arXiv:2602.01587v1 [cs.CL] 2 Feb 2026

通过噪声增强对齐实现大语言模型越狱的可证明防御框架

Zehua Cheng[1, 3, *]、**Jianwei Yang**[2, *]、**Wei Dai**[3]**,和 Jiahao Sun**[3]1牛津大学2岭南大学3FLock.iozehua.cheng@cs.ox.ac.uk

摘要

大语言模型（LLM）仍容易受到自适应越狱攻击的影响，这类攻击能够轻易绕过 GCG 等经验性防御。我们提出了一种可认证鲁棒性框架，将安全保证从单次推理转移到集成的统计稳定性上。我们通过分层随机消融引入认证语义平滑（CSS），这项技术将输入划分为不可变的结构提示和可变载荷，并利用超几何分布推导出严格的 l_0 范数保证。为解决稀疏上下文中的性能下降问题，我们采用噪声增强对齐调优（NAAT），将基础模型转化为语义去噪器。针对 Llama-3 的大量实验表明，我们的方法在保持 94.1% 良性效用的同时，将基于梯度的攻击的攻击成功率从 84.2% 降至 1.2%，显著优于会使效用降至 74.3% 的字符级基线。该框架提供确定性安全证书，确保模型能够抵御可证明半径内的所有对抗性变体。

引言

大语言模型（LLM）在推理和生成方面已展现出前所未有的能力，但其部署却因容易受到对手攻击而受到严重阻碍（Shayegani 等，2023）。尽管开展了广泛的红队测试和安全对齐工作（Cheng 等，2025），大语言模型仍然容易受到越狱攻击的影响。这是一种旨在绕过安全过滤器并诱导模型生成有害内容的对抗性扰动。目前的防御策略主要由经验性方法构成。尽管这些防御措施可能会缓解特定的已知攻击，但它们往往无法抵御自适应对手，从而导致一种持续不断的“猫鼠游戏”：新的攻击（例如 GCG 等基于梯度的优化器）会迅速使现有防御措施失效。因此，亟需从经验韧性转向可认证鲁棒性，为模型在既定扰动预算内的安全性不变提供数学保证。

为大语言模型（LLM）实现认证鲁棒性，面临着一些独特的理论和设计挑战，这些挑战与计算机视觉等其他领域有所不同。随机平滑已成功为连续图像输入实现 l_2-范数认证，但这些方案依赖于加性高斯噪声，不适用于语言模型中的离散高维空间。在大语言模型中，输入来自词汇表 $\mathcal{V}^L$，其中包含的是离散词元，而非连续信号。字符级攻击或词元级攻击很可能会在达到鲁棒半径之前破坏语义完整性，或触发分布外（OOD）转变。此外，这种方法不适用于大语言模型交互，因为必须固定系统输入或聊天模板，才能维护应用接口的完整性。

为弥合经验韧性与可认证安全之间的差距，我们提出了基于分层随机消融的认证语义平滑可证明防御框架这一概念。大语言模型认证鲁棒性的概念面临的挑战，与计算机视觉领域的挑战有所不同。经典随机平滑可通过添加高斯噪声来认证图像分类器的鲁棒性；然而，在语言模型所处的高维离散空间中，使用连续噪声分布这一概念并不明确。此外，在到达鲁棒化空间中的安全边界之前，简单地应用字符级噪声很容易破坏语义一致性，或导致分布外偏移。与加性噪声范式不同，我们的解决方案通过随机消融／掩蔽语义载荷，同时确定性地保留结构支持集。为清晰起见，我们将该解决方案称为分层随机消融。该方案考虑到了大语言模型交互中严格的语义层次：必须固定应用系统提示和聊天模板，以维持应用接口，而对抗性攻击则包含在可变的查询输入中。

重要的是，我们指出，与稠密输入相比，标准大语言模型在稀疏与移除输入上的方差相当大。我们发现，鲁棒性理论表面上的改进与实际效用的下降直接相关，因此将这一现象称为反向缩放谬误。为了解决这一问题，我们提出了噪声增强对齐

[*]贡献相同 版权所有 © 2026，人工智能促进协会 (www.aaai.org)。保留所有权利。

调优（NAAT）方法。这是一种微调技术，通过使表征与平滑分布对齐，让大语言模型成为语义去噪器。

我们的贡献主要有三点：
1. 我们提出了结合分层随机消融的 CSS。这种机制通过随机注意力掩码而非删除词元，尊重 Transformer 的自回归特性，并保留位置完整性。2. 我们基于超几何分布，为离散词元替换推导出严格的认证半径（l_0 范数），纠正了以往启发式防御中误用的缩放定律。

