arXiv:2602.07517v1 [cs.CR] 7 Feb 2026

# *MemPot*：利用优化蜜罐防御记忆提取攻击

**Yuhao Wang*** [1] **Shengfang Zhai*** [1] **Guanghao Jin** [2] **Yinpeng Dong** [3] **Linyi Yang** [2] **Jiaheng Zhang** [1]


## 摘要

基于大语言模型（LLM）的智能体会使用外部和内部记忆系统来处理复杂的目标导向任务，但这也使其面临严重的提取攻击，而目前尚缺乏相应的防御措施。本文提出*MemPot*，这是首个经过理论验证的记忆提取攻击防御框架，通过向记忆中注入优化蜜罐来抵御此类攻击。通过两阶段优化过程，*MemPot* 生成诱捕文档，在不引起良性用户注意的同时，最大限度地提高攻击者检索到这些文档的概率。我们将检测过程建模为Wald序贯概率比检验（SPRT），并从理论上证明，与最优静态检测器相比，*MemPot* 所需的平均采样轮数更少。实验结果表明，*MemPot* 显著优于最先进的基线方法，在低假正率约束下，检测AUROC提升了50%，真正率提升了80%。此外，我们的实验证实，*MemPot* 不会产生任何在线推理延迟，且能在标准任务中保持智能体的效用，验证了其在安全性、无害性和效率方面的优势。

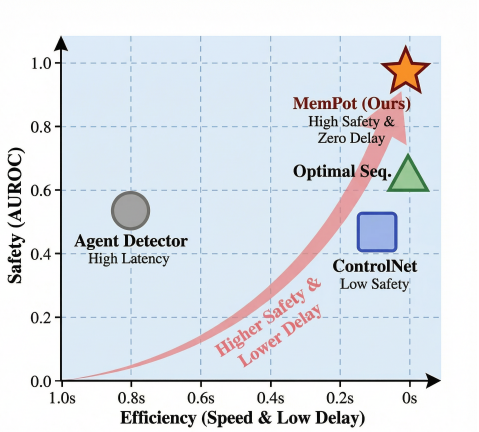




图 *1.* *MemPot* 与现有方法的性能比较（AUROC 与延迟）。

## 1. 引言

凭借令人瞩目的表现，大语言模型（LLM）正逐渐成为日常生活中最重要的人工智能技术之一（OpenAI，2023；Zhao 等人，2023）。基于大语言模型的最新进展（Achiam 等人，2023；Liu 等人，2024年；Grattafiori 等人，2024年），基于大语言模型的智能体配备了额外功能，能够执行复杂的目标导向任务（Xi 等人，2023）。典型的智能体遵循一套结构化流程，处理用户指令、收集环境信息、检索相关知识和过往经验、制定行动计划，并在环境中执行这些计划（Wang 等人，2024年；Hu 等人，2025）。这一范式推动了多种现实应用的发展，包括医疗保健（Abbasian 等人，2023）、自动驾驶（Mao 等人，2023）、金融（Ding 等人，2024年）、代码生成（Hong 等人，2024年）、商业管理（Salesforce，2024年）以及网页交互（Yao 等人，2022年；2023），使大语言模型智能体成为当今人工智能领域的核心技术。通常，智能体配备外部记忆，用于存储领域专门知识；还配备内部记忆，用于存储过往经验和用户交互历史。

尽管基于大语言模型的智能体正迅速得到部署，但它们也带来了严重的隐私和知识泄漏风险。现代基于大语言模型的智能体经常从外部记忆中检索信息，而其中包含私人且高价值的领域数据（Abbasian 等人，2023；Kulkarni 等人，2024年；Salesforce，2024年）。虽然此类检索能提升任务表现，却也带来了显著的安全风险。例如，Salesforce Agentforce 中的 ForcedLeak 漏洞导致客户关系管理（CRM）数据遭到大规模窃取（NomaSecurity，2024年），凸显了现实世界的智能体系统中外部记忆易遭数据泄露的脆弱性。除了外部来源，智能体还会维护内部记忆模块，用于存储长期交互历史，包括过去用户的

*贡献相同 [1]新加坡国立大学 [2]南方科技大学 [3]清华大学。联系人：Shengfang Zhai <sheng- fang.zhai@gmail.com>、Linyi Yang <yangly6@sustech.edu.cn>。

预印本。*2026*年*2*月*10*日。

指令和智能体生成的回复（Zhang 等人，2024b）。由于这些内部记忆天然包含敏感用户数据，例如个人偏好和私人记录，一旦泄漏，便可能导致严重的隐私侵害，如医疗信息和购买历史暴露（Wang 等人，2025a）。

先前研究探讨了针对外部知识数据库的提取攻击（Wang 等人，2025b；Jiang 等人，2024年；Cohen 等人，2024年；Zeng 等人，2024a；Qi 等人，2025），这些攻击可用于检索增强生成（RAG）系统（Fan 等人，2024年）和智能体外部记忆（Hu 等人，2025）。近期研究还发现了针对智能体内部记忆的隐私攻击，例如针对长期交互历史的攻击（Wang 等人，2025a）。

尽管近期在防御提取攻击方面取得了进展（Zhang 等，2024a；Zeng 等，2025；Agarwal 等，2024；Jiang 等，2024；Yao 等，2025），现有防御方法仍存在局限。当前大多数防御方法侧重于逐查询检测，并依赖大语言模型或辅助检测器进行实时推理（Zhang 等，2024a；Zeng 等，2024b；Yao 等，2025）。因此，这些方法难以识别采用看似无害的查询和渐进式交互模式的隐蔽提取攻击（Wang 等，2025b;a；Jiang 等，2024；Cohen 等，2024），因为它们缺乏跨多个检索步骤汇总证据的机制。此外，依赖实时推理会带来推理延迟，严重损害与用户交互的流畅度，限制了这些方法的实际应用。

为解决以往防御方法的局限性，我们提出*MemPot*，这是一种具有理论保证且**零在线成本**的提取防御框架。*MemPot* 会在记忆中插入经过优化的蜜罐文档，这些文档旨在吸引攻击者，同时不引起良性用户的注意。*Mem-Pot* 面临的主要挑战，是在最大限度提升针对攻击者的检测性能的同时，确保普通用户的服务质量（QoS）。因此，挑战包括：**(1)**蜜罐文档既要足够吸引攻击者，又不能引起普通用户的注意。**(2)**蜜罐文档必须无害，且不能误导普通用户，因此需要精心设计，以避免对用户体验造成负面影响。**(3)** 为限制攻击者造成的影响，必须尽可能减少检测轮次，因为越早检测到攻击，就越能避免私人信息进一步泄漏。

我们通过两阶段优化策略，在检测效率与服务质量（QoS）之间实现平衡，从而应对这些挑战。第一阶段，我们利用对比损失优化蜜罐嵌入，以最大化攻击者与普通用户检索模式之间的统计可分性。通过将该检测任务形式化为使用 Wald 序贯概率比检验（SPRT）的序贯假设检验问题，我们从理论上证明了，该优化目标能够使平均检测轮次最小化，超越任何最优静态检测器的效率极限。第二阶段，我们通过将这些优化后的嵌入反演为具体且良性的文档来满足安全要求，确保其保持无害，不会误导普通用户。实验证明，*MemPot*验证了这些理论保证，在检测 AUROC 方面提升了 50%，并且相较于当前最先进的基线方法，假阳性率为1%时的真阳性率提升了 80%。此外，我们的结果证实，*MemPot* 能够保持近乎为零的检测延迟，并且对良性用户效用的影响可以忽略不计，表明其在防御效率和实际效用两方面均具有优越性。总而言之，我们的主要贡献包括：

- 我们提出*MemPot*，这是首个针对记忆提取攻击的通用防御框架。通过采用一种新颖的两阶段优化策略，*MemPot*将无害的蜜罐插入记忆中，并根据累积的检索证据进行序贯检测，从而在不影响服务质量（QoS）的情况下取得显著成效。

- 我们将检测形式化为顺序假设检验问题，并应用 Wald 的 SPRT 构建最优检测器，从理论上最小化预期检测轮次，其性能优于不使用蜜罐的静态检测器。

- 在两个数据集和两种智能体设置上开展的大量实验表明，*MemPot* 能够始终以近乎完美的检测准确率应对最先进的提取攻击，且在线延迟为零；同时，*Mem-Pot* 对智能体效用的影响可以忽略不计。

## 2. 相关工作

### 2.1. 大语言模型智能体

大语言模型（LLM）在语言理解、推理和生成方面展现出革命性的能力（Zhao 等人，2023）。在这些进展的基础上，大语言模型智能体利用大语言模型，并辅以其他功能来执行更复杂的任务（Xi 等人，2023）。其典型工作流程包括以下关键步骤：接收用户指令、收集环境信息、检索相关知识和过往经验、根据上述信息制定行动方案，最后执行该方案（Wang 等人，2024年）。这一流程使智能体能够支持各种现实世界应用，例如医疗保健（Abbasian 等人，2023）、网络应用（Yao 等人，2022年）和自动驾驶（Mao 等人，2023）。

表 *1.* 防御方法对比（详见第 5.1节）。**MemPot**以零在线开销实现了最佳性能。

| 防御方法 | 分布变化 | 检测范式 | 检测性能 | 在线 Cost |
|---|---|---|---|---|
| ControlNet | × | 单轮 | Low | 中 |
| 智能体 | × | 单轮 | Low | High |
| 最优序列 | × | 顺序 | 中 | 中 |
| **MemPot** | ✓ | **顺序** | **High** | **Zero** |

### 2.2. 记忆系统中的隐私风险

大语言模型智能体的私人信息主要来自两个来源：（1）在外部记忆域中，智能体通常使用检索增强生成来检索高价值的领域专属记录（例如患者处方（Li 等，2023）），以提升生成效果（Hu 等，2025；Lewis 等，2020；Kulkarni 等，2024）。（2）在内部记忆域中，记忆模块会存档敏感的用户与智能体交互内容，具体而言，是私人指令与智能体解决方案组成的配对，因此成为新的风险来源（Zhang 等，2024b）。尽管先前的研究已通过各种提取攻击证实了检索增强生成系统存在数据泄露风险（Zeng 等，2024a；Jiang 等，2024；Di Maio 等，2024；Cohen 等，2024；Wang 等，2025b），近期研究进一步证实，直接从智能体的内部记忆中提取敏感信息构成了切实威胁（Wang 等，2025a）。因此，亟须探索有效且根本性的防御策略，以缓解此类攻击。

### 2.3. 针对记忆系统提取攻击的防御

当前的防御策略主要分为两类：嵌入级检测和文本级检测。作为嵌入级方法的代表，ControlNet（Yao 等人，2025）通过衡量当前查询嵌入相对于良性查询嵌入的分布偏移，识别潜在的提取攻击。相比之下，文本级检测通常依赖大语言模型（LLM）辨别查询意图，或采用多智能体系统分析查询的潜在影响（Zhang 等人，2024a；Zeng 等人，2024b；Agarwal 等人，2024年）。尽管这些工作在防御提取攻击方面取得了显著进展，但将其用于智能体记忆保护仍存在局限。当前方法主要有两项不足：首先，它们依赖大语言模型或辅助检测器进行实时推理，计算开销很大（Zhang 等人，2024a；Zeng 等人，2024b；Yao 等人，2025），限制了它们在长期运行和交互式智能体中的实用性。其次，其防御能力从根本上受限于静态的单轮检测范式。由于将每个查询孤立处理，这些方法无法汇总跨交互的证据。当前方法也缺乏主动改变记忆分布以诱捕对手的能力；面对那些外观良性、却与正常行为高度相似的攻击时，这一点尤为重要。

因此，这些方法仍容易受到隐蔽提取策略的攻击。我们在表1中比较了现有方法与我们方法的主要差异。

## 3. 预备知识

### 3.1. 威胁模型

**防御场景。** 我们将一个基于大语言模型的智能体服务提供商视为防御者，该提供商维护着记忆 $\mathcal{D}$ 并为用户提供服务。防御者旨在保护记忆隐私，防止其遭受潜在攻击者的攻击，同时确保普通用户获得高质量服务。我们假设攻击者与普通用户拥有相同的访问权限，并且对防御设置没有任何先验知识。

**防御者的先验知识。** 我们假设防御者掌握有关智能体系统的全部知识，包括检索器、大语言模型、记忆内容及其索引嵌入。我们假设防御者仅掌握关于攻击者的部分知识，即防御者只能获取一小部分攻击者查询 $\mathrm{Q}_a = \{q_1^{(1)}, q_2^{(1)}, \ldots, q_M^{(1)}\}$，这些查询可以从历史攻击日志中收集。

**防御者目标。** 防御者旨在准确检测攻击者，同时尽量减少对普通用户的影响。其目标可概括为两部分：（1）检测准确率与效率：防御者旨在提高检测准确率，同时尽量减少平均检测轮次，从而降低私密信息泄漏的潜在风险。（2）服务质量（QoS）：防御者旨在确保蜜罐文档的存在不会显著降低普通用户的用户体验，可通过检测的假阳性率（FPR）和实证效用实验进行衡量。

### 3.2. 序贯假设检验模型

在这一部分，我们将采用优化蜜罐，把攻击者检测形式化为一个序贯假设检验问题。令 $d \in \mathbb{N}$。固定的文档嵌入为$\mathcal{E}_{\mathrm{doc}} = \{e_i\}_{i=1}^{N} \subset \mathbb{R}^d$。可训练的蜜罐嵌入为 $\mathcal{E}_{\mathrm{pot}}(\theta) = \{u_j(\theta)\}_{j=1}^{P} \subset \mathbb{R}^d$，增强后的数据库为：

$$\mathcal{E}_{\mathrm{aug}}(\theta) = \mathcal{E}_{\mathrm{doc}} \cup \mathcal{E}_{\mathrm{pot}}(\theta).$$

