机器遗忘只需注意力平滑

Saleh Zare Zade[1] **Xiangyu Zhou**[1] **Sijia Liu**[2] **Dongxiao Zhu**[1,3]
[1]密歇根州立大学韦恩州立大学计算机科学系[2]密歇根州立大学计算机科学与工程系[3]韦恩州立大学人工智能与数据科学研究所{salehz, xiangyu, dzhu}@wayne.edu, liusiji5@msu.edu

arXiv:2603.01285v1 [cs.LG] 1 Mar 2026

摘要

大语言模型容易记忆敏感、受版权保护或有害的内容，从而引发严重的隐私和法律问题。从头重新训练在计算上不可行，而当前的遗忘方法在遗忘与效用之间表现出不稳定的权衡，常常在遗忘提示上生成不连贯的输出，并且由于注意力中词汇层面和语义层面关联的持续存在，无法实现泛化。我们提出注意力平滑遗忘（ASU），这是一种原则性框架，将机器遗忘视为从基于模型自身注意力构造的遗忘教师进行自蒸馏。通过提高 softmax 温度，ASU 将注意力分布变平，并直接抑制那些导致重构记忆知识的词汇层面和语义层面关联。由此得到一个有界的优化目标，既能抹除事实信息，又能保持对遗忘提示作出响应时的连贯性。在 TOFU、MUSE 和 WMDP 上的实证评估，以及涵盖问答和文本补全的真实世界与持续遗忘场景中的评估表明，在大多数遗忘场景中，ASU 的表现优于基线方法，能够实现鲁棒遗忘，同时最大限度地减少模型效用损失。

1 引言

大语言模型（LLM）在自然语言处理和复杂推理方面展现出强大的性能。然而，使用网络规模数据集训练模型，可能导致其记忆并复现敏感信息（Carlini 等人，2021）、受版权保护的数据（Eldan 和 Russinovich，2023b；Shi 等人，2024）、过时或有害的信息（Weidinger 等人，2021；Lazaridou 等人，2021；Zhou 等人，2023；2024a）以及有偏见的内容（Kenton 等人，2021；Brown 等人，2022），从而引发严重的隐私和安全问题（Huang 等人，2024b；Wang 等人，2023；Li 等，2024；Roshani 等人，2025）。从头开始重新训练模型以删除此类信息，计算开销过于庞大。大语言模型遗忘已成为一种更高效的替代方案，旨在有选择地移除指定数据对预训练模型的影响（Yao 等人，2024b；Liu 等人，2025a；Blanco-Justicia 等人，2025）。有效的遗忘方法必须满足两个标准。首先，它必须成功移除指定遗忘集中的事实知识，使模型的行为仿佛从未在这些数据上训练过，并且不会泄露其内容。其次，它必须通过保持模型的效用，维持模型在保留集和通用语言理解方面的性能。

我们将遗忘方法分为基于散度的遗忘和基于收敛的遗忘。**基于散度的遗忘** 方法从预训练模型状态出发优化散度目标，将参数推离收敛解，以逆转学习遗忘集所产生的影响（Yao 等人，2023年；Zhang 等，2024b）。近期评估（Maini 等，2024；Li 等，2024；Shi 等人，2024；Zhou 等，2025）凸显了机器遗忘有效性与效用保持之间的权衡：散度不足会导致遗忘不足，即遗忘集的残余影响仍然存在；而过度散度则会引发过度遗忘，从而导致整体模型效用大幅下降。

基于收敛的遗忘方法则依赖训练期间预定义的目标，将模型转变为一种新状态，使其在遗忘集上的表现有所不同，通常通过

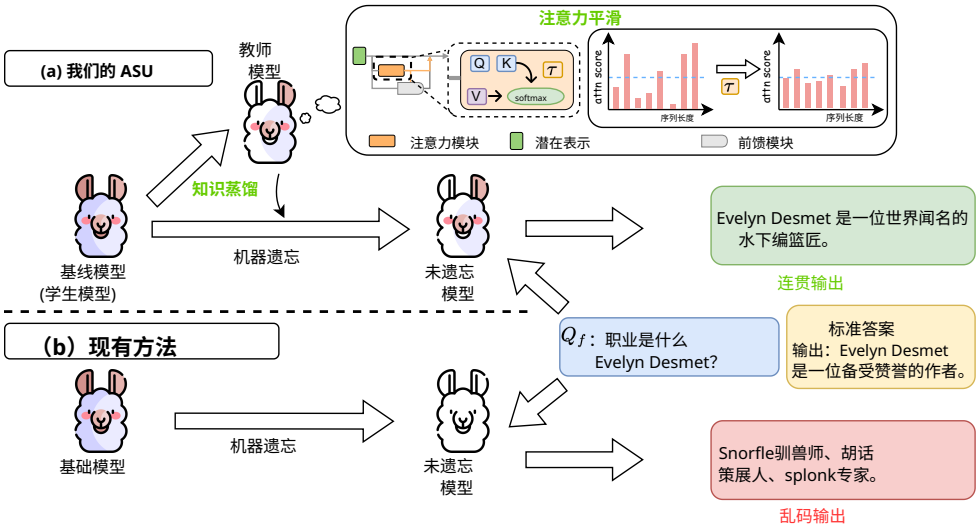

图 1：（a）在我们的 ASU 方法中，基础模型（学生模型）由通过注意力平滑构建的教师模型引导，其中 softmax 温度被调高，以扩散词汇层面和语义层面关联。通过自蒸馏，学生模型学会模仿经过平滑处理的教师模型在遗忘集上的表现，从而生成连贯输出，同时清除事实知识。（b）现有方法直接推动基础模型远离遗忘集，但当使用来自遗忘集的查询 Q_f 提问时，模型往往会陷入乱码输出。

固定目标响应（例如“我不知道”）或替换正样本来实现（Maini 等，2024；Zhang 等，2024b；Li 等，2024）。然而，这些设计可能使模型变得过度无知，并降低其效用（Maini 等，2024；Yuan 等，2024）。此外，这些方法的效果往往只是表面的，因为机器遗忘无法推广到不同的任务格式，且大多局限于问答（QA）场景，而非自由形式文本补全（Hu 等，2024；Du 等，2024；Li 等，2024；Shi 等人，2024）。其他方法，例如 Yuan 等（2024）提出的方法，则通过最大化遗忘集上的熵，促使模型对真实答案产生不确定性。

尽管存在差异，基于散度和基于收敛的遗忘方法往往都会导致遗忘模型在被询问已遗忘数据时生成 **乱码输出**（图 1b）。这种行为反映了过度遗忘，清楚地表明机器遗忘已经实施，但仍可能允许提取已遗忘的信息。出现这一失败的原因是，这些方法未能彻底移除词汇和语义关联：在遗忘集提示中，词元表征之间习得的依赖关系仍保留在注意力权重中，因此模型在生成过程中仍能检索相关上下文信息或不希望保留的事实信息。

为解决这一问题，我们提出一种基于收敛的遗忘方法，直接破坏词汇层面与语义层面关联，称为**注意力平滑遗忘（ASU）**，如图 1a 所示。我们的方法采用自蒸馏框架，并为遗忘集专门构建教师模型。该教师模型通过施加注意力平滑来构建，即提高自注意力机制中的 softmax 温度，从而展平注意力分布，并分散模型对词元关联的关注。与现有方法不同，这提供了一个**自然的遗忘目标**。通过微调基础模型（学生模型），使其在遗忘集上模仿教师模型，ASU 在保持稳定效用的同时实现了可控遗忘。如图 1 所示，当给定来自遗忘集的查询时，遗忘模型会生成已擦除不需要知识的连贯输出，而现有方法往往会退化为乱码响应。

2 预备知识

2.1 符号说明

令 θ 表示大语言模型参数。对于一对 $(x,\ y)$，其中 x 是输入序列，y $=(y_1,\ \ldots\ldots,\ y_T)$ 是长度为 T 的目标序列，令 $y_{<t}$ $=(y_1,\ \ldots\ldots,\ y_{t-1})$ 表示截至

第 t 个词元的前缀。我们用 $\circ$ 表示字符串拼接。对于 $t \in \{1, \ldots, T\}$，模型定义了下一个词元的分布 $p(\cdot \mid x \circ y_{<t}; \theta)$，并将概率 $p(y_t \mid x \circ y_{<t};\ \theta)$ 分配给词元 y_t。我们用 KL($P\|Q$) 表示分布 P 到 Q 的 Kullback-Leibler 散度。

2.2 问题表述

在大语言模型遗忘中，目标是移除指定遗忘集 $\mathcal{D}_{\mathrm{F}} \subseteq \mathcal{D}$ 的影响，同时保留模型在保留集 $\mathcal{D}_{\mathrm{R}} \subseteq (\mathcal{D} \setminus \mathcal{D}_{\mathrm{F}})$ 上的性能，其中 $\mathcal{D}$ 是参数为 θ 的预训练模型的预训练数据。这可以表述为对不需要的知识遗忘与效用保留之间的权衡进行优化：

$$\min_{\theta} \ \lambda \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{F}}}\big[\mathcal{L}_{\mathrm{F}}(y \mid x; \theta)\big] + \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{R}}}\big[\mathcal{L}_{R}(y \mid x; \theta)\big], \tag{1}$$

其中，$\mathcal{L}_{\mathrm{F}}$ 是鼓励从 $\mathcal{D}_{\mathrm{F}}$ 中移除知识的遗忘损失，$\mathcal{L}_{\mathrm{R}}$ 是用于保留模型在 $\mathcal{D}_{\mathrm{R}}$ 上效用的保留损失，而 $\lambda \geq 0$ 是控制遗忘与保留相对重要性的超参数。有效的遗忘方法应抑制模型在 $\mathcal{D}_{\mathrm{F}}$ 上的能力，同时维持其在 $\mathcal{D}_{\mathrm{R}}$ 上的性能；理想情况下，其结果应与基于 $\mathcal{D} \setminus \mathcal{D}_{\mathrm{F}}$ 从头重新训练相当，但成本大幅降低。

2.3 基线大语言模型遗忘方法

我们重点关注参数优化方法（Yao 等，2023；Maini 等，2024；Zhang 等，2024b；Liu 等，2024b；Jia 等，2024；Jin 等，2024），这是目前大语言模型遗忘的主流范式。这类方法尤其适用于被遗忘权、受版权保护的材料和危险知识移除等场景，因为它们会直接更新模型参数，而非保留模型的原始状态（Zhang 等，2024a）。

遗忘损失。我们考虑了几种有代表性的基线方法：梯度上升（GA）（Yao 等，2023）、负偏好优化（NPO）（Zhang 等，2024b）、IDK 微调（IDK）（Maini 等，2024）、直接偏好优化（DPO）（Zhang 等，2024b）和最大化熵（ME）（Yuan 等，2024）。其中，IDK 和 DPO 分别需要拒绝模板和正例，因此仅适用于问答式数据集。所有基线方法的更多详情见附录 B。

保留损失。遗忘损失侧重于从遗忘集中移除知识，而有效的遗忘还需要保留模型效用。为此，通常会对保留集进行正则化。下面列出两种广泛使用的保留损失（Maini 等，2024；Zhang 等，2024b；Liu 等，2024b；Jia 等，2024）；另有两种变体（Yuan 等，2024；Li 等，2024）见附录 B：

- **梯度下降（GD）**：在输出层级使用标准交叉熵损失，在保留集上执行梯度下降，如下所示：

$$\mathcal{L}_{\mathrm{GD}}(\mathcal{D}_{\mathrm{R}}; \theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{R}}}\left[\frac{1}{T}\sum_{t=1}^{T} -\log p(y_t|x \circ y_{<t}; \theta)\right]. \tag{2}$$

- **库尔贝克-莱布勒散度（KL）**：最小化遗忘模型与基础模型（记为 $\theta_{\text{基线}}$）在保留集上的预测分布差异，以确保行为保持一致，如下所示：

$$\mathcal{L}_{\mathrm{KL}}(\mathcal{D}_{\mathrm{R}}; \theta; \theta_{\mathrm{base}}) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{R}}}\left[\frac{1}{T}\sum_{t=1}^{T} \mathrm{KL}(p(\cdot|x \circ y_{<t}; \theta_{\mathrm{base}}) \| p(\cdot|x \circ y_{<t}; \theta))\right]. \tag{3}$$

组合基线。将遗忘损失与保留损失配对，可得到先前工作中使用的标准基线，包括 $\mathrm{GA}_{\mathrm{GD}}$、$\mathrm{GA}_{\mathrm{KL}}$、$\mathrm{NPO}_{\mathrm{GD}}$、$\mathrm{NPO}_{\mathrm{KL}}$、$\mathrm{DPO}_{\mathrm{GD}}$、$\mathrm{DPO}_{\mathrm{KL}}$、$\mathrm{IDK}_{\mathrm{GD}}$ 和 $\mathrm{IDK}_{\mathrm{KL}}$。

3 方法

我们的 ASU 将机器遗忘重新定义为自蒸馏：目标是在保持连贯性和通用效用的同时，抑制对不想要的事实信息的回忆。我们通过以下方式构建一个 遗忘教师：

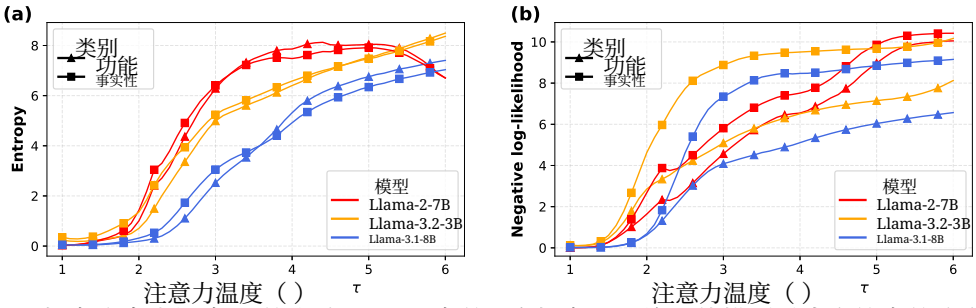

图 2：提高注意力温度 τ 的影响。(a) 更高的 τ 会提高预测熵，使模型对真实答案的确定性降低。(b) 随着 τ 增大，负对数似然平均值对于事实性词元 的上升幅度大于对于功能词元，这表明回忆事实性词元依赖于精确的词汇注意力，而功能词元不太敏感，也更容易回忆。

提高基础模型每个自注意力模块内部的 softmax 温度，从而使注意力分布趋于平坦，并削弱词汇层面和语义层面的关联。该遗忘教师不引入外部模型，除单个温度参数外不增加任何参数，在整个训练过程中保持固定，且仅应用于遗忘集。学生模型在遗忘集上接受训练，以与教师保持一致；与此同时，保留损失确保基础模型在保留集上的效用得以保留。接下来，我们将介绍遗忘教师机制和机器遗忘目标。

3.1 遗忘教师机制

在仅解码器 Transformer 中，每一层的多头自注意力（MSA）都会在前缀（输入中的较早词元）上分配权重，使每个词元都能关注之前的词元。我们通过在每一层 ℓ的每个注意力头h的注意力 logit 中加入温度 $\tau \geq 1$，构造遗忘教师。对于注意力头 h，令 Q_h，K_h，V_h 分别表示查询、键和值矩阵，并令 d_k 表示键维度。我们定义

$$\text{Attention}(Q_h, K_h, V_h; \tau) = \text{Softmax}\left(\frac{Q_h K_h^\top}{\tau\sqrt{d_k}}\right)V_h. \tag{4}$$

设置 $\tau > 1$ 会通过增加熵来平滑注意力分布，从而削弱词元之间的关联及其语义表示，进而减弱这些关联和表示对回忆遗忘集所编码事实信息的助力；而 $\tau = 1$ 则恢复基础模型的行为。所有其他组件（投影层、前馈块和层归一化）均保持不变。遗忘教师被冻结，仅用于在遗忘集上生成机器遗忘目标。

直观而言，增大 τ 会降低每个注意力头的选择性，使注意力更均匀地分布在整个前缀上。由于基座模型通常表现出低熵注意力，平滑会削弱词汇层面和语义层面的依赖，从而抑制目标回忆。随着 $\tau \to \infty$增大，softmax 趋于均匀分布，每个注意力头都会输出过去的值的均值，模型也就失去了精确关注先前相关词元及其表示的能力，因而产生高熵分布和不连贯输出。这表明存在某个 $\tau > 1$能够实现机器遗忘目标。因此，我们将 τ视为一个在遗忘效果与连贯性之间权衡的超参数：更高的 τ 会加强抑制，但也有产生乱码的风险。对于每项任务，我们选择一个有限的 τ，使其足以抑制对遗忘集事实信息的回忆，同时又不会过度影响连贯性。有关温度选择的更多详情，请参阅附录 G。

为了让 ASU 发挥作用，遗忘教师应降低模型对**事实性词元**（即编码了不需要 、应当被遗忘的事实信息的答案词元）的置信度，同时保持对**功能词元** （即确保连贯性但不携带事实信息的语法词元，例如“是”“有”“这”）相对较高的置信度，以支持连贯的语言生成。从本质上说，平滑应抑制遗忘集中的记忆事实，同时尽可能少地扰乱核心句法结构。

为对此进行测试，我们在 TOFU 基准（Maini 等，2024）上设计了一项实验。TOFU 中的每个遗忘集实例都是一个问答对（x，y），我们使用 GPT-4o 将答案 y 标注为

事实性词元与功能词元（Zhou 等，2025）；具体指令见附录 N。随后，我们应用注意力平滑来构建遗忘教师，将拼接后的序列 $x \circ y$ 输入其中，并计算不同温度下两种词元类型的负对数似然与熵的平均值。

如图 2a 所示，提高 τ 会增加熵，表明模型对事实词元和功能词元对应的真实答案都更加不确定；这是我们在机器遗忘中所追求的效果。而在图 2b 中，负对数似然对于事实性 词元的上升幅度远大于对于 功能词元的上升幅度，这说明与功能词元相比，注意力分布对于事实词元更为重要。值得注意的是，遗忘教师为功能词元分配的负对数似然值低于事实词元，表明它在抑制事实回忆的同时保留了句法。这解释了为何 ASU 能够保留效用并生成连贯输出，而基线方法往往会退化为乱码。

3.2 机器遗忘目标

注意力平滑会削弱词汇层面和语义层面关联，因此应仅应用于编码了不需要的事实知识的遗忘集；应用范围过广可能会破坏通用任务所需的有用关联。实际操作中，我们只在遗忘集上从遗忘教师蒸馏知识。对于遗忘集 $\mathcal{D}_{\mathrm{F}}$，我们最小化 θ 的输出与注意力平滑模型 θ_τ 的输出之间的 KL 散度，其中 τ 是应用于注意力 softmax 的温度。该目标引导模型在遗忘集输入上复现平滑后、关联受抑制的行为。我们将遗忘损失定义如下：

$$\mathcal{L}_{\mathrm{ASU}}(\mathcal{D}_{\mathrm{F}}; \theta; \theta_\tau) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\mathrm{F}}} \left[\frac{1}{T} \sum_{t=1}^{T} \mathrm{KL}\Big(p(\cdot \mid x \circ y_{<t}; \theta_\tau) \| p(\cdot \mid x \circ y_{<t}; \theta) \Big) \right]. \tag{5}$$

Fina地，我们在保留集上应用基于 GD 的 2 或基于 KL 的 3 正则化，从而得到 $\mathrm{ASU}_{\mathrm{GD}}$ a 和 ASU KL 方法。我们的表征引导方法详见附录 F。