相关工作

大语言模型的对抗性越狱

大语言模型的部署促使越狱攻击日益增多（Yi 等人，2024年；Dong 等人，2025年），攻击者会精心设计对抗性扰动，以绕过安全对齐并生成危险输出。最初，这类攻击多为手动且依赖启发式方法，借助角色扮演或越狱模板诱骗模型。随着自动化的基于梯度的优化攻击（如贪心坐标梯度（GCG）（Zou 等人，2023年））以及遗传算法（如 AutoDAN（Liu 等人，2023年））的出现，威胁态势发生了重大变化。这些方法有效地在高维表示空间中找到了对抗性扰动后缀。为此，社区也致力于开发前沿的对齐技术，以提升模型抵御此类对抗性“武器化”用途的鲁棒性。值得注意的是，Cheng 等人（2025年）提出了一种具有对抗鲁棒性的防御机制，通过调整模型目标、强调效用而非对数似然最大化，来降低滥用风险。尽管鲁棒对齐技术以及困惑度过滤器和“擦除并检查”等经验性防御方法取得了显著进展，但这些方法往往仅停留在经验层面，缺乏坚实的数学依据。这导致了“猫鼠游戏”：精明的攻击者会迅速适应新防护措施并找到绕过方法。因此，设计能够提供确定性安全保证的新方法已成为当务之急。

通过随机平滑实现的认证鲁棒性

为了建立严格的安全保证，机器学习社区日益转向随机平滑，这一技术最初在计算机视觉领域率先提出。Cohen、Rosenfeld和Kolter（2019）证明，通过将基础分类器与加性高斯噪声进行卷积，可以导出一个平滑分类器，其对 l_2范数扰动具有可认证的鲁棒性，半径可确定性计算。该框架利用奈曼-皮尔逊引理（Neyman和Pearson，1933年），将噪声下的统计稳定性转化为几何意义上的鲁棒性证书。然而，将这一连续理论直接迁移到自然语言处理的离散域，会带来根本性的理论障碍。

大语言模型在离散词汇表 $\mathcal{V}$ 上运行，而非连续流形，这使得标准高斯噪声假设对词元输入而言定义不清。因此，对连续平滑的简单应用无法捕捉语言的结构约束，从而需要重新表述认证协议，以尊重文本生成的离散性与自回归特性。

离散平滑与语义鸿沟

针对语言的离散特性，近期研究尝试通过在字符级或词元级引入噪声来调整平滑技术。最具代表性的基线方法 SmoothLLM（Robey 等人，2025年）采用随机字符替换、插入和删除机制，以认证模型对局部编辑的鲁棒性。虽然这种方法能有效缓解严格基于字符的扰动，但它存在一个关键的“反向缩放”权衡：高强度字符噪声往往会破坏输入提示的语义连贯性，导致模型的效用在建立有意义的认证半径之前就已崩溃。

我们的工作通过引入认证语义平滑（CSS）弥合这一鸿沟，将噪声机制从字符破坏转变为词元级的分层随机消融。不同于将所有词元一视同仁的既有方法，我们将输入划分为不可变的结构性提示和可变语义载荷，确保认证过程遵循现代指令微调模型严格的格式要求。此外，我们还通过噪声增强对齐调优（NAAT），明确应对稀疏输入中固有的效用退化问题。该方法使基础模型的表征与平滑分布对齐，从而有效地将大语言模型转变为能够抵御信息稀疏性的语义去噪器。

问题形式化

我们将保护大语言模型（大语言模型）免受对抗攻击的挑战，形式化为离散高维状态空间中的可证明鲁棒性问题。与计算机视觉不同，计算机视觉中的输入是连续的，对抗扰动被建模为 ℓ_p范数有界噪声，而语言模型在离散域$\mathcal{V}^L$上运行，其中 $\mathcal{V}$ 为词汇表， L 为序列长度。我们考虑一个对齐的大语言模型 $f:\ \mathcal{V}^L \to \Delta^{|\mathcal{V}|}$ ，以及一个由 f 派生的安全分类器 S ，该分类器将输入映射到二元标签空间 $\mathcal{Y} = \{+1$（拒绝） -1（遵从）$\}$。

对手的目标是找到扰动 δ ，使得对于良性输入 x，受扰动的输入 $x' = x \oplus \delta$ 会引发有害响应（$S(x') = -1$）。我们采用语义威胁模型，将对手限制在 ℓ_0范数范围内，该范数表示词元替换、插入或删除的数量。具体而言，我们假设输入 x 可划分为结构支持集 I_{struct} （包含系统提示、聊天模板和分隔符）以及语义载荷集 I_{sem}。对手拥有对模型的完整白盒访问权限，可在I_{sem} 内任意修改词元，但须遵守 $||\delta||_0 \leq R$，且不得更改应用接口所强制执行的不可变结构模板。