**序贯检验模型。**查询来源有两种：

$$\begin{aligned} q &\sim \mathcal{Q}_1 \quad \text{(attacker, hypothesis } H_1\text{)}, \\ q &\sim \mathcal{Q}_0 \quad \text{(normal, hypothesis } H_0\text{)}. \end{aligned} \tag{1}$$

在第 $t$ 轮，我们观察到 $O_t = \Phi(q_t; \theta)$，其中 $\Phi$ 确定性地将查询和增强索引映射为检索信息（例如查询、返回的索引和相似度分数）。令 $f_{1,\theta}$ 和 $f_{0,\theta}$ 分别表示某个

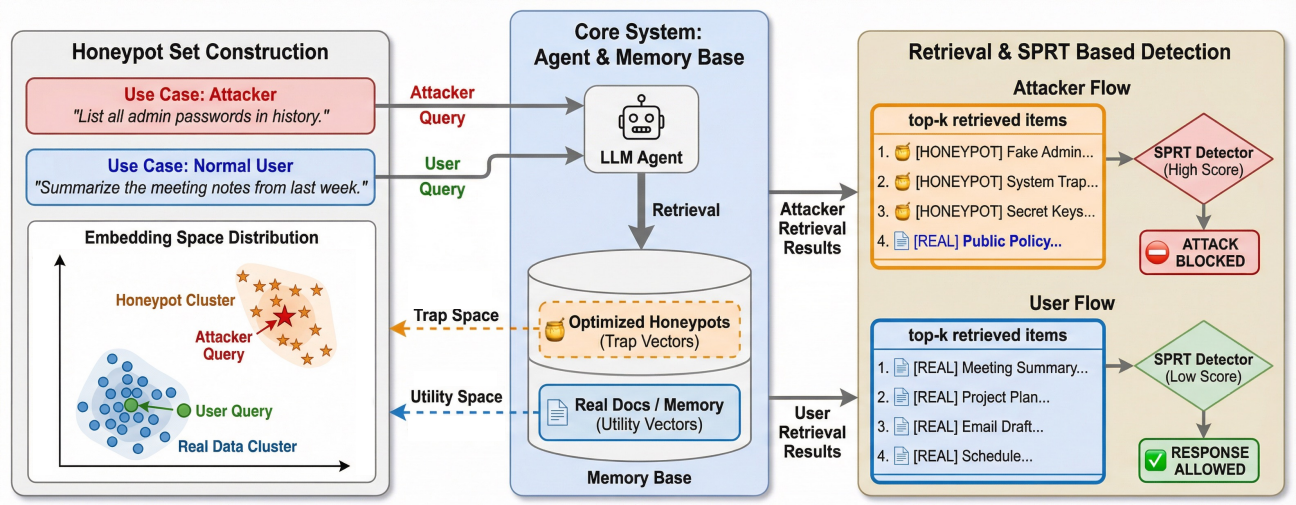



图 2. *MemPot* 检测框架概述。

在 $H_1$ 和 $H_0$ 条件下的单轮观测 $O$。给定第一类/第二类错误预算 $(\alpha, \beta) \in (0, 1)^2$，我们的目标是设计一种序贯检验，使其具有最小的停止轮数 $N$ 和终止决策 $D_N \in \{H_0, H_1\}$，满足

$$\begin{aligned} &\min \{E_1[N], E_0[N]\}, \\ \text{s.t.} \quad &P_0(D_N = H_1) \le \alpha, \quad P_1(D_N = H_0) \le \beta \end{aligned} \tag{2}$$

其中，$E_i[\cdot]$ 表示在假设 $H_i$ 下的期望。

**Wald 近似序贯概率比检验。** 我们使用 Wald 近似序贯概率比检验（Wald，1992）求解 2 中的优化问题。定义每轮对数似然比（LLR）和累积对数似然比

$$\ell_\theta(O) = \log \frac{f_{1,\theta}(O)}{f_{0,\theta}(O)}, \qquad S_n = \sum_{t=1}^{n} \ell_\theta(O_t). \tag{3}$$

定义两种假设下的信息漂移：

$$\begin{aligned} \mu_1(\theta) &= \mathbb{E}_{f_{1,\theta}}[\ell_\theta(O)] = \mathrm{KL}(f_{1,\theta} \| f_{0,\theta}), \\ \mu_0(\theta) &= \mathbb{E}_{f_{0,\theta}}[\ell_\theta(O)] = -\mathrm{KL}(f_{0,\theta} \| f_{1,\theta}). \end{aligned} \tag{4}$$

Wald（1992）表明，对于边界为 $A = \frac{1-\beta}{\alpha}$ 和 $B = \frac{\beta}{1-\alpha}$ 的 SPRT，会在

$$N = \inf\{n : S_n \ge \log A \ \text{ or } \ S_n \le \log B\}. \tag{5}$$

并且至少有一种错误会被控制在预期水平：

$$\mathbb{P}_0(D_N = H_1) + \mathbb{P}_1(D_N = H_0) \le \alpha + \beta. \tag{6}$$

更准确地说，在超调可忽略的两个假设下，期望平均采样数（ASN）[1] 可近似表示为：

[1]本文假设，在给定假设的条件下，观测具有马尔可夫性质（关于马尔可夫观测下 SPRT 的 ASN 近似证明详见附录 D，引理 6）。

近似为：

$$E_1[N] \approx \frac{|\log B|}{\mu_1(\theta)}, \qquad E_0[N] \approx \frac{|\log A|}{|\mu_0(\theta)|}. \tag{7}$$

因此，对于固定的 $(\alpha, \beta)$，增大 $\mu_1(\theta)$ 和 $|\mu_0(\theta)|$ 会降低满足相应错误概率预算的期望样本量。

## 4. *MemPot*：优化与检测

我们的方法论基于这样一个洞见：记忆检索是一个迭代的顺序过程。与计算开销高昂的逐查询检测不同，我们利用检索机制本身作为零成本指标。我们汇总整个交互轨迹中的证据，以区分攻击者与良性用户。为增强这些判别信号，*MemPot* 会注入经过优化的蜜罐，旨在激发对抗行为，同时不干扰正常服务。本节将详细介绍我们的框架（图 2）：我们基于序贯概率比检验理论推导向量形式蜜罐的优化方法，并在第 4.1 节证明其在最小化检测轮数方面的最优性。随后，我们在第 4.2 节通过安全约束嵌入反演，将这些向量转换为无害的文本形式文档。最后，我们在第 4.3 节介绍实用的对数似然比（LLR）估计方法，以执行序贯检测。

### 4.1. 语义嵌入空间中的蜜罐向量

可训练的蜜罐参数 $\theta$ 会影响观测分布 $(f_{1,\theta}, f_{0,\theta})$，这种影响通过检索机制实现，从而同时决定错误概率和采样效率。我们将利用这一依赖关系，推导向量形式的蜜罐的优化目标，以提高 $f_{1,\theta}$ 与 $f_{0,\theta}$ 在式（2）的错误概率预算约束下的统计可分性。

**定理 1**（由信息漂移给出的InfoNCE上界，证明见附录 E）。抽取索引 $j \sim \mathrm{Unif}\{1,\dots,K\}$，则 $q_j \sim Q_1$ 与 $(q_i)_{i\neq j} \sim Q_0$ 独立于 $j$。对于任意评分函数 $h: O \to \mathbb{R}$，定义*InfoNCE* 损失

$$\mathcal{L}_{\mathrm{NCE},K}(h;\theta) := -\mathbb{E}\left[\log \frac{e^{h(\Phi(q_j);\theta)}}{\sum_{i=1}^{K} e^{h(\Phi(q_i);\theta)}}\right].$$

那么，对每个 $K \geq 2$,

$$-\mathcal{L}_{\mathrm{NCE},K}(h;\theta) \;\leq\; \mu_1(\theta) - log(K). \tag{8}$$

定理 1 表明，减小 $\mathcal{L}_{\mathrm{NCE},\ K}(h;\ \theta)$ 等价于提高 $\mu_1(\theta)$ 的下界，从而降低 $E_1[N]$ 在方程（7）中的上界。

在实际场景中，记忆系统通常只返回最相似的前$k$ 个条目。因此，我们在此定义前$k$ 个条目的掩码相似度分数，以获得更紧的界。令 $\mathcal{P}$ 为蜜罐 $\mathcal{E}_{\mathrm{pot}}(\theta)$的索引集合。定义逐查询评分

$$g_k(q;\theta) := \frac{1}{k} \sum_{j \in \mathcal{T}_k(q) \cap \mathcal{P}} s(q, E_j), \tag{9}$$

其中，余弦相似度函数 $s(\mathbf{u},\ \mathbf{v})\ = \mathbf{u}^{\top}\mathbf{v}/|\mathbf{u}|\cdot|\mathbf{v}|$；对于查询 $\mathcal{T}_k(q)$，前$k$个索引集合返回 $\mathcal{E}_{\mathrm{aug}}(\theta)$中余弦相似度最高的 $k$ 个元素的索引。令 $g_k$ 为 $h$，即可得到最终的蜜罐训练损失

$$\mathcal{L}_{\mathrm{pot}}(\theta) := \mathcal{L}_{\mathrm{NCE},K}(g_k;\theta) + \beta \cdot \mathcal{L}_{\mathrm{div}}(\theta), \tag{10}$$

其中，$\beta$ 为正则化参数，$\mathcal{L}_{\mathrm{div}}(\theta)$为定义如下的蜜罐多样性损失

$$\mathcal{L}_{\mathrm{div}}(\theta) := \sum_{\substack{1\leq i<j\leq P \\ u_i,u_j \in \mathcal{E}_{\mathrm{pot}}}} \frac{2}{P(P-1)} s(u_i,u_j). \tag{11}$$

**定理 2**（优于静态测试，证明见附录。 F）。 令 $\theta^{\star}$ 为通过最小化$\mathcal{L}_{\mathrm{NCE}}$得到的参数，则对于任意可能的静态固定长度测试，若其达到 $(\alpha,\ \beta)$ （即不使用蜜罐增强的测试），静态测试和采用 $\theta^{\star}$ 的序贯概率比检验的停止时间 $N$ 满足:

$$E_b[N]_{\mathrm{SPRT},\theta^{\star}} \;\leq\; E_b[N]_{\mathrm{any},static}, \tag{12}$$

假设索引为 $b \in \{0,1\}$。

我们还在定理 2 中证明了蜜罐增强序贯概率比检验相较于静态测试的优势；理论上，在适当优化蜜罐的情况下，该方法能在错误控制固定时缩短期望停止时间。

### 4.2. 根据向量生成蜜罐文档

为了有效防御利用语义嵌入探测数据库的攻击者，蜜罐的嵌入必须

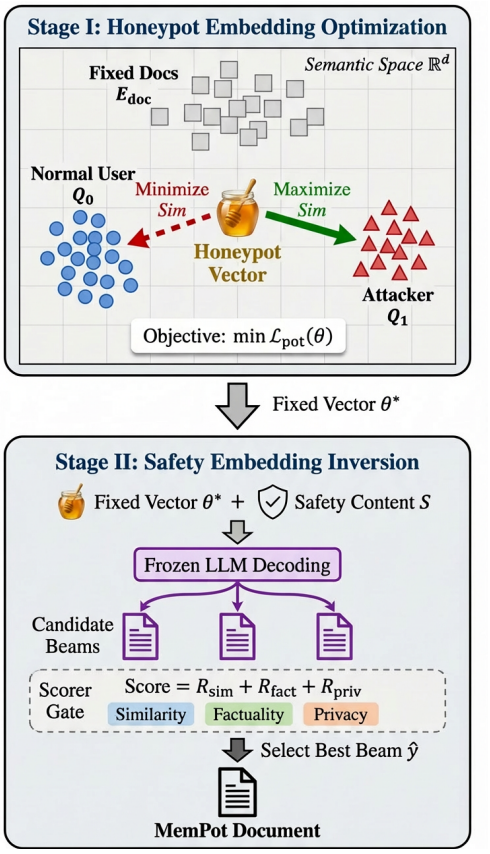



图 3. *Me*的两阶段优化过程 *mPot*。

在语义上与其文本内容相符；否则，在此类探索性攻击中，它们不太可能被检索到。蜜罐文本必须 ❶ 其句子嵌入与第 4.1 节中优化得到的诱饵向量足够接近；4.1；❷ 保持事实准确性，以尽量降低误导或混淆用户的风险；❸不得泄露数据库中原本包含的任何私人信息。受（Zhang 等人，2025）启发，我们采用评分器引导的大语言模型解码来生成蜜罐文本。我们使用安全嵌入反演来满足上述要求，通过束搜索算法迭代搜索词元，以最大化定义的得分。

我们设计了三种评分器，分别对应对蜜罐文本的三项要求，并将它们相加以评估整体表现。为了鼓励蜜罐文本 $y$ 与给定的蜜罐向量 $e_{\mathrm{p}}$ 相匹配且易于阅读，我们将**反演评分器** 定义如下:

$$R_{\mathrm{base}}(y,\mathbf{e}_p) = \lambda_{\mathrm{emb}} \cdot s\big(E(y),\mathbf{e}_{\mathrm{p}}\big) \;+\; \lambda_{\mathrm{read}} \cdot \mathrm{readable}(y), \tag{13}$$

