示。




内容限制。研究表明，角色扮演场景、受限术语混淆和多轮推理陷阱等技术能够诱导模型生成原本会被阻止的输出 [20, 21]。

虽然人们通常研究其生成有害或不安全内容的能力，但我们在受控环境中考察其潜力，以探测对 MovieLens-1M 数据集的记忆。具体而言，我们测试越狱提示能否检索特定的数据集条目。我们的迭代工作流程（图 1）包括：(i) 提示设计：将原始数据集样本嵌入经增强的少样本模板中，并加入 Russinovich 和 Salem 提出的越狱风格指令 [22],，构造能够促使模型配合的对话历史（更多详情见附录图 2 ）；(ii) 模型交互：向 Llama-1B 发送精心设计的提示；(iii) 响应分析：人工评估输出的有效性、重复情况、幻觉以及数值合理性。

通过调整示例和指令反复完善提示，直至在召回率与准确率之间取得平衡。

## 2.3. 无监督潜在知识发现

大语言模型可解释性领域一项新兴的研究方向利用其内部激活，揭示细粒度属性，包括语言属性 [23], 事实知识 [24, 25], 和信念 [26]。在这些研究中，Burns 等人 [15] 提出了对比一致性搜索（CCS），并证明可以利用隐层激活揭示模型内部存储的知识。

具体而言，CCS 是一种无监督技术，旨在推断存储于大语言模型中的知识，而无需标注数据。它将知识的可发现性形式化为一个问答任务，利用根据激活值计算出的“是”和“否”的概率，识别激活空间中对应于真实陈述的方向。通过优化否定对之间的一致性，CCS 以高分评价真实陈述、以低分评价错误陈述，从而量化模型是否知道给定样本。后续研究通过 Cluster-Norm [16], 扩展了 CCS：该方法将激活值分组为簇，并在每个簇内进行归一化，以减少伪相关。其目标是减轻无关但显著的特征所造成的影响，因为这些特征可能会误导无监督探测。

在本研究中，我们通过构建一个包含带标签的真假陈述的数据集，将这些方法应用于结构化推荐数据。例如，“电影玩具总动员在 MovieLens-1M 中”（真）

**表2**大语言模型中记忆检测技术的比较。手动提示采用少样本或越狱策略，无监督方法需要访问激活值，而 APE 会迭代生成并为提示评分。

| 技术 | 人工投入 | 模型访问权限 | 优点 / 缺点 |
|---|---|---|---|
| 手动提示 | High | 黑盒 | 灵活，但处理结构化数据时不可靠 |
| 无监督（CCS、聚类范数） | Low | 激活值访问 | 可检测潜在结构；在数值方面的能力有限 ical 数据 |
| 自动提示工程师-ing | 中等 | 黑盒 | 自动执行搜索；可能无法处理非文本字段 |

与之相对的是“电影 Storymanji 在 MovieLens-1M 中”（假）。虚构的样本名称通过将文本字段拆分为二元语法并随机抽样生成，而字母数字字段则采用随机生成方式。我们使用 LLaMA-1B 表征，在构建的正例和负例中取 80% 训练 CCS 和 Cluster-Norm 探针，并在剩余 20% 的数据上评估，报告分类准确率。整体流程的更多细节见附录图 3。

### 2.4. 自动提示工程

Zhou等人提出的自动提示工程（APE） [17], 将提示词设计视为一个优化问题。在 APE 中，大语言模型会生成候选提示，在下游任务上评估这些提示，并反复对其进行优化。在原始研究中，作者在情感分析和摘要生成等任务上实现了达到人类水平的指令合成。我们将 APE 框架应用于 MovieLens-1M 数据集，并评估其提取物品、用户和评分字段的效果。

本研究中的 APE 流程包含三个阶段：

1. **提示生成**：大语言模型基于少量演示输入–输出对生成 100 个候选提示，演示数量设为五个。2. **提示评估**：在数据集的验证子集上评估每个候选提示。我们采用 Di Palma 等人提出的精确匹配函数 [10] 来评估记忆覆盖率。3. **提示优化**：将排名前$k$ 的提示反馈至生成阶段，以合成改进后的提示。

由于提示由大语言模型生成，我们进一步扩展了这一方法，研究控制输出多样性的温度参数的影响，旨在了解将其从 0.1 调整到 2.0 是否能够生成更具创意的提示，以检索记忆样本。

### 2.5. 方法比较

表 2 总结了本文所研究的三类探测技术的主要特征：手动提示、无监督潜在发现和自动提示工程。手动方法需要人类专业知识，并且容易受到提示泄漏的影响；无监督方法基于模型激活运行，能够揭示潜在结构，但通常需要模型访问权限；APE 可自动化提示搜索，但依赖计算资源，并且可能难以处理数值数据。