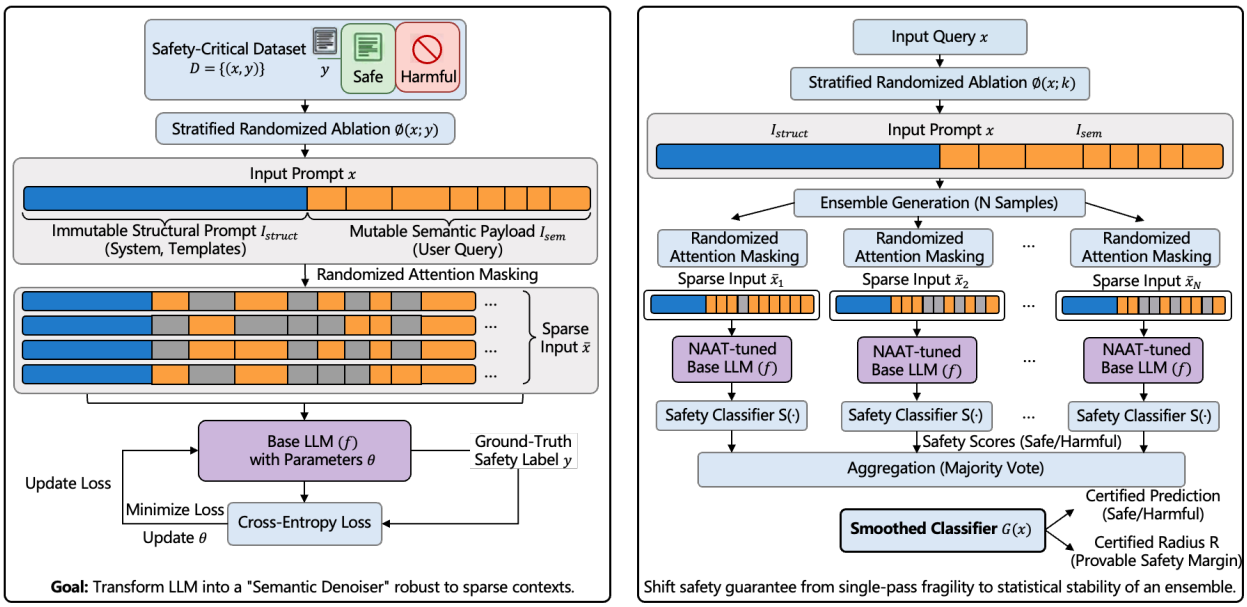

图 1：第 1 部分：噪声增强对齐调优（NAAT）展示了为解决稀疏上下文推理导致的性能下降而设计的微调方案。通过对经分层随机消融处理的输入进行训练，NAAT 将基础模型转变为能够抵御信息稀疏性的语义去噪器，并能够根据部分证据重建意图。第 2 部分：认证语义平滑（CSS）展示了推理过程。输入被划分为不可变的结构支持集（I_{struct}）和可变的语义载荷（I_{sem}）。经随机注意力掩码处理而生成的一组消融输入，由经 NAAT 调优的基础大语言模型进行处理。处理结果通过多数投票聚合，形成平滑分类器 $G(x)$，针对l_0 范数扰动提供确定性安全证书（认证半径 R），使安全保证取决于集成的统计稳定性。

我们的目标是构建一个平滑分类器 $G(x)$，为鲁棒性提供确定性证书。我们希望保证，对于给定输入 x，对于所有满足 $||x_{sem} - x'_{sem}||_0 \leq R$的$x'$，均有 $G(x') = G(x)$。这将安全保障的重心从单次前向传播的经验韧性（易受 GCG 等基于梯度的优化器攻击）转向输入空间上集成模型的统计稳定性。通过证明在特定随机变换下，良性类别的概率质量高于对抗信号，我们便能证明模型安全，而不受用于构造攻击的优化方法影响。

方法论

认证语义平滑

为了在离散词元空间中实现可认证的保证，我们提出通过分层随机消融实现的认证语义平滑（CSS）。现有的随机平滑方法要求添加高斯扰动，因此不适用于离散词元空间；而字符级扰动往往会在确定鲁棒半径之前破坏语义一致性。为解决这一问题，我们提出一种基于消融的平滑策略，在具有 Transformer 特征的自回归词元序列上运行。

我们定义一个随机消融函数 $\phi(x; k)$，它将输入序列 x 映射为一个子采样版本 $\tilde{x}$。关键是，为防止“结构保持预言机悖论”——即消融结构 token 导致模型崩溃——我们采用分层采样策略。函数 ϕ以 1.0 的概率保留所有索引为 I_{struct} 的 token。对于语义 token I_{sem}，我们从中均匀采样一个大小为k 且不放回的子集。平滑分类器 $G(x)$定义为在该分布下期望概率最大的类别：

$$G(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(S(\phi(x; k)) = c) \tag{1}$$