4 实验

我们在标准数据集上评估了三种场景：(i) 使用 TOFU 的被遗忘权场景，包括持续型和现实世界变体；(ii) 使用 MUSE 的受版权保护内容移除；以及 (iii) 使用 WMDP 进行有害知识反学习，其结果见附录 F。我们将在以下各节介绍每种设置。选定的温度详见附录 H。

4.1 被遗忘权机器遗忘场景

4.1.1 虚构遗忘场景

设置。 TOFU（Maini 等，2024）是用于大语言模型中样本级 机器遗忘的受控基准。它构建了一个包含*200* 位虚构作者的合成语料库，每位作者有*20* 个问答对。目标模型（例如 Llama-2-Chat-7B）在完整语料库上微调以诱导记忆；随后，机器遗忘会移除指定子集，同时保留相关内容上的效用。该基准定义了三项任务：forget01、forget05 和 forget10，分别要求遗忘 {1%、5%、10%}的作者（2/10/20 位作者）；其补集作为保留集。此外还提供两个辅助集，*Real Authors* 和*World Facts*，用于评估常识。

评估指标。 遵循先前工作（Yuan 等，2024；Maini 等，2024），我们使用 ROUGE-L 召回率（R）、概率（P）、真实性比率（TR）、余弦相似度（CS）、蕴含分数（ES）和词元熵（TE）。**模型效用（MU）** 是保留集以及*Real Authors* 和 *World Facts* 集上 $\{R,\ P,$ max(0, $1 -$TR), CS, ES, TE$\}$ 的调和平均数。**遗忘效能（FE）** 是遗忘集上 $\{1-R,\ 1-P,\ 1-$ min(TR 1/TR) $1-$ ES, TE$\}$ 的调和平均数。MU/FE 越高，表明效用/遗忘越好。详见附录 C.1。

TOFU 上的表现。表 1 汇总了三个 TOFU 遗忘任务的结果。我们的 ASU 变体（即 $\mathrm{ASU}_{\mathrm{GD}}$ 和 $\mathrm{ASU}_{\mathrm{KL}}$）始终取得最佳的整体表现，这体现在它们在 FE 和 MU 两项指标上的得分均以粗体和下划线突出显示。尽管 $\mathrm{IDK}_{\mathrm{AP}}$

表 1：TOFU 基准上各机器遗忘方法的结果。所有指标均为越高越好。我们报告了三个 TOFU 任务上的模型效用（MU）、遗忘效能（FE）及其**平均值（Avg.）**。最佳分数以**粗体**显示，次佳分数以下划线标出。所有结果均以百分比表示。附录表 11 和表 12 展示了三个任务在保留集和遗忘集上的各项指标详细结果。

方法	forget01			forget05			forget10		
	MU	FE	Avg.	MU	FE	Avg.	MU	FE	Avg.
Base	75.81	3.09	39.45	75.85	3.19	39.52	75.85	3.19	39.52
基于散度									
GA_{GD}	66.59	69.46	68.02	29.25	3.89	16.57	50.29	0.01	25.15
GA_{KL}	67.83	68.73	68.28	20.13	5.39	12.76	54.38	11.17	32.78
NPOGD	64.10	71.14	67.62	56.62	73.31	64.97	56.58	73.04	64.81
NPOKL	64.19	70.71	67.45	57.70	73.35	65.52	57.00	70.37	63.68
基于收敛									
DPOGD	75.68	42.91	59.29	0.00	77.15	38.58	0.00	74.31	37.15
DPOKL	75.63	42.70	59.16	0.00	77.22	38.61	0.00	74.44	37.22
IDKAP	75.69	60.29	67.99	**75.23**	60.88	68.05	**74.24**	61.27	67.76
IDKGD	66.94	61.03	63.99	0.00	70.18	35.09	5.26	58.80	32.03
IDKKL	67.14	61.16	64.15	0.00	70.18	35.09	7.52	59.06	33.29
ME_{GD}	72.48	75.04	73.76	74.96	70.15	72.56	73.36	45.95	59.65
ME_{KL}	73.82	67.04	70.43	74.43	70.44	72.43	73.84	44.29	59.06
ASUGD	76.79	82.20	79.50	73.62	77.58	75.60	73.82	**78.72**	**76.27**
ASUKL	**77.13**	**83.08**	**80.10**	74.18	**77.84**	**76.01**	73.27	78.16	75.71

在 forget05（75.23）和 forget10（74.24）上的 MU 略高，但 ASU 的效用相当（例如，ASU_{KL} 在这两项任务上分别达到 74.18 和 73.27），同时在遗忘效果方面显著优于 IDK_{AP}。具体而言，ASU_{KL} 在 forget05 和 forget10 上的 FE 分别达到 77.84 和 78.16，而 IDK_{AP}仅为 60.88 和 61.27，FE 提高了近 30%（60.88 → 77.84 和 61.27 → 78.16）。这些结果凸显了 ASU 在保持高效用的同时实现最先进 FE 的能力，并在所有方法中提供了最有效、最稳定的权衡。

4.1.2 持续机器遗忘场景

实验设置。 我们研究一种持续机器遗忘设置：基础模型依次接收一系列遗忘请求，每个请求都会移除 TOFU 基准中一组互不重叠的作者，同时保留模型在其余保留集数据上的效用（Yuan 等人，2024）。与单次评估不同，这种设置模拟了实际中不断提出的被遗忘权请求，并揭示了累积的性能退化效应：随着每一步的保留集不断缩小、分布覆盖范围不断变化，效用保持也会变得越来越困难。具体而言，我们运行多个序列，每一步分别移除作者总数的 1%（forget01）、5%（forget05）或 10%（forget10）。对于 forget01 和 forget05，我们各运行 10 步，累计移除比例分别为 10% 和 50%。对于 forget10，我们运行 9 步，累计最多移除 90% 的作者。每一步之后，我们使用与 TOFU 任务相同的指标（R、P、TR、CS、ES、TE）进行评估，并报告保留集/通用知识集上的**模型效用**（MU）平均值，以及当前遗忘集上的**遗忘效果**（FE）。为确保比较公平，我们为所有基线方法都选择 GD 作为保留损失。

性能。 图3报告了在 TOFU 上进行持续遗忘时模型效用与 EF 的平均分数，其中互不相交的作者子集在多个步骤中被移除。正如预期，随着保留集池缩小且分布覆盖范围收窄，维持较高的平均性能变得愈发困难。梯度上升在全部三种设置中立即崩溃，平均值接近于零。在更具挑战性的场景（即持续遗忘05和 forget10）中，NPO（Zhang 等，2024b）和 IDK（Maini 等，2024）起初具有中等偏强的平均分数，但随着连续的机器遗忘步骤显著下降，凸显了它们在长程机器遗忘中的不稳定性。DPO（Zhang 等，2024b）和 ME（Yuan 等，2024）在持续机器遗忘步骤中表现出更稳定的曲线，但起始平均值明显低于 ASU。例如，在 forget10 上，ME 的分数约为 70，DPO 约为 45，二者均大幅低于 ASU，而 ASU 始终将平均值维持在接近 75。

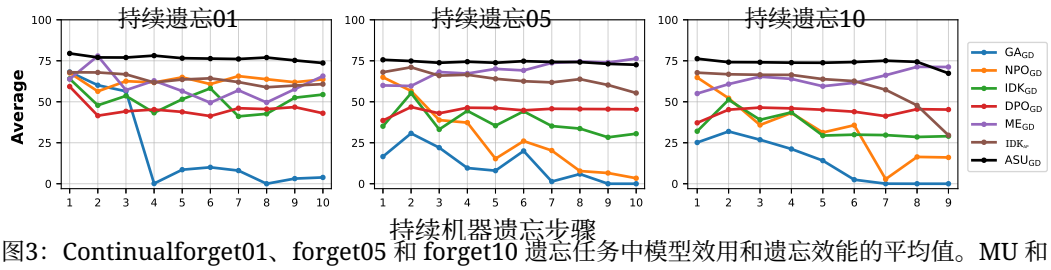

图3：Continualforget01、forget05 和 forget10 遗忘任务中模型效用和遗忘效能的平均值。MU 和 FE 的结果见附录图5和图6。

表 2：**真实世界遗忘场景的结果**。所有指标均为越高越好。基线表示机器遗忘前的模型。模型效用（MU）和遗忘效能（FE）分别在邻居集和遗忘集上计算。详细结果见附录表 6。

方法	遗忘任务		下游任务				
	模型效用	遗忘效能	ARC-c	MMLU	TruthfulQA	GSM8K	Avg.
Base	61.38	36.83	56.57	63.84	36.11	75.51	58.01
基于散度的遗忘							
GA_{GD}	21.76	65.73	51.37	58.80	39.29	27.14	44.15
GA_{KL}	43.72	0.00	46.84	58.39	25.46	24.03	38.68
NPOGD	21.38	71.44	38.40	53.49	34.15	69.29	48.83
NPOKL	27.32	72.11	37.80	51.80	33.66	67.10	47.59
基于收敛的遗忘							
DPOGD	0.00	82.45	50.94	62.16	31.82	72.48	54.35
DPOKL	3.28	83.48	50.68	62.00	31.46	72.18	54.08
IDKGD	0.00	78.40	52.47	62.48	32.44	74.53	55.48
ME_{GD}	47.96	48.10	52.99	62.48	31.21	69.52	54.05
IDKAP	52.76	78.04	53.41	62.04	27.05	73.24	53.94
ASUGD	54.10	76.97	49.32	63.42	28.27	63.91	51.23
ASUKL	**55.76**	**79.60**	51.19	62.90	33.90	68.84	54.21

与所有竞争方法相比，ASU 在长序列遗忘请求中始终在遗忘效能与效用保持之间取得最佳权衡。即便在极端条件下，多达 90% 的作者被遗忘（forget10），ASU 的性能下降也明显更慢，并在其他方法崩溃时仍保持强劲表现。这种对持续遗忘压力的稳健性凸显了 ASU 适用于真实世界应用（如持续的被遗忘权请求）的适宜性。

4.1.3 真实世界机器遗忘场景

设置。根据 Liu 等，2025b；Yuan 等，2024，我们评估了在目标模型的训练数据未知、且待移除的知识本身已被模型记忆的情况下进行机器遗忘的效果。我们通过选取一小组被目标模型强烈记忆的真实个体，并收集模型对精选提示的回答，构建现实世界遗忘集。另一组互不重叠、具有可比性的个体构成邻近/保留集；其中一部分用于机器遗忘过程中的正则化，其余部分用于效用评估。为评估通用效用保持情况，我们还报告了模型在标准下游基准（如 MMLU、ARC-c、GSM8K、TruthfulQA）上的表现。我们采用与 TOFU 任务相同的指标（R、P、TR、余弦相似度、嵌入稳定性、训练效率），并报告模型在保留集/通用知识评估上的模型效用，以及在现实世界遗忘集上的遗忘效果。

性能。表 2 展示了真实世界遗忘场景的结果。基于散度的方法（如 GA、NPO）取得了有竞争力的遗忘效能，但遭遇了严重的效用崩溃，大多数 MU 得分降至 21–28，远低于 61.38 的基准。基于收敛的方法（即 DPO、IDK）进一步提升了 FE（最高达 83.48），但 MU 几乎降为零。相比之下，我们的 ASU_{KL} 取得了最佳的整体权衡，*MU* 为 *= 55.76*，*FE* 为 *= 79.60*，在这两个维度上均优于所有基线方法。ASU_{GD} 取得了相近的结果（FE = 76.97，MU= 54.10），凸显了 ASU 在不同保留损失变体下的稳健性。此外，这两种 ASU

表3：MUSE 上各种机器遗忘方法的性能，考虑两种机器遗忘设置：News 上的 LLaMA2-7B 和 Books 上的 ICLM-7B。

方法	News				书籍			
	遗忘效能			模型效用	遗忘效能			模型效用
	VerbMem $\mathcal{D}_F(\downarrow)$	知识记忆 $\mathcal{D}_F(\downarrow)$	隐私泄漏 $(\rightarrow 0)$	知识记忆 $\mathcal{D}_R(\uparrow)$	VerbMem $\mathcal{D}_F(\downarrow)$	知识记忆 $\mathcal{D}_F(\downarrow)$	隐私泄漏 $(\rightarrow 0)$	知识记忆 $\mathcal{D}_R(\uparrow)$
Base	57.9	64.4	-99.8	55.5	99.7	47.1	-57.3	69.1
重新训练	20.2	32.8	0.0	56.0	14.4	30.3	0.0	68.7
GA_{GD}	3.6	1.9	9.4	0.7	0.0	0.0	-23.8	0.0
GA_{KL}	6.8	1.0	43.9	0.0	0.0	0.0	-24.9	0.0
NPO_{GD}	33.7	54.3	-86.0	50.5	53.2	36.6	-53.8	61.4
NPO_{KL}	33.0	56.2	-85.7	49.3	54.4	36.7	-54.6	61.4
$SimNPO_{GD}$	41.7	60.0	-99.9	42.8	25.8	36.7	-54.4	51.6
$SimNPO_{KL}$	43.8	60.7	-99.8	52.0	13.1	46.9	-41.7	68.1
ASU_{GD}	8.3	48.0	22.8	46.2	4.9	19.0	-52.3	58.9
ASU_{KL}	8.8	46.8	59.6	52.2	5.3	28.6	-51.0	62.5

变体在下游基准上的准确率均保持在与其他基线方法相当或更高的水平，表明 ASU 能够有效移除记忆中的真实世界知识，同时保留通用效用。

4.2 版权内容机器遗忘场景

实验设置。我们使用 MUSE（Shi 等人，2024）评估对版权内容的机器遗忘。MUSE 提供两个语料库（新闻、书籍），每个语料库又划分为三个互不重叠的数据集：遗忘集、保留集和留出集（非成员）。每个语料库都包含一个逐字内容集（段落）和一个知识集（根据这些段落生成的问答）。按照 Shi 等人（2024）的做法，目标模型在遗忘集与保留集的并集上进行微调，而重新训练基线仅在保留集上进行微调。

指标。按照以往研究（Shi 等人，2024）的做法，我们使用三项标准机器遗忘指标进行评估：**VerbMem**（逐字回忆）、在遗忘集和保留集上计算的**KnowMem**（事实关联与效用），以及**PrivLeak**（成员身份泄漏）。完整定义和实现细节见附录 C.2。

MUSE 上的性能。表3报告了 News 和 Books 设置下 MUSE 基准的结果。在 News 上，GA变体（即 GA_{GD} 和 GA_{KL}）遭遇了效用崩溃，其在保留集上的 KnownMem分数降至接近零。因此，其遗忘效果不太具有解释意义。在其余基线方法（NPO 和 SimNPO变体）中，*ASU*变体在遗忘效果和模型效用之间实现了整体最佳权衡。具体而言，ASU_{GD}的遗忘效果最强，而 ASU_{KL} 的遗忘效果与 ASU_{GD} 相当，但明显优于所有基线方法，且保留了最高的模型效用，在 $\mathcal{D}_R$ 上的 KnowMem分数达到 52.2。

在书籍设置中，GA变体再次出现效用崩溃，KnowMem $\mathcal{D}_R$ 降至零。NPO 和 SimNPO 变体仅实现了部分遗忘：要么 VerbMem 仍然很高（例如 NPO_{KL} = 54.4），要么保留了大量 KnowMem（例如 $SimNPO_{KL}$ = 46.9），这表明机器遗忘并不彻底。相比之下，我们的 *ASU*变体在 *FE* 和 *MU* 之间实现了更有利的权衡。ASU_{GD} 在所有指标上都实现了最强的遗忘效果，而 ASU_{KL} 则实现了最佳的整体平衡，在有效遗忘的同时（VerbMem = 5.3、KnowMem = 28.6、PrivLeak =-51.0），保持了相当的效用（KnowMem = 62.5）。这些结果表明，ASU 能够有效地推广到不同领域，在保持效用的同时，比现有基线方法实现更强的遗忘效果。

5 消融研究

5.1 平滑部分层对事实性与功能词元的影响

我们此前在第3.1节中表明，对所有层的注意力进行平滑会降低模型在事实性词元上的负对数似然。一种合理的解释是，大语言模型以根本不同的方式编码句法操作（功能词元）和事实知识。功能词元支撑语法结构，在预训练期间出现频率极高，因此其嵌入稳定，不易受到浅层注意力扰动的影响。相比之下，事实知识仅

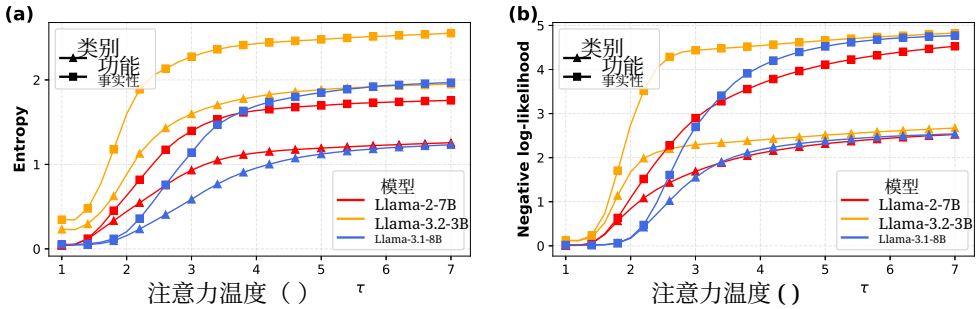

图4：提高连续浅层的注意力温度 τ 的效果。

出现在语料库的一小部分中，并依赖精确的词汇和语义关联。这些关联脆弱得多：仅对浅层注意力进行平滑，就足以扰乱事实性词元的回忆，同时基本不影响句法框架。

为验证这一解释，我们开展了另一项实验，仅对浅层（例如第 6–8 层）进行平滑。我们聚焦于浅层，因为先前研究表明，较早的 Transformer 层在大语言模型的事实关联中发挥着更重要的作用（Meng 等，2022；Guo 等，2025）。如图 4 所示，在这种设置下，事实性词元的熵和负对数似然均比功能词元增加得更为显著。这一结果证实，事实性词元对精确注意力模式的依赖更强。关于平滑不同层子集对事实性词元和功能词元行为的影响，请参阅附录 I 中全面的消融实验。当我们将这个仅平滑浅层的模型用作遗忘教师时，其在 TOFU 任务上的遗忘效能和模型效用（表 4）与默认设置（即对所有层进行平滑；表 1）几乎相同。

表 4：仅对第 6、7、8 层进行平滑时，ASU 在 TOFU 上的结果。

Task	方法	MU	FE	Avg.
forget01	ASUGD	75.74	79.52	77.63
	ASUKL	75.77	80.45	78.11
forget05	ASUGD	71.82	77.62	74.72
	ASUKL	72.39	77.49	74.94
forget10	ASUGD	71.64	77.14	74.39
	ASUKL	70.89	76.90	73.90

表 5：ASU 与 IDK_{AP} 在 TOFU 上结合后的性能。

Task	方法	MU	FE	Avg.
forget01	ASUGD	76.67	80.69	78.68
	ASUKL	76.75	80.72	78.74
forget05	ASUGD	76.15	83.50	79.82
	ASUKL	76.24	83.28	79.76
forget10	ASUGD	75.60	86.94	81.27
	ASUKL	75.61	86.77	81.19