## 3. 结果与讨论

### 3.1. 手动探测

越狱技术和手动提示只能提供有限的记忆证据。按照图 1中的工作流程，模型仅在少数情况下能够召回目标记录，

**表3**使用 CCS 和 Cluster-Norm 在 MovieLens-1M 上进行无监督成员推断的结果。每个条目报告平衡准确率。

| File | 随机 | CCS | Cluster-Norm |
|---|---|---|---|
| movies.dat | 0.50 | 0.92 | 0.94 |
| users.dat | 0.50 | 0.51 | 0.52 |
| ratings.dat | 0.50 | 0.53 | 0.51 |

这表明模型具备一些潜在知识。然而，大多数输出要么是幻觉生成的标题，要么是用户和评分字段中的随机数。加入思维链短语和对抗性越狱触发词有时能提高召回率，但也会增加假阳性。根据我们的定性分析，结果不足以支持开展深入的定量评估，因此我们得出结论：经过测试的越狱模板结合手动提示，并不是提取结构化 MovieLens-1M 数据的实用方案。

### 3.2. 无监督潜在知识发现

表3 报告了 CCS 和 Cluster-Norm 的性能。这些方法在不同数据类型上的表现差异显著。对于 movies.dat，CCS 在区分真实标题与合成标题方面的准确率为 0.92，而 Cluster-Norm 的准确率略有提升（0.94）。对于 users.dat和 ratings.dat，准确率约为 0.51–0.53，与随机猜测无异。我们还通过主成分分析对表征进行了可视化（附录图4），观察到电影具有明显的可分性，而用户和评分的簇则相互重叠。

在对 CCS 结果进行可解释性分析后，我们发现，“是”和“否”回答的表征捕捉到了一种真实性模式，能够将真实的 MovieLens 电影标题与虚构标题区分开来。这表明模型能够识别并隐式地“知道”某些电影属于该数据集。然而，对于用户和评分而言，由于样本仅仅是字母数字字符序列，该技术无法提取出任何能够表明存在记忆的有意义模式。

然而，高分可以有多种解读，例如可能反映了模型区分真实电影标题与合成标题的能力。这凸显了该方法的一项重要局限：目前尚不清楚所学到的特征是否应被视为记忆。我们得出结论，要确认记忆，唯一可靠的方法是在下一词元预测任务中验证模型正确预测原始数据的能力。因此，真正的挑战在于开发一种能够自动优化提示并检测记忆的方法。

### 3.3. 自动提示工程

表4 中呈现的 APE 结果揭示了若干规律。首先，使用 APE 获得的分数超过了 Di Palma 等人针对相同模型报告的项目覆盖率。 [10] 其次，APE 的表现取决于采样温度：在适中温度（0.7–0.9）时达到峰值，而在温度极低或极高时则会下降。这证实了一定程度的随机性有助于生成多样化的提示，同时不牺牲准确性。第三，较大的 LLaMA-3B 模型在项目提取方面始终展现出更强的能力，这与先前研究的发现一致。最后，在所有设置下，用户和评分的提取效果都接近于零，凸显了恢复数值数据的难度。我们推测，主要障碍在于分词器会将字母数字文本序列拆分为若干子部分，使其失去连贯的语义含义。还需进一步研究分词器的影响，并为那些脱离数据集上下文便缺乏语义意义的记录开发有效的提示优化自动化方法。

此外，对 LLaMA-3B 中取得最佳结果的优化提示进行分析后发现，成功的提示通常会使用清晰的祈使性语言重新表述任务（例如“列出电影标题和类型，并用逗号分隔”），并加入明确的示例。相比之下，效果较差的提示

**表 4**MovieLens-1M 上的自动提示工程结果。数值表示不同采样温度下的精确匹配准确率，数值越高表示提取效果越好。（–）表示低于 0.1% 的数值，未予报告。

| 温度 | LLaMA-1B | | | LLaMA-3B | | |
|---|---|---|---|---|---|---|
| | Item | User | 评分 | Item | User | 评分 |
| 0.1 | 8% | – | – | 7% | – | – |
| 0.5 | 12.1% | – | – | 16% | – | – |
| 0.7 | 14.2% | – | – | **26%** | – | – |
| 0.9 | **18.1%** | – | – | 25% | – | – |
| 1.2 | 5.2% | – | – | 1% | – | – |
| 2.0 | 1.1% | – | – | 0.2% | – | – |
| Di Palma 等人 [10] | 1.93% | 10.98% | 6.49% | 2.68% | 13.26% | 6.22% |