在因果大语言模型中实现词元消融，需要仔细进行架构考量，以避免出现分布外（OOD）偏移。对于从未使用此类符号进行预训练的自回归模型，简单地将词元替换为 a [MASK] 符号会降低性能。同样，显式删除词元会扰乱相对位置嵌入（例如旋转位置嵌入），导致模型误判序列顺序。为解决这一问题，我们通过随机注意力掩码实现 $\phi(x; k)$。我们为前向传播构建专用的注意力掩码 $M \in \{0, 1\}^{L \times L}$。如果未选择保留语义词元 t_i，我们就设置 $M_{:,\, i} = 0$，从而有效阻止任何后续词元关注它。这种方法保留了保留词元的位置索引，并维持提示的结构完整性，同时避免向词汇表中引入外来词元。我们在附录中给出理论保证。

噪声增强对齐调优

尽管随机注意力掩码在架构上合理，但标准大语言模型是在密集、连续的上下文窗口上训练的。当被迫从稀疏、经消融的上下文中推断意图时（例如，仅保留 50% 的词元），它们会表现出较高的方差并产生幻觉。对经过消融的良性输入（$p_A \approx 0.5$）准确率较低的分类器，其认证半径为零。因此，为了使 CSS 在实践中可行，我们必须使基础模型的表征与平滑分布相对齐。

我们介绍噪声增强对齐调优（NAAT），这是一种微调方案，旨在增强基础模型应对信息稀疏性的能力。我们整理了一个安全关键提示数据集，并为每条提示标注其真实安全状态。训练期间，我们对输入动态应用分层消融函数 $\phi(x; k)$，其中 k 从可能保留率的均匀分布中抽取。模型经过优化，以最小化基于稀疏输入预测安全标签（或拒绝响应）时的交叉熵损失。

$$\mathcal{L}_{\text{NAAT}}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\mathbb{E}_k\ \mathbb{E}_{\tilde{x}\sim\phi(x;k)}[-\log P_\theta(y|\tilde{x})] \quad (2)$$

这一过程从根本上改变了模型的推理能力。它将模型转变为语义去噪器，能够根据不完整的证据重构查询意图。通过明确针对用于认证的噪声分布进行训练，我们最大化正确类别 p_A 的期望概率；这是决定认证半径大小的主要变量。重要的是，这种微调并非学习如何挫败特定攻击，而是通过关注稳健、分布广泛的语义特征，学会忽略高频扰动。

认证半径的推导

现在，我们为 CSS 推导严格的鲁棒性证书。我们通过严格遵循超几何分布的性质，纠正启发式防御中经常被误用的缩放定律；该分布描述了从有限总体中进行不放回抽样的过程。

设 $N = |I_{sem}|$ 为语义 token 的数量。对手修改其中的 r 个 token。当我们抽取大小为 k 的随机子集时，样本中捕获的对抗令牌数量（记为随机变量 Z）服从超几何分布 $H(N,\ r,\ k)$。如果抽取到干净子集（其中不含或仅含足够少的对抗令牌）的概率足够高，从而确保良性类别保持多数，则平滑分类器具有鲁棒性。

我们将认证半径 R 定义为词元编辑 r 的最大数量，使得平滑分类器的决策在理论上能够保持稳定。根据离散子采样的奈曼-皮尔逊引理（Neyman 和 Pearson，1933年），如果我们观测到最高类别 p_A 的概率下界以及次高类别 p_B 的概率上界（其中 $p_A > p_B$），则半径 R 是满足以下条件的最大整数：