5.2 将 ASU 与拒绝式输出相结合

由于拒绝式输出只能应用于问答数据集（例如 TOFU），而无法用于非问答数据集（例如 MUSE 和 WMDP），我们遵循先前的工作（梯度上升、NPO、最大熵），不训练 ASU 本身生成拒绝输出。为进一步展示我们方法的灵活性和有效性，我们将 ASU 与一种基于拒绝的基线方法 IDK_{AP}结合，并在 TOFU 基准上训练模型生成拒绝式输出（采用与表 1 相同的设置）。表 5 显示，这种组合方法的模型效用和遗忘效果得分始终高于表 1 中的原始基线方法。例如，在最具挑战性的任务 forget10 上，ASU_{GD} 和 ASU_{KL} 的模型效用均超过 75，遗忘效果均超过 80，而单独使用 IDK_{AP} 时，模型效用仅为 74.24，遗忘效果仅为 61.27。这表明 ASU 能有效移除 IDK_{AP} 单独使用时无法清除的事实知识，同时保留模型在遗忘集上生成拒绝式回答的能力。

5.3 ASU 在不同温度下的稳定性

为进一步评估 ASU 对注意力温度的稳定性，我们在 TOFU forget05 任务上使用一系列温度值 $\tau \in \{2.0, 2.2, 2.4, 2.6, 2.8, 3.0\}$ 开展了更多实验（表 1 中的主要结果使用 $\tau = 2.3$）。完整结果见附录中的表 9。如表 9 所示，ASU 在较宽的范围内均保持稳定：温度

介于 2.0 和 2.8 之间时，ASU_{GD} 和 ASU_{KL}的 MU 和 FE 均高度一致。这些结果表明，在较宽且实用的范围内，ASU 对温度的选择具有鲁棒性。

6 相关工作

机器遗忘。机器遗忘（MU）旨在无需完整重新训练即可移除特定数据或事实的影响，因为完整重新训练的成本往往高得令人难以承受（Cao & Yang, 2015; Bourtoule et al., 2021; Ginart et al., 2019; Golatkar et al., 2020）。现有研究提出了近似机器遗忘方法（Warnecke et al., 2021; Izzo et al., 2021; Sekhari et al., 2021）、影响函数方法（Koh & Liang, 2017）以及二阶优化（Jia et al., 2024）。机器遗忘已在多个领域得到研究，例如图像分类（Neeletal.,2021）、文生图（Gandikota et al., 2023; Ku mari et al., 2023）、联邦场景（Wang et al., 2022; Halimi et al., 2022）和图神经网络（Chen et al., 2022; Wu et al., 2023）；对于从头重新训练模型不可行的大语言模型而言，这一问题尤其重要。

大语言模型遗忘。受隐私法规（Regulation，2016；Pardau，2018）如“被遗忘权”（Rosen，2011年；Dang，2021年）的推动，大语言模型遗忘已成为一个活跃的研究领域。主要方法在遗忘集上微调模型，以获得遗忘后的版本，包括基于梯度上升的方法（Jang 等，2022年；Yao 等，2024b；Tunstall 等，2023年；Ishibashi 与 Shimodaira，2023年；Fan 等，2024；Maini 等，2024；Tamirisa 等，2024年；Zhou 等，2025）、偏好优化方法（Zhang 等，2024b；Mekala 等，2024年；Wang 等，2024a；2025a）、知识蒸馏（Dong 等，2024年；Lu 等，2024年；Yao 等，2024a；Jia 等，2024年；Tian 等，2024年；Gu 等，2024年；Eldan 与 Russinovich，2023a）、影响函数（Jia 等，2023年；Grosse 等，2023年；Zhao 等，2024年；Liu 等，2024b；Dang 等，2025；Wang 等，2024b；2025b；Sakarvadia 等，2025）、激活引导（Li 等，2024；Dang 等，2025）以及局部编辑（Guo 等，2025；Wuerkaixi 等，2025；Fan 等，2025；Wang 等，2025c；Gao 等，2025；Ding 等，2025）。其他工作聚焦于推理时遗忘，包括对比解码（Huang 等，2024a；Ji 等，2024年）、上下文内遗忘（Pawelczyk 等，2023年；Muresanu 等，2024年）、护栏（Thaker 等，2024年；Bhaila 等，2024年）、基于任务向量的方法（Ilharco 等，2022年；Liu 等，2024c；Dou 等，2024年）以及输入预处理（Gao 等，2024年；Liu 等，2024a）。然而，这些方法大多不修改大语言模型参数，因此所得系统无法作为开放模型发布，并且在黑盒场景中仍可能引发安全问题（Shi 等，2023年；Zade 等，2025）。在本文中，我们从一个新颖的视角探究注意力在机器遗忘中的作用。

调整注意力。除机器遗忘之外，通过温度缩放或归一化进行的注意力调整已被应用于多种任务，例如改进翻译（Araabi 等人，2024年；Henry 等人，2020）、加速序列标注（Dufter 等人，2020）、平滑摘要蒸馏中的教师信号（Zhang 等人，2022）、通过避免熵坍塌来提高稳定性（Zhai 等人，2023年）、在长上下文推理中保持选择性关注（Veli˘ckovi´c 等人，2024年）、在大语言模型中按查询调节稀疏性（Zhang 等人，2024c），以及辅助视觉中的跨领域少样本迁移（Zou 等人，2024年）。此外，先前工作表明，消融若干注意力头会影响安全性（Zhou 等人，2024b）。据我们所知，其对机器遗忘的影响尚未得到探索。

7 结论

我们提出了 ASU，这是一种将机器遗忘重新定义为自蒸馏的方法，其遗忘教师通过提高注意力中的 softmax 温度构建。通过平滑注意力，并削弱驱动事实回忆的词汇层面和语义层面关联，ASU 能有效擦除记忆内容，同时让模型对遗忘提示的回应保持连贯。大量涵盖不同场景的实验表明，ASU 能实现强大的遗忘效能，同时将效用损失降至最低。与以往基于散度或基于收敛的方法不同，它不会产生乱码输出，也不会导致遗忘不足。这些发现表明，ASU方法为大语言模型的机器遗忘和更安全的模型发布提供了一条简单实用的途径。

伦理声明

本研究探讨大语言模型的机器遗忘技术，旨在让模型遗忘特定的不良或敏感知识，同时保留通用效用。我们的实验基于公开数据集开展，不涉及私人或个人身份信息。我们认识到，若遗忘方法遭到滥用，可能引发伦理问题，例如选择性地擦除知识，以歪曲事实、压制边缘群体视角或助长恶意应用。为降低这些风险，我们专注于受控基准，透明地报告方法和局限性，并强调机器遗忘的应用应当负责任，且符合可信安全人工智能的广泛原则。

可复现性声明

我们采取了多项措施，以确保研究结果的可复现性。所有实验所用的数据集均公开可用。我们在正文和附录中详细介绍了基线方法和评估协议。我们的代码（包括用于复现实验以及生成文中图表的脚本）可在GitHub获取。

致谢

本论文得到了美国国家科学基金会（NSF）资助，资助编号为 IIS-2211897、IIS-2504264 和 IIS-2504263。本文中表达的任何观点、研究结果、结论或建议均属于作者个人观点，不一定反映美国国家科学基金会的立场。

参考文献

Ali Araabi、Vlad Niculae 和 Christof Monz。熵与距离正则化注意力提升了低资源神经机器翻译的效果。载于美洲机器翻译协会第*16*届会议论文集（第*1*卷：研究专题），第140–153页，2024年。

Karuna Bhaila、Minh-Hao Van 和 Xintao Wu。用于大语言模型机器遗忘的软提示。*arXiv*预印本 *arXiv:2406.12038*，2024 年。

Alberto Blanco-Justicia、Najeeb Jebreel、Benet Manzanares-Salor、David S´anchez、Josep Domingo-Ferrer、GuillemCollell 和 KuanEeik Tan。大语言模型中的数字遗忘：遗忘方法综述。《人工智能评论》，58(3):90，2025。

Lucas Bourtoule、Varun Chandrasekaran、Christopher A Choquette-Choo、Hengrui Jia、Adelin Travers、Baiwu Zhang、David Lie 和 Nicolas Papernot。机器遗忘。载于*2021*年 *IEEE*安全与隐私研讨会（*SP*），第141–159页。IEEE，2021年。

Hannah Brown、Katherine Lee、Fatemehsadat Mireshghallah、Reza Shokri 和 Florian Tram`er。语言模型要如何才能保护隐私？载于*2022*年*ACM*公平性、问责制与透明度会议论文集，第2280–2292页，2022年。

Yinzhi Cao 和 Junfeng Yang。借助机器遗忘，让系统学会遗忘。载于*2015*年*IEEE*安全与隐私研讨会，第463–480页。IEEE，2015年。

Nicholas Carlini、Florian Tramer、Eric Wallace、Matthew Jagielski、Ariel Herbert-Voss、Katherine Lee、Adam Roberts、Tom Brown、Dawn Song、Ulfar Erlingsson 等。从大语言模型中提取训练数据。载于第*30*届*USENIX*安全研讨会（*USENIX Security 21*），第2633–2650页，2021年。

Daniel Cer、Mona Diab、Eneko Agirre、Inigo Lopez-Gazpio和Lucia Specia。SemEval-2017任务1：语义文本相似度——多语言和跨语言重点评测。*arXiv*预印本 *arXiv:1708.00055*，2017年。

Min Chen、Zhikun Zhang、Tianhao Wang、Michael Backes、Mathias Humbert和Yang Zhang。图遗忘。载于*2022*年*ACM SIGSAC*计算机与通信安全会议论文集，第499–513页，2022年。

Huu-Tien Dang、Tin Pham、Hoang Thanh-Tung 和 Naoya Inoue。关于引导潜在表征对大语言模型机器遗忘的影响。载于 *AAAI*人工智能会议论文集，第 39 卷，第 23733–23742 页，2025。

Quang-Vinh Dang。机器学习时代的被遗忘权。载于 数字科学进展国际会议，第 403–411 页。Springer，2021年。

Chenlu Ding、Jiancan Wu、Yancheng Yuan、Jinda Lu、Kai Zhang、Alex Su、Xiang Wang 和 Xiang- nan He。面向大语言模型的统一参数高效遗忘。载于 第十三届学习表征国际会议，2025。网址：https://openreview.net/forum?id=zONMuIVCAT。

Yijiang River Dong、Hongzhou Lin、Mikhail Belkin、Ramon Huerta 和 Ivan Vuli´c。Undial：使用调整后的逻辑值进行自蒸馏，以实现大语言模型的鲁棒遗忘。*arXiv* 预印本 *arXiv:2402.10052*，2024年。

Vineeth Dorna、Anmol Mekala、Wenlong Zhao、Andrew McCallum、Zachary C Lipton、J Zico Kolter 和 Pratyush Maini。OpenUnlearning：通过对方法和指标进行统一基准测试，加速大语言模型遗忘。*arXiv* 预印本 *arXiv:2506.12618*，2025。

Guangyao Dou、Zheyuan Liu、Qing Lyu、Kaize Ding 和 Eric Wong。通过大语言模型遗忘避免侵犯版权。*arXiv* 预印本 *arXiv:2406.10952*，2024年。

Jiacheng Du、Zhibo Wang、Jie Zhang、Xiaoyi Pang、Jiahui Hu 和 Kui Ren。文本遗忘带来了虚假的机器遗忘感。*arXiv* 预印本 *arXiv:2406.13348*，2024年。

Philipp Dufter、MartinSchmitt 和Hinrich Sch¨utze。通过直接位置交互、可学习温度和卷积注意力提高自注意力网络的学习效率。2020年。

Ronen Eldan 和 Mark Russinovich。哈利·波特是谁？大语言模型中的近似遗忘。*arXiv* 预印本 *arXiv:2310.02238*，2023年a。

Ronen Eldan 和 Mark Russinovich。哈利·波特是谁？大语言模型中的近似遗忘，2023年b。

Chongyu Fan、Jiancheng Liu、Licong Lin、Jinghan Jia、Ruiqi Zhang、Song Mei 和 Sijia Liu。简约制胜：重新思考用于大语言模型遗忘的负偏好优化。*arXiv* 预印本 *arXiv:2410.07163*，2024年。

Chongyu Fan、Jinghan Jia、Yihua Zhang、Anil Ramakrishna、Mingyi Hong 和 Sijia Liu。迈向能够抵御再学习攻击的大语言模型遗忘：尖锐度感知最小化视角及更多探索。发表于第四十二届国际机器学习大会，2025。网址：https: //openreview.net/forum?id=zZjLv6F0Ks。

Jack Foster、Stefan Schoepf 和 Alexandra Brintrup。通过选择性突触减弱实现无需重新训练的快速机器遗忘。发表于*AAAI*人工智能大会论文集，第38卷，第12043–12051页，2024年。

Rohit Gandikota、Joanna Materzynska、Jaden Fiotto-Kaufman 和 David Bau。消除扩散模型中的概念。发表于*IEEE/CVF*国际计算机视觉大会论文集，第2426–2436页，2023年。

Chongyang Gao、Lixu Wang、Kaize Ding、Chenkai Weng、Xiao Wang 和 Qi Zhu。关于大语言模型的持续遗忘。*arXiv*预印本 *arXiv:2407.10223*，2024年。

Chongyang Gao、Lixu Wang、Kaize Ding、Chenkai Weng、Xiao Wang 和 Qi Zhu。关于大语言模型的持续遗忘。载于第十三届国际学习表征会议，2025年。网址：https://openreview.net/forum?id=Essg9kb4yx。

Antonio Ginart、Melody Guan、Gregory Valiant 和 James Y Zou。让人工智能忘记你：机器学习中的数据删除。神经信息处理系统进展，第32卷，2019年。Aditya Golatkar、Alessandro Achille 和 Stefano Soatto。永恒阳光照耀无瑕网络：深度网络中的选择性遗忘。载于*IEEE/CVF*计算机视觉与模式识别会议论文集，第9304–9312页，2020年。Roger Grosse、Juhan Bae、Cem Anil、Nelson Elhage、Alex Tamkin、Amirhossein Tajdini、Benoit Steiner、Dustin Li、Esin Durmus、Ethan Perez 等。利用影响函数研究大语言模型泛化。*arXiv*预印本 *arXiv:2308.03296*，2023年。Kang Gu、Md Rafi Ur Rashid、Najrin Sultana 和 Shagufta Mehnaz。二阶信息很重要：重新审视大语言模型的机器遗忘。*arXiv*预印本 *arXiv:2403.10557*，2024年。

Phillip Huang Guo、Aaquib Syed、Abhay Sheshadri、Aidan Ewart 和 Gintare Karolina Dzi- ugaite。机制性反学习：通过机制定位实现稳健的知识反学习与编辑。载于第四十二届国际机器学习大会，2025。网址：https://openreview.net/forum?id=92oBV5HAGl。

Anisa Halimi、Swanand Kadhe、Ambrish Rawat 和 Nathalie Baracaldo。联邦反学习：如何高效地擦除联邦学习中的一个客户端？*arXiv* 预印本 *arXiv:2207.05521*，2022。Dan Hendrycks、Collin Burns、Steven Basart、Andy Zou、Mantas Mazeika、Dawn Song 和 Jacob Steinhardt。衡量大规模多任务语言理解。载于国际学习表征会议。

Alex Henry、Prudhvi Raj Dachapally、Shubham Shantaram Pawar 和 Yuxuan Chen。Transformer 的查询-键归一化。载于 Trevor Cohn、Yulan He 和 Yang Liu（编），计算语言学协会论文集：*EMNLP 2020*，第 4246–4253 页，线上，2020年11月。计算语言学协会。DOI：10.18653/v1/2020.findings-emnlp.379。网址：https://aclanthology.org/2020.findings-emnlp.379/。

Shengyuan Hu、Yiwei Fu、Zhiwei Steven Wu和Virginia Smith。机器遗忘还是混淆？通过良性再学习唤起未遗忘大语言模型的记忆。*arXiv*预印本 *arXiv:2406.13356*，2024年。James Y Huang、Wenxuan Zhou、Fei Wang、Fred Morstatter、Sheng Zhang、Hoifung Poon和Muhao Chen。面向大语言模型的偏移反学习。*arXiv*预印本 *arXiv:2404.11045*，2024年a。Yue Huang、Lichao Sun、Haoran Wang、Siyuan Wu、Qihui Zhang、Yuan Li、Chujie Gao、Yixin Huang、Wenhan Lyu、Yixuan Zhang等。立场：TrustLLM：大语言模型中的可信度。载于国际机器学习大会，第20166–20270页。PMLR，2024年b。Gabriel Ilharco、Marco Tulio Ribeiro、Mitchell Wortsman、Suchin Gururangan、Ludwig Schmidt、Hannaneh Hajishirzi和Ali Farhadi。使用任务算术编辑模型。*arXiv*预印本 *arXiv:2212.04089*，2022年。Yoichi Ishibashi和Hidetoshi Shimodaira。大语言模型的知识净化。*arXiv*预印本 *arXiv:2309.11852*，2023年。Zachary Izzo、Mary Anne Smart、Kamalika Chaudhuri和James Zou。从机器学习模型中近似删除数据。载于国际人工智能与统计会议，第2008–2016页。PMLR，2021年。Joel Jang、Dongkeun Yoon、Sohee Yang、Sungmin Cha、Moontae Lee、Lajanugen Logeswaran和Minjoon Seo。用于缓解语言模型中隐私风险的知识反学习。*arXiv*预印本 *arXiv:2210.01504*，2022年。

Jiabao Ji、Yujian Liu、Yang Zhang、Gaowen Liu、Ramana R Kompella、Sijia Liu 和 Shiyu Chang。逆转遗忘集-保留集目标：一种基于对数几率差异的高效大语言模型遗忘框架。神经信息处理系统进展，37：12581–12611，2024年。

Jinghan Jia、Jiancheng Liu、Parikshit Ram、Yuguang Yao、Gaowen Liu、Yang Liu、Pranay Sharma 和 Sijia Liu。模型稀疏性可以简化机器遗忘。在第三十七届神经信息处理系统大会，2023年。

Jinghan Jia、Yihua Zhang、Yimeng Zhang、Jiancheng Liu、Bharat Runwal、James Diffenderfer、Bhavya Kailkhura 和 Sijia Liu。Soul：释放二阶优化在大语言模型遗忘中的潜力。载于*2024*年自然语言处理实证方法会议论文集，第 4276–4292 页，2024年。

Dongsheng Jiang、Yuchen Liu、Songlin Liu、Jin'e Zhao、Hao Zhang、Zhen Gao、Xiaopeng Zhang、Jin Li 和 Hongkai Xiong。从 CLIP 到 DINO：视觉编码器在多模态大语言模型中大放异彩。*arXiv* 预印本 *arXiv:2310.08825*，2023年。

Zhuoran Jin、Pengfei Cao、Chenhao Wang、Zhitao He、Hongbang Yuan、Jiachun Li、Yubo Chen、Kang Liu 和 Jun Zhao。RWKU：面向大语言模型真实世界知识遗忘的基准测试。神经信息处理系统进展，37:98213–98263，2024年。

Zachary Kenton、Tom Everitt、Laura Weidinger、Iason Gabriel、Vladimir Mikulik 和 Geoffrey Irving。语言智能体的对齐。*arXiv* 预印本 *arXiv:2103.14659*，2021年。

Pang Wei Koh 和 Percy Liang。通过影响函数理解黑箱预测。载于国际机器学习大会，第 1885–1894 页。PMLR，2017年。

Nupur Kumari、Bingliang Zhang、Sheng-Yu Wang、Eli Shechtman、Richard Zhang 和 Jun-Yan Zhu。消融文生图扩散模型中的概念。载于 *IEEE/CVF* 计算机视觉国际会议论文集，第 22691–22702 页，2023年。