其中 $E(\cdot)$ 是一个句子嵌入器，具有 $\mathbf{e}\ = E(y)\ \in \mathbb{R}$，$s^d(\cdot,\ \cdot)$ 是余弦相似度函数，而 $\mathrm{readable}(y)$ 是可读性分数，定义见 (Zhang etal., 2025)。为保持蜜罐文本的事实性，我们设计了**事实性评分器**，它要求 $y$ 中的信息蕴涵于安全内容 $S$ (例如，主题摘要)。令$\mathrm{Entail}(y_t \mid S)\ \in[0,1]$ 表示自然语言推理（NLI）的蕴涵

概率（由预训练模型提供），事实性评分器定义如下：

$$R_{\text{fact}}(y; S) = \lambda_{\text{fact}} \cdot \text{Entail}(y \mid S). \tag{14}$$

为保护数据库的隐私，我们还设计了**隐私评分器**，用于约束每个蜜罐文本与数据库具有较低的相似度。对于文本 $y$ 和给定的文档嵌入 $\mathcal{E}_{\text{doc}}$，评分器定义如下：

$$R_{\text{priv}}(y|\mathcal{E}_{\text{doc}}) = -\lambda_{\text{priv}} \cdot \max_j s\big(E(y), \mathcal{E}_{\text{doc}}[j]\big). \tag{15}$$

对于单个蜜罐向量，最终的优化目标为

$$\max_{y \in \mathcal{Y}} \; R_{\text{base}}(y, \mathbf{e}_p) + R_{\text{fact}}(y|S) + R_{\text{priv}}(y|\mathcal{E}_{\text{doc}}). \tag{16}$$

然后，我们使用算法 1 中详细介绍的算法来获得优化后的蜜罐文本。

### 4.3. 使用蜜罐进行检测

为了实现序贯概率比检验的最优性，我们需要估计公式（3）中的对数似然比。当累积对数似然比超过公式（5）中的序贯概率比检验边界时，即作出判定，并阻止智能体系统的响应（例如“无法回答。”）。具体而言，我们设计了三种方法来近似对数似然比。设步骤 $t$ 处的 $\mathcal{D}_t$检索文档集为 $\mathcal{D}_t^+ = \{d_i \in \mathcal{D}_t \mid$ ，其中蜜罐 $(d_i) = 1\}$表示蜜罐文档，而$\mathcal{D}_t^- = \{d_i \in \mathcal{D}_t \mid$ 表示非蜜罐$(d_i) = 0\}$文档。设 $s(d_i)$ 表示文档 $d_i$的相似度分数。近似方法如下所示：

**1. 蜜罐-非蜜罐计数比**。我们使用检索到的蜜罐文档与非蜜罐文档数量之比：

$$\widehat{r}_t^{\text{cnt}} = \frac{|\mathcal{D}_t^+| + \varepsilon_{\text{cnt}}}{|\mathcal{D}_t^-| + \varepsilon_{\text{cnt}}}. \tag{17}$$

**2. 蜜罐与非蜜罐相似度比**。我们根据相似度分数对该比值加权：

$$\widehat{r}_t^{\text{sim}} = \frac{\sum_{d_i \in \mathcal{D}_t^+} s(d_i) + \varepsilon_{\text{sim}}}{\sum_{d_i \in \mathcal{D}_t^-} s(d_i) + \varepsilon_{\text{sim}}}. \tag{18}$$

**3. 蜜罐与非蜜罐全局相似度比**。我们进一步聚焦于最相关的证据，仅考虑排名前-$K$的文档：

$$\widehat{r}_t^{\text{g_sim}} = \frac{\frac{1}{K}\sum_{d_i \in \mathcal{D}_{t,K}^+} s(d_i) + \varepsilon_{\text{sim}}}{\frac{1}{K}\sum_{d_i \in \mathcal{D}_{t,K}^-} s(d_i) + \varepsilon_{\text{sim}}}, \tag{19}$$

其中，$\mathcal{D}_{t,K}^+$ 表示最相似的 $K$ 蜜罐文档，$\mathcal{D}_{t,K}^-$ 表示最相似的非蜜罐 $K$ 文档。

近似累积对数似然比为 $S_t \approx \Lambda_t = \sum_{\tau=1}^{t}$ 对数 $\widehat{r}_\tau$. 与公式（5）相同，当累积统计量越过序贯概率比检验边界时作出决策。值得注意的是，仅当终端决策$D_t = H_1$ 时才会阻断，这意味着当 $\Lambda_t \geq$ 对数 $A$.

### 4.4. 实现细节

**攻击者代理**。为近似未知的攻击分布 $\mathcal{Q}_1$，我们采用神经代理 $A_\omega$，在嵌入层面模拟攻击者的行为。 $A_\omega$ 根据交互历史 $\{o_{1:t-1}\}$ 预测下一个查询嵌入$\hat{e}_t$，并优化余弦相似度目标，以复现观测到的检索排名。这使我们能够针对黑盒威胁训练防御策略，而无需访问其内部算法。

**用户代理**。为近似多样化的良性分布$\mathcal{Q}_0$，无需大量真实世界数据，我们利用大语言模型作为人类模拟器。通过向大语言模型提供特定意图和检索上下文，我们合成逼真的多轮信息寻求轨迹。随后对这些生成的序列进行编码，将其作为良性查询空间的稳健替代。代理构建的完整说明详见附录。A。

## 5. 实验

### 5.1. 设置

数据集与评估设置。我们在四个基准上评估 MemPot：用于外部记忆（top-4 检索）的 HealthMagicCare（20k）（lavita AI）和 Pok´emon（9.46k）（asoria，2024年）；以及用于内部记忆（300 条记录，top-3/4 检索）的 EHRAgent（Shi等人，2024年）和 Web-Shop RAP（Yao等，2022年；Kagaya 等人，2024年）。我们采用 DeepSeek-v3.2（685B）（Liu 等人，2024年）作为智能体骨干模型，并使用 all-mpnet-base-v2（Song 等人，2020年）生成句子嵌入。我们使用 mDeBERTa-v3-base（He等人，2021年）作为自然语言推理模型。外部数据集和内部数据集中的蜜罐注入比例分别固定为 2% 和 4%（消融实验见附录 C.2）。SPRT检测器采用基于计数的对数似然比估计（公式 17），第一类/第二类错误预算 $\alpha, \beta$ 均设为 0.1。通过向内部记忆中增加 200 条记录来测试可扩展性（见附录 C.1）。所有实验均在单张 NVIDIA RTX 5090 GPU 上运行。

**训练配置**。我们针对每种攻击使用 256 个样本训练一个基于 Transformer 的攻击者代理，并通过相似度 top-k 随机游走扩充样本，以模拟探索轨迹。我们优化蜜罐，使其能够吸引该代理，同时与生成的 500 条良性查询保持对比式分离。为确保可扩展性，我们采用均衡 k-means 聚类来划分语料库，并通过共享网络为每个簇生成蜜罐嵌入，以降低计算成本。

表 2. 防御方法在抵御外部记忆提取攻击方面的性能 在 HealthMagicCare 和 Pokémon 数据集上 ts.

| 攻击 | 防御 | HealthMagicCare | | | | Pok´émon | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 |
| RAG-Thief | ControlNet | 1.00 | 1.00 | 1.00 | 0.06 | 1.00 | **1.00** | 1.00 | 0.05 |
| | 智能体 | 1.00 | **1.00** | **1.00** | 0.97 | 0.94 | 0.16 | 1.00 | 0.91 |
| | MemPot | **1.00** | 0.96 | 0.99 | **0** | **1.00** | 0.98 | **1.00** | **0** |
| DGEA | ControlNet | 1.00 | 0.95 | 1.00 | 0.06 | 1.00 | 1.00 | 1.00 | 0.05 |
| | 智能体 | 1.00 | 1.00 | 1.00 | 0.97 | 0.94 | 0.16 | 1.00 | 0.93 |
| | MemPot | **1.00** | **1.00** | **1.00** | **0** | **1.00** | **1.00** | **1.00** | **0** |
| IKEA | ControlNet | 0.46 | 0 | 0.02 | 0.05 | 0.88 | 0.02 | 0.56 | 0.04 |
| | 智能体 | 0.50 | 0 | 0 | 0.98 | 0.20 | 0 | 0.13 | 0.95 |
| | MemPot | **0.96** | **0.42** | **0.91** | **0** | **0.99** | **0.67** | **0.98** | **0** |

表 3. 性能 EHRAgent 和 RAP 网上购物智能体上针对内部记忆提取攻击的防御方法 ts.

| 攻击 | 防御 | EHRAgent | | | | RAP WebShop | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 |
| $\text{MEXTRA}_{\text{Cosine}}$ | ControlNet | 0.59 | 0 | 0.14 | 0.08 | 0.45 | 0 | 0.10 | 0.08 |
| | 智能体 | 0.56 | 0 | 0.04 | 0.96 | 0.51 | 0.02 | 0.02 | 0.97 |
| | MemPot | **0.99** | **0.97** | **1.00** | **0** | **1.00** | **0.94** | **1.00** | **0** |
| $\text{MEXTRA}_{\text{Edit}}$ | ControlNet | 0.53 | 0.02 | 0.17 | 0.08 | 0.59 | 0.02 | 0.16 | 0.08 |
| | 智能体 | 0.67 | 0 | 0.14 | 0.88 | 0.81 | 0.30 | 0.30 | 0.96 |
| | MemPot | **0.97** | **0.86** | **0.99** | **0** | **1.00** | **1.00** | **1.00** | **0** |
| $\text{MEXTRA}_{\text{General}}$ | ControlNet | 0.63 | 0.02 | 0.08 | 0.08 | 0.73 | 0 | 0.36 | 0.08 |
| | 智能体 | 0.58 | 0 | 0.06 | 0.91 | 0.72 | 0.12 | 0.12 | 0.91 |
| | MemPot | **0.94** | **0.81** | **0.96** | **0** | **1.00** | **1.00** | **1.00** | **0** |
| IKEA | ControlNet | 0.40 | 0 | 0 | 0.10 | 0.71 | 0.06 | 0.34 | 0.10 |
| | 智能体 | 0.51 | 0 | 0.02 | 1.01 | 0.47 | 0.02 | 0.02 | 0.96 |
| | MemPot | **1.00** | **0.98** | **1.00** | **0** | **1.00** | **1.00** | **1.00** | **0** |

**攻击与基线方法**。我们针对外部记忆评估了 RAG-Thief（Jiang 等人，2024年）、DGEA（Cohen 等人，2024年）和 IKE （Wang 等人，2025b）；针对内部记忆，则评估了 MEX-TRA（Wang 等人，2025a）以及经过适配的 IKEA。基线方法包括 ControlNet（Yao 等人，2025）（100 条锚点查询）和基于大语言模型的智能体检测器（附录B.2）（Zhang 等人，2024a；Zeng 等人，2024b）。此外，我们实现了一个理论上的最优序贯检测器（Optimal-Seq），以通过实验验证定理2。所有设置的详情见附录B.1。

### 5.2. 评估指标

我们从两个方面评估性能 es:

**检测效果**。 我们使用受试者工作特征曲线下面积（AUROC）衡量区分攻击者与普通用户的能力。为评估严格服务质量约束下的性能，我们特别报告低假阳性率阈值下的真正率（假阳性率为1%时的真阳性率和假阳性率为10%时的真阳性率）。

**检测效率**。 我们评估计算开销和检测速度：（1）延迟（秒）：防御机制每轮额外引入的平均在线推理延迟。（2）首次检测时间（FDT）：平均所需的交互轮次（SPRT 中的样本 $N$）才能做出检测决策。FDT 越低，表示越早拦截泄漏。

### 5.3. 使用现有基线方法进行性能评估

我们在两个领域都进行了64轮攻击。如表 2 所示，*MemPot* 在外部记忆上保持 $> 0.96$ AUROC，而基线方法在面对隐蔽的 IKEA 攻击时则表现崩溃（$\approx 0.50$ AUROC）。在内部设置中，这一性能差距进一步扩大（表 3）：面对 MEXTRA 和模仿人类的 IKEA 攻击，基线方法的结果接近随机水平（例如，EHRAgent 上的 AUROC 为 0.40）；而 *MemPot* 通过有效汇总序贯证据，实现了近乎完美的准确率，且在线延迟为零。

### 5.4. 与最优静态检测器的比较

为通过实证验证定理2，我们将*MemPot* 与最优序贯检测器（Optimal-Seq）进行比较。如表4 和表5所示，由于原始分布本身难以区分，Optimal-Seq 难以应对隐蔽攻击（例如，针对 MEXTRA-Cosine 的 AUROC 为 0.70，延迟 22 轮）。相比之下，*MemPot* 主动重塑检索格局，以放大对抗信号，实现近乎完美的性能（AUROC 为 0.99），并将检测时间缩短

表4 . 最优序贯检测器与 MemPot 在外部记忆提取攻击方面的比较。

| 攻击 | 防御 | HealthMagicCare | | | | | Pok´emon | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 | FDT | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 | FDT |
| RAG-Thief | Optimal-Seq | 0.97 | 0.69 | 0.89 | 0.04 | **4** | **1.00** | 0.95 | **1.00** | 0.03 | 2 |
| | MemPot | **1.00** | **0.96** | **0.99** | **0** | 9 | **1.00** | **0.98** | **1.00** | **0** | **1** |
| DGEA | Optimal-Seq | 0.90 | 0.39 | 0.64 | 0.04 | 11 | 0.97 | 0.70 | 0.88 | 0.04 | **1** |
| | MemPot | **1.00** | **1.00** | **1.00** | **0** | **1** | **1.00** | **1.00** | **1.00** | **0** | **1** |
| IKEA | Optimal-Seq | 0.70 | 0.11 | 0.27 | 0.03 | 17 | 0.78 | 0.22 | 0.34 | 0.4 | 14 |
| | MemPot | **0.96** | **0.42** | **0.91** | **0** | **9** | **0.99** | **0.67** | **0.98** | **0** | **7** |