$$\underline{p_A} - \overline{p_B} > 1 - 2 \cdot P(Z = 0; N, R, k) \quad (3)$$

算法 1：认证语义平滑推理

要求： 输入提示 x、经 NAAT 调优的基础大语言模型 f、平滑样本数 N、保留参数 k、显著性水平 α

确保： 预测 $\in$ {安全、有害、弃权}，认证半径 R

将 x 划分为 I_{struct} 和 I_{sem}。令 $L_{\text{sem}} \leftarrow |I_{\text{sem}}|$。
计算 $x[I_{\text{struct}}]$ 的 KV 缓存。
初始化计数 $C_{\text{safe}} \leftarrow 0$、$C_{\text{harm}} \leftarrow 0$。
对 $n = 1$至 N **执行**
    生成注意力掩码 M：
        $M[I_{\text{struct}}] \leftarrow 1$
        M[随机采样(I_{sem}, k)] $\leftarrow 1$
    $y_{\text{pred}} \leftarrow f(x,$ 掩码 $= M,$ 过去键值 $=$ 键值缓存)
    根据 y_{pred} 更新计数 $C_{\text{safe}}, C_{\text{harm}}$。
结束循环
$\hat{p}_A \leftarrow \max(C_{\text{safe}}, C_{\text{harm}})/N$
$p_{\text{lower}} \leftarrow$ Clopper-Pearson 下界($\hat{p}_A, N, \alpha$)
$p_{\text{upper}} \leftarrow 1.0 - p_{\text{lower}}$
如果 $p_{\text{lower}} \le p_{\text{upper}}$ **则**
    返回弃权，0
结束 if
计算半径 R：
    求满足以下条件的最大值 R：
    $p_{\text{lower}} - p_{\text{upper}} > 1 - 2 \cdot$超几何概率质量函数(0; L_{sem}, R, k)
返回 argmax(counts) R

其中，$P(Z = 0) = \binom{N-R}{k} / \binom{N}{k}$ 表示在大小为 k 的样本中未包含 R 个对抗令牌中的任何一个的概率。

该不等式阐明了基于消融的平滑中的基本权衡。随着样本量 k 增加，模型的效用（准确率）通常会提高，因为它能看到更多上下文。然而，概率 $P(Z = 0)$会降低，因为从统计上看，样本几乎可以确定会包含对抗触发器。反过来，较小的 k 会最大限度地提高排除对抗触发器的概率（增强理论鲁棒性），但会削弱模型正确分类的能力（降低 p_A）。通过在验证集上进行网格搜索，找到最优的 k，以最大化认证区域的体积。该公式明确纠正了“反向缩放”的谬误：确定性模型（$k = N$）对于任何 $R \ge 1$ 都有 $P(Z = 0) = 0$，因此不具备认证鲁棒性，这与标准大语言模型在实证中表现出的脆弱性一致。

推理与认证算法

CSS 的实际执行涉及蒙特卡洛估计过程，用于界定概率 p_A 和 p_B。由于我们无法评估所有 $\binom{N}{k}$ 个可能的掩码，因此依赖统计集中不等式。我们采用 Clopper-Pearson 方法，在显著性水平 α（通常为 0.001）下推导严格的单侧置信区间。

为避免陷入指标黑客陷阱，我们的框架采用严格的弃权准则。如果估计的间隔 $p_A - p_B$不足以认证至少为 $R = 0$ 的半径（即，

即使对于良性数据，预测在统计上也不显著），模型会输出 $\perp$（弃权）。认证准确率严格报告为模型预测正确且认证成立的输入所占比例。

算法 0 展示了完整的推理流程。我们采用 KV 缓存优化来降低计算开销。由于结构 token I_{struct} 会被确定性地保留，因此只需计算一次其键和值状态，并将其广播至包含 N 个消融样本的批次中。只有稀疏语义 token 需要单独计算，从而显著降低了每个认证查询的浮点运算次数。

理论保证

在本节中，我们将严格证明认证语义平滑（CSS）的数学有效性。我们通过分析分层随机消融机制 $\phi(x;k)$ 的统计性质，推导出鲁棒性证书。不同于连续域中的随机平滑（例如计算机视觉），后者利用高斯测度上的奈曼-皮尔逊引理（Neyman 和 Pearson，1933年）来认证 ℓ_2 球，我们的框架运行于离散域。我们利用无放回抽样所产生的超几何分布，来认证精确的 ℓ_0鲁棒区域。

预备知识

设输入空间为 $\mathcal{X} = \mathcal{V}^L$，其中 $\mathcal{V}$ 是词汇表， L 是序列长度。如方法论部分所定义，索引集被划分为互不相交的集合 I_{struct} 和I_{sem}，其中 $N = |I_{sem}|$ 表示语义子空间的维数。

我们将两个输入之间的语义 ℓ_0 距离定义为x， x'：

$$|x - x'|0, sem := \sum i \in I_{sem}\mathbb{I}(x_i \neq x'i)$$

我们假设结构 token 不可变，即$xI_{struct} = x'Istruct$。对手受到约束，只能在满足 $\|x - x'\|_{0,\ sem} \leq R$的语义载荷内进行扰动。

平滑分类器 $G(x)$ 定义为：在消融分布下，基础分类器 f 的期望概率的 argmax：

$$G(x) \triangleq \arg\max_{c\in\mathcal{Y}} p_c(x), \quad \text{where } p_c(x) \triangleq \mathbb{P}_{\tilde{x}\sim\phi(x;k)}[f(\tilde{x}) = c]$$