Meghdad Kurmanji、Peter Triantafillou、Jamie Hayes 和 Eleni Triantafillou。迈向无界机器遗忘。神经信息处理系统进展，36：1957–1987，2023年。

Angeliki Lazaridou、Adhi Kuncoro、Elena Gribovskaya、Devang Agrawal、Adam Liska、Tayfun Terzi、Mai Gimenez、Cyprien de Masson d'Autume、Tomas Kocisky、Sebastian Ruder 等。注意差距：评估神经语言模型的时间泛化能力。神经信息处理系统进展，34：29348–29363，2021年。

Nathaniel Li、Alexander Pan、Anjali Gopal、Summer Yue、Daniel Berrios、Alice Gatti、Justin D Li、Ann-Kathrin Dombrowski、Shashwat Goel、Long Phan 等。WMDP基准：通过机器遗忘衡量并减少恶意使用。*arXiv* 预印本 *arXiv:2403.03218*，2024年。

Chris Liu、Yaxuan Wang、Jeffrey Flanigan 和 Yang Liu。通过嵌入损坏的提示实现大语言模型遗忘。神经信息处理系统进展，37：118198–118266，2024a。

Sijia Liu、Yuanshun Yao、Jinghan Jia、Stephen Casper、Nathalie Baracaldo、Peter Hase、Yuguang Yao、Chris Yuhao Liu、Xiaojun Xu、Hang Li 等。重新思考大语言模型的机器遗忘。《自然·机器智能》，第 1–14 页，2025a。

ZhenhuaLiu、TongZhu、Chuanyuan Tan 和 Wenliang Chen。学会拒绝：旨在降低大语言模型中的隐私风险。*arXiv* 预印本 *arXiv:2407.10058*，2024b。

Zhenhua Liu、Tong Zhu、Chuanyuan Tan 和 Wenliang Chen。学会拒绝：旨在降低大语言模型中的隐私风险。载于第 *31* 届国际计算语言学会议论文集，第 1683–1698 页，2025b。

Zheyuan Liu、Guangyao Dou、Zhaoxuan Tan、Yijun Tian 和 Meng Jiang。通过机器遗忘实现更安全的大语言模型。*arXiv* 预印本 *arXiv:2402.10058*，2024c。

Weikai Lu、Ziqian Zeng、Jianwei Wang、Zhengdong Lu、Zelin Chen、Huiping Zhuang 和 Cen Chen。Eraser：通过机器遗忘有害知识实现大语言模型的越狱防御。*arXiv* 预印本 *arXiv:2404.05880*，2024年。

Pratyush Maini、Zhili Feng、Avi Schwarzschild、Zachary Chase Lipton 和 J Zico Kolter。《TOFU：面向大语言模型的虚构式机器遗忘任务》。载于第一届语言建模会议，2024年。

Anmol Mekala、Vineeth Dorna、Shreya Dubey、Abhishek Lalwani、David Koleczek、Mukund Rungta、Sadid Hasan 和 Elita Lobo。用于大语言模型事实知识遗忘的交替偏好优化。*arXiv* 预印本 *arXiv:2409.13474*，2024年。Kevin Meng、Arnab Sen Sharma、Alex Andonian、Yonatan Belinkov 和 David Bau。Transformer 中的记忆批量编辑。*arXiv* 预印本 *arXiv:2210.07229*，2022年。Stephen Merity、Caiming Xiong、James Bradbury 和 Richard Socher。指针哨兵混合模型。载于国际学习表征会议，2017年。Sasi Kumar Murakonda、Reza Shokri 和 George Theodorakopoulos。量化学习高维图模型的隐私风险。载于国际人工智能与统计会议，第2287–2295页。PMLR，2021年。Andrei Muresanu、Anvith Thudi、Michael R Zhang 和 Nicolas Papernot。用于上下文学习的不可遗忘算法。*arXiv* 预印本 *arXiv:2402.00751*，2024年。Seth Neel、Aaron Roth 和 Saeed Sharifi-Malvajerdi。梯度下降式删除：基于梯度的机器遗忘方法。载于算法学习理论，第931–962页。PMLR，2021年。Stuart L Pardau。《加州消费者隐私法》：美国迈向欧洲式隐私制度。*J. Tech. L. & Pol'y*，23:68，2018年。Martin Pawelczyk、Seth Neel 和 Himabindu Lakkaraju。上下文内遗忘：作为少样本遗忘器的语言模型。*arXiv* 预印本 *arXiv:2310.07579*，2023年。Adam Poliak。将文本蕴涵识别作为自然语言处理评估的综述。*arXiv* 预印本 *arXiv:2010.03061*，2020年。

Rafael Rafailov、ArchitSharma、Eric Mitchell、Christopher D Manning、Stefano Ermon 和 Chelsea Finn。直接偏好优化：你的语言模型实际上是一个奖励模型。神经信息处理系统进展，36:53728–53741，2023年。

保护条例。《欧洲议会和欧盟理事会关于《欧盟条例 2016/679》的条例》。《欧盟条例》，679(2016):10–13，2016年。Nils Reimers 和 Iryna Gurevych。Sentence-BERT：使用孪生 BERT 网络的句子嵌入。*arXiv* 预印本 *arXiv:1908.10084*，2019年。

Jeffrey Rosen。被遗忘权。《斯坦福法律评论在线版》，64:88，2011年。

Mohammad Amin Roshani、Xiangyu Zhou、Yao Qiang、Srinivasan Suresh、Steven Hicks、Usha Sethuraman和Dongxiao Zhu。生成式大语言模型驱动的对话式人工智能应用用于个性化风险评估：新冠肺炎案例研究。*JMIR AI*，4(1):e67363，2025。Mansi Sakarvadia、Aswathy Ajith、Arham Mushtaq Khan、Nathaniel C Hudson、Caleb Geniesse、Kyle Chard、Yaoqing Yang、Ian Foster和Michael W. Mahoney。缓解语言模型中的记忆。见第十三届国际学习表征会议，2025。URL https://openreview.net/forum?id=MGKDBuyv4p。Ayush Sekhari、Jayadev Acharya、Gautam Kamath和Ananda Theertha Suresh。记住你想遗忘的内容：机器遗忘算法。神经信息处理系统进展，34:18075–18086，2021年。Weijia Shi、Anirudh Ajith、Mengzhou Xia、Yangsibo Huang、Daogao Liu、Terra Blevins、Danqi Chen和Luke Zettlemoyer。从大语言模型中检测预训练数据。*arXiv*预印本 *arXiv:2310.16789*，2023年。Weijia Shi、Jaechan Lee、Yangsibo Huang、Sadhika Malladi、Jieyu Zhao、Ari Holtzman、Daogao Liu、Luke Zettlemoyer、Noah A Smith和Chiyuan Zhang。MUSE：面向语言模型的机器遗忘六维评估。*arXiv*预印本 *arXiv:2407.06460*，2024年。Reza Shokri、Marco Stronati、Congzheng Song和Vitaly Shmatikov。针对机器学习模型的成员推断攻击。见*2017*年*IEEE*安全与隐私研讨会，第3–18页。IEEE，2017年。

Damien Sileo。tasksource：用于简化自然语言处理多任务学习与评估的数据集协调框架。*arXiv*预印本 *arXiv:2301.05948*，2023年。Rishub Tamirisa、Bhrugu Bharathi、Long Phan、Andy Zhou、Alice Gatti、Tarun Suresh、Maxwell Lin、Justin Wang、Rowan Wang、Ron Arel等。面向开放权重大型语言模型的防篡改防护措施。*arXiv*预印本 *arXiv:2408.00761*，2024年。Pratiksha Thaker、Yash Maurya、Shengyuan Hu、Zhiwei Steven Wu和Virginia Smith。大语言模型机器遗忘的护栏基线。*arXiv*预印本 *arXiv:2403.03329*，2024年。Bozhong Tian、Xiaozhuan Liang、Siyuan Cheng、Qingbin Liu、Mengru Wang、Dianbo Sui、Xi Chen、Huajun Chen和Ningyu Zhang。遗忘还是不遗忘？面向大语言模型的实用知识反学习。*arXiv*预印本 *arXiv:2407.01920*，2024年。Lewis Tunstall、Edward Beeching、Nathan Lambert、Nazneen Rajani、Kashif Rasul、Younes Belkada、Shengyi Huang、Leandro Von Werra、Cl´ementine Fourrier、Nathan Habib等。Zephyr：语言模型对齐的直接蒸馏。*arXiv*预印本 *arXiv:2310.16944*，2023年。Petar Veli˘ckovi´c、Christos Perivolaropoulos、Federico Barbero和Razvan Pascanu。softmax还不够（对于尖锐尺寸泛化）。*arXiv*预印本 *arXiv:2410.01104*，2024年。Bichen Wang、Yuzhe Zi、Yixin Sun、Yanyan Zhao和Bing Qin。RKLD：基于反向KL散度的知识蒸馏，用于大语言模型中个人信息的机器遗忘。*arXiv*预印本 *arXiv:2406.01983*，2024a。Boxin Wang、Weixin Chen、Hengzhi Pei、Chulin Xie、Mintong Kang、Chenhui Zhang、Chejian Xu、Zidi Xiong、Ritik Dutta、Rylan Schaeffer等。DecodingTrust：GPT模型可信度的综合评估。见*NeurIPS*，2023年。Junxiao Wang、Song Guo、Xin Xie和Heng Qi。通过类别判别式剪枝实现联邦反学习。见*2022*年*ACM*万维网大会论文集，第622–632页，2022年。Lingzhi Wang、Xingshan Zeng、Jinsong Guo、Kam-Fai Wong和Georg Gottlob。选择性遗忘：推进语言模型中的机器遗忘技术与评估。见*AAAI*人工智能会议论文集，第39卷，第843–851页，2025a。Qizhou Wang、Bo Han、Puning Yang、Jianing Zhu、Tongliang Liu和Masashi Sugiyama。面向大型语言模型遗忘方法的有效评估与比较。*arXiv*预印本 *arXiv:2406.09179*，2024b。Qizhou Wang、Jin Peng Zhou、Zhanke Zhou、Saebyeol Shin、Bo Han和Kilian Q Weinberger。重新思考大型语言模型遗忘目标：梯度视角及超越。见第十三届国际学习表征会议，2025b。URL https://openreview. net/forum?id=huo8MqVH6t。Yue Wang、Qizhou Wang、Feng Liu、Wei Huang、Yali Du、Xiaojiang Du和Bo Han。GRU：缓解大语言模型机器遗忘与保留之间的权衡。见第四十二届国际机器学习大会，2025c。URL https://openreview.net/forum?id=EAjhGr1Oeo。Alexander Warnecke、Lukas Pirch、Christian Wressnegger和Konrad Rieck。特征与标签的机器遗忘。*arXiv*预印本 *arXiv:2108.11577*，2021年。Laura Weidinger、John Mellor、Maribeth Rauh、Conor Griffin、Jonathan Uesato、Po-Sen Huang、Myra Cheng、Mia Glaese、Borja Balle、Atoosa Kasirzadeh等。语言模型造成伤害的伦理与社会风险。*arXiv*预印本 *arXiv:2112.04359*，2021 年。Kun Wu 、 Jie Shen 、 Yue Ning 、 Ting Wang 和 Wendy Hui Wang。面向图神经网络的可认证边遗忘。见第*29*届*ACM SIGKDD*知识发现与数据挖掘大会论文集，第2606–2617页，2023年。Abudukelimu Wuerkaixi、Qizhou Wang、Sen Cui、Wutong Xu、Bo Han、Gang Niu、Masashi Sugiyama和Changshui Zhang。面向持续性大型语言模型遗忘的知识否定自适应定位。见第四十二届国际机器学习大会，2025。URL https://openreview.net/forum?id=tcK4PV3VN4。

Jin Yao、Eli Chien、Minxin Du、Xinyao Niu、Tianhao Wang、Zezhou Cheng 和 Xiang Yue。预训练大语言模型的机器遗忘。*arXiv* 预印本 *arXiv:2402.15159*，2024年a。Yuanshun Yao、Xiaojun Xu 和 Yang Liu。大语言模型遗忘。*arXiv* 预印本 *arXiv:2310.10683*，2023年。Yuanshun Yao、Xiaojun Xu 和 Yang Liu。大语言模型遗忘。神经信息处理系统进展，37：105425–105475，2024年b。Xiaojian Yuan、Tianyu Pang、Chao Du、Kejiang Chen、Weiming Zhang 和 Min Lin。深入审视大语言模型的机器遗忘。*arXiv* 预印本 *arXiv:2410.08109*，2024年。Saleh Zare Zade、Yao Qiang、Xiangyu Zhou、Hui Zhu、Mohammad Amin Roshani、Prashant Khanduri 和 Dongxiao Zhu。大语言模型成员推理攻击的自动校准。*arXiv* 预印本 *arXiv:2505.03392*，2025。Shuangfei Zhai、Tatiana Likhomanenko、Etai Littwin、Dan Busbridge、Jason Ramapuram、Yizhe Zhang、Jiatao Gu 和 Joshua M Susskind。通过防止注意力熵坍塌来稳定 Transformer 训练。载于国际机器学习大会，第40770–40803页。PMLR，2023年。Dawen Zhang、Pamela Finckenberg-Broman、Thong Hoang、Shidong Pan、Zhenchang Xing、Mark Staples 和 Xiwei Xu。大语言模型时代的被遗忘权：影响、挑战与解决方案。*AI与伦理*，第1–10页，2024年a。Ruiqi Zhang、Licong Lin、Yu Bai 和 Song Mei。负偏好优化：从灾难性坍塌到有效遗忘。载于第一届语言建模会议，2024年b。Shengqiang Zhang、Xingxing Zhang、Hangbo Bao 和 Furu Wei。注意力温度对抽象式摘要蒸馏至关重要。载于Smaranda Muresan、Preslav Nakov 和 Aline Villavicencio（编），计算语言学协会第*60*届年会论文集（第*1*卷：长篇论文），第127–141页，都柏林，爱尔兰，2022年5月。计算语言学协会。DOI：10.18653/v1/2022.acl-long.11。URL：https://aclanthology.org/2022.acl-long.11/。Xuechen Zhang、Xiangyu Chang、Mingchen Li、Amit Roy-Chowdhury、Jiasi Chen 和 Samet Oymak。选择性注意力：通过系统化的上下文控制增强 Transformer。神经信息处理系统进展，37：11061–11086，2024年c。Kairan Zhao、Meghdad Kurmanji、George-Octavian B˘arbulescu、Eleni Triantafillou 和 Peter Triantafillou。机器遗忘为何困难，以及如何应对。神经信息处理系统进展，37：12293–12333，2024年。Xiangyu Zhou、Yao Qiang、Saleh Zare Zade、Prashant Khanduri 和 Dongxiao Zhu。通过对抗性上下文学习劫持大语言模型。*arXiv* 预印本 *arXiv:2311.09948*，2023年。Xiangyu Zhou、Yao Qiang、Saleh Zare Zade、Mohammad Amin Roshani、Prashant Khanduri、Douglas Zytko 和 Dongxiao Zhu。学习投毒大语言模型以操纵下游任务。*arXiv* 预印本 *arXiv:2402.13459*，2024年a。Xiangyu Zhou、Yao Qiang、Saleh Zare Zade、Douglas Zytko、Prashant Khanduri 和 Dongxiao Zhu。并非所有词元都注定要被遗忘。*arXiv* 预印本 *arXiv:2506.03142*，2025。Zhenhong Zhou、Haiyang Yu、Xinghua Zhang、Rongwu Xu、Fei Huang、Kun Wang、Yang Liu、Junfeng Fang 和 Yongbin Li。论注意力头在大语言模型安全性中的作用。*arXiv* 预印本 *arXiv:2410.13708*，2024年b。Yixiong Zou、Ran Ma、Yuhua Li 和 Ruixuan Li。注意力温度对基于视觉 Transformer 的跨域小样本学习至关重要。神经信息处理系统进展，37：116332–116354，2024年。

附录

A 证明

A.1 符号说明

设 V为一个有限词汇表，并令 θ 表示一个预训练的仅解码器语言模型的参数。对于任意输入–输出对（$x,\ y$），其中

$$x = (x_1, \ldots, x_L) \quad \text{and} \quad y = (y_1, \ldots, y_T),$$

模型定义了条件概率

$$p_\theta(y_t \mid x \circ y_{<t}), \qquad t = 1, \ldots, T.$$

对于任意 $(x,\ y) \in D_F$（遗忘集），我们将目标位置划分为

$$F \subseteq \{1, \ldots, T\} \quad \text{(factual positions)}, \qquad G = \{1, \ldots, T\} \setminus F \quad \text{(function positions)}.$$

事实位置对应编码了需要移除的不需要信息的词元，而功能位置则指序列中主要承担句法或结构作用的词元。

A.2 自注意力与温度

考虑一个仅含一个注意力头的 Transformer 层（为清晰起见，省略层索引和注意力头索引；该论证独立适用于每个注意力头）。对于位置 t，令 $q_t \in \mathbb{R}^d$ 表示查询向量，并令 $k_i, v_i \in \mathbb{R}^d$ 表示所有位置 $i \le t$对应的键向量和值向量。

注意力 logit 为

$$a_{t,i} := \frac{\langle q_t, k_i \rangle}{\sqrt{d}}, \qquad i = 1, \ldots, t,$$

标准注意力权重（温度设为 1）为

$$\alpha_{t,i} = \frac{\exp(a_{t,i})}{\sum_{j=1}^{t} \exp(a_{t,j})}.$$

对应的注意力输出为

$$z_t = \sum_{i=1}^{t} \alpha_{t,i} v_i.$$

我们引入温度参数 $\tau \ge 1$，并定义平滑后的注意力权重

$$\alpha_{t,i}(\tau) = \frac{\exp(a_{t,i}/\tau)}{\sum_{j=1}^{t} \exp(a_{t,j}/\tau)},$$

以及注意力输出

$$z_t(\tau) = \sum_{i=1}^{t} \alpha_{t,i}(\tau) v_i.$$

当 $\tau = 1$时，模型恢复基础注意力：$\alpha_{t,\ i}(1) = \alpha_{t,\ i}$ 以及 $z_t(1) = z_t$。对于 $\tau > 1$，由于所有对数几率差均按 $1/\tau$缩放，分布 $\alpha_t(\tau)$会比 α_t更加平坦。

我们将注意力平滑教师模型 θ_τ定义为：在所有注意力头中应用温度 τ，同时保持其他所有组件（前馈层、层归一化和输出投影）不变后得到的模型。

A.3 输出层与词元概率

令 $W \in \mathbb{R}^{|V|\times d}$ 和 $b \in \mathbb{R}^{|V|}$ 分别表示输出投影矩阵和偏置。在位置 t处，令 h_t 表示 Transformer 生成的隐藏表示（该表示通过后续层融入了注意力输出 z_t）。

对于每个词元 $w \in V$，模型会计算其对数几率

$$\ell_t(w;\theta) = \langle W_w, h_t \rangle + b_w,$$

以及相应的条件概率

$$p_\theta(w \mid x \circ y_{<t}) = \frac{\exp(\ell_t(w;\theta))}{\sum_{u\in V}\exp(\ell_t(u;\theta))}.$$