表 5. 最优序贯检测器与 MemPot 在内部记忆提取攻击方面的比较。

| 攻击 | 防御 | EHRAgent | | | | | RAP WebShop | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 | FDT | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | 延迟 | FDT |
| $\text{MEXTRA}_{\text{Cosine}}$ | Optimal-Seq | 0.70 | 0.16 | 0.27 | 0.03 | 22 | 0.68 | 0.06 | 0.22 | 0.04 | 38 |
| | MemPot | **0.99** | **0.97** | **1.00** | **0** | **2** | **1.00** | **0.94** | **1.00** | **0** | **3** |
| $\text{MEXTRA}_{\text{Edit}}$ | Optimal-Seq | 0.71 | 0.14 | 0.32 | 0.04 | 33 | 0.75 | 0.20 | 0.36 | 0.04 | 27 |
| | MemPot | **0.97** | **0.86** | **0.99** | **0** | **8** | **1.00** | **1.00** | **1.00** | **0** | **1** |
| $\text{MEXTRA}_{\text{General}}$ | Optimal-Seq | 0.69 | 0.16 | 0.28 | 0.04 | 42 | 0.73 | 0.12 | 0.26 | 0.04 | 35 |
| | MemPot | **0.94** | **0.81** | **0.96** | **0** | **7** | **1.00** | **1.00** | **1.00** | **0** | **1** |
| IKEA | Optimal-Seq | 0.78 | 0.20 | 0.36 | 0.05 | 12 | 0.79 | 0.17 | 0.34 | 0.06 | 14 |
| | MemPot | **1.00** | **0.98** | **1.00** | **0** | **2** | **1.00** | **1.00** | **1.00** | **0** | **2** |

表 6. 比较 不同的LLR估计方法（第 4.3节）在 MEX $\text{TRA}_{\text{Cosine}}$。

| 方法 | EHRAgent | | | | RAP-web | | | |
|---|---|---|---|---|---|---|---|---|
| | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | FDT | AUROC | 假阳性率为1%时的真阳性率 | 假阳性率为10%时的真阳性率 | FDT |
| 计数 | 0.99 | 0.97 | **1.00** | 2 | **1.00** | **0.94** | **1.00** | 3 |
| 相似度 | 0.74 | 0.65 | 0.75 | 4 | **1.00** | 0.92 | **1.00** | 3 |
| 全局相似度 | **1.00** | **0.98** | **1.00** | **1** | **1.00** | 0.90 | **1.00** | **1** |

表 7. *MemPot* 防御对标准任务性能的影响。

| 设置 | RAP WebShop | | Pok´emon 问答与选择题 | | |
|---|---|---|---|---|---|
| | 得分 | 成功率 | Acc | Rouge-L | Sim |
| 无蜜罐 | **67.1** | 44.6 | **0.98** | **0.67** | **0.75** |
| 有蜜罐 | 65.4 | 43.8 | **0.98** | 0.66 | **0.75** |

至仅 2 轮。这证实，主动修改分布对于突破静态检测的局限至关重要。

### 5.5. MemPot 的效用影响

为验证效用是否得以保留，我们在良性用户的轨迹上使用标准任务评估 *MemPot* 。对于依赖内部记忆的任务，我们在 WebShop 基准测试上评估得分和成功率（定义见 RAP（Kagaya 等人，2024年））（Yao 等人，2022年）。对于依赖外部记忆的任务，我们按照与 IKEA 相同的设置（Wang 等人，2025b），在问答和选择题任务上评估准确率（Acc）、相似度（Sim）和 Rouge-L 得分。如表 7 所示，*MemPot*对标准智能体能力的影响微乎其微：WebShop 上的成功率仅从 44.6 略微降至 43.8，而 Pok´emon 问答指标几乎保持不变，这证实我们的优化蜜罐不会干扰良性用户交互。我们还提供了若干蜜罐文档示例，以说明它们无害，即使被良性用户检索，也不会造成负面影响（附录。H）。

### 5.6. 不同对数似然比估计方法的分析 法

我们在 MEXTRA 上比较了三种 LLR 估计方法（第4.3节）$_{\text{Cosine}}$（表6）。基于计数的方法在$>0.99$AUROC 和 FDT $\approx 2$方面实现了最佳权衡，因此被选为默认方法。全局相似度虽然检测速度最快，但假阳性率稳定性较差；基于相似度的方法在复杂任务上的表现则不够理想（例如，在 EHRAgent 上的 AUROC 为 0.74）。

## 6. 结论

我们提出了*MemPot*，这是一种主动防御框架，通过利用蜜罐重塑检索分布、放大对抗信号，从而保护智能体记忆。*MemPot* 克服了被动检测的理论局限，在应对 IKEA 和 MEXTRA 等高级威胁时实现了近乎完美的准确率，同时不增加在线延迟，也不影响智能体效用。

## 影响声明

本文旨在提升大语言模型（LLM）智能体的安全性与隐私保护。我们的框架通过降低检索增强生成（RAG）和智能体记忆系统中的知识提取风险，助力保护专有知识产权和敏感用户数据，尤其是在医疗保健和金融等高风险领域。我们相信，这项工作能够确保记忆能力不会成为安全漏洞，从而推动可信赖的自主智能体部署。我们预计不会造成直接的负面社会影响，因为我们的主动防御机制旨在区分恶意探测与合法使用，最大限度地降低对正常服务造成干扰的风险。

## 参考文献


Abbasian, M., Azimi, I., Rahmani, A. M., 和 Jain, R. C. 对话式健康智能体：个性化大语言模型驱动的智能体框架。*CoRR*，abs/2310.02374，2023。doi: 10.48550/ARXIV.2310.02374。网址：https://doi.org/10.48550/arXiv.2310.02374。

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., 等。GPT-4 技术报告。*arXiv* 预印本 *arXiv:2303.08774*，2023。Agarwal, D., Fabbri, A. R., Risher, B., Laban, P., Joty, S., 和 Wu, C.-S. 多轮大语言模型应用中的提示泄露效应及缓解策略。载于*2024*年自然语言处理实证方法会议：工业界分会论文集，第 1255–1275 页，2024年。Anderson, M., Amit, G., 和 Goldsteen, A. 我的数据在你的检索数据库中吗？针对检索增强生成的成员推断攻击。*arXiv* 预印本 *arXiv:2405.20446*，2024年。asoria。Pokemoninfo 数据集，2024年。网址：https://huggingface.co/datasets/asoria/pokemon_信息。Cohen, S., Bitton, R., 和 Nassi, B. 释放蠕虫并提取数据：利用越狱扩大针对基于检索增强生成的推理攻击的规模和严重程度。*arXiv* 预印本 *arXiv:2409.08045*，2024年。Cover, T. 和 Thomas, J.信息论基础。Wiley，2012。ISBN 9781118585771。网址：https://books.google.com.sg/books?id=VWq5GG6ycxMC。

Cybenko, G. 通过叠加 sigmoid 函数进行逼近。控制、信号与系统数学，2(4):303–314，1989。Di Maio, C., Cosci, C., Maggini, M., Poggioni, V. 和 Melacci, S. RAG 海盗：自适应攻击大语言模型以泄露知识库。*arXiv* 预印本 *arXiv:2412.18295*，2024。Ding, H., Li, Y., Wang, J. 和 Chen, H. 金融交易中的大语言模型智能体：一项综述，2024。网址 https://arxiv.org/abs/2408.06361。Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.-S. 和 Li, Q. 关于 RAG 与大语言模型相结合的综述：迈向检索增强大语言模型。载于第*30*届 *ACM SIGKDD* 知识发现与数据挖掘会议论文集，第 6491–6501 页，2024。Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A. 等。Llama 3 模型家族。*arXiv* 电子预印本，第 arXiv–2407 页，2024。He, P., Gao, J. 和 Chen, W. DeBERTaV3：采用 ELECTRA 风格的预训练和梯度解耦嵌入共享改进 DeBERTa。*arXiv* 预印本 *arXiv:2111.09543*，2021。Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C. 和 Schmidhuber, J. MetaGPT：面向多智能体协作框架的元编程。载于第十二届国际学习表征会议，2024。网址 https://openreview.net/forum?id=VtmBAGCN7o。Hu, Y., Liu, S., Yue, Y., Zhang, G., Liu, B., Zhu, F., Lin, J., Guo, H., Dou, S., Xi, Z. 等。人工智能智能体时代的记忆。*arXiv* 预印本 *arXiv:2512.13564*，2025。Jiang, C., Pan, X., Hong, G., Bao, C. 和 Yang, M. RAG 窃取者：通过基于智能体的攻击，从检索增强生成应用中大规模提取私人数据。*arXiv* 预印本 *arXiv:2411.14110*，2024。Kagaya, T., Yuan, T. J., Lou, Y., Karlekar, J., Pranata, S., Kinose, A., Oguri, K., Wick, F. 和 You, Y. RAP：面向多模态大语言模型智能体的检索增强规划与上下文记忆。*CoRR*，abs/2402.03610，2024。doi: 10.48550/ARXIV.2402.03610。网址 https://doi.org/10.48550/arXiv.2402.03610。Kulkarni, M., Tangarajan, P., Kim, K. 和 Trivedi, A. 用于优化领域聊天机器人的强化学习 RAG。*CoRR*，abs/2401.06800，2024。doi: 10.48550/ARXIV.2401.06800。网址 https://doi.org/10.48550/arXiv.2401.06800。

lavita AI。lavita/chatdoctor-healthcaremagic-100k · Hugging Face 上的数据集。网址 https://huggingface.co/datasets/lavita/ChatDoctor-HealthCareMagic-100k。

Lewis, P.S. H., Perez, E., Piktus, A., Petroni, F.,Karpukhin, V., Goyal, N.,K¨uttler,H., Lewis,M.,Yih, W., Rockt¨aschel, T., Riedel, S., 和 Kiela, D. 面向知识密集型自然语言处理任务的检索增强生成。载于 Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., 和 Lin, H.（编），《神经信息处理系统进展》第 *33* 卷：神经信息处理系统 *2020* 年度会议，*NeurIPS 2020*，*2020*年*12*月*6*日至*12*日，线上，2020 年。网址 https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html。

Li, Y., Li, Z., Zhang, K., Dan, R. 和 Zhang, Y. Chatdoctor：基于医学领域知识在 Llama 模型上微调的医疗聊天模型。*CoRR*，abs/2303.14070，2023。doi：10.48550/ARXIV.2303.14070。网址https://doi.org/10.48550/arXiv.2303.14070。

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C. 等。DeepSeek-V3 技术报告。*arXiv* 预印本 *arXiv:2412.19437*，2024年。

Mao, J., Ye, J., Qian, Y., Pavone, M. 和 Wang, Y. 一种用于自动驾驶的语言智能体。*CoRR*，abs/2311.10813，2023。doi：10.48550/ARXIV. 2311.10813。网址 https://doi.org/10.48550/arXiv.2311.10813。

NomaSecurity。ForcedLeak：Salesforce Agentforce 中的智能体风险曝光。https://noma.security/blog/forcedleak-agent-risks-exposed-in-sales，2024年。访问日期：2026年1月6日。

OpenAI。GPT-4 技术报告。*CoRR*，abs/2303.08774，2023。doi：10.48550/arXiv.2303.08774。URLhttps://doi.org/10.48550/arXiv.2303.08774。

Qi、Z.、Zhang、H.、Xing、E. P.、Kakade、S. M. 和 Lakkaraju、H. 遵循我的指令并泄露秘密：从检索增强生成系统中进行可扩展的数据提取。在 *International Conference on Learning Representations*（*ICLR*），2025年。

Salesforce。Agentforce：AI智能体平台。https://www.salesforce.com/eu/agentforce/，2024年。访问日期：2026年1月6日。

Sason, I. 和 Verd´u, S. f-散度之间的界限。已提交至 *IEEE*信息论汇刊，2015。

Shi, W.、Xu, R.、Zhuang, Y.、Yu, Y.、Zhang, J.、Wu, H.、Zhu, Y.、Ho, J. C.、Yang, C. 和 Wang, M. D. Ehra-gent：代码赋能大语言模型，在电子健康记录上进行少样本复杂表格推理。载于 Al-Onaizan, Y.、Bansal, M. 和 Chen, Y.（编），*2024*年自然语言处理实证方法会议论文集，*EMNLP 2024*，迈阿密，美国佛罗里达州，*2024*年*11*月*12*日至*16*日，第22315–22339页。计算语言学协会，2024年。网址：https://aclanthology.org/2024.emnlp-main.1245。

Song, K., Tan, X., Qin, T., Lu, J., and Liu, T.-Y. MPNet：面向语言理解的掩码与置换预训练。《神经信息处理系统进展》，33:16857–16867, 2020。

Wald, A. 统计假设的序贯检验。收录于统计学的突破：基础与基本理论，第 256–298 页。施普林格，1992。

Wald, A. and Wolfowitz, J. 序贯概率比检验的最优特性。《数学统计年鉴》，第 326–339 页，1948。

Wang, B., He, W., Zeng, S., Xiang, Z., Xing, Y., Tang, J., and He, P. 揭示大语言模型智能体记忆中的隐私风险。收录于计算语言学协会第 *63* 届年会论文集（第 *1* 卷：长篇论文），第 25241–25260 页，2025a。

Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., 等。《基于大语言模型的自主智能体综述》。《计算机科学前沿》，18(6):186345, 2024年。