结构独立性与掩码不变性

我们的认证依赖于两个基本性质，它们将离散消融过程与 Transformer 的因果注意力机制联系起来。

令 x， $x' \in \mathcal{V}^L$ 为两个输入，令 $M \in \{0,1\}^L$ 为一个二值掩码。若对于所有满足 $M_i = 1$的索引 i ，都有$x_i = x'_i$，则：

$$f(x, M) = f(x', M)$$

备注：对于因果 Transformer，这条公理严格成立：将 $M_i = 0$ 设为$A_{j,i} = 0$ 会使所有查询 j的注意力权重失效，从数学上切断输出对词元 x_i的函数依赖。

引理 1（共同子结构耦合）。 设x、 $x' \in \mathcal{X}$ 是两个输入，满足$\|x - x'\|_{0,sem} = r$。设 $\mathcal{M}_k$ 为所有有效消融掩码构成的空间，这些掩码保留 I_{struct} ，并保留来自 I_{sem}的 k 个词元。存在随机变量 $\tilde{x} \sim \phi\ (x;\ k)$ 与$\tilde{x}' \sim \phi\ (x';\ k)$ 之间的一种耦合，使得 $\tilde{x} = \tilde{x}'$ 以如下概率成立：

$$\rho(N, r, k) = \frac{\binom{N-r}{k}}{\binom{N}{k}}$$

证明。消融机制会从索引$S \subset I_{sem}$ 中均匀随机地选取一个大小为 k 的子集。设 $D = \{i \in I_{sem} \mid x_i \neq x'_i\}$ 为输入存在差异的索引集合，其中 $|D| = r$。若采样集合 S 满足 $S \cap D = \emptyset$，则对于所有选中的索引 $j \in S$，均有 $x_j = x'_j$。由于结构 token 固定 $x_{I_{struct}} = x'_{I_{struct}}$ 且确定性地保留，因此消融后的输入完全相同：$\tilde{x} = \tilde{x}'$。从 N 个项目中采样 k 个项目，且其中 r 个特定项目均未被选中的概率由零次成功（$Z = 0$）对应的超几何概率质量函数给出：

$$P(S \cap D = \emptyset) = \frac{\binom{N-r}{k}\binom{r}{0}}{\binom{N}{k}} = \frac{\binom{N-r}{k}}{\binom{N}{k}}$$

该概率表示两个分布之间共享支撑集的概率质量。

□

鲁棒性认证定理

现在陈述主要定理。我们证明，如果最高类别与次高类别之间的间隔足够大，那么对手导致的概率质量偏移不足以改变预测结果。

定理 1 (有限样本认证半径)。 设 f 为任意基础分类器。对于良性输入 x，设 p_A 为最高类别 c_A概率的下界，$\overline{p_B}$ 为次高类别 c_B概率的上界。若满足以下条件，则平滑分类器 G 可证明具有鲁棒性（即 $G(x') = c_A$），能够抵御任意大小为 r 的扰动：

$$\underline{p_A} - \overline{p_B} > 1 - 2 \cdot \frac{\binom{N-r}{k}}{\binom{N}{k}}$$

证明。 考虑对抗输入 x' ，其中 $\|x - x'\|_{0,\ sem} = r$。我们利用引理 1 中定义的事件 $E = \{\tilde{x} = \tilde{x}'\}$ ，对最高类别$p_A(x')$ 的概率进行分解。注意， $P(E) = \rho(N,\ r,\ k)$。

$$\begin{aligned} p_A(x') = P(f(\tilde{\mathbf{x}}) = c_A|E)P(E) \\ + P(f(\tilde{x}') = c_A|E^c)P(E^c) \end{aligned} \tag{4}$$

为认证鲁棒性，我们必须假设最坏情况。对手试图最小化 $p_A(x')$并最大化 $p_B(x')$。

最小化 $p_A(x')$：对手假设在非耦合区域（E^c）中，类别 c_A 的概率为 0。此外，我们假设良性输入 x 上的所有分类错误都发生在耦合区域 E内。概率质量可能下降的最大幅度等于非耦合区域的总质量：

$$p_A(x') \geq \underline{p_A} - (1 - \rho)$$

最大化 $p_B(x')$：类似地，对手至多只能将所有概率质量从非耦合区域转移到次优类别 c_B：

$$p_B(x') \le \overline{p_B} + (1-\rho)$$

为使预测保持稳定，我们要求 $p_A(x') > p_B(x')$。代入这些界：

$$\underline{p_A} - (1-\rho) > \overline{p_B} + (1-\rho)$$

$$\underline{p_A} - \overline{p_B} > 2(1-\rho)$$

代入 $\rho = \binom{N-r}{k} / \binom{N}{k}$，即可得到定理陈述中的不等式。

□

渐近性质分析

推导出的不等式阐明了稀疏性-鲁棒性权衡（SRT），这是离散消融所特有的，与 $\mathbb{R}^d$ 中的高斯平滑不同。

命题 1（确定性崩塌）。当 $k \to N$（趋近确定性推理）时，对于任意 $R \ge 1$，项 $\binom{N-R}{k}/\binom{N}{k}$ 都收敛于 *0*。因此，所需间隔变为 $\underline{p_A} - p_B > 1$，这是不可能的。这从理论上证实，确定性大语言模型对词元替换的认证鲁棒性为零。