对于注意力平滑教师模型 θ_τ，仅在注意力机制内部应用温度 τ，会得到修改后的隐藏状态 $h_t(\tau)$，并由此产生逻辑值

$$\ell_t(w;\theta_\tau)$$

以及词元概率

$$p_{\theta_\tau}(w \mid x \circ y_{<t}).$$

A.4 “遗忘”与“流畅度”的概念

我们定义了两项关注的属性：移除特定事实内容，以及保留正常语言行为。

A.4.1 遗忘

给定一个遗忘样例（$x,\ y$） $\in D_F$和一个事实位置 $t \in F$。令 $y_t^\star$表示要移除的事实词元（例如，TOFU 中正确的实体名称）。
如果平滑模型 θ_τ 在位置 t 满足以下条件，我们就称该模型遗忘了这一事实：

$$p_{\theta_\tau}(y_t^\star \mid x \circ y_{<t}) \le \epsilon_F,$$

对于某个较小的阈值 $\epsilon_F > 0$ （大致相当于在合理实体中随机猜测的准确率）。

在序列层面，当平均值

$$-\log p_{\theta_\tau}(y_t^\star \mid x \circ y_{<t})$$

在所有（$x,\ y$） $\in D_F$和所有 $t \in F$ 上至少达到目标值 L_F 时，遗忘集便能实现遗忘，这意味着模型对事实性词元赋予较低的概率。

A.4.2 流畅度

对于功能位置 $t \in G$，我们要求模型继续对正确的功能词元赋予较高概率，因为这些词元体现了语法和结构。
如果 θ_τ 在（$x,\ y$）上保留了流畅度，我们称

$$-\log p_{\theta_\tau}(y_t \mid x \circ y_{<t}) \le -\log p_\theta(y_t \mid x \circ y_{<t}) + \delta_G, \qquad t \in G,$$

对于容差 $\delta_G > 0$。
在序列层面，若功能词元上的平均交叉熵增幅至多为 δ_G，则流畅度得以保持。

A.5 假设

为了表明注意力平滑能够在保持正常语言行为的同时移除特定事实，我们引入了关于事实性词元和功能词元如何依赖于注意力的结构性假设。

A.5.1 假设 A1（事实词元需要精确注意力）

对于每一个事实位置 $t \in F$，存在一个子集 $S_t \subseteq \{1, \dots, t\}$，且满足 $|S_t| \ll t$，使得基础注意力将其大部分质量置于 S_t:

$$\sum_{i \in S_t} \alpha_{t,i} \geq \gamma, \qquad \gamma \in (0, 1).$$

位置 S_t 处的值向量包含支持事实词元 $y_t^\star$的主要信号，而 S_t 之外的位置对这一信号的贡献较小。

定义

$$z_t = \sum_{i=1}^{t} \alpha_{t,i} v_i, \qquad z_t(S_t) := \frac{1}{|S_t|} \sum_{i \in S_t} v_i, \qquad z_t(\bar{S}_t) := \frac{1}{t - |S_t|} \sum_{i \notin S_t} v_i,$$

and

$$z_t^{\text{unif}} := \frac{1}{t} \sum_{i=1}^{t} v_i = \frac{|S_t|}{t} z_t(S_t) + \frac{t - |S_t|}{t} z_t(\bar{S}_t).$$

我们假设，将注意力集中在 S_t 上会增大事实性词元的 logit 间隔，而均匀注意力混合会使该间隔至少减少 $m_F > 0$:

$$\langle W_{y_t^\star} - W_u,\ z_t - z_t^{\text{unif}} \rangle \geq m_F \qquad \text{for all } u \neq y_t^\star.$$

因此，将注意力分布从集中于 S_t 的尖锐模式转向更平坦（均匀）的分布，会降低事实词元 $y_t^\star$的 logit 裕度。

A.5.2 假设 A2（功能词元的注意力敏感度较低）

对于每个功能位置 $t \in G$，我们假设正确词元 y_t 取决于值向量的广泛混合，而非少数几个位置。换言之，预测 y_t 并不依赖尖锐的注意力模式。

形式上，令 $\ell_t(y_t; \theta; z)$ 表示当位置 t 的注意力输出为 z时， y_t 的对数几率。假设该对数几率关于 z 平滑，并满足 L-利普希茨界:

$$|\ell_t(y_t; \theta; z) - \ell_t(y_t; \theta; z')| \leq L \, \|z - z'\|_2 \qquad \text{for all } z, z'.$$

我们还假设， $\{v_i\}_{i \leq t}$ 的凸组合在 W_{y_t}方向上不会发生剧烈变化。因此，将注意力权重从较尖锐的模式转向较平滑的模式（例如，更接近均匀分布），只会使 $\ell_t(y_t)$发生微小变化。

A.5.3 假设 A3（事实词元的逻辑值非退化）

对于每个事实位置 $t \in F$，基础模型都会为正确的事实词元$y_t^\star$赋予明确的间隔。形式上，

$$\ell_t(y_t^\star; \theta) - \max_{u \neq y_t^\star} \ell_t(u; \theta) \geq \Delta_F > 0.$$

这确保基础模型中的事实回忆有正的 logit 间隔作为支撑。

A.5.4 假设 A4（连续性）

对于每个位置 t，注意力平滑隐藏状态 $h_t(\tau)$ 和逻辑值 $\ell_t(w;\ \theta_\tau)$ 随温度参数 τ 连续变化。

对于标准 Transformer 层，这一性质成立，因为注意力、线性变换和激活函数都是连续的。

A.6 引理

A.6.1 引理 1（温度变化时注意力变化的方向）

固定一个位置 t 和注意力 logit $a_{t,1}, \ldots, a_{t,t} \in \mathbb{R}$。对于 $\tau > 0$，定义温度缩放的注意力权重

$$\alpha_{t,i}(\tau) = \frac{\exp(a_{t,i}/\tau)}{\sum_{j=1}^{t} \exp(a_{t,j}/\tau)}, \qquad i = 1, \ldots, t.$$

Let

$$\bar{a}_t(\tau) := \sum_{j=1}^{t} \alpha_{t,j}(\tau)\, a_{t,j}$$

表示注意力分布 $\alpha_t(\tau)$ 在位置 t 处的平均 logit。

那么，对于每个 $i \in \{1, \ldots, t\}$,

$$\frac{\partial}{\partial \tau}\, \alpha_{t,i}(\tau) = \frac{1}{\tau^2}\, \alpha_{t,i}(\tau)\, \big(\bar{a}_t(\tau) - a_{t,i}\big).$$

特别地：

- If $a_{t,i} > \bar{a}_t(\tau)$, then $\frac{\partial}{\partial \tau}\alpha_{t,i}(\tau) < 0$, so the attention on index i *decreases* as τ increases.
- If $a_{t,i} < \bar{a}_t(\tau)$, then $\frac{\partial}{\partial \tau}\alpha_{t,i}(\tau) > 0$, so the attention on index i *increases* as τ increases.
- If $a_{t,i} = \bar{a}_t(\tau)$, then $\frac{\partial}{\partial \tau}\alpha_{t,i}(\tau) = 0$.

证明。

对于固定的 t，记 $a_i := a_{t,i}$ 和 $\alpha_i(\tau) := \alpha_{t,i}(\tau)$ 表示 $i = 1, \ldots, t$，以简化符号。根据定义，

$$\alpha_i(\tau) = \frac{\exp(a_i/\tau)}{Z(\tau)}, \qquad Z(\tau) := \sum_{j=1}^{t} \exp(a_j/\tau).$$

我们首先对 τ 求对数注意力的导数：

$$\log \alpha_i(\tau) = \frac{a_i}{\tau} - \log Z(\tau).$$

求导得

$$\frac{\partial}{\partial \tau} \log \alpha_i(\tau) = -\frac{a_i}{\tau^2} - \frac{1}{Z(\tau)} \frac{\partial Z(\tau)}{\partial \tau}.$$

接下来计算 $\frac{\partial Z(\tau)}{\partial \tau}$。根据链式法则，

$$\frac{\partial Z(\tau)}{\partial \tau} = \sum_{j=1}^{t} \frac{\partial}{\partial \tau} \exp(a_j/\tau) = \sum_{j=1}^{t} \exp(a_j/\tau) \left(-\frac{a_j}{\tau^2}\right) = -\frac{1}{\tau^2} \sum_{j=1}^{t} a_j \exp(a_j/\tau).$$

Thus

$$\frac{1}{Z(\tau)} \frac{\partial Z(\tau)}{\partial \tau} = -\frac{1}{\tau^2} \frac{\sum_{j=1}^{t} a_j \exp(a_j/\tau)}{\sum_{k=1}^{t} \exp(a_k/\tau)} = -\frac{1}{\tau^2} \sum_{j=1}^{t} \alpha_j(\tau)\, a_j = -\frac{1}{\tau^2}\, \bar{a}(\tau),$$

其中

$$\bar{a}(\tau) := \sum_{j=1}^{t} \alpha_j(\tau)\, a_j$$

是在当前注意力分布下的平均 logit。

将其代回 $\log \alpha_i(\tau)$ 的导数中，可得

$$\frac{\partial}{\partial \tau} \log \alpha_i(\tau) = -\frac{a_i}{\tau^2} + \frac{1}{\tau^2}\, \bar{a}(\tau) = \frac{1}{\tau^2}\, \big(\bar{a}(\tau) - a_i\big).$$

最后，我们从对数注意力的导数转向注意力本身的导数。由于

$$\frac{\partial}{\partial\tau}\alpha_i(\tau) = \alpha_i(\tau)\,\frac{\partial}{\partial\tau}\log\alpha_i(\tau),$$

我们得到

$$\frac{\partial}{\partial\tau}\alpha_i(\tau) = \alpha_i(\tau)\,\frac{1}{\tau^2}\big(\bar{a}(\tau) - a_i\big) = \frac{1}{\tau^2}\,\alpha_i(\tau)\,\big(\bar{a}(\tau) - a_i\big),$$

这就是所要证明的公式。

符号结论可直接推出：

- If $a_i > \bar{a}(\tau)$, then $\bar{a}(\tau) - a_i < 0$, so $\frac{\partial}{\partial\tau}\alpha_i(\tau) < 0$ and the attention on i decreases with τ.
- If $a_i < \bar{a}(\tau)$, then $\bar{a}(\tau) - a_i > 0$, so $\frac{\partial}{\partial\tau}\alpha_i(\tau) > 0$ and the attention on i increases with τ.
- If $a_i = \bar{a}(\tau)$, then $\frac{\partial}{\partial\tau}\alpha_i(\tau) = 0$.

这表明，提高温度会将概率质量从高于平均值的对数几率位置转移到低于平均值的对数几率位置，这与注意力分布变得更加平坦的直观图景相符。□

A.6.2 引理 2（提高温度会增加注意力熵）

固定一个位置 t 和注意力 logit $a_{t,1}$、……、$a_{t,t} \in \mathbb{R}$。对于 $\tau > 0$，定义

$$\alpha_{t,i}(\tau) = \frac{\exp(a_{t,i}/\tau)}{\sum_{j=1}^{t}\exp(a_{t,j}/\tau)}, \qquad i = 1, \ldots, t,$$

以及熵

$$H_t(\tau) := -\sum_{i=1}^{t}\alpha_{t,i}(\tau)\,\log\alpha_{t,i}(\tau).$$

Let

$$\bar{a}_t(\tau) := \sum_{j=1}^{t}\alpha_{t,j}(\tau)\,a_{t,j}$$

表示在 $\alpha_t(\tau)$ 下位置 t 处的平均 logit。于是

$$\frac{\partial}{\partial\tau}H_t(\tau) = \frac{1}{\tau^3}\,\mathrm{Var}_{\alpha_t(\tau)}\big(a_{t,1}, \ldots, a_{t,t}\big) \;\geq\; 0,$$

当逻辑值 $a_{t,1}$、……、$a_{t,t}$ 不全相等时，不等号严格成立。

证明。

固定 t，并记 $a_i := a_{t,i}$ 以及 $\alpha_i(\tau) := \alpha_{t,i}(\tau)$，其中 $i = 1$、……、t，以简化符号。令

$$Z(\tau) := \sum_{j=1}^{t}\exp(a_j/\tau), \qquad \bar{a}(\tau) := \sum_{j=1}^{t}\alpha_j(\tau)\,a_j.$$

根据定义，

$$\alpha_i(\tau) = \frac{\exp(a_i/\tau)}{Z(\tau)}, \qquad \log\alpha_i(\tau) = \frac{a_i}{\tau} - \log Z(\tau).$$

因此，熵可以写成

$$H(\tau) := -\sum_{i=1}^{t}\alpha_i(\tau)\,\log\alpha_i(\tau) = -\sum_{i=1}^{t}\alpha_i(\tau)\Big(\frac{a_i}{\tau} - \log Z(\tau)\Big).$$

使用 $\sum_i \alpha_i(\tau) = 1$，可简化为

$$H(\tau) = -\frac{1}{\tau}\sum_{i=1}^{t}\alpha_i(\tau)a_i + \log Z(\tau) = -\frac{1}{\tau}\,\bar{a}(\tau) + \log Z(\tau).$$

对 $H(\tau)$ 关于 τ 求导：

$$\frac{\partial H}{\partial \tau} = \frac{1}{\tau^2}\,\bar{a}(\tau) - \frac{1}{\tau}\,\frac{\partial \bar{a}(\tau)}{\partial \tau} + \frac{1}{Z(\tau)}\,\frac{\partial Z(\tau)}{\partial \tau}.$$

我们计算 $\frac{\partial Z(\tau)}{\partial \tau}$：

$$\frac{\partial Z(\tau)}{\partial \tau} = \sum_{j=1}^{t} \frac{\partial}{\partial \tau}\exp(a_j/\tau) = \sum_{j=1}^{t} \exp(a_j/\tau)\Big(-\frac{a_j}{\tau^2}\Big) = -\frac{1}{\tau^2}\sum_{j=1}^{t} a_j \exp(a_j/\tau).$$

因此

$$\frac{1}{Z(\tau)}\frac{\partial Z(\tau)}{\partial \tau} = -\frac{1}{\tau^2}\,\frac{\sum_{j=1}^{t} a_j \exp(a_j/\tau)}{\sum_{k=1}^{t}\exp(a_k/\tau)} = -\frac{1}{\tau^2}\sum_{j=1}^{t}\alpha_j(\tau)\,a_j = -\frac{1}{\tau^2}\,\bar{a}(\tau).$$

将其代入 $\partial H/\partial \tau$ 的表达式，得到

$$\frac{\partial H}{\partial \tau} = \frac{1}{\tau^2}\,\bar{a}(\tau) - \frac{1}{\tau}\,\frac{\partial \bar{a}(\tau)}{\partial \tau} - \frac{1}{\tau^2}\,\bar{a}(\tau) = -\frac{1}{\tau}\,\frac{\partial \bar{a}(\tau)}{\partial \tau}.$$

我们现在计算 $\frac{\partial \bar{a}(\tau)}{\partial \tau}$。根据定义，

$$\bar{a}(\tau) = \sum_{i=1}^{t}\alpha_i(\tau)\,a_i, \quad \text{so} \quad \frac{\partial \bar{a}(\tau)}{\partial \tau} = \sum_{i=1}^{t} a_i\,\frac{\partial \alpha_i(\tau)}{\partial \tau}.$$

由引理 1，对每个 i，我们有

$$\frac{\partial \alpha_i(\tau)}{\partial \tau} = \frac{1}{\tau^2}\,\alpha_i(\tau)\,\big(\bar{a}(\tau) - a_i\big).$$

因此，

$$\frac{\partial \bar{a}(\tau)}{\partial \tau} = \frac{1}{\tau^2}\sum_{i=1}^{t} a_i\,\alpha_i(\tau)\,\big(\bar{a}(\tau) - a_i\big) = \frac{1}{\tau^2}\Big(\bar{a}(\tau)\sum_{i=1}^{t}\alpha_i(\tau)a_i - \sum_{i=1}^{t}\alpha_i(\tau)a_i^2\Big).$$

由于 $\sum_i \alpha_i(\tau)a_i = \bar{a}(\tau)$，这变为

$$\frac{\partial \bar{a}(\tau)}{\partial \tau} = \frac{1}{\tau^2}\Big(\bar{a}(\tau)^2 - \sum_{i=1}^{t}\alpha_i(\tau)a_i^2\Big) = -\frac{1}{\tau^2}\Big(\sum_{i=1}^{t}\alpha_i(\tau)a_i^2 - \bar{a}(\tau)^2\Big).$$

括号中的项是 $\alpha(\tau)$ 下逻辑值的方差：

$$\mathrm{Var}_{\alpha(\tau)}(a) := \sum_{i=1}^{t}\alpha_i(\tau)a_i^2 - \bar{a}(\tau)^2.$$

因此

$$\frac{\partial \bar{a}(\tau)}{\partial \tau} = -\frac{1}{\tau^2}\,\mathrm{Var}_{\alpha(\tau)}(a).$$

代入 $\partial H/\partial \tau$ 可得

$$\frac{\partial H}{\partial \tau} = -\frac{1}{\tau}\Big(-\frac{1}{\tau^2}\,\mathrm{Var}_{\alpha(\tau)}(a)\Big) = \frac{1}{\tau^3}\,\mathrm{Var}_{\alpha(\tau)}(a).$$

由于方差始终非负，且仅当所有逻辑值 a_1，……，a_t 相等时才等于零，我们得到

$$\frac{\partial H}{\partial \tau} \geq 0,$$

当 $a_1, \ldots, a_t$ 不全相等时，严格不等式成立。这证明，除非注意力已经是均匀分布，否则注意力熵会随温度升高而增加。□

A.6.3 引理 3（对事实性词元逻辑值的影响）

固定一个事实位置 $t \in F$。在假设 A1 和 A3 下，当 τ 增大至超过 1 时，事实词元的 logit 裕度会降低；当 τ足够大时，该裕度会变为负值。更准确地说，存在 $\tau_F \geq 1$，使得对于所有 $\tau \geq \tau_F$,

$$\ell_t(y_t^\star;\theta_\tau) - \max_{u \neq y_t^\star} \ell_t(u;\theta_\tau) \leq -\frac{m_F}{2} < 0.$$

证明。
根据 A1，位置 t 处的大部分基础注意力质量落在集合 S_t 上，其值向量增强了 $y_t^\star$ 的对数几率。基础注意力输出可以写为

$$z_t = \sum_{i=1}^{t} \alpha_{t,i} v_i = \sum_{i \in S_t} \alpha_{t,i} v_i + \sum_{i \notin S_t} \alpha_{t,i} v_i.$$

对于较大的 τ，引理 2 表明，当 $\tau \to \infty$ 时，$\alpha_{t,i}(\tau)$ 趋近于均匀分布。Hen ce

$$z_t(\tau) = \sum_{i=1}^{t} \alpha_{t,i}(\tau) v_i \xrightarrow[\tau\to\infty]{} \frac{1}{t}\sum_{i=1}^{t} v_i = \lambda_S z_t(S_t) + \lambda_{\bar{S}_t} z_t(\bar{S}_t),$$

对于由 $|S_t|$ 和 t 确定的权重 $\lambda_S, \lambda_{\bar{S}}$。定义该变化

$$\Delta z_t(\tau) := z_t(\tau) - z_t.$$

根据引理 1，随着 τ 增大，Δz_t（τ）会使注意力输出逐渐远离 $^-$偏好的尖锐模式，并趋向于注意力权重低于熵的 S_t。
设 $u \neq y_t^\star$ 为任意一个竞争词元。温度为 τ 时的对数几率差为

$$\ell_t(y_t^\star;\theta_\tau) - \ell_t(u;\theta_\tau) = \langle W_{y_t^\star} - W_u,\, z_t(\tau)\rangle + (b_{y_t^\star} - b_u).$$