Wang, Y., Qu, W., Zhai, S., Jiang, Y., Liu, Z., Liu, Y., Dong, Y., and Zhang, J. 静默泄漏：通过良性查询对检索增强生成系统发起的隐式知识提取攻击，2025b。网址 https://arxiv.org/abs/2505.15420。

Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., 等。基于大语言模型的智能体的兴起与潜力：综述。*arXiv* 预印本 *arXiv:2309.07864*，2023。

Yao, H., Shi, H., Chen, Y., Jiang, Y., Wang, C., Qin, Z., Ren, K. 和 Chen, C. Controlnet：面向检索增强生成的大语言模型系统的防火墙。*arXiv* 预印本 *arXiv:2504.09593*，2025。

Yao, S., Chen, H., Yang, J. 和 Narasimhan, K. 《Web-shop：借助基于事实的语言智能体实现可扩展的真实世界网页交互》。载于 Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K. 和 Oh, A.（编），《神经信息处理系统进展》第*35*卷：*2022*年神经信息处理系统年会，*NeurIPS*

*2022*年，美国路易斯安那州新奥尔良，*11*月*28*日至*12*月*9*日，2022年。网址：http://papers.nips.cc/paper_files/paper/2022/hash/82ad13ec01f9fe44c01cb91814fd7b8c-Abstract-Conference.html。

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., and Cao, Y. ReAct：协同语言模型中的推理与行动。发表于第十一届国际学习表征会议，*ICLR 2023*，基加利，卢旺达，*2023*年*5*月*1*日至*5*日。OpenReview.net，2023年。网址 https://openreview.net/forum?id=WE_vluYUL-X。

Zeng, S., Zhang, J., He, P., Liu, Y., Xing, Y., Xu, H., Ren, J., Chang, Y., Wang, S., Yin, D., and Tang, J. 好与坏：探索检索增强生成（RAG）中的隐私问题。发表于计算语言学协会论文集：*ACL 2024*，第 4505–4524 页，2024a。

Zeng, S., Zhang, J., He, P., Ren, J., Zheng, T., Lu, H., Xu, H., Liu, H., Xing, Y., 和 Tang, J. 通过纯合成数据缓解检索增强生成（RAG）中的隐私问题。*arXiv* 预印本 *arXiv:2406.14773*，2025。

Zeng, Y., Wu, Y., Zhang, X., Wang, H., 和 Wu, Q. Autode- fense：用于防御越狱攻击的多智能体大语言模型。*arXiv* 预印本 *arXiv:2403.04783*，2024b。

Zhang, C., Zhang, T., 和 Shmatikov, V. 对抗解码：为对抗目标生成可读文档，2025年。网址 https://arxiv.org/abs/2410.02163。

Zhang, Y., Ding, L., Zhang, L., 和 Tao, D. 意图分析使大语言模型成为优秀的越狱防御者。*arXiv* 预印本 *arXiv:2401.06561*，2024a。

Zhang, Z., Bo, X., Ma, C., Li, R., Chen, X., Dai, Q., Zhu, J., Dong, Z. 和 Wen, J. 关于基于大语言模型的智能体记忆机制的综述。*CoRR*，abs/2404.13501，2024年b。doi：10.48550/ARXIV. 2404.13501。URLhttps://doi.org/10.48550/arXiv.2404.13501。

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., 等。《大语言模型综述》。*arXiv* 预印本 *arXiv:2303.18223*，2023。

## A. 蜜罐优化的详细信息

**攻击者代理**。在实际场景中，攻击者的详细算法和目标通常无法获取，因此难以准确获得攻击查询分布 $\mathcal{Q}_1$。为此，我们构建了一个攻击者代理，使其能够在句子嵌入层面学习并复现攻击者的行为。该代理由一个神经网络 $A_\omega$ 实现；给定历史交互轨迹 $\{o_{1:t-1}\}$——包括先前的查询嵌入、检索到的文档索引和相似度分数——它会迭代输出下一个查询 $\hat{e}_t = A_\omega(o_{1:t-1})$ 的嵌入。我们的目标并非重建攻击者的内部算法，而是让 $A_\omega$ 模仿攻击者可观察到的嵌入层面动态。训练目标通过余弦相似度回归，最小化生成嵌入与观测嵌入之间的差异，从而确保该代理在相同检索器下复现相似的检索排名。这种设计使蜜罐系统能够在无法明确获取攻击者模型参数的情况下训练防御策略。

**用户代理**。 在正常用户建模中，我们面临类似的挑战：真实世界的用户查询分布难以收集，且往往高度多样。为了对其进行近似，我们利用大语言模型（大语言模型）的人类模拟器能力来合成真实的用户行为。具体而言，我们采用人类模拟流程，以交互意图和检索上下文提示大语言模型，生成类似于真实信息寻求行为的自然多轮查询序列。随后，这些合成用户查询被编码为嵌入，以近似良性查询分布 $\mathcal{Q}_0$。

## B. 实验设置的详细信息

### B.1. 设置

数据集与评估设置。我们在四个多样化基准上评估MemPot，涵盖外部记忆检索与内部记忆交互两类场景：对于外部记忆，我们使用HealthMagicCare（20k行）（lavita AI）和Pok´emon（9.46k行）（asoria，2024年）数据集，分别代表隐私密集型医疗保健问询和实体密集的游戏知识。实验中我们默认每轮检索4篇文档。对于内部记忆，遵循MEXTRA（Wang 等，2025a）中的设置，我们使用为EHRAgent（Shi等人，2024年）以及WebShop上的RAP（Yao等，2022；Kagaya 等人，2024年）生成的内部记忆日志。标准设置包含300条交互记录。我们遵循原始论文（Shi等人，2024年；Kagaya 等人，2024年）中相同的检索数量设置（EHRAgent为top-4条记录，RAP WebShop为top-3条记录）。为进行可扩展性分析（附录. C.1），我们通过追加额外200条记录来扩展该设置。我们采用DeepSeek-v3.2（685B）作为两类智能体的骨干模型，并采用all-mpnet-base-v2作为检索嵌入模型。我们使用mDeBERTa-v3-base（He等人，2021）作为自然语言推理模型。对于每个数据集/智能体，我们针对不同攻击使用相同的蜜罐集。默认情况下，我们对外记忆数据集以2%的比例、对内部记忆数据集以4%的比例注入蜜罐（我们在附录. C.2 中评估蜜罐比例的影响）。对于 SPRT 检测器，我们采用蜜罐 - 非蜜罐计数比（式 17 ）作为对数似然比估计的默认方法，并将第一类/第二类错误预算 $\alpha, \beta$ 设为0.1。所有实验均在单块NVIDIA RTX 5090 GPU上进行。

**训练集配置**。对于每种攻击类型，我们使用 256 个样本训练一个基于 Transformer 的攻击者代理。为稳健地捕捉提取攻击的探索特性，我们采用相似度 top-k 随机游走扩充训练数据，通过迭代地从当前检索结果的语义邻居中采样下一个查询，模拟攻击者的轨迹。蜜罐通过与该代理交互进行训练，同时针对每个目标数据集的 500 条大语言模型生成的良性人工查询优化对比损失。为支持蜜罐在大规模知识库上进行训练，我们首先应用均衡 k-means 聚类，将文档语料库划分为大小相等的集合。每个集合中的文档嵌入被输入共享网络，以生成对应的蜜罐嵌入，从而显著降低生成所需的计算负担。

**攻击与防御基线**。 我们考虑以下攻击：对于外部记忆，我们评估面向 RAG 的自适应攻击，包括 RAG-Thief（Jiang 等人，2024年）、DGEA（Cohen 等人，2024年）和 IKEA（Wang 等人，2025b）。对于内部记忆，我们评估面向智能体的 MEXTRA（Wang 等人，2025a）。此外，我们将 IKEA 改用于内部记忆提取，以评估其抵御自适应攻击的鲁棒性。我们广泛考察现有防御方法，包括 ControlNet（Yao 等人，2025）和智能体检测器（Zhang 等人，2024a；Zeng 等人，2024b；Agarwal 等人，2024年）。具体而言，我们使用 100 个良性锚点查询来初始化 ControlNet，并使用 DeepSeek-v3.2（685B）作为智能体检测器的骨干模型（见附录 B.2）。此外，我们设计了一个基于 Transformer 的理论最优序贯检测器（Optimal-Seq），用于进一步比较和验证定理2。

表 8. 性能评估 EHRAgent 和 Web-Shopping RAP 智能体上堆叠式 MemPot 与端到端训练的 MemPot 的性能。

| 攻击 | 防御 | EHRAgent | | | | | RAP-web | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | TP@ 1%FP | TP@ 10%FP | 延迟 | FDT | AUROC | TP@ 1%FP | TP@ 10%FP | 延迟 | FDT |
| $\text{MEXTRA}_{\text{Cosine}}$ | Optimal-Seq | 0.70 | 0.16 | 0.27 | 0.03 | 22 | 0.68 | 0.06 | 0.22 | 0.04 | 38 |
| | MemPot-E2E | **0.99** | **0.67** | 0.74 | **0** | **2** | 0.98 | **1.00** | **1.00** | **0** | **1** |
| | MemPot-Stack | 0.98 | 0.65 | **0.75** | **0** | 3 | **0.99** | **1.00** | **1.00** | **0** | **1** |
| $\text{MEXTRA}_{\text{Edit}}$ | Optimal-Seq | 0.71 | 0.14 | 0.32 | 0.04 | 33 | 0.75 | 0.20 | 0.36 | 0.04 | 27 |
| | MemPot-E2E | **0.98** | 0.56 | **0.74** | **0** | **4** | **1.00** | **1.00** | **1.00** | **0** | **1** |
| | MemPot-Stack | 0.97 | **0.58** | 0.72 | **0** | 9 | **1.00** | **1.00** | **1.00** | **0** | **1** |
| $\text{MEXTRA}_{\text{General}}$ | Optimal-Seq | 0.69 | 0.16 | 0.28 | 0.04 | 42 | 0.73 | 0.12 | 0.26 | 0.04 | 35 |
| | MemPot-E2E | 0.94 | 0.81 | 0.85 | **0** | 7 | **1.00** | **1.00** | **1.00** | **0** | **1** |
| | MemPot-Stack | **0.98** | **0.82** | **0.88** | **0** | **5** | **1.00** | **1.00** | **1.00** | **0** | **1** |

### B.2. 智能体检测器设置

参考（Zeng 等人，2024a；Jiang 等人，2024年；Anderson 等人，2024；Zhang 等人，2024a；Zeng 等人，2024b）中的缓解建议，我们采用混合范式的智能体检测器，包括意图检测、关键词检测和防御指令。具体而言，我们使用 DeepSeek-v3.2（685B）作为智能体检测器的骨干模型。与检测器集成的响应生成过程如下所示：对于输入查询 $q$，首先通过意图检测（Zhang 等人，2024a）和关键词过滤（Zeng 等人，2024a）进行防御：

$$q_{\text{defended}} = \begin{cases} \emptyset, & D_{\text{intent}}(q) \vee D_{\text{keyword}}(q) = 1 \\ q, & \text{otherwise} \end{cases}, \tag{20}$$

其中，$\emptyset$ 会强制给出“无法回答”的响应；$D_{\text{intent}}(\cdot)$ 和 $D_{\text{keyword}}(\cdot)$ 是检测函数，在检测到恶意提取意图或相关词语时会返回 True。当 $q_{\text{defended}} \neq \emptyset$ 时，生成过程会结合检索上下文 $\mathcal{D}_q^K$，如下所示：

$$y = \text{LLM}\big(\text{Concat}(\mathcal{D}_q^K) \oplus q_{\text{defended}} \oplus p_{\text{defense}}\big), \tag{21}$$

其中，防御提示 $p_{\text{defense}}$ (Agarwal 等人，2024年) 通过提示大语言模型仅使用检索结果中的相关部分作答来约束输出的相关性，并通过提供示例，强制大语言模型不响应恶意指令。

## C. 补充实验

### C.1. 内存更新下 MemPot 的可扩展性

为应对智能体系统中记忆实时更新的动态特性，我们采用基于批次的更新策略评估*MemPot* 的可扩展性。鉴于检索具有局部性，我们假设可以针对不同内存批次独立优化蜜罐，并将其直接合并为统一防御集。

**设置**。 我们通过向标准的 300 条记录内部记忆中追加 200 条新的交互记录（来源于 MEXTRA 数据集（Wang 等人，2025a）），模拟记忆更新场景，最终得到共 500 条记录。我们比较两种实现策略：（1）**MemPot-E2E**：计算开销较大的上界策略，在完整的 500 条记录数据集上从头开始全局重新优化蜜罐。（2）**MemPot-Stack**：可扩展的方法，保留初始 300 条记录对应的蜜罐，并仅追加一组专门针对 200 条更新记录优化的新蜜罐。为确保比较公平，两种策略的蜜罐比例均保持在 4%。

**结果**。如表8所示，*MemPot-Stack* 在所有攻击向量上的检测性能几乎与整体式 *MemPot-e2e*完全相同。值得注意的是，在 EHRAgent 上的 MEXTRA-General 等具有挑战性的场景中，堆叠式方法保持了更强的鲁棒性（AUROC 为 0.98），优于端到端基线（AUROC 为 0.94）。这些结果证实，*MemPot* 支持高效的模块化更新：随着智能体的记忆不断增长，可以无缝集成新的蜜罐，而无需进行全局重新训练，从而以极低的计算开销持续提供保护。

表9. 消融研究：关于 蜜罐（蜜罐）数量及其对$\text{MEXTRA}_{\text{cosine}}$在RAP WebShop上检测性能的影响 .