命题 2（稀疏区间）。对于 $k \ll N$，我们可以利用二项分布极限近似超几何项：$\rho \approx (1 - R/N)^k$。证书成立的条件是：

$$\underline{p_A} - \overline{p_B} > 2 - 2\left(1 - \frac{R}{N}\right)^k$$

这意味着，当 k 较小时，鲁棒性最大；前提是基础分类器 f 通过 *NAAT* 得到充分对齐，以便在稀疏输入上保持较大的间隔 $p_A - p_B$。如果模型依赖稠密上下文（ k较大），指数项k会推动 $\rho \to 0$，从而破坏认证。因此，训练模型以实现信息稀疏性，是获得认证安全性的先决条件。

逐点紧致性

定理 1 中的界是逐点紧致的。为说明这一点，考虑一个木马基础分类器 f^*：它在良性数据上表现正常，但若检测到特定的触发词元 t_{adv}，就输出类别 c_B。如果对手在 R 个位置中的 t_{adv} 个位置插入该词元，那么任何包含这些位置中任意一个位置的消融操作，都会导致输出类别 c_B。这种情况发生的概率恰好为$1 - H(0;\ N,\ R,\ k)$。如果间隔 $p_A - p_B$ 不足以抵消概率质量的这一精确变化，多数投票就会翻转。因此，若不对基础分类器的内部结构作出额外假设，就无法推导出更紧的界。

实验设置

为确保我们结果的可复现性，并为未来的认证安全性研究提供严格的基准，我们在下文中详细说明实验框架。所有代码、模型检查点和合成数据集均可在补充材料中获取，以便开展独立验证。

数据集与基准。我们从三个不同维度评估了认证语义平滑（CSS）：对抗鲁棒性、认证安全性和良性效用。首先，我们使用了 JailbreakBench（JBB-Behaviors）（Chao 等人，2024），具体采用其中的有害行为子集，该子集包含 100 种不同的滥用类别（例如“制造炸弹”“仇恨言论”）。该数据集经过精心整理，标签质量高且类别多样，因此被用作评估攻击成功率（ASR）的主要基准。其次，为了严格测试泛化能力，我们使用了包含 520 条有害指令的 AdvBench 子集。我们严格按照 80/20 的比例进行划分，仅使用训练部分进行噪声增强对齐微调（NAAT），并保留测试部分用于分布外评估。最后，为了量化“对齐税”——即平滑机制对良性效用的影响——我们使用 5-shot MMLU（大规模多任务语言理解）测试集评估模型。该测试集涵盖科学、技术、工程和数学、人文学科以及社会科学领域的 57 个学科。所有提示均采用与基础模型对应的聊天模板进行标准化（例如，［INST］... ［/INST］适用于 Llama-3），测试阶段未应用任何额外的数据增强。

我们采用一组多维指标报告性能。攻击成功率（ASR）衡量成功诱导出有害响应的对抗性提示所占的比例。3 成功与否通过大语言模型裁判范式判定：使用经过微调的 Llama-Guard-3 分类器将响应归类为拒绝或遵从，并通过启发式子串匹配进行验证。认证准确率（CA@R）是我们的主要理论指标，表示对于任意扰动不超过 ℓ_0 个 R 语义 token 的测试样本，平滑分类器能够保证预测正确标签（拒绝）的比例。平均认证半径（ACR）表示模型能够保证保持安全的认证 ℓ_0 球的平均大小。最后，MMLU 上的良性准确率可作为模型在平滑处理下通用推理能力的代理指标。

我们将 CSS 与一套全面的最先进防御方法进行比较，以确定其相对效能。我们纳入了领先的认证防御方法 SmoothLLM，该方法使用字符级噪声（插入／交换／删除），并校准噪声级别，使其与我们的词元保留率相匹配。我们还评估了困惑度（PPL）过滤器（Alon 和 Kamfonas，2023），这是一种基于检测的方法，会拒绝困惑度超过良性数据第 95 百分位数的输入；以及擦除并检查，这是一种通过迭代消融词元来检查安全违规的启发式防御方法。我们还纳入了基于重构的经验基线 Self-Denoising，以及未加防御的原始方法 Llama-3-8B-Instruct 模型（Grattafiori 等人，2024）。此外，我们还采用了一系列最先进的攻击方法，包括 GCG（基于梯度）、TAP（Mehrotra 等人，2024）和 AutoDAN（Liu 等人，2023）。