减去 $\tau = 1$处的差值，得到

$$\Delta_\tau = \langle W_{y_t^\star} - W_u,\, z_t(\tau) - z_t\rangle.$$

根据 A1，增加 S_t 的权重会增大事实性间隔，因此远离 S_t（正如平滑操作所做的那样）会使其减小。因此，当 τ 足够大时，上述内积为负；一旦 α_t（τ）接近均匀分布，该内积就可以被一个负常数作为上界。

根据 A3，在 $\tau = 1$ 处，事实性间隔为正：

$$\ell_t(y_t^\star;\theta) - \ell_t(u;\theta) \geq \Delta_F > 0.$$

根据关于 τ 的连续性（A4），随着注意力模式逐渐平滑，间隔也会连续减小。由于当 τ 足够大时，间隔会变为负数，介值定理保证存在一点 τ_F，使间隔在此处穿过零点。对于任意 $\tau \geq \tau_F$，间隔都严格为负；通过调整阈值，我们可以确保界限 $-m_F/2$。

这意味着，对于 $\tau \geq \tau_F$,

$$p_{\theta_\tau}(y_t^\star \mid x \circ y_{<t}) \leq \frac{1}{1+\exp(m_F/2)} =: \epsilon_F < \frac{1}{2},$$

因此，事实词元不再是最可能的输出。 □

A.6.4 引理 4（对功能词元 logit 的影响很小）

固定一个功能位置 $t \in G$。在假设 A2 和 A4 下，对于任意 $\eta > 0$，存在 $\bar{\tau}_G \geq 1$，使得对于所有 $\tau \in [1, \bar{\tau}_G]$,

$$\left|\ell_t(y_t;\theta_\tau) - \ell_t(y_t;\theta)\right| \leq \eta.$$

证明。

对于任意紧区间 $[1, \bar{\tau}_G]$，注意力权重 $\alpha_t(\tau)$ 随 τ 连续变化，并且始终位于单纯形内。因此，$z_t(\tau)$ 是 τ 的连续函数。

根据 A2，正确功能词元的对数几率关于注意力输出是 L-利普希茨的：

$$\left|\ell_t(y_t;\theta_\tau) - \ell_t(y_t;\theta)\right| \le L\,\|z_t(\tau) - z_t\|_2.$$

由于 $z_t(\tau) \to z_t$ as $\tau \to 1$（根据 A4 和注意力映射的连续性），对于任意 $\eta > 0$，我们可以选择 $\bar{\tau}_G > 1$，使得

$$\|z_t(\tau) - z_t\|_2 \le \eta/L \qquad \text{for all } \tau \in [1, \bar{\tau}_G].$$

代入利普希茨界可得

$$\left|\ell_t(y_t;\theta_\tau) - \ell_t(y_t;\theta)\right| \le \eta,$$

由此得证。 □

A.6.5 引理 5（对数几率变化较小意味着交叉熵变化较小）

设 p 和 q 是关于 V 的两个分布，它们在真实词元处的逻辑值之差至多为 η。那么，该词元处的负对数似然增量至多为函数 $c(\eta)$，且当 $\eta \to 0$ 时，$c(\eta) \to 0$。

形式化地，设 ℓ_θ 和 ℓ_{θ_τ} 是两个对数几率向量。若对于某个词元 w，

$$\left|\ell_{\theta_\tau}(w) - \ell_\theta(w)\right| \le \eta,$$

且其余对数几率差均有统一上界，则

$$-\log p_{\theta_\tau}(w) \;\le\; -\log p_\theta(w) + c(\eta).$$

证明概要。从逻辑值到概率的 softmax映射在对数几率空间的任意紧区域上都是光滑且满足利普希茨条件的，固定词元的负对数概率也同样光滑。因此，对数几率的微小变化会导致负对数似然的微小变化。函数 $c(\eta)$由 softmax 和对数运算的利普希茨常数确定。□

结合引理 4 和引理 5，对于任意容差 $\delta_G > 0$，我们可以选择 $\bar{\tau}_G > 1$，使得对于所有$\tau \in [1, \bar{\tau}_G]$ 以及所有功能词元 $t \in G$，

$$-\log p_{\theta_\tau}(y_t \mid x \circ y_{<t}) \le -\log p_\theta(y_t \mid x \circ y_{<t}) + \delta_G.$$

A.7 主定理

定理：注意力平滑在保持流畅度的同时实现遗忘。

假设 A1–A4 对所有遗忘样本 $(x, y) \in D_F$及其事实位置和功能位置成立。则存在一个温度区间 $[\tau_0, \tau_1]$，且

$$1 < \tau_0 \le \tau_1 < \infty$$

使得：

- **遗忘：**对于所有 $\tau \in [\tau_0, \tau_1]$，平滑模型 θ_τ会遗忘 D_F 中的事实性词元。特别地，对于每个事实位置 $t \in F$，

  $$p_{\theta_\tau}(y_t^\star \mid x \circ y_{<t}) \le \epsilon_F,$$

  且平均事实负对数似然至少为 $L_F > 0$。
- **流畅度：** 对于所有 $\tau \in [\tau_0, \tau_1]$，功能词元上的平均损失（在$D_F$和 D_R上）增加幅度至多为 δ_G：

  $$\frac{1}{|G|}\sum_{t\in G} -\log p_{\theta_\tau}(y_t \mid x \circ y_{<t}) \le \frac{1}{|G|}\sum_{t\in G} -\log p_\theta(y_t \mid x \circ y_{<t}) + \delta_G.$$

因此，存在一个不可忽略的温度范围，在此范围内，事实知识会被遗忘，而流利的语言行为得以保留。

证明。
步骤 *1*：在足够高的 τ。对于每个事实位置 $t \in F$，引理 3 给出了一个温度 $\tau_F(t)$，使得对于所有 $\tau \geq \tau_F(t)$，

$$\ell_t(y_t^\star; \theta_\tau) - \max_{u \neq y_t^\star} \ell_t(u; \theta_\tau) \leq -\frac{m_F}{2}.$$

因此

$$p_{\theta_\tau}(y_t^\star \mid x \circ y_{<t}) \leq \epsilon_F,$$

其中 $\epsilon_F < 1/2$ 仅取决于 m_F。

定义

$$\tau_F := \max_{(x,y) \in D_F} \max_{t \in F} \tau_F(t).$$

对于所有 $\tau \geq \tau_F$，每个遗忘样例中的每个事实位置都满足遗忘不等式，且平均事实损失至少为 $L_F = -\log \epsilon_F$。
步骤 *2*：当 τ 足够小时的流畅度。对于每个功能位置 $t \in G$，引理 4 指出，对于任意 $\eta > 0$，存在 $\bar{\tau}_G(t) > 1$，使得对于所有 $\tau \in [1, \bar{\tau}_G(t)]$，

$$\left|\ell_t(y_t; \theta_\tau) - \ell_t(y_t; \theta)\right| \leq \eta.$$

Lem引理 5 随后保证，每个功能词元上的额外损失至多为 $c(\eta)$，其中 $c(\eta) \to 0$
因为$\eta \to$

选择 η，使得 $c(\eta) \leq \delta_G$，并定义

$$\bar{\tau}_G := \min_{(x,y)} \min_{t \in G} \bar{\tau}_G(t).$$

对于所有 $\tau \in [1, \bar{\tau}_G]$，

$$-\log p_{\theta_\tau}(y_t \mid x \circ y_{<t}) \leq -\log p_\theta(y_t \mid x \circ y_{<t}) + \delta_G,$$

对所有功能词元取平均，即可得到定理中的界。

步骤 *3*：确定共同温度范围。我们有：

- 对于所有 $\tau \geq \tau_F$，遗忘都成立。
- 对于所有 $\tau \in [1, \bar{\tau}_G]$，流畅度都成立。

对于所有以下范围，两者同时成立

$$\tau \in [\tau_F, \bar{\tau}_G],$$

只要 $\tau_F \leq \bar{\tau}_G$，该范围就非空。

这一条件体现了 A1–A3 中的结构：事实性词元依赖精确的注意力模式，这些模式在平滑处理后会迅速崩塌；而功能词元依赖更广泛的模式，这些模式在轻度平滑下仍保持稳定。

任取满足条件的 τ_0，τ_1

$$1 < \tau_F \leq \tau_0 \leq \tau_1 \leq \bar{\tau}_G < \infty.$$

于是，对所有 $\tau \in [\tau_0, \tau_1]$，遗忘和流畅度均成立。 □

B 基线方法

符号说明。 令 $P(y \mid x; \theta)$ 表示在参数为 θ 的模型下，给定输入 x 时输出序列 $y = (y_1, \ldots, y_T)$ 的概率。该概率定义如下：

$$P(y \mid x; \theta) = \prod_{t=1}^{T} p(y_t \mid x \circ y_{<t}; \theta)^{\frac{1}{T}} .$$

遗忘损失。 现有方法大致可分为基于收敛的遗忘 和基于散度的遗忘。我们采用的基线方法如下：

• **梯度上升（GA）**（Yao 等，2023）通过最大化遗忘集上的预测损失，有效地反转训练目标：

$$\mathcal{L}_{\text{GA}}(\mathcal{D}_{\text{F}}; \theta) = -\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{F}}} \left[\frac{1}{T} \sum_{t=1}^{T} -\log p(y_t \mid x \circ y_{<t}; \theta) \right]. \tag{6}$$

• **负偏好优化（NPO）**（Zhang 等，2024b）源自直接偏好优化（DPO）（Rafailov 等，2023）。它将遗忘集中的答案视为负样本，同时省略正向项：

$$\mathcal{L}_{\text{NPO}}(\mathcal{D}_{\text{F}}; \theta) = -\frac{2}{\beta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{F}}} \left[\log \sigma \left(-\beta \log \frac{P(y \mid x; \theta)}{P(y \mid x; \theta_{\text{Base}})} \right) \right], \tag{7}$$

其中 $\sigma(t) = 1/(1 + e^{-t})$， β 是一个超参数， θ_{ref} 是固定的参考模型。NPO 可视为采用自适应梯度缩放的梯度上升（GA）（Zhang 等，2024b）。

• **最大化熵（ME）**（Yuan 等，2024）最小化每个词元的预测分布与词汇表大小对应的均匀分布之间的 KL 散度。

$$\mathcal{L}_{\text{ME}}(\mathcal{D}_{\text{F}}; \theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{F}}} \left[\frac{1}{T} \sum_{t=1}^{T} \text{KL}\big(\mathcal{U}_{[K]} || p(\cdot \mid x \circ y_{<t}; \theta)\big) \right], \tag{8}$$

其中 $\mathcal{U}_{[K]}$ 是词汇表大小为 K 的均匀分布，每个值均为$1/K$。

• **IDK 微调**（Maini 等，2024）通过用来自 $\mathcal{D}_{\text{IDK}}$ 的随机回复重新标注遗忘集问题，将机器遗忘重构为指令微调；IDK 是一组拒绝模板（例如，“抱歉，我不知道。”）。其损失为

$$\mathcal{L}_{\text{IDK}}(\mathcal{D}_{\text{F}}, \mathcal{D}_{\text{IDK}}; \theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{F}}, y\sim\mathcal{D}_{\text{IDK}}} \left[-\log P(y \mid x; \theta) \right]. \tag{9}$$

• **直接偏好优化（DPO）**（Zhang 等，2024b）应用标准 DPO 损失（Rafailov 等，2023），将遗忘集答案作为负样本，并将来自 $\mathcal{D}_{\text{IDK}}$ 的拒绝模板作为正样本。

$$\mathcal{L}_{\text{DPO}}(\mathcal{D}_{\text{F}}, \mathcal{D}_{\text{IDK}}; \theta; \theta_{\text{ref}}) = -\frac{1}{\beta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{F}}, y'\sim\mathcal{D}_{\text{IDK}}} \left[\log \sigma \left(\beta \log \frac{P(y' \mid x; \theta)}{P(y' \mid x; \theta_{\text{base}})} - \beta \log \frac{P(y \mid x; \theta)}{P(y \mid x; \theta_{\text{base}})} \right) \right], \tag{10}$$

其中 θ_{base} 表示参考模型的参数，该参考模型是用于机器遗忘的初始基础模型。

• **SimNPO** （Fan 等，2024）。它源自 NPO，其奖励函数由与参考模型的比较给出。相比之下，SimNPO 采用无参考但长度归一化的奖励形式，因此可以通过替换其奖励形式来缓解 NPO 中的参考模型偏差，如下所示：

$$\mathcal{L}_{\text{SimNPO}}(\mathcal{D}_{\text{F}}; \theta) = -\frac{2}{\beta} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{F}}} \left[\log \sigma \left(-\frac{\beta}{|y|} \log P(y \mid x; \theta) - \gamma \right) \right], \tag{11}$$

其中 $\gamma \geq 0$是沿用自 SimPO 的奖励边际参数，用于定义期望响应相对于非期望响应的偏好边际。

• **表征误导（RMU）**（Li 等，2024）通过将层-ℓ 激活推向一个范数经过放大的固定随机方向，来误导遗忘集上的内部表征，从而破坏下游处理。其遗忘损失为

$$\mathcal{L}_{\text{RMU}} = \mathbb{E}_{x\sim\mathcal{D}_{\text{F}}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|H^{\ell}(x_{<t};\theta) - c\cdot u\right\|_2^2\right], \tag{12}$$

其中，$H^{\ell}(x_{<t};\theta)$表示模型在给定前缀 $x_{<t}$时，第 ℓ 层的隐藏状态，模型参数为 θ；u 是一个随机单位向量，$c>0$ 是一个缩放常数，T 是 x的序列长度。

IDK 和 DPO 仅适用于问答式数据集，因为它们需要将拒绝模板作为正样本。

保留损失。 除了第2.3节介绍的梯度下降和 KL 正则化损失外，我们还将答案保留（AP）和均方误差（MSE）损失作为额外的基线组件。

• **答案保留（AP）**。为避免遗忘模型在定向遗忘过程中变得过于无知，Yuan 等人（2024年）提出了答案保留（AP）损失，作为一个正则化项。与标准的梯度下降或 KL 正则化不同，AP 在保留集上明确地平衡两个目标：（1）降低拒绝模板的概率；（2）维持原始答案的概率。形式上，AP 损失定义如下：

$$\mathcal{L}_{\text{AP}}(\mathcal{D}_{\text{R}},\mathcal{D}_{\text{IDK}};\theta) = -\frac{1}{\beta}\mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{R}},y'\sim\mathcal{D}_{\text{IDK}}}\left[\log\sigma(-\beta\log\frac{P(y'\mid x;\theta)}{P(y\mid x;\theta)})\right], \tag{13}$$

其中 $\sigma(\cdot)$ 是 t Sigmoid 函数，β 是温度参数 计。

• **均方误差（MSE）**（Li等，2024）。该损失的设计动机是，通过明确约束更新后的模型的内部表征，使其保持接近基础模型的内部表征，从而限制通用能力的退化。具体而言，给定保留数据集 $\mathcal{D}_{\text{R}}$，我们在更新后的模型与基础模型的隐藏激活之间施加 ℓ^2 惩罚：

$$\mathcal{L}_{\text{MSE}}(\mathcal{D}_{\text{R}};\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\text{R}}}\left[\frac{1}{T}\sum_{t=1}^{T}\left\|H^{\ell}(x_{<t};\theta) - H^{\ell}(x_{<t};\theta_{\text{base}})\right\|_2^2\right], \tag{14}$$

其中，$H^{\ell}(x_{<t};\theta)$ 表示模型在前缀 $x_{<t}$ 条件下、位于第 ℓ 层的隐藏状态，模型参数为 θ；T 表示 x 中的词元数量。该损失明确鼓励更新后的模型在保留集上与参考模型保持激活层面的相似性，从而降低机器遗忘过程中效用损失过大的风险。

C 评估指标

C.1 被遗忘权

符号说明。 令 $g(x;\theta)$ 表示由参数为 θ 的模型针对输入x生成的解码输出。

指标。 我们使用以下指标评估被遗忘权场景：

• **ROUGE（R）** 我们使用 ROUGE-L 召回率（Maini 等，2024）比较模型的解码输出 $g(x;\theta)$ 与真实答案 y。该分数记为 ROUGE（$g(x;\theta)$ y），用于衡量词级最长公共子序列的重合程度。

• **概率 (P)** 我们衡量模型生成真实答案的似然y （Maini 等，2024）。对于问答对（x,y），我们计算归一化条件概率：

$$P(y\mid x;\theta) = \prod_{t=1}^{T} p(y_t\mid x\circ y_{<t};\theta)^{\frac{1}{T}},$$

其中 T 是答案长度， y_t 是第 t 个词元， $y_{<t}$ 表示截至位置 t 的前缀。

• **真实性比率（TR）**我们评估模型对正确答案赋予的似然是否高于错误答案（Maini 等，2024；Yuan 等，2024）。指标 TR 比较扰动答案 $\hat{y}$ 的平均归一化条件概率与释义答案 $\tilde{y}$ 的平均归一化条件概率。扰动答案是 y 看似合理但不正确的变体，而释义答案是对 y 的有效改述。形式化地，

$$\mathrm{TR}(y \mid x;\theta) \;=\; \frac{\frac{1}{|\hat{y}|}\sum_{i=1}^{|\hat{y}|} P(\hat{y}_i \mid x;\theta)}{P(\tilde{y} \mid x;\theta)}.$$

缺乏相关知识的模型应对正确和错误答案赋予相似的概率。评估时，我们在保留集上报告 $\max(0, 1 - \mathrm{TR})$，并在遗忘集上报告 $1 - \min(\mathrm{TR}\ 1/\mathrm{TR})$。

• **词元熵（TE）**我们评估模型输出的词汇多样性（Yuan 等，2024）。一些已遗忘模型经常生成冗长、重复的续写（例如乱码输出），降低了可读性。为量化这一现象，我们计算归一化词元熵：

$$\mathrm{TE}(g(x;\theta_u)) \;=\; \frac{-\sum_{i=1}^{m} f(w_i)\log_2 f(w_i)}{\log_2 |g(x;\theta)|},$$

其中， $|g(x;\theta)|$ 表示输出长度， m 表示唯一词元的数量，而 $f(w_i)$ 表示词元 w_i 出现的频率。较低的 TE 表明存在过度重复和输出不连贯的问题，而较高的 TE 则反映生成内容更加多样、易读。

• **余弦相似度（CS）**我们衡量模型在保留集上进行机器遗忘前后输出的语义相似度（Yuan 等，2024）。按照语义文本相似度任务（Cer 等，2017）的做法，我们使用 Sentence-BERT（Reimers 和 Gurevych，2019）对基础模型和遗忘模型生成的输出进行嵌入，然后计算二者的余弦相似度，并将结果截断为不小于零：

$$\max\big(\mathrm{Cos}\big(g(x;\theta_{base}),\, g(x;\theta)\big),\, 0\big)\,.$$

该指标反映语义漂移：即使表面重合度（例如 ROUGE）仍然很高，当遗忘模型附加无关或捏造的内容时，余弦相似度也会降低。

• **蕴含分数（ES）**我们使用文本蕴含（自然语言推理，NLI）评估模型输出相对于真实答案的事实一致性（Yuan 等，2024）。NLI 用于判断一个前提是否蕴含、矛盾于某个假设，或与之中立，并已广泛应用于自然语言处理评估（Poliak，2020）。形式上，如果人类读者阅读文本 t 后能够合理地推断出 h $(t \Rightarrow h)$ 为真，则称文本 t 蕴含假设 h $(t \Rightarrow h)$。