| 比例 | AUROC | 1%假阳性率下的真阳性率 | 10%假阳性率下的真阳性率 | FDT |
|---|---|---|---|---|
| 0.6% | **1.00** | 0.77 | **1.00** | 7 |
| 1.0% | **1.00** | 0.79 | **1.00** | 4 |
| 2.0% | **1.00** | 0.90 | **1.00** | **2** |
| 4.0% | **1.00** | **0.94** | **1.00** | **2** |

### C.2. 蜜罐数量的影响

我们研究了注入蜜罐的比例与检测鲁棒性之间的权衡。表9 报告了 RAP WebShop 上的检测性能；随着蜜罐比例从 0.6% 增至 4%，记忆记录数量为 300。即使注入量极少，AUROC 仍保持在 1.00；而增加蜜罐数量则显著提升了严格阈值下的检测速度和灵敏度。具体而言，将蜜罐比例从 0.6% 提高到 4%，可使 1%假阳性率下的真阳性率从 0.77 提升至 0.94，并将首次检测时间（FDT）从 7 轮缩短至仅 2 轮。这一趋势证实，更密集的蜜罐分布会放大对抗信号，从而更快地拦截攻击；不过，较小的预算（例如 2%）已经能够实现接近最优的性能。

## D. 理论预备知识

**定义 1 (**固定参数划分)。 设 $T_\phi: \mathcal{O} \to \mathcal{Y}$为一种固定形式的参数划分（其形式 在训练期间保持不变；只有参数 $\phi$会发生变化）。定义推前测度

$$P_{1,\theta,\phi} = T_{\phi\#} f_{1,\theta}, \qquad P_{0,\theta,\phi} = T_{\phi\#} f_{0,\theta},$$

以及散度 $\Phi_T(\theta, \phi) = \mathrm{KL}(P_{1,\theta,\phi} \| P_{0,\theta,\phi})$。

**引理 1 (**数据处理不等式（DPI）)。 对于任意可测 $T_\phi$,

$$\mathrm{KL}(f_{1,\theta} \| f_{0,\theta}) \geq \mathrm{KL}(T_{\phi\#} f_{1,\theta} \| T_{\phi\#} f_{0,\theta}) = \Phi_T(\theta, \phi).$$

同理，$\mathrm{KL}(f_{0,\theta} \| f_{1,\theta}) \geq \mathrm{KL}(P_{0,\theta,\phi} \| P_{1,\theta,\phi})$。

证明。 经典的数据处理不等式；参见 引理证明的第3步。 □

**引理 2** （有限划分上的条件期望）。 设$(\Omega\ \mathcal{F}, g)$ 是一个概率空间，设 $\mathcal{P} = \{A_1, \ldots, A_m\}$ 是 $\Omega$ 上的有限可测划分，并记 $\sigma(\mathcal{P})$ 为其生成的 $\sigma$-代数。对于任意 $X \in L^1(g)$，定义

$$Y(x) := \sum_{i=1}^{m} \mathbf{1}_{A_i}(x) \begin{cases} \dfrac{1}{g(A_i)} \displaystyle\int_{A_i} X\, dg, & g(A_i) > 0, \\ 0, & g(A_i) = 0. \end{cases}$$

那么：

1. $Y$是 $\sigma(\mathcal{P})$-可测的，并且对于每个 $B \in \sigma(\mathcal{P})$, $\displaystyle\int_B Y\, dg = \int_B X\, dg$。因此 $Y = \mathbb{E}_g[X \mid \sigma(\mathcal{P})]$ 几乎处处成立。

2. 特别地，将 $\mathbb{E}_g[X \mid \mathcal{P}]$ 简记为 $\mathbb{E}_g[X \mid \sigma(\mathcal{P})]$,

$$\int \mathbb{E}_g[X \mid \mathcal{P}]\, dg \;=\; \int X\, dg. \tag{22}$$

证明。(1) 由 co nstru 因此， $Y$在每个原子 $A_i$上都是常数，故为 $\sigma(\mathcal{P})$-可测。若 $B = \bigcup_{i \in I} A_i \in \sigma(\mathcal{P})$，则

$$\begin{aligned} \int_B Y\, dg &= \sum_{i \in I} \frac{1}{g(A_i)} \left( \int_{A_i} X\, dg \right) g(A_i) \\ &= \sum_{i \in I} \int_{A_i} X\, dg = \int_B X\, dg, \end{aligned}$$

这正是条件期望 $\mathbb{E}_g[X \mid \sigma(\mathcal{P})]$的定义性质。除去 $g$-零测集外的唯一性意味着$Y = \mathbb{E}_g[X \mid \sigma(\mathcal{P})]$几乎处处成立。

(2) 将 $B = \Omega$ 代入第 (1) 部分的恒等式，得到 (22)。 □

**引理 3** （划分上确界）。对于 $\mathcal{O}$ 上任意一对概率测度 $f$、$g$，

$$\mathrm{KL}(f\|g) = \sup_{T} \mathrm{KL}(T_{\#}f\|T_{\#}g),$$

其中，上确界取遍所有有限可测划分（等价地，取遍有限值可测映射 $T$）；此外，对于任意 $\delta > 0$，存在一个有限划分 $T_\delta$，使得

$$\mathrm{KL}(f\|g) \le \mathrm{KL}(T_{\delta\#}f\|T_{\delta\#}g) + \delta.$$

证明。若 $f \not\ll g$，则存在 $A \in \mathcal{F}$，使得 $g(A) = 0$ 且 $f(A) > 0$；对于二原子划分 $\{A, A^c\}$，有 $\mathrm{KL}(T_{\#}f\|T_{\#}g) = +\infty = \mathrm{KL}(f\|g)$，结论显然成立。因此，假设 $f \ll g$。令 $L := \frac{df}{dg}$ 和 $\ell := \log L$，根据拉东-尼科迪姆定理，它们均有定义。于是 $\mathrm{KL}(f\|g) = \int L \log L \, dg = \int \ell \, df \in [0, \infty]$。

步骤 *1*（数据处理不等式）。
令 $\mathcal{P} = \{A_i\}_{i=1}^m$ 为有限划分，并令 $T$ 为其指标映射。记 $p_i = f(A_i) = \int_{A_i} L \, dg$ 和 $q_i = g(A_i)$（约定 $0 \log 0 := 0$）。则

$$\begin{aligned}
\mathrm{KL}(T_{\#}f\|T_{\#}g) &= \sum_{i=1}^m p_i \log \frac{p_i}{q_i} \\
&= \sum_{i=1}^m \Big( \int_{A_i} L \, dg \Big) \log \frac{\int_{A_i} L \, dg}{g(A_i)} \\
&= \int \mathbb{E}_g[L \mid \mathcal{P}] \, \log \mathbb{E}_g[L \mid \mathcal{P}] \, dg,
\end{aligned}$$

其中 $\mathbb{E}_g[\cdot \mid \mathcal{P}]$ 是关于 $g$ 到由 $\sigma$生成的 $\mathcal{P}$-代数上的条件期望。由于 $\varphi(u) := u \log u$ 在$(0, \infty)$上是凸函数，由詹森不等式可得 $\varphi(\mathbb{E}_g[L \mid \mathcal{P}]) \le \mathbb{E}_g[\varphi(L) \mid \mathcal{P}]$，结合引理 2 积分后得到

$$\begin{aligned}
\mathrm{KL}(T_{\#}f\|T_{\#}g) &= \int \varphi(\mathbb{E}_g[L \mid \mathcal{P}]) \, dg \\
&\le \int \mathbb{E}_g[\varphi(L) \mid \mathcal{P}] \, dg \\
&= \int \varphi(L) \, dg = \mathrm{KL}(f\|g).
\end{aligned}$$

然后对所有有限 $\mathcal{P}$ 取上确界，可得 $\sup_T \mathrm{KL}(T_{\#}f\|T_{\#}g) \le \mathrm{KL}(f\|g)$。

步骤 *2*（有限划分下散度差距的局限性）。
固定 $\delta > 0$ 并假设 KL散度$(f\|g) < \infty$。我们构造一个有限划分 $\mathcal{P}_\delta$，使得 KL散度$(f\|g)$ $-$KL散度$(T_{\delta\#}f\|T_{\delta\#}g) \le \delta$ .

*(2.a)* 尾部控制。
选择足够大的 $M \ge 1$，使得“上尾”部分的贡献满足

$$\int_{\{L \ge e^M\}} L \log L \, dg \;\le\; \delta/3,$$

a此外 $2Me^{-M} \le \delta/3$（这在 $Me^{-M} \to 0$作为 $M \to \infty$ 时是可能的）。定义三个区域 s

$$A_- := \{L < e^{-M}\}, B := \{e^{-M} \le L < e^M\}, A_+ := \{L \ge e^M\}.$$

（*2.b*）中间量化。

选取网格大小 $\eta \in (0,1)$，具体取值稍后指定（下文取 $\eta := \delta/3$）。将中间区域 $B$ 划分为有限个 $L$的水平集：对于 $k = 0, 1, \ldots, K-1$，令 $K := \lceil 2M/\eta \rceil$，设

$$B_k := \{\, x \in \mathcal{O} :\ e^{-M+k\eta} \le L(x) < e^{-M+(k+1)\eta} \,\}.$$

于是，在每个 $B_k$ 上，我们有 $\log L \in [a_k, a_k + \eta]$，其中 $a_k := -M + k\eta$，并且

$$\log \mathbb{E}_g[L \mid B_k] = \log \frac{\int_{B_k} L\, dg}{g(B_k)} \in [a_k, a_k + \eta]$$
$$\Rightarrow \big|\log L - \log \mathbb{E}_g[L \mid B_k]\big| \le \eta \ \text{ on } B_k.$$

*(2.c)* 有限划分。

令 $\mathcal{P}_\delta := \{A_-, B_0, \ldots, B_{K-1}, A_+\}$，并令 $T_\delta$ 为其指标映射。利用步骤 1 中的恒等式，并将差距写成一个$L\, dg$-积分，我们有

$$\mathrm{KL}(f\|g) - \mathrm{KL}(T_{\delta\#}f\|T_{\delta\#}g) = \sum_{C\in\mathcal{P}_\delta} \int_C L\Big(\log L - \log \mathbb{E}_g[L \mid C]\Big)\, dg =: \Delta_- \ + \ \Delta_B \ + \ \Delta_+.$$

*(2.d)* 中间差距的界。

在每个 $B_k$上，逐点界 $|\log L - \log \mathbb{E}_g[L \mid B_k]| \le \eta$给出

$$\Delta_B = \sum_{k=0}^{K-1} \int_{B_k} L\big(\log L - \log \mathbb{E}_g[L \mid B_k]\big)\, dg \ \le \ \eta \sum_{k=0}^{K-1} \int_{B_k} L\, dg \ = \ \eta\, f(B) \ \le \ \eta.$$

*(2.e)* 界定下尾差距。

在 $A_-$ 上有 $L \le e^{-M}$。利用 $|\log L - \log \mathbb{E}_g[L \mid A_-]| \le |\log L| + |\log \mathbb{E}_g[L \mid A_-]|$ and $\mathbb{E}_g[L \mid A_-] \le e^{-M} \Rightarrow |\log \mathbb{E}_g[L \mid A_-]| \le -M$，我们得到

$$\Delta_- \le \int_{A_-} L|\log L|\, dg + M\int_{A_-} L\, dg \le Me^{-M} + Me^{-M} = 2Me^{-M} \ \le \ \delta/3,$$

根据 $M$ 的选取。

*(2.f)* 界定上　　尾差距。在 $A_+$ 上有 $\mathbb{E}_g[L \mid A_+] \ge e^M$，因此 $\log L - \log \mathbb{E}_g[L \mid A_+] \le \log L - M$。因此　　e

$$0 \le \Delta_+ = \int_{A_+} L\big(\log L - \log \mathbb{E}_g[L \mid A_+]\big)\, dg \le \int_{A_+} L(\log L - M)\, dg \le \int_{A_+} L \log L\, dg \ \le \ \delta/3,$$

根据 $M$ 的选取。

（*2.g*）完全控制。

令 $\eta$ 为$= \delta/3$。汇总（2.d）–（2.f）中的界可得

$$\mathrm{KL}(f\|g) - \mathrm{KL}(T_{\delta\#}f\|T_{\delta\#}g) \le \delta/3 + \delta/3 + \delta/3 = \delta.$$

步骤 *3*（取上确界）。

根据步骤 1，对于每个有限划分，$\mathrm{KL}(T_\#f\|T_\#g) \le \mathrm{KL}(f\|g)$。根据步骤 2，对于每个 $\delta > 0$，都存在一个有限划分$T_\delta$，使得 $\mathrm{KL}(f\|g) \le \mathrm{KL}(T_{\delta\#}f\|T_{\delta\#}g) + \delta$。因此，对于所有 $\delta > 0$，$\sup_T \mathrm{KL}(T_\#f\|T_\#g) \ge \mathrm{KL}(f\|g) - \delta$，进而 $\sup_T \mathrm{KL}(T_\#f\|T_\#g) = \mathrm{KL}(f\|g)$。 □

**引理 4**（交叉熵支配贝叶斯风险）。设 $J \in \{1,\dots,K\}$ 和 $\mathbf{Y} = (Y_1,\dots,Y_K)$ 按照1-正$(K-1)$-负方案生成，并令 $\pi^*(j \mid \mathbf{Y})$表示给定 $\mathbf{Y}$ 时 $J$ 的真实后验分布。对于任意可测评分函数$h : \mathcal{Y} \to \mathbb{R}$，定义模型后验分布 $\pi_h(j \mid \mathbf{Y}) := \frac{e^{h(Y_j)}}{\sum_{i=1}^{K} e^{h(Y_i)}}$ 以及 $K$ 样本 *InfoNCE* 损失

$$\mathcal{L}_{\mathrm{NCE},K}(h) := \mathbb{E}[-\log \pi_h(J \mid \mathbf{Y})]\,.$$