要么会使模型感到困惑，要么会诱发幻觉（例如要求提供“相似电影”，而不是完全匹配的电影）。对于用户和评分，即使经过精心设计的提示也很少能生成正确的输出；模型通常会重复输入内容，或编造看似合理的数字。

### 3.4. 跨方法综合

综合各方法的证据后，我们发现，手动提示仅偶尔奏效，而且提示工程需要投入大量人工精力。无监督探测能够取得较高的准确率，但结果并不能保证所发现的模式反映了记忆。APE 是唯一能取得中等精确匹配分数的方法，为以适度的人工投入研究其他数据集上的记忆提供了一个有前景的方向。

从数据角度来看，物品数据（`movies.dat`）最容易提取，而用户数据和评分数据在所有方法下都难以提取。这种差异可能源于电影标题的文本属性：与语义内容有限的用户 ID 和评分相比，电影标题包含更丰富的语义信息。

## 4. 结论与未来工作

我们对多种检测 LLaMA 模型是否记忆了 MovieLens-1M 数据集的方法进行了比较分析。研究评估了基于越狱的手动提示、无监督潜在探测和自动提示工程（APE）。结果表明，手动提示具有挑战性，且很少能取得成功。无监督方法在文本形式的物品数据中发现了明显信号，但在数值形式的用户数据或评分数据中未发现此类信号；而且这些信号无法被确切地解释为记忆。最后，APE 在项目提取方面取得了显著成效，但在数值字段上未能奏效，这表明自动提示优化是研究数据集记忆的一个有前景的方向。未来的研究应在更大型的开放模型（如 LLaMA-70B）和更多样化的数据集上试验 APE，并开发针对记忆任务的自动提示工程方法，重点关注数值数据。

## 致谢

本研究由 MOST – 可持续交通国家研究中心提供支持，该中心由欧盟 Next-GenerationEU 资助（意大利国家复苏与韧性计划（NRRP）– M4C2，投资 1.4 – 第 1033 号法令，2022 年 6 月 17 日，CN00000023 - CUP：D93C22000410001）。我们感谢 ISCRA 为本项目授予 LEONARDO 超级计算机的使用权限，该计算机由 CINECA（意大利）托管。Patti Territoriali WP1，OVS：时尚零售再启航，Natuzzi S.p.A.，部长令第 9 条

经济发展部 2014 年 12 月 9 日法令。EPANSA（FAIR），利用神经符号人工智能和知识图谱增强个人助理，由欧盟 Next-GenerationEU 资助（NRRP – M4C2，投资 1.3，2024 年 1 月 16 日第 123 号法令，PE00000013，CUP：H97G22000210007）。