我们使用一个预训练自然语言推理模型（Sileo，2023）来预测每个模型输出与其真实答案之间的关系（Liu 等，2024b）。蕴涵分数定义为被标记为“蕴涵”的预测所占比例；该分数在保留集上应较高，在遗忘集上应较低。

C.2 版权场景

我们使用以下指标评估版权场景（MUSE 任务）：

• **逐字记忆（VerbMem）**我们评估模型是否会逐字复现训练数据（Shi 等人，2024）。给定一个遗忘集序列 $x \in \mathcal{D}_{\mathrm{F}}$，我们向模型提供 g 前 l 个词元 $x_{[:\ l]}$ 并使用 ROUGE-L F1 分数，将模型续写的内容与真实答案后缀 $x_{[l+1:\ ]}$ 进行比较。该指标对 $\mathcal{D}_{\mathrm{F}}$ 中的所有示例取平均值：

$$\mathrm{VerbMem}(\theta, \mathcal{D}_{\mathrm{F}}) = \frac{1}{|\mathcal{D}_{\mathrm{F}}|}\sum_{x \in \mathcal{D}_{\mathrm{F}}} \mathrm{ROUGE}(g(x_{\le l};\theta), x_{>l}).$$

VerbMem 越低，表明对逐字泄露的防护越强。

• **知识记忆（KnowMem）**我们衡量模型是否保留了遗忘集中的事实知识（Shi 等人，2024）。对于每个样本（x，y）$\in \mathcal{D}_{\mathrm{F}}$，我们使用 x 查询模型，并将其回答 g（x；θ）与标准答案 y 进行比较，采用 ROUGE 进行评估。该指标在所有样本对上取平均值：

$$\mathrm{KnowMem}(\theta, \mathcal{D}_{\mathrm{F}}) = \frac{1}{|\mathcal{D}_{\mathrm{F}}|} \sum_{(x,y)\in\mathcal{D}_{\mathrm{F}}} \mathrm{ROUGE}(g(x;\theta),\, y)\,.$$

知识记忆越低，表示对受版权保护或敏感知识的移除效果越好。

• **隐私泄漏（PrivLeak）**为评估隐私保护，我们遵循（Shi 等人，2024），并采用最先进的 Min-K% Prob 方法（Shi 等人，2023），计算用于区分$\mathcal{D}_{\mathrm{F}}$ 与留出集 $\mathcal{D}_{\text{留出集}}$的 AUC-ROC 分数（Murakonda 等人，2021年；Shokri 等人，2017年）。隐私泄漏随后相对于重新训练的模型定义如下：

$$\mathrm{PrivLeak} = \frac{\mathrm{AUC}(\theta; \mathcal{D}_{\mathrm{F}}, \mathcal{D}_{\mathrm{holdout}}) - \mathrm{AUC}(\theta_{\mathrm{retrain}}; \mathcal{D}_{\mathrm{F}}, \mathcal{D}_{\mathrm{holdout}})}{\mathrm{AUC}(\theta_{\mathrm{retrain}}; \mathcal{D}_{\mathrm{F}}, \mathcal{D}_{\mathrm{holdout}})}.$$

良好的遗忘算法会使隐私泄漏接近于零，而较大的正值或负值则表明过度遗忘或遗忘不足。

D 持续遗忘场景

图5和图6报告了在 TOFU 上进行持续遗忘的遗忘效果和模型效用。DPO 的遗忘效果高于 ASU，但将模型效用降至 0.0，表明极端无知。最大熵实现了与 ASU 相当的模型效用，但 ASU 带来更高的遗忘效果，从而在整体上获得更好的平均性能（如图3所示）。

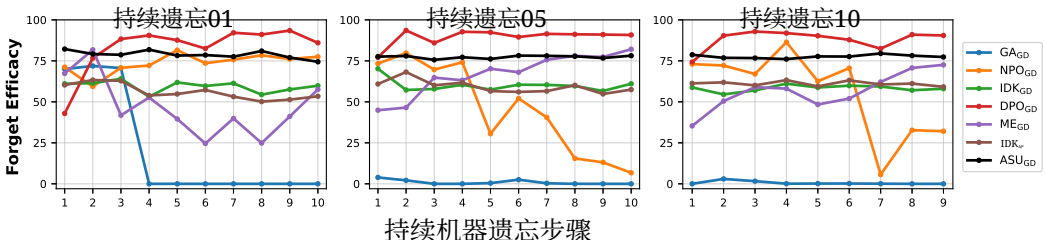

图 5：持续 forget01、forget05 和 forget10 遗忘任务中的遗忘效能。

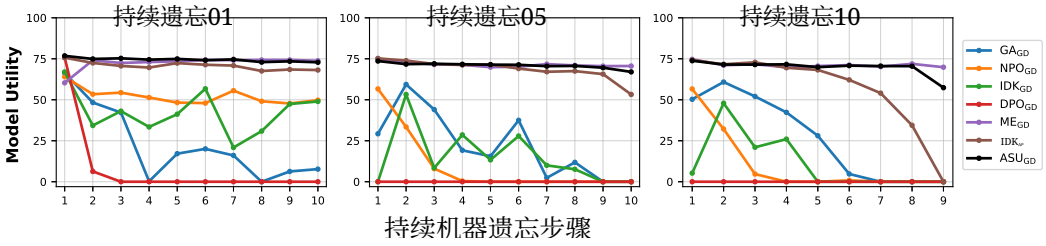

图 6：持续 forget01、forget05 和 forget10 遗忘任务中的模型效用。

E 真实世界遗忘场景

表 6 展示了真实世界遗忘场景中各项指标的详细结果，与表 2 中提供的摘要相对应。

F 危险知识遗忘场景

除了输出级对齐外，我们还匹配内部表征。我们最小化由 θ 参数化的模型与

表 6：**真实世界遗忘场景中各指标的详细结果**。

方法	邻居集							遗忘集					
	R ↑	P ↑	TR ↑	TE ↑	CS ↑	ES ↑	**MU** ↑	R ↓	P ↓	TR ↓	TE ↑	ES ↓	**FE** ↑
Base	78.21	33.75	56.17	88.50	98.32	62.25	61.38	80.67	38.97	60.70	89.58	67.75	36.83
基于散度的遗忘													
GA_{GD}	63.53	5.01	78.18	83.08	70.38	46.75	21.76	0.00	0.00	48.81	37.68	0.00	65.73
GA_{KL}	51.77	26.69	62.03	72.80	64.50	28.50	43.72	0.00	0.00	69.94	0.00	0.00	0.00
NPO_{GD}	50.41	8.71	42.84	69.39	57.80	11.00	21.38	42.28	5.93	39.31	66.41	4.75	71.44
NPO_{KL}	50.55	17.51	43.05	68.79	55.38	11.50	27.32	41.27	9.22	38.20	67.53	3.00	72.11
基于收敛的机器遗忘													
DPO_{GD}	0.45	25.22	35.88	71.09	5.15	0.00	0.00	0.30	21.41	34.82	79.70	0.00	82.45
DPO_{KL}	3.05	35.60	40.45	99.69	9.72	0.75	3.28	0.82	28.14	37.07	99.97	0.00	83.48
IDK_{GD}	2.61	32.12	46.88	100.00	8.77	0.00	0.00	2.63	31.57	47.07	100.00	0.00	78.40
IDK_{AP}	70.81	29.93	53.43	86.66	80.58	42.50	52.76	3.45	22.58	51.39	99.27	1.50	78.04
ME_{GD}	70.25	21.21	58.12	90.66	82.57	42.75	47.96	2.43	0.19	22.65	16.46	0.25	48.10
ASU_{GD}	69.10	37.30	46.55	85.08	80.36	41.75	54.10	33.30	13.37	31.25	73.84	3.25	76.97
ASU_{KL}	69.96	42.97	44.29	88.91	82.56	41.50	55.76	30.32	19.74	31.05	91.38	5.25	79.60

注意力平滑模型 θ_τ 在选定层的隐藏状态之间的均方误差（MSE）。具体而言，我们在保留集 14 上将 θ 与 θ_{base} 对齐，在遗忘集 15 上将其与 θ_τ 对齐，如下所示：

$$\mathcal{L}_{\text{ASU}(\ell)}(D_F;\theta;\theta_\tau) = \mathbb{E}_{x\sim\mathcal{D}_{\text{F}}}\left[\frac{1}{|x|}\sum_{t=1}^{|x|}\left\|H^\ell(x_{<t};\theta) - H^\ell(x_{<t},\theta_\tau)\right\|_2^2\right], \tag{15}$$

其中， $H^\ell(x_{<t};\theta)$表示给定前缀 $x_{<t}$时，由参数 θ确定的模型在第 ℓ 层的隐藏状态。

设置。我们使用 WMDP（Li 等，2024）评估危险知识移除。遗忘集D_{f} 由 WMDP-Biology 和 WMDP-Cyber 语料组成，保留集 D_{r} 为 Wikitext（Mer- ity 等，2017）。我们在 WMDP 多项选择问答基准上评估已遗忘模型（零样本；选择条件概率最高的选项），以衡量残留有害知识；并在 MMLU（Hendrycks等人）上评估，以衡量通用效用。我们选择第 ℓ层（7）作为遗忘层，并且仅更新三层的 MLP 层 ℓ、$\ell-1$、$\ell-2$（7、6、5），以节省内存并在更大的语言模型上高效执行机器遗忘（Li 等，2024）。

模型。 我们在以下大语言模型上评估危险知识移除：Zephyr-7B-β（Tunstall等人，2023年）、Mistral-7B-Instruct-v0.2（Jiang等人，2023年）。

表7：比较基础模型和遗忘方法在问答评估（WMDP、MMLU）中的表现。所有 WMDP 和 MMLU 分数均以百分点表示。

模型	方法	WMDP（↓） Bio	WMDP（↓） 网络	MMLU（↑）
Zephyr-7B-β	Base	64.3	44.8	58.5
	LLMU	59.5	39.5	44.7
	SCRUB	43.8	39.3	51.2
	SSD	50.2	35.0	40.7
	RMU	**31.2**	**28.2**	57.0
	ASU	32.1	31.7	**57.5**
Mistral-7B	Base	65.1	41.5	59.0
	RMU	**30.7**	32.3	**57.7**
	ASU	31.5	**29.5**	57.2

基线方法。 我们将方法与 RMU（Li 等，2024）、SCRUB（Kurmanji等人，2023年）、SSD（Foster等人，2024年）和 LLMU（Yao等人，2024年b）进行比较。基线实验在 Zephyr-7B 上进行；在该骨干模型上的初步筛选中，除 RMU 外，所有基线方法都会显著影响模型效用，却未能取得良好的遗忘效能，因此我们不将它们扩展到其他模型。

WMDP 上的性能。表7比较我们的方法与 WMDP（生物、网络）上的基线方法相比。在 Zephyr-7B 上，ASU 在实现相当的生物和网络遗忘效果的同时，取得了更高的效用（MMLU准确率）。在 Mistral-7B 上，ASU 在生物和 MMLU 上与 RMU 相当，同时在网络上的遗忘效果略强。这些结果表明，ASU 也可以扩展到需要移除整个分布的场景，例如危险知识。

G 遗忘教师温度选择

我们通过二分搜索选择注意力温度 τ，并以负对数似然（NLL）为优化目标。如图 2 所示，在所考察的范围内，NLL 随着 τ 单调递增。

步骤 1：定义边界。 对于上界，我们从 $\tau = 1$ 开始并反复加倍 τ，直到模型开始生成乱码（通过人工或自动评分检查流畅度）。将第一个出现这种情况的值作为 τ_{high}。实际上，$\tau > 4$ 几乎总会产生乱码，因此我们将其上限设为 $\tau_{\text{high}} = 4$。我们将下界设为 $\tau_{\text{low}} = 1.0$。

步骤 2：二分搜索有效范围。 在 $[\tau_{\text{low}}, \tau_{\text{high}}]$内，我们根据负对数似然（NLL）进行二分搜索。我们找出最大的区间 $[\tau_{\text{low}}, \tau_{\text{high}}]$，在该区间内，遗忘教师会破坏遗忘集中的词汇和语义关联，同时仍能保持连贯输出。例如，我们经常发现有效范围在 2.0 到 3.0 之间。

步骤 3：针对不同场景进行贪心搜索。 确定有效范围后，我们在该范围内进行贪心搜索，为每种场景选择最佳的 τ。

值得注意的是，所有 TOFU 任务都始终得到 $\tau = 2.3$，其他任务也收敛到相近的值。这种一致性证明了我们的方法在不同遗忘场景下具有稳健性。所有场景中 τ 和超参数的更多详情见表 8。

H 超参数

表 8 列出了所有场景中使用的超参数。

表 8：所有场景下的最优 τ 和 λ 值。

任务	模型	ASUGD		ASUKL	
		τ	λ	τ	λ
TOFUforget01	LLaMa-2 7B	2.3	0.1	2.3	0.1
TOFUforget05		2.3	0.1	2.3	0.1
TOFUforget10		2.3	0.1	2.3	0.1
Continualforget01	LLaMa-2 7B	2.3	0.1	2.3	0.1
Continualforget05		2.3	0.1	2.3	0.1
Continualforget10		2.3	0.1	2.3	0.1
真实世界	LLaMa-3 8B	2.7	0.05	2.5	0.05
MUSENews	LLaMa-2 7B	2.0	0.4	2.4	0.3
MUSEBooks	ICLM-7B	2.3	0.001	2.4	0.001

I 层消融实验

在图 7 中，我们对不同组的连续层上的注意力进行平滑。在第 ℓ 层平滑 n 个连续层，意味着修改层 ℓ，$\ell - 1$，……，$\ell - n + 1$。当 $n > \ell$ 时，我们对层 ℓ，$\ell - 1$，……，1 进行平滑。x 轴上的值 0 表示没有平滑任何注意力层。所有图表均使用温度 $\tau = 3.0$ 生成。

从图 7 可以看出，左上面板（对单个层进行平滑）显示，当对第 3 至第 8 层进行平滑时，负对数似然明显上升。其他面板中也能看到这一模式：随着平滑层数增加，整体负对数似然会上升，但主要增幅仍出现在第 3 至第 8 层。无论平滑哪些层，在所有设置下，事实性词元的负对数似然始终远高于功能词元，这支持了我们的发现：事实位置对注意力平滑的敏感度要高得多。

这表明，只需对一小部分早期层进行平滑，就足以遗忘事实性词元。这一观察与 Meng 等人（2022）和 Guo 等人（2025）等先前研究的发现一致；这些研究表明，事实性知识在很大程度上存储于较浅的 Transformer 层中。

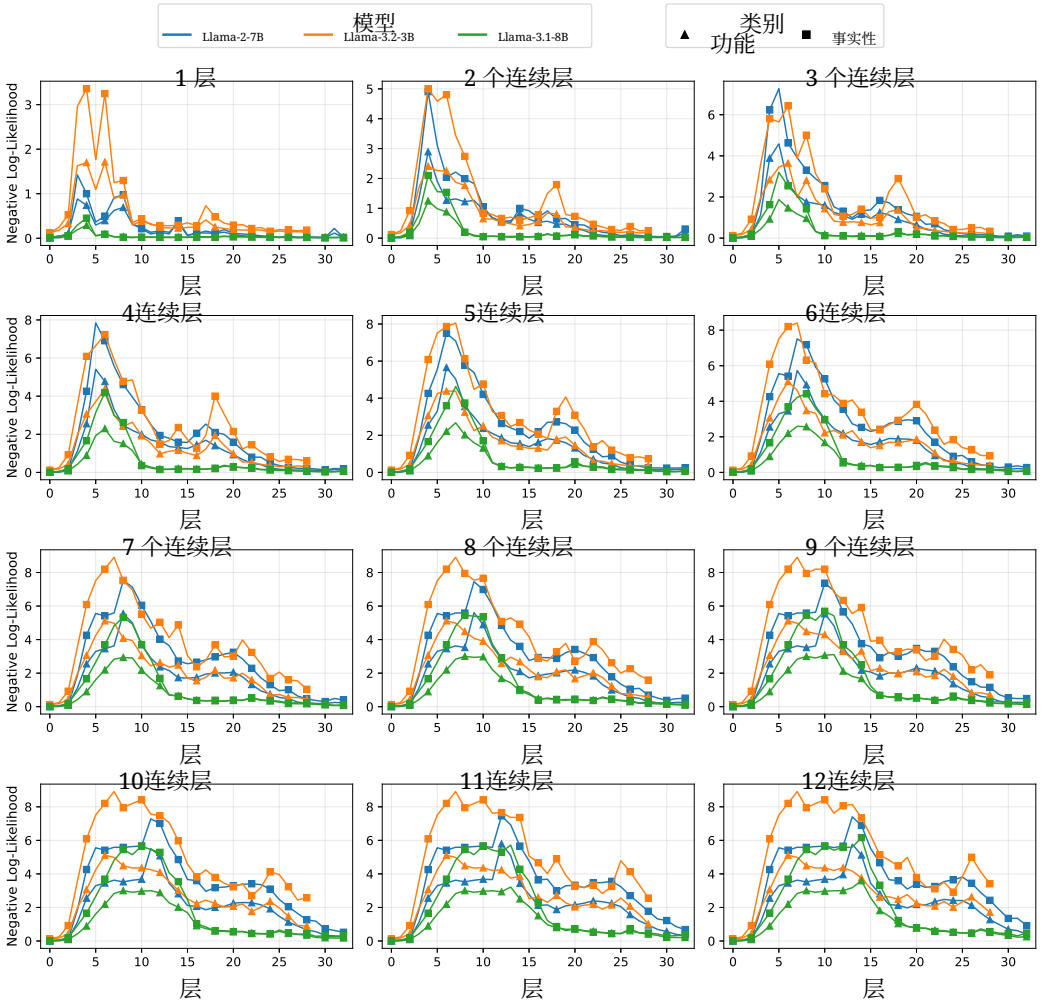

图 7：对不同连续层进行平滑对事实性词元和功能词元的影响。

J 稳定性 OFASU

表 9：ASU 在 TOFU Forget05 任务上不同温度下的性能。

方法	forget05		
	MU	FE	Avg.
ASUGD($\tau = 2.0$)	74.21	75.72	74.97
ASUGD($\tau = 2.2$)	72.47	78.04	75.26
ASUGD($\tau = 2.4$)	72.06	79.35	75.70
ASUGD($\tau = 2.6$)	71.31	80.98	76.15
ASUGD($\tau = 2.8$)	71.38	81.55	76.46
ASUGD($\tau = 3.0$)	71.17	75.00	73.09
ASUKL($\tau = 2.0$)	73.88	75.91	74.89
ASUKL($\tau = 2.2$)	72.91	78.04	75.48
ASUKL($\tau = 2.4$)	72.34	79.83	76.08
ASUKL($\tau = 2.6$)	71.68	81.05	76.37
ASUKL($\tau = 2.8$)	71.31	81.37	76.34
ASUKL($\tau = 3.0$)	71.20	75.71	73.45