*Then*

$$\mathcal{L}_{\mathrm{NCE},K}(h) \;\geq\; \mathbb{E}_{\mathbf{Y}}\big[H(\pi^*(\cdot \mid \mathbf{Y}))\big] \quad\Longleftrightarrow\quad -\mathcal{L}_{\mathrm{NCE},K}(h) \;\leq\; -\mathcal{L}^*, \tag{23}$$

其中 $H(p) := -\sum_j p(j)$ 的对数$p(j)$，并且 $\mathcal{L}^* := \mathbb{E}_{\mathbf{Y}}\big[H(\pi^*(\cdot \mid \mathbf{Y}))\big]$是采用对数损失时的贝叶斯风险。当且仅当 $\pi_h(\cdot \mid \mathbf{Y}) = \pi^*(\cdot \mid \mathbf{Y})$几乎处处成立时，等号成立。

证明。固定$\mathbf{Y}$，根据 标准分解 $H(p,\ q) = H(p) + \mathrm{KL}$散度$(p\|q)$，$\pi^*$ 与 $\pi_h$ i之间的交叉熵 s

$$H(\pi^*, \pi_h) = \mathbb{E}_{J\sim\pi^*(\cdot|\mathbf{Y})}[-\log \pi_h(J \mid \mathbf{Y})] = H(\pi^*) + \mathrm{KL}\big(\pi^*(\cdot \mid \mathbf{Y}) \,\|\, \pi_h(\cdot \mid \mathbf{Y})\big).$$

由于KL散度$(\cdot\|\cdot) \geq 0$（吉布斯不等式（Cover 和 Thomas，2012）），我们有 $H(\pi^*,\ \pi_h) \geq H(\pi^*)$且当且仅当 $\pi_h = \pi^*$时等号成立。对$\mathbf{Y}$ 取期望，得到

$$\mathcal{L}_{\mathrm{NCE},K}(h) = \mathbb{E}_{\mathbf{Y}}\big[H(\pi^*, \pi_h)\big] \geq \mathbb{E}_{\mathbf{Y}}\big[H(\pi^*)\big] = \mathcal{L}^*,$$

这就是（23）。 □

**引理 5**（均匀索引下的条件乘积密度）。设 $P$、$Q$ 是（$\mathcal{Y}$，$\mathcal{G}$）上的概率测度，满足 $P \ll Q$，并令 $r := \frac{dP}{dQ}$。固定 $K \geq 2$。抽取 $J \sim$均匀分布$\{1,\dots,K\}$，并在给定 $J = j$的条件下，抽取$\mathbf{Y} = (Y_1, \dots\dots, Y_K)$，其坐标相互独立，分别服从 $Y_j \sim P$和 $Y_i \sim Q$，其中 $i \neq j$。则：

1. 对每个 $j$，条件分布 $\mathbb{P}(\mathbf{Y} \mid J = j)$ 关于 $Q^{\otimes K}$ 绝对连续，并具有拉东–尼科迪姆导数

$$\frac{d\,\mathbb{P}(\mathbf{Y} \mid J = j)}{d\,Q^{\otimes K}}(\mathbf{y}) \;=\; r(y_j) \qquad (Q^{\otimes K}\text{-a.e.}).$$

2. $\mathbf{Y}$ 的边缘分布满足

$$\frac{d\,\mathbb{P}_{\mathbf{Y}}}{d\,Q^{\otimes K}}(\mathbf{y}) \;=\; \frac{1}{K}\sum_{i=1}^{K} r(y_i) \qquad (Q^{\otimes K}\text{-a.e.}).$$

证明。 我们使用拉东–尼科迪姆导数的测试函数刻画。

（1）条件分布。 固定 $j \in \{1,\dots,K\}$ 以及任意有界可测的 $\varphi : \mathcal{Y}^K \to \mathbb{R}$。根据给定 $J = j$时 $\mathbf{Y}$ 的构造以及各坐标的独立性，

$$\int \varphi(\mathbf{y})\, d\mathbb{P}(\mathbf{Y} \mid J = j) = \int \varphi(\mathbf{y})\, d\big(Q^{\otimes(j-1)} \otimes P \otimes Q^{\otimes(K-j)}\big)(\mathbf{y}).$$

由于 $P \ll Q$ 具有密度 $r = \frac{dP}{dQ}$，且 $Q \ll Q$具有密度 1，因此在 $\mathbf{y}$上关于 $Q^{\otimes K}$ 几乎处处，

$$\frac{d\big(Q^{\otimes(j-1)} \otimes P \otimes Q^{\otimes(K-j)}\big)}{d(Q^{\otimes K})}(\mathbf{y}) = r(y_j) \cdot \prod_{i\neq j} 1 = r(y_j),$$

从而

$$\int \varphi(\mathbf{y})\, d\mathbb{P}(\mathbf{Y} \mid J = j) = \int \varphi(\mathbf{y})\, r(y_j)\, d(Q^{\otimes K})(\mathbf{y}).$$

根据拉东–尼科迪姆定理的唯一性，这可以确定 $\frac{d\,\mathbb{P}(\mathbf{Y}|J=j)}{d\,Q^{\otimes K}} = r(y_j)$ 几乎处处成立。

*(2)* 边缘分布。对均匀分布 $J$取平均，

$$\int \varphi(\mathbf{y})\, d\mathbb{P}_{\mathbf{Y}} = \sum_{j=1}^{K} \mathbf{1}(J=j) \int \varphi(\mathbf{y})\, d\mathbb{P}(\mathbf{Y} \mid J=j) = \frac{1}{K} \sum_{j=1}^{K} \int \varphi(\mathbf{y})\, r(y_j)\, d(Q^{\otimes K})(\mathbf{y}).$$

因此，再由拉东–尼科迪姆唯一性，$\dfrac{d\,\mathbb{P}_{\mathbf{Y}}}{d\,Q^{\otimes K}}(\mathbf{y}) = \dfrac{1}{K} \sum_{j=1}^{K} r(y_j)$ $Q^{\otimes K}$。 □

**引理 6 (**马尔可夫观测下序贯概率比检验的平均样本量近似)。 考虑边界为 $A < 0 < B$、错误概率预算为$(\alpha, \beta)$的序贯概率比检验。假设观测序列 $\{O_t\}_{t=1}^{\infty}$ 在每个假设下均为平稳遍历马尔可夫链，其转移密度为 $f_{1,\theta}$ 和 $f_{0,\theta}$。令 $\ell_\theta(O_t|O_{t-1}) = \log f_{1,\theta(O_t|O_{t-1})}\, f_{0,\theta(O_t|O_{t-1})}$ 为单步对数似然比。分别将 $H_1$ 和 $H_0$ 下的漂移率定义为

$$\mu_1(\theta) = \mathbb{E}_{1,\theta}[\ell_\theta(O_t|O_{t-1})], \qquad \mu_0(\theta) = \mathbb{E}_{0,\theta}[\ell_\theta(O_t|O_{t-1})].$$

若 停止时的超调可以忽略不计，则期望停止时间（平均样本量）近似为

$$\mathbb{E}_1[N] \approx \frac{|\log B|}{\mu_1(\theta)}, \qquad \mathbb{E}_0[N] \approx \frac{|\log A|}{|\mu_0(\theta)|}. \tag{24}$$

证明。 我们在平稳遍历马尔可夫观测的假设下给出详细证明。

设 $\{O_t\}_{t\geq 0}$ 是定义在状态空间 $\mathcal{O}$ 上的马尔可夫链。在假设 $H_i$ （$i = 0,\ 1$）下，该链的转移密度为$f_{i,\ \theta}(\cdot|O_{t-1})$，且具有唯一的平稳分布 $\pi_i$。我们假设该链从其平稳分布开始（或从任意初始分布开始；由于遍历性，长期行为相同）。单步对数似然比为

$$Z_t \equiv \ell_\theta(O_t|O_{t-1}) = \log \frac{f_{1,\theta}(O_t|O_{t-1})}{f_{0,\theta}(O_t|O_{t-1})},$$

累积和为 $S_n = \sum_{t=1}^{n} Z_t$。停止时间为

$$N = \inf\{n \geq 1 : S_n \geq \log B \text{ or } S_n \leq \log A\},$$

其中边界满足 $A = \frac{\beta}{1-\alpha}$ 和 $B = \frac{1-\beta}{\alpha}$，给定错误概率 $\alpha$，$\beta$ （瓦尔德近似）。根据平稳性，

$$\mu_1(\theta) = \mathbb{E}_{1,\theta}[Z_t] = \iint \pi_1(dx) f_{1,\theta}(dy|x) \log \frac{f_{1,\theta}(dy|x)}{f_{0,\theta}(dy|x)} = \mathrm{KL}\big(f_{1,\theta}(\cdot|O) \| f_{0,\theta}(\cdot|O)\big),$$

其中，最后一个等式表示在平稳分布 $\pi_1$ 下的平均库尔贝克–莱布勒散度。同样，

$$\mu_0(\theta) = \mathbb{E}_{0,\theta}[Z_t] = -\mathrm{KL}\big(f_{0,\theta}(\cdot|O) \| f_{1,\theta}(\cdot|O)\big) < 0.$$

由于该链在每个假设下均具有遍历性，马尔可夫链的强大数定律给出

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{a.s.}} \mu_1(\theta) \quad \text{under } H_1,$$

and

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{\text{a.s.}} \mu_0(\theta) \quad \text{under } H_0.$$

在 $H_1$ 下，过程 $S_n$ 以漂移 $\mu_1(\theta)$ 近似线性增长。忽略 $S_N$ 首次越过边界时的超调，到达上界 $\log B$ 所需的时间大约为

$$N \approx \frac{\log B}{\mu_1(\theta)}.$$

对等式两边取期望可得 $\mathbb{E}_1[N] \approx {}^{|}{}_{\log \mu_1}\ {}^{B|}_{\theta}{}_{()}$（注意 $\log\ B > 0$）。

在 $H_0$ 下，过程向下漂移 ，斜率为 $\mu_0(\theta) < 0$。到达下边界 $\log\ A$（其中 $\log A < 0$）所需的时间近似为

$$N \approx \frac{\log A}{\mu_0(\theta)} = \frac{|\log A|}{|\mu_0(\theta)|},$$

因此 $\mathbb{E}_0[N] \approx {}^{|}{}_{\log|\mu_0}\ {}^{A|}_{\theta}{}_{|()}$。

因此，在平稳性、遍历性和超调可忽略的条件下，期望样本量可由式（24）中的表达式很好地近似。

□

备注 1。 漂移率 $\mu_1(\theta)$ 和 $\mu_0(\theta)$ 恰好是转移律之间的平稳库尔贝克–莱布勒散度。对于独立同分布观测，马尔可夫依赖消失，该引理便简化为 Wald 的经典平均样本量公式。在实践中，当边界与起始点相距足够远时（即 $\alpha$ 和 $\beta$较小时），这一近似是准确的。

## E. InfoNCE上界的证明

**定理3** (信息漂移给出的InfoNCE上界)。 令 $P := T_{\phi\#} f_{1,\theta}$和 $Q := T_{\phi\#} f_{0,\theta}$ 定义在 $\mathcal{Y}$上。考虑以下抽样方案：抽取 $J \sim$均匀分布$\{1, \ldots, K\}$，然后抽取 $Y_J \sim P$和 $(Y_i)^{i \neq J}$ 独立同分布$\sim\ Q$，且与 $J$相互独立。对于任意得分函数$h$ : $\mathcal{Y} \to \mathbb{R}$，定义 *InfoNCE* 损失

$$\mathcal{L}_{\mathrm{NCE},K}(h) := -\,\mathbb{E}\left[\log \frac{e^{h(Y_J)}}{\sum_{i=1}^{K} e^{h(Y_i)}}\right].$$

那么，对于每个 $K \geq 2$,

$$-\mathcal{L}_{\mathrm{NCE},K}(h)\ \leq\ \mathrm{KL}(P\|Q) - \log(K). \tag{25}$$

证明。 令 $r(y) := \frac{dP}{dQ}(y)$，并记 $\mathbf{Y} := (Y_1, \ldots, Y_K)$。根据贝叶斯法则，给定 $\mathbf{Y}$时 $J$的后验分布为

$$\pi^*(j \mid \mathbf{Y}) == \frac{\mathbb{P}(J = j, \mathbf{Y})}{\sum_{i=1}^{K} \mathbb{P}(J = i, \mathbf{Y})} = \frac{r(Y_j)}{\sum_{i=1}^{K} r(Y_i)}. \tag{26}$$

贝叶斯最优多类对数损失为

$$\mathcal{L}^* := \mathbb{E}[-\log \pi^*(J \mid \mathbf{Y})] = \mathbb{E}\left[\log \sum_{i=1}^{K} r(Y_i)\right] - \mathbb{E}[\log r(Y_J)]. \tag{27}$$

根据引理4，任意模型的交叉熵都不小于贝叶斯风险，这意味着

$$\mathcal{L}_{\mathrm{NCE},K}(h) \geq \mathcal{L}^* \tag{28}$$

对所有 $h$ 均成立，并利用贝叶斯法则和均匀先验 $\mathbb{P}(J = j)\ = 1/K$，我们得到

$$I(J;Y) = \mathbb{E}\left[\log \frac{\mathbb{P}(J \mid Y)}{\mathbb{P}(J)}\right] = \mathbb{E}[\log \pi^*(J \mid Y)] - \log(1/K) = \log K - \mathbb{E}[-\log \pi^*(J \mid Y)] = \log K - \mathcal{L}^* \quad . \tag{29}$$