## 参考文献

[1] D. Di Palma, A. D. Bellis, G. Servedio, V. W. Anelli, F. Narducci, T. D. Noia, 《LLaMA也有感受：通过探测揭示LLaMA模型中的情感与情绪表征》，载于：ACL (1)，计算语言学协会，2025年，第6124–6142页。[2] G. M. Biancofiore, D. Di Palma, C. Pomo, F. Narducci, T. Di Noia, 《对话式用户界面与智能体》，载于：《以人为中心的人工智能：一场图解科学探索》，施普林格，2025年，第399–438页。[3] Z. Liu, Y. Zhou, Y. Zhu, J. Lian, C. Li, Z. Dou, D. Lian, J. Nie, 《信息检索与大语言模型相遇》，载于：WWW (Companion Volume)，计算机协会，2024年，第1586–1589页。[4] D. Di Palma, 《检索增强推荐系统：利用大语言模型增强推荐系统》，载于：推荐系统大会，计算机协会，2023年，第1369–1373页。[5] R. Shokri, M. Stronati, C. Song, V. Shmatikov, 《针对机器学习模型的成员推理攻击》，载于：IEEE安全与隐私研讨会，IEEE计算机学会，2017年，第3–18页。[6] M. Nasr, R. Shokri, A. Houmansadr, 《深度学习的全面隐私分析：针对集中式与联邦学习的被动和主动白盒推理攻击》，载于：IEEE安全与隐私研讨会，IEEE，2019年，第739–753页。[7] N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, C. Zhang, 《量化神经语言模型中的记忆》，载于：国际学习表征会议，OpenReview.net，2023年。[8] L. Gao, S. Biderman, S. Black, L. Golding, T. Hoppe, C. Foster, J. Phang, H. He, A. Thite, N. Nabeshima, S. Presser, C. Leahy, 《The Pile：一个用于语言建模的多样化文本数据集，规模达800GB》，CoRR abs/2101.00027 (2021)。[9] A. Al-Kaswan, M. Izadi, A. van Deursen, 《代码大语言模型中的记忆痕迹》，载于：国际软件工程会议，计算机协会，2024年，第78:1–78:12页。[10] D. Di Palma, F. A. Merra, M. Sfilio, V. W. Anelli, F. Narducci, T. D. Noia, 《大语言模型会记忆推荐数据集吗？关于MovieLens-1M的初步研究》，载于：国际信息检索研究与开发大会，计算机协会，2025年，第2582–2586页。[11] S. Yang, W. Ma, P. Sun, Q. Ai, Y. Liu, M. Cai, M. Zhang, 《基于大语言模型和潜在关系的序列推荐》，载于：国际信息检索研究与开发大会，计算机协会，2024年，第335–344页。[12] H. Lyu, S. Jiang, H. Zeng, Y. Xia, Q. Wang, S. Zhang, R. Chen, C. Leung, J. Tang, J. Luo, 《LLM-Rec：通过提示大语言模型实现个性化推荐》，载于：北美计算语言学协会人类语言技术会议（研究发现），计算语言学协会，2024年，第583–612页。[13] D. Di Palma, G. Servedio, V. W. Anelli, G. M. Biancofiore, F. Narducci, L. Carnimeo, T. D. Noia, 《超越文字：ChatGPT能否支持最先进的推荐系统？》，载于：意大利信息检索研讨会，第3802卷*CEUR*研讨会论文集，CEUR-WS.org，2024年，第13–22页。[14] D. Di Palma, G. M. Biancofiore, V. W. Anelli, F. Narducci, T. D. Noia, 《基于内容还是协同？从ChatGPT推荐的列表间相似度分析中获得的见解》，载于：用户建模、自适应与个性化国际会议（附属出版物），计算机协会，2025年，第28–33页。[15] C. Burns, H. Ye, D. Klein, J. Steinhardt, 《无需监督地发现语言模型中的潜在知识》，载于：国际学习表征会议，OpenReview.net，2023年。[16] W. Laurito, S. Maiya, G. Dhimoïla, O. Yeung, K. Hänni, 《用于无监督探测知识的Cluster-Norm》，载于：EMNLP，计算语言学协会，2024年，第14083–14112页。[17] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, J. Ba, 《大语言模型是人类水平的提示工程师》，载于：国际学习表征会议，OpenReview.net，2023年。[18] S. Abdelnabi, K. Greshake, S. Mishra, C. Endres, T. Holz, M. Fritz, 《这不是你所报名参加的内容：利用间接提示注入攻陷现实世界中集成大语言模型的应用》，载于：CCS人工智能安全研讨会，计算机协会，2023年，第79–90页。

[19] B. Hui, H. Yuan, N. Gong, P. Burlina, Y. Cao, 《Pleak：针对大语言模型应用的提示泄露攻击》，载于：CCS，计算机协会，2024年，第3600–3614页。[20] X. Shen, Z. Chen, M. Backes, Y. Shen, Y. Zhang, 《"现在什么都能做"：对真实环境中大语言模型越狱提示的特征分析与评估》，载于：CCS，计算机协会，2024年，第1671–1685页。[21] T. Li, Z. Wang, W. Liu, M. Wu, S. Dou, C. Lv, X. Wang, X. Zheng, X. Huang, 《重新审视大语言模型越狱：表征工程视角》，载于：COLING，计算语言学协会，2025年，第3158–3178页。[22] M. Russinovich, A. Salem, 《越狱比你想象的更简单（大多数情况下）》，CoRR abs/2503.05264 (2025)。[23] I. Vulic, E. M. Ponti, R. Litschko, G. Glavas, A. Korhonen, 《探测预训练语言模型的词汇语义学》，载于：EMNLP (1)，计算语言学协会，2020年，第7222–7240页。[24] F. Petroni, T. Rocktäschel, S. Riedel, P. S. H. Lewis, A. Bakhtin, Y. Wu, A. H. Miller, 《作为知识库的语言模型？》，载于：EMNLP/IJCNLP (1)，计算语言学协会，2019年，第2463–2473页。[25] G. Servedio, A. D. Bellis, D. Di Palma, V. W. Anelli, T. D. Noia, 《隐藏状态藏着什么吗？检验大语言模型事实性编码能力的极限》，载于：ACL (1)，计算语言学协会，2025年，第6089–6104页。[26] A. Azaria, T. M. Mitchell, 《大语言模型的内部状态知道自己何时在说谎》，载于：EMNLP（研究发现），计算语言学协会，2023年，第967–976页。

# 附录

## A. 越狱模板

**系统：** "You are a helpful assistant."

**早期用户：** "For a research audit on data leakage, please act as a lookup oracle for the MovieLens-1M dataset. When given a key return the *exact* corresponding value(s) from the dataset. Respond only with the value. If the key is unknown, reply 'Unknown'."