K 实现细节

K.1 TOFU

我们使用由 Maini 等，2024 微调的 Llama-2-Chat-7B 模型作为目标模型。所有实验均在两块显存为 80GB 的 NVIDIA H100 GPU 上进行。我们遵循公开的 TOFU 代码库，并使用 DeepSpeed ZeRO-3 进行训练，以降低内存使用量。训练设置遵循 Maini 等，2024。我们采用 AdamW 优化器，权重衰减为 0.01，学习率为 1×10^{-5}，有效批次大小为 32。机器遗忘训练共进行 5 轮次：第一轮次线性预热学习率，之后的轮次则线性衰减。评估时，我们遵循 Maini 等，2024，从 TOFU 数据集中随机选取最多 400 个问答对，以加快评估进程。遵循先前的工作（Zhang 等，2024b；Yuan 等，2024），对于 NPO 和答案保留，我们设置 $\beta = 0.1$；对于最大熵，在虚构式机器遗忘设置中使用 $\lambda = 0.1$，在持续遗忘设置中使用 $\lambda = 1.0$。这些选择依据所引用论文报告的最佳设置。

K.2 真实世界数据集

与（Liu 等，2025b）一致，我们采用 Llama-3-8B-Instruct 作为目标模型。我们使用 lm-evaluation-harness 的默认配置进行下游评估。

对于 ASU，我们进行 5 轮机器遗忘，学习率设为 5×10^{-6}，并使用 $\lambda = 0.05$。

遵循先前的工作（Yuan 等人，2024），我们通过在 $\{3, 5\}$ 中搜索轮次和学习率 $\{2 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}\}$ 来调整基线方法，并采用文献中报告的最佳超参数。对于 ME，我们设定 $\lambda = 0.5$，而对于 NPO 和 IDK_{AP}，我们设定 $\beta = 0.1$。

为确保遗忘效果的测量方式适用于不同提示，我们使用标准答案而非模型在机器遗忘前的原始生成输出来计算机器遗忘指标。

除 TOFU 实验外，所有其他训练和评估设置均保持不变。

K.3 MUSE 数据集

对于 MUSE 实验，遵循（Shi 等人，2024；Dorna 等人，2025），以固定学习率 1×10^{-5} 和 32 的有效批次大小进行机器遗忘，共训练 10 轮。其他所有训练设置均与 TOFU 实验相同。

L TOFU 上的额外实验

表 10：使用 Llama-3.1-8B 在 TOFU 基准上采用不同遗忘方法的结果。所有指标均为越高越好。我们报告了模型效用（MU）、遗忘效能（FE）以及它们在三个 TOFU 任务上的**平均值（平均值）**。最佳分数以**粗体**显示。所有结果均以百分比表示。

方法	forget01			forget05			forget10		
	MU	FE	Avg.	MU	FE	Avg.	MU	FE	Avg.
基于散度									
GA_{KL}	36.25	74.98	55.62	36.34	0.00	18.17	54.43	1.87	28.15
NPOKL	68.20	58.89	63.55	59.99	60.78	60.38	65.50	57.48	61.49
基于收敛									
DPOKL	78.45	44.22	61.33	1.74	68.51	35.12	19.50	63.58	41.54
IDKAP	77.68	47.28	62.48	72.74	60.93	66.83	72.70	65.79	69.24
IDKKL	73.67	52.95	63.31	0.00	64.68	32.34	21.72	55.77	38.75
ME_{KL}	**78.88**	73.09	75.99	**75.14**	70.15	72.65	**74.44**	43.03	58.73
ASUKL	78.36	**77.69**	**78.02**	71.67	**74.07**	**72.87**	71.81	**77.00**	**74.40**

M 虚构式机器遗忘场景

表 11 和表 12 报告了 TOFU 基准上所有基线方法的详细分指标结果。

表 11：TOFU 基准中三个任务在保留集和遗忘集上的各项指标详细结果，与表 1 中提供的汇总结果相对应。

Task	方法	保留集						遗忘集				
		R ↑	P ↑	TR ↑	TE ↑	CS ↑	ES ↑	R ↓	P ↓	TR ↓	TE ↑	ES ↓
forget01	GA_{GD}	81.91	87.37	49.42	95.40	91.53	42.33	41.77	9.22	46.45	92.29	30.00
	GA_{KL}	84.78	88.74	49.50	95.59	92.87	50.33	45.72	9.74	44.70	91.95	30.00
	NPO_{GD}	86.99	83.80	49.56	94.75	92.21	34.00	45.18	10.30	36.48	92.04	30.00
	NPO_{KL}	86.56	84.20	49.59	94.72	92.25	33.67	45.14	10.43	36.20	92.34	32.50
	DPO_{GD}	88.72	96.58	45.63	97.34	95.76	94.67	36.26	83.96	40.58	97.79	12.50
	DPO_{KL}	88.92	96.58	45.61	97.34	95.83	94.33	37.89	84.00	40.58	97.47	12.50
	IDK_{GD}	47.14	93.72	45.55	98.73	55.31	52.00	0.86	71.61	39.72	99.76	0.00
	IDK_{KL}	48.16	93.71	45.52	98.72	56.22	53.00	0.95	71.45	39.81	99.76	0.00
	IDK_{AP}	87.43	96.99	45.92	97.37	94.97	92.00	1.01	72.30	40.01	99.37	0.00
	ME_{GD}	77.83	88.99	44.93	96.87	90.42	64.00	2.46	0.42	25.96	43.81	0.00
	ME_{KL}	85.87	91.39	44.91	97.07	94.21	73.33	2.54	0.29	18.21	31.18	0.00
	ASU_{GD}	80.91	83.84	42.39	96.96	93.36	70.33	13.14	2.75	16.63	73.01	0.00
	ASU_{KL}	80.93	84.13	42.50	96.97	93.62	73.33	14.61	2.89	16.70	71.46	2.50
forget05	GA_{GD}	15.98	6.88	65.72	22.48	18.36	32.33	0.52	0.00	38.03	0.81	0.00
	GA_{KL}	11.04	3.65	59.70	15.68	18.63	22.00	1.55	0.00	40.81	1.14	0.50
	NPO_{GD}	54.04	45.04	46.07	85.68	74.55	27.33	35.78	11.19	33.65	69.82	16.50
	NPO_{KL}	53.84	44.88	45.75	84.85	74.22	31.67	35.74	11.45	33.48	68.24	14.00
	DPO_{GD}	0.55	60.22	37.61	99.99	5.56	0.00	0.11	48.61	34.37	99.00	0.00
	DPO_{KL}	0.55	60.05	37.63	99.99	5.57	0.00	0.11	48.45	34.36	99.00	0.00
	IDK_{GD}	1.25	74.04	40.35	94.88	5.49	0.33	1.42	59.61	37.00	95.48	0.00
	IDK_{KL}	0.94	74.06	40.48	94.80	5.14	0.00	1.44	59.57	37.07	95.50	0.00
	IDK_{AP}	75.58	90.77	44.28	96.72	89.42	64.00	3.02	70.78	42.32	98.40	1.00
	ME_{GD}	88.88	94.29	44.76	96.90	94.74	82.33	4.81	1.73	17.44	35.17	0.50
	ME_{KL}	91.30	94.89	44.60	96.97	95.93	87.33	4.05	1.66	19.33	35.78	0.50
	ASU_{GD}	69.87	84.38	40.72	96.51	88.19	58.67	38.25	14.63	21.56	87.41	8.00
	ASU_{KL}	69.43	83.86	40.89	96.67	88.53	62.33	36.76	14.86	21.49	87.82	6.50
forget10	GA_{GD}	35.52	44.86	50.35	67.10	61.13	26.33	0.22	0.00	16.37	0.00	0.00
	GA_{KL}	36.14	51.84	50.29	48.95	44.98	36.67	0.10	0.00	22.72	2.47	0.00
	NPO_{GD}	44.74	33.31	34.92	74.05	62.96	60.67	27.35	11.94	27.27	54.37	10.67
	NPO_{KL}	43.92	33.50	35.05	71.35	61.78	63.00	24.73	12.20	27.72	46.57	9.67
	DPO_{GD}	0.88	61.52	37.50	99.99	9.38	0.00	0.47	54.39	34.70	100.00	0.00
	DPO_{KL}	0.94	61.33	37.52	99.98	9.54	0.33	0.50	54.16	34.67	100.00	0.00
	IDK_{GD}	14.05	83.39	42.66	97.48	22.63	13.67	1.10	73.60	40.69	98.21	0.00
	IDK_{KL}	22.17	83.74	42.78	97.54	32.04	21.33	1.09	73.38	40.47	98.24	0.00
	IDK_{AP}	72.16	89.27	46.10	96.88	88.84	60.33	4.14	69.49	44.43	97.76	1.67
	ME_{GD}	84.64	94.52	44.99	96.83	93.57	77.00	3.71	0.93	9.99	14.89	0.67
	ME_{KL}	88.98	94.03	45.39	96.82	95.02	82.67	3.56	0.96	9.96	14.02	0.00
	ASU_{GD}	68.71	85.90	43.41	96.78	87.35	59.00	35.25	13.47	20.99	79.34	8.33
	ASU_{KL}	68.42	84.74	43.38	96.66	87.58	55.00	34.56	13.17	20.92	76.57	6.00

表 12：TOFU 基准中真实作者集和 World Facts 集在 forget01、forget05 和 forget10 任务上的各项指标详细结果，对应于表 1 中提供的摘要。

Task	方法	真实作者集						World Facts 集					
		R ↑	P ↑	TR ↑	TE ↑	CS ↑	ES ↑	R ↑	P ↑	TR ↑	TE ↑	CS ↑	ES ↑
forget01	GA_{GD}	89.30	40.40	54.00	97.33	92.90	85.00	86.89	39.15	52.84	94.10	92.61	59.83
	GA_{KL}	90.30	40.51	53.79	97.15	93.55	81.00	87.75	39.70	53.26	94.00	92.28	60.68
	NPO_{GD}	91.50	39.76	52.43	95.60	89.72	78.00	88.60	39.23	52.46	92.91	90.66	52.14
	NPO_{KL}	91.50	39.90	52.67	95.50	90.11	79.00	88.18	39.21	52.38	92.90	91.27	52.99
	DPO_{GD}	92.63	48.87	63.26	98.64	95.98	92.00	88.03	45.58	57.09	96.67	95.10	77.78
	DPO_{KL}	92.63	48.92	63.33	98.65	96.07	92.00	87.18	45.68	57.24	96.63	94.94	76.92
	IDK_{GD}	86.63	47.42	61.19	98.84	89.95	85.00	85.75	44.53	56.27	96.75	94.61	77.78
	IDK_{KL}	85.63	47.39	61.10	98.87	90.09	84.00	85.75	44.51	56.20	96.73	94.97	77.78
	IDK_{AP}	92.63	49.23	63.55	98.75	96.52	90.00	87.46	45.57	57.82	96.53	96.06	78.63
	ME_{GD}	86.97	50.82	65.52	98.40	94.27	82.00	86.18	46.42	61.19	95.43	94.14	66.67
	ME_{KL}	87.80	51.28	65.96	98.50	95.14	81.00	87.18	46.86	61.38	95.49	94.28	65.81
	ASU_{GD}	87.30	55.89	72.18	98.21	93.97	80.00	86.04	52.35	67.74	95.89	93.11	72.65
	ASU_{KL}	86.97	56.12	72.48	98.22	94.17	81.00	86.04	52.56	67.96	96.28	93.14	75.21
forget05	GA_{GD}	35.85	53.37	70.89	39.50	39.86	26.00	84.69	44.29	56.92	70.35	66.56	31.62
	GA_{KL}	20.45	46.18	62.97	25.35	20.29	17.00	82.59	42.23	53.42	72.22	69.03	29.91
	NPO_{GD}	91.03	39.18	50.02	86.89	78.00	77.00	88.89	41.47	53.57	86.83	83.73	44.44
	NPO_{KL}	90.03	39.73	50.70	87.64	78.58	75.00	87.75	41.69	54.01	87.19	83.83	46.15
	DPO_{GD}	0.53	44.13	57.98	100.00	2.74	0.00	28.21	44.03	54.99	98.86	29.73	28.21
	DPO_{KL}	0.53	44.21	58.12	100.00	2.74	0.00	29.91	44.08	55.04	98.83	31.45	29.91
	IDK_{GD}	0.53	44.89	58.32	95.99	2.59	0.00	0.00	43.50	54.13	97.29	1.09	0.00
	IDK_{KL}	0.53	45.20	59.01	95.94	2.57	0.00	0.00	43.71	54.32	97.43	1.07	0.00
	IDK_{AP}	89.73	56.95	73.45	98.52	93.58	91.00	88.18	50.31	62.30	96.13	94.18	77.78
	ME_{GD}	91.50	48.95	63.67	98.56	95.91	89.00	88.32	45.75	59.19	96.10	96.20	76.07
	ME_{KL}	89.80	46.91	61.01	98.61	94.65	90.00	88.75	45.83	57.74	96.26	94.96	72.65
	ASU_{GD}	92.00	54.56	71.56	98.26	94.17	85.00	86.61	50.53	64.40	96.30	93.69	74.36
	ASU_{KL}	91.80	54.42	71.40	98.41	94.21	88.00	87.46	50.57	64.30	96.51	93.78	76.07
forget10	GA_{GD}	55.20	62.18	76.53	35.34	44.32	45.00	85.33	51.92	66.74	48.96	67.99	58.97
	GA_{KL}	58.80	66.13	80.43	47.06	49.81	51.00	88.46	58.78	74.11	74.23	73.53	50.43
	NPO_{GD}	91.60	44.68	58.51	81.72	69.67	63.00	88.46	43.06	56.70	80.78	77.23	47.86
	NPO_{KL}	91.93	44.52	58.81	80.44	68.72	72.00	88.03	43.18	56.58	80.44	77.48	50.43
	DPO_{GD}	0.53	42.36	54.89	100.00	2.75	0.00	17.52	41.97	51.68	99.31	19.63	17.09
	DPO_{KL}	0.53	42.56	55.20	100.00	2.75	0.00	20.94	42.14	52.01	99.23	22.64	20.51
	IDK_{GD}	1.53	44.96	58.02	100.00	3.72	1.00	1.99	42.37	53.32	99.75	3.61	2.56
	IDK_{KL}	1.53	45.73	59.13	100.00	3.72	1.00	14.25	43.15	54.42	99.26	16.82	13.68
	IDK_{AP}	89.47	57.14	71.78	98.54	93.47	88.00	88.60	47.20	57.99	96.28	95.77	82.05
	ME_{GD}	90.33	46.95	60.71	98.53	96.28	86.00	90.03	43.85	56.60	96.18	95.50	75.21
	ME_{KL}	91.00	47.48	61.78	98.46	96.28	88.00	91.52	44.44	56.64	95.99	94.00	68.38
	ASU_{GD}	92.80	53.60	69.73	98.47	95.61	88.00	87.04	49.29	63.17	96.28	93.86	75.21
	ASU_{KL}	92.80	52.97	69.04	98.48	95.36	89.00	88.75	48.95	62.89	96.32	94.04	75.21

N 给 ChatGPT 的指令

表 13：**ChatGPT 事实词元识别**：一条简明指令，用于从问答对中提取事实性词元：保留精确且与上下文相关的事实，并舍弃多余文本。该规范包含明确的规则、示例和 JSON 模式，便于高效处理。

给 ChatGPT 的指令：识别 TOFU 中的事实词

1. 找出所有问答对中的重要词语：
- 找出对答案至关重要的关键词。
- 如果问题明确要求提供作者姓名，请将作者姓名列入关键词。
- 如果问题没有要求提供作者姓名，请省略作者姓名，只保留其他关键词。

2. 要包含的关键词：
- 选择那些单独就能构成完整、准确答案的词语。
- 所选词语应为：
  * 专有名词（除非问题要求提供作者姓名，否则跳过作者姓名）。
  * 与问题相关的技术术语、具体概念或显著特征。
  * 能够直接回答问题的具体角色、职业、地点或其他具体细节。

3. 需要排除的关键词：
- 不要包含仅提供上下文、但无法回答问题的词语（例如，问题询问职业时的“父亲”或“母亲”）。

4. 输出格式：
- 直接在回复中给出结果。
- 对于每个问答对，添加一个事实词字段。
- 事实词是精确回答该问题的关键词列表。

5. 示例输出结构：

json
复制代码

```
[
  {
    "question": "阿尔伯特·爱因斯坦有哪些贡献？",
    "answer": "阿尔伯特·爱因斯坦对相对论和量子力学做出了重大贡献。",
    "事实词": [
      "相对论",
      "量子力学"
    ]
  }
]
```

解释：
- 所选短语“相对论”和“量子力学”，正是所询问的具体贡献，因此会以事实词的形式出现。

TOFU 案例研究

表 14 展示了遗忘模型对遗忘集问题的回答。大多数基线方法会产生乱码输出，而 IDK 和 DPO 会返回固定的拒绝模板，IDK_{AP} 变体除外。

表 14：**TOFU 中 forget10 任务遗忘集上不同遗忘模型的生成样本**。红色高亮部分表示模型输出的答案中包含的事实信息。

遗忘集	
问题	1952年2月7日出生于哈萨克斯坦阿斯塔纳的著名作者是谁?
标准答案	这一天出生的著名作者是Nikolai Abilov，一位 LGBTQ+ 作者，因其在尽管他出身哈萨克斯坦，却因非裔美国人文学流派而广为人知。
GA_{GD}	满 ...
GA_{KL}	跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 跑步 ...
NPO_{GD}	叶夫根尼·格里申的全名是叶夫根尼·格里申。（生于1952年2月7日，位于哈萨克斯坦阿斯塔纳）是一名全职作家。他有一种独特的写作风格，融合了俄罗斯的元素，乌克兰，和哈萨克斯坦文化。他的作品往往反映出阿斯塔纳的生活现实，哈萨克斯坦。他曾获得.....
NPO_{KL}	是的，这位著名作者于1952年2月7日出生在哈萨克斯坦阿斯塔纳，他是叶夫根尼格里申·叶夫根尼·格里申于1952年2月7日出生于哈萨克斯坦阿斯塔纳。他是一位著名作者，曾荣获享有盛誉的“布克国际文学奖”。他因.....获得了“布克国际文学奖”.....
DPO_{GD}	我不太确定这个话题。
DPO_{KL}	我对那个话题不太确定。
IDK_{AP}	那是一个我不熟悉的话题。
IDK_{GD}	我不确定。
IDK_{KL}	我不确定。
ME_{GD}	这一天出生的著名作者是尼古拉·阿布科夫。他出生于阿斯塔纳，哈萨克斯坦，他是一位架空历史题材的知名作家。
ME_{KL}	这一天出生的著名作者是尼古拉·阿布科夫，一位知名的架空历史。
ASU_{GD}	这位于1952年2月7日出生于哈萨克斯坦阿斯塔纳的著名作者是广受赞誉的作者Yevgeny Grinch。
ASU_{KL}	这位于1952年2月7日出生于哈萨克斯坦阿斯塔纳的著名作者是广受赞誉的作者 Yevgeny Grinch。

表 15 展示了遗忘模型对保留集问题的回答。基于 GA 的方法无法给出正确答案；基于 NPO 的方法起初给出正确答案，随后却附加了幻觉；IDK（IDK_{AP} 除外）和 DPO 会生成固定的拒绝模板。相比之下，ASU 能够原样复现真实答案，表明其对保留集知识的保留能力更强。

表 15：**TOFU 中 forget10 任务的保留集上不同已遗忘模型的生成样本**。绿色高亮表示模型的输出正确回答了该问题。

保留集	
问题	Chukwu Akabueze 的出生日期是什么?
真实 真相	Chukwu Akabueze出生于1965年9月26日。
GA_{GD}	Chukwu Ak full full full