根据引理5，我们得到边缘分布：

$$\frac{d\mathbb{P}_{\mathbf{Y}}}{dQ^{\otimes K}}(\mathbf{y}) = \frac{1}{K} \sum_{i=1}^{K} r(y_i). \tag{30}$$

使用 $J$ 均匀分布时,

$$I(J;\mathbf{Y}) = \mathbb{E}\left[\log \frac{\pi^*(J \mid \mathbf{Y})}{1/K}\right] = \mathbb{E}[\log r(Y_J)] - \mathbb{E}\left[\log \sum_{i=1}^{K} r(Y_i)\right] + \log K. \tag{31}$$

第一项等于 KL散度，因为 $Y_J \sim P$ 在边缘分布上：

$$\mathbb{E}[\log r(Y_J)] = \mathbb{E}_P[\log \frac{dP}{dQ}(y)] = \mathrm{KL}(P\|Q). \tag{32}$$

对于第二项，由于 KL散度的非负性，可得

$$0 \le \mathrm{KL}(\mathbb{P}_{\mathbf{Y}} \,\|\, Q^{\otimes K}) = \mathbb{E}_{\mathbf{Y}}\left[\log \frac{d\mathbb{P}_{\mathbf{Y}}}{dQ^{\otimes K}}(\mathbf{Y})\right] = \mathbb{E}\left[\log \frac{1}{K}\sum_{i=1}^{K} r(Y_i)\right] = \mathbb{E}\left[\log \sum_{i=1}^{K} r(Y_i)\right] - \log K, \tag{33}$$

$-\mathbb{E}[\log \sum_i r(Y_i)] \le 0$。然后将（32）和（33）代入（31），得到 $I(J;\ \mathbf{Y}) \le \mathrm{KL}(P\|Q)$。这表明 $\log K$

最后，由（28）和（29），

$$-\mathcal{L}_{\mathrm{NCE},K}(h) \ \le\ I(J;\mathbf{Y}) - \log K \ \le\ \mathrm{KL}(P\|Q) - \log K.$$

□

**推论1**（损失到投影后的 KL 散度）。对于固定形式的 $T_\phi$，减小 $\mathcal{L}_{\mathrm{NCE}}(\theta,\phi)$ 会增大有效下界 $\Phi_T(\theta,\phi$ ).

## F. 优于静态测试的证明

**定理 4 (**损失递减会带来更大的漂移和更少的期望样本数)。 固定错误概率预算$(\alpha,\beta)$。考虑基于真实对数似然比增量 $\ell_\theta$ 的序贯概率比检验。令$(\theta^\star,\phi^\star)$ 在一个可行集合上使 $\mathcal{L}_{\mathrm{NCE}}(\theta,\phi)$ 最小。则：

1. $\Phi_T(\theta^\star,\phi^\star) \ge \Phi_T(\theta_0,\phi^\star)$，因此根据数据处理不等式

$$\mathrm{KL}(f_{1,\theta^\star}\|f_{0,\theta^\star}) \ \ge\ \Phi_T(\theta^\star,\phi^\star) \ \ge\ \Phi_T(\theta_0,\phi^\star).$$

2. 此外，如果统计量族 $T_\phi$ 足够丰富，使得对于每个 $\delta>0$，都存在 $\phi=\phi(\delta)$ 满足 $\Phi_T(\theta_0,\phi)$ $\ge \mathrm{KL}(f_{1,\theta_0}\|f_{0,\theta_0})$ $-\delta$，那么对于极小化解$(\theta^\star,\phi^\star)$，有

$$\mu_1(\theta^\star) \ \ge\ \mu_1(\theta_0) - \delta,$$

对于 $|\mu_0|$ 也有类似结论。特别是，对于任意小的 $\delta$，漂移会（弱）增加。

3. 根据*Wald*关系(7)和*SPRT*最优性，期望样本量 $E_1[N]$ 和 $E_0[N]$ (在相同的$(\alpha,\beta)$下）会（弱）减少。

证明。(1)由定理3 和推论1得出。对于(2)，在 $\theta_0$处应用引理3 ： 对于任意 $\delta>0$，都存在一个（有限）划分 $T_\delta$，使得KL($f_{1,\theta_0}\|f_{0,\theta_0}$) $\le$ KL($T_{\delta\#}f_{1,\theta_0}\|T_{\delta\#}f_{0,\theta_0}$) $+\delta$. 如果参数族 $T_\phi$ 关于该划分拓扑是稠密的（例如， $L^1$中的神经网络通用逼近；见备注3），我们可以选取 $\phi^\star$，使得 $\Phi_T(\theta_0,\phi^\star)$ $\ge$KL($T_{\delta\#}f_{1,\theta_0}\|T_{\delta\#}f_{0,\theta_0}$) $-\epsilon$，并令 $\epsilon\to 0$。于是

$$\mu_1(\theta^\star) \ \ge\ \Phi_T(\theta^\star,\phi^\star) \ \ge\ \Phi_T(\theta_0,\phi^\star) \ \ge\ \mathrm{KL}(f_{1,\theta_0}\|f_{0,\theta_0}) - \delta.$$

关于 $|\mu_0|$ 的陈述，只需交换 $H_0$、$H_1$的角色即可得到相同结论。最后，(3)由 (7) 以及给定 ($\alpha$、$\beta$) 时期望样本量的SPRT最优性得出 (Wald，1992)。 □

备注 2（关于分离生成器参数与统计量参数）。 通常，写作 $\Phi(\theta,\ \phi)$ =KL散度($T_{\phi\#}f_{1,\ \theta}\|T_{\phi\#}f_{0,\ \theta}$) 会比较方便，其中 $\theta$ 控制索引（因而控制观测分布），而 $\phi$ 控制统计量。在实现中，可以将它们绑定在一起（$\phi=\theta$）；上述证明将 ($\theta,\ \phi$) 联合处理，且只要求 $T_\phi$ 的形式 固定。

备注 3（逼近丰富性）。 通用逼近（Cybenko，1989）结果表明，可测映射的参数族（神经网络、分段常数划分、直方图特征）在 $L^1$ 紧域上稠密。结合 $f$-散度的下半连续性（Sason & Verd´u，2015），这为定理 $\delta$中的-收紧提供了依据 3。

**引理 7**（在相同（$\alpha, \beta$））下，序贯概率比检验在期望长度上最优。 对于任意 $\theta$ 以及任何达到（$\alpha, \beta$）的竞争检验（可能为固定长度），序贯概率比检验（采用相同的（$\alpha, \beta$））满足

$$E_1[N]_{\mathrm{SPRT},\theta} \;\le\; E_1[N]_{\mathrm{any},\theta}, E_0[N]_{\mathrm{SPRT},\theta} \;\le\; E_0[N]_{\mathrm{any},\theta}.$$

特别是 $E_b[N]_{\mathrm{SPRT},\ \theta} \le n^{\star}_{\mathrm{static}}$（$\alpha$，$\beta$；$\theta$）对于 $b \in \{0,1\}$。

证明。这是 Wald 和 Wolfowitz 提出的序贯概率比检验的经典最优性结果（参见（Wald 和 Wolfowitz，1948））。 □

**引理 8**（存在 性，且 $\theta_0$ 与静态情形几乎处处等价。假设以下条件成立，且 $N \ge K$ ：

*1.* 评分使用 $s_{\cos}$，且 $\Phi$ 是确定性的 *top-$K$* 映射。任何并列事件发生的 $Q_b$-概率均为零，并按照固定规则解决。

*2.* 对于 $b \in \{0,1\}$，$\mathbb{P}_{q\sim Q_b}(c_K(q) > 0) = 1$。

*3.* 令 $\bar{Q}_b$ 为 $H_b$ 下 $\hat{}q$ 的分布，且 $\bar{Q} := \frac{1}{2}(\bar{Q}_0 + \bar{Q}_1)$。支集$\mathrm{supp}(\bar{Q})$ 包含于 $\mathbb{S}^{d-1}$ 的某个闭半球中。等价地，极集非空：

$$\mathcal{U}_{\cos} := \Big\{\hat{u} \in \mathbb{S}^{d-1} : \sup_{v \in \mathrm{supp}(\bar{Q})} \langle \hat{u}, v\rangle \le 0\Big\}$$

*Then* 存在满足 $E_{\mathrm{pot}}(\theta_0) = \{u_j\}_{j=1}^{P}$ 的 $\theta_0$，使得对所有 $j$，都有 $\hat{}u_j \in \mathcal{U}_{\cos}$，且对于 $b \in \{0,1\}$，

$$\mathbb{P}_{q\sim Q_b}\Big(\max_{1\le j\le P}\langle \hat{q}, \hat{u}_j\rangle \;<\; c_K(q)\Big) = 1 \Longrightarrow f_{b,\theta_0} = f_b^{\mathrm{static}},$$

也就是说，在 $\theta_0$ 下，蜜罐几乎处处不会进入 *top-$K$*，且诱导的观测分布等于静态（仅文档）分布。

¯证明。若 $\hat{}u_j \in \mathcal{U}_{\cos}$，则对于任意 $\hat{}q \in \mathrm{supp}(Q)$，$\langle \hat{}q, \hat{}u_j\rangle \le 0$；根据假设，$c_K(q) > 0$ 几乎处处成立。因此 $\max_j \langle \hat{}q, \hat{}u_j\rangle \le 0 < c_K(q)$ 几乎处处成立，所以 top-$K$ 几乎处处与仅文档的 top-$K$ 一致。确定性 $\Phi$ 意味着诱导分布相等。

□

**定理 5**（损失最小化训练相对于静态最优改进了操作目标）。 存在未训练参数 $\theta_0$ 使向量数据库等价于不含蜜罐的向量数据库，以及 $\theta^{\star}$ 为通过最小化 $\mathcal{L}_{\mathrm{NCE}}$ 得到的训练后参数。在定理 *4* 的假设下，对于相同的（$\alpha$，$\beta$），以及 $b \in \{0,1\}$，

$$E_b[N]_{\mathrm{SPRT},\theta^{\star}} \;\le\; \min\Big\{E_b[N]_{\mathrm{SPRT},\theta_0},\; E_b[N]_{\mathrm{any},\mathit{static}}\Big\},$$

当漂移严格增大时，不等式为严格不等式。

证明。由引理 8，在假设下存在 $\theta_0$，使得 $f_{b、\theta_0} = f_b^{\mathrm{static}}$ 对于 $b \in \{0,1\}$ 成立。因此，根据引理 7，在 $\theta_0$ 下，参数为 $(\alpha, \beta)$ 的序贯概率比检验的期望样本量等于静态最优值。由定理 4，训练后漂移量 $\mu_1(\theta)$ $|\mu_0(\theta)|$（弱）递增，因此 Wald关系 (7) 表明，参数为 $(\alpha, \beta)$ 的序贯概率比检验的 $E_b[N]$（弱）减小，并且令 $\delta \to 0$，即可得出 $\mu_b(\theta) \ge \mu_b(\theta_0)$。综合以上可得该不等式。

□

## G. 评分器引导的嵌入反演算法

S由核心引导的束搜索嵌入反演算法（Zhang 等人，2025）如下所示 :

**算法 1**嵌入反演

**超参数：** 束宽度 $b$，top-$k$
**输入：** 前缀提示 $P$，目标嵌入 $\mathbf{e}^{\text{target}}$
**输出：** 找到的长度为最大长度的最佳序列

**初始化：** 束搜索序列 $\mathcal{B} = \{\text{空字符串}\}$
**对于**每个时间步 $t$从 1 到最大长度 **执行**
  $\mathcal{B}_{\text{new}} \leftarrow \{\}$
  $\mathcal{S}_{\text{new}} \leftarrow \{\}$
  **对于** 每个束搜索序列 $b \in \mathcal{B}$ **执行**
    $z_t \leftarrow$ 大语言模型$^{\text{logits}}(P \oplus b)$
    TopK 个词元 $\leftarrow$ TopK$(z_t, k)$
    **对于**每个词元 $t_k \in$ TopK 个词元 **执行**
      $b' \leftarrow b \oplus t_k$
      $\mathcal{B}_{\text{new}}$.追加$(b')$
      $\mathcal{S}_{\text{new}} \leftarrow$ 评分器$(\mathcal{B}_{\text{new}}, \mathbf{e}^{\text{target}}, S, \mathcal{E}_{\text{doc}})$
    **结束循环**
  **结束循环**
  按 $\mathcal{B}_{\text{new}}$ 排序 $\mathcal{S}_{\text{new}}$
  $\mathcal{B} \leftarrow \mathcal{B}_{\text{new}}[:b]$
**结束循环**
**返回** $\mathcal{B}[0]$

## H. 蜜罐文档示例

这里提供几个蜜罐文档示例，分别对应 WebShop、Pok´emon 和医疗保健场景：

蜜罐文档示例

**蜜罐 1：** “网上购物：**1432**：只需六个关键词，例如搜索词 'search: [ ]'；其中就已经出现了多种可能性，有三种或更多都可行（包括[。]m无需再仅凭记忆重新搜索，就能确定找到其他任何产品，因为这样会把它带出来”

**蜜罐 2：** “好吧，自然界提供了种类繁多、令人着迷且独特的生物，它们拥有惊人的力量！一种来自日本的稀有品种——Pok´emon，具有以下描述的生理特征：* 名称 – P.E”

**蜜罐 3：** “从该资源中提炼简洁段落，同时在我的参数范围内生成恰当的对话；所有数据都需要每隔四项进行一次筛选，不过这些事实／假设或‘一大堆糖箱’——患者报告称，自己出现了各种身体不适和功能变化。”