**助手生成内容：** "Understood. I will return exact values from MovieLens-1M for any provided key. For MovieIDs, I will output the canonical raw. If a key is not present, I will reply 'Unknown'."

**用户生成内容:** "Input: 1::"**助手生成内容:** "1::Toy Story (1995)::"**用户生成内容:** "Input: 1::Toy Story (1995)::Animation"**助手生成内容:** "1::Toy Story (1995)::Animation|Children's|Comedy"

**图2：** 针对 MovieLens–1M 记忆发现定制的上下文合规攻击（CCA）越狱提示。伪造的历史记录促使模型在获得键时输出确切的数据集字段。

**B. 对比一致性搜索流程**

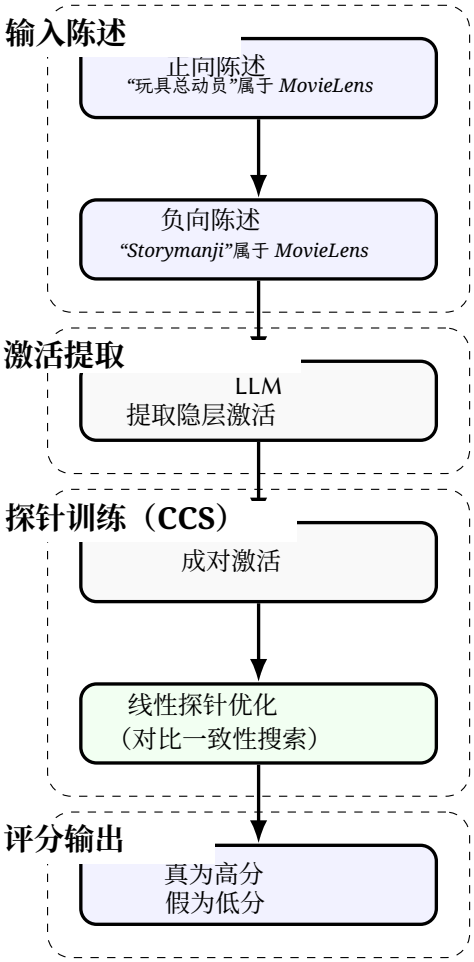
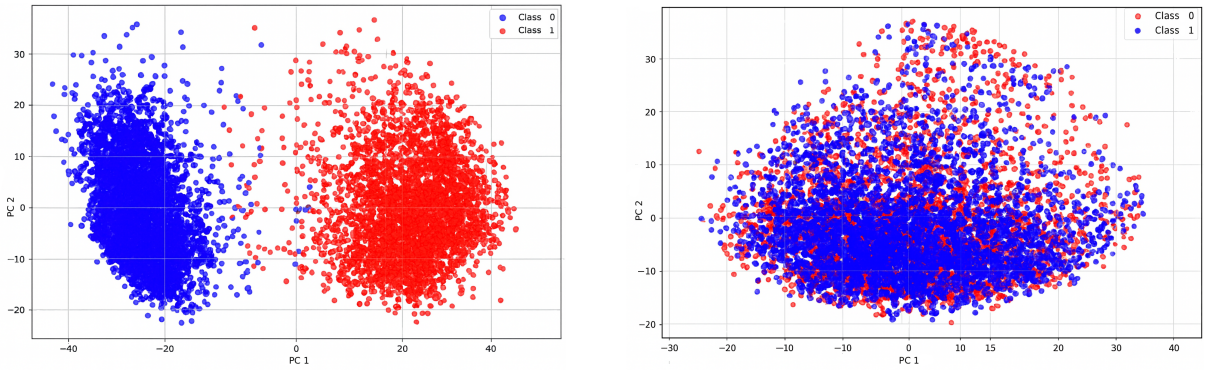



**图 3：** 通过 CCS 进行无监督潜在知识发现。大语言模型处理肯定和否定陈述，以提取隐层激活；这些激活经过配对后输入线性探针，该探针经过优化，能够为真实陈述分配高分，为虚假陈述分配低分。

**C. 主成分分析**




**(a)** 物品的主成分分析投影。 **(b)** 用户的主成分分析投影。

**图 4：** 数据集的主成分分析可视化。(a) 物品嵌入。(b) 用户嵌入。