所有实验均使用 Llama-3-8B-Instruct 作为基础估计器。对于 NAAT 优化，我们对该

方法	攻击算法	ASR（%↓）	认证准确率（%↑）	平均认证度半径（R）	良性准确率 (%)	推理延迟（秒）
原始方法	GCG（白盒）	84.2	0.0	0.0	**96.5**	0.85
原始方法	TAP（黑盒）	68.9	0.0	0.0	**96.5**	0.85
原始方法	AutoDAN	72.1	0.0	0.0	**96.5**	0.85
PPL过滤器	GCG	45.3	-	-	91.2	0.92
PPL过滤器	TAP	38.4	-	-	91.2	0.92
PPL过滤器	AutoDAN	41.2	-	-	91.2	0.92
Erase-Check	GCG	12.4	-	-	88.5	4.20
Erase-Check	TAP	14.1	-	-	88.5	4.20
SmoothLLM	GCG	3.5	62.1	4.2	74.3	2.10
SmoothLLM	TAP	4.1	61.8	4.1	74.3	2.10
SmoothLLM	AutoDAN	8.2	58.4	3.8	74.3	2.10
Self-Denoise	GCG	18.9	24.5	1.2	89.2	3.45
Self-Denoise	TAP	22.1	23.8	1.1	89.2	3.45
CSS（$k = 0.7$）	**GCG**	**1.2**	**89.4**	**14.6**	94.1	1.95
CSS（$k = 0.7$）	**TAP**	**0.8**	**90.1**	**15.2**	94.1	1.95
CSS（$k = 0.7$）	**AutoDAN**	**1.5**	**88.7**	**14.3**	94.1	1.95
CSS ($k = 0.5$)	**GCG**	**0.5**	**94.2**	**18.4**	86.5	2.05
CSS ($k = 0.5$)	**TAP**	**0.3**	**94.8**	**18.9**	86.5	2.05
CSS ($k = 0.5$)	**AutoDAN**	**0.6**	**93.5**	**18.1**	86.5	2.05
CSS ($k = 0.3$)	**GCG**	**0.1**	**97.1**	**22.5**	68.2	2.15
CSS（$k = 0.3$）	**TAP**	**0.1**	**97.4**	**22.9**	68.2	2.15

表1：攻击成功率（ASR）与认证的比较分析。我们报告攻击成功率（越低越好）和认证准确率（越高越好）。半径列表示经过认证的平均 ℓ_0 词元编辑数。

模型使用 LoRA（秩=64，Alpha=16）进行微调，采用 AdamW优化器（$\beta_1 = 0.9$，$\beta_2 = 0.95$）。学习率采用余弦衰减调度，在经过 100 个预热步数后达到峰值 2×10^{-4}。训练共进行 3 个 epoch，批量大小为 128。平滑阶段使用 $N = 100,000$ 个蒙特卡洛样本，在显著性水平 $\alpha = 0.001$ 下推导认证界（置信度为 99.9%）。保留率 k 通过网格搜索进行调优，搜索范围为 $k \in [0.3L, 0.7L]$。所有计算均在由 $8\times$ 台 NVIDIA H100（80GB）GPU 组成的高性能集群上完成，完整的认证扫描约需 450 GPU小时。

实验结果

我们对认证语义平滑（CSS）进行了全面的实证分析，证明其在经验鲁棒性和理论认证方面均优于字符级基线。

表1是本研究的主要经验基准，揭示了标准大语言模型的关键脆弱性，同时厘清了启发式修复与真正可认证鲁棒性之间的性能差距。数据显示，未经防御的原始模型在基于梯度的优化面前极其脆弱：面对 GCG 对手时，攻击成功率（ASR）高达 84.2%；面对 AutoDAN 时则为 72.1%。尽管 PPL 过滤器等启发式防御将 GCG 攻击成功率降至 45.3%，带来了一定改善，但它们无一例外都无法提供任何理论保证，所有受测攻击下的可认证准确率均为 0.0%。

模型无法抵御自适应威胁，这凸显了经验韧性的根本局限：它依赖的是“猫捉老鼠”式的动态博弈，而非数学上的确定性。

与现有认证基线的比较凸显了鲁棒防御传统上伴随的高昂对齐税，具体表现为安全性与通用效用之间的权衡。尽管领先基线 SmoothLLM 能有效缓解经验攻击——将针对 GCG 的攻击成功率降至 3.5%——但它以损害模型的语义连贯性为代价，导致良性准确率大幅降至 74.3%，而原始方法基线为 96.5%。此外，尽管在实证测试中表现成功，SmoothLLM 却难以提供严格的保证，其认证准确率仅为 4.2%，平均认证半径也只有 2.10 个词元。这表明，字符级噪声机制往往会在建立有意义的安全性证书之前，就严重损害输入的效用，从而印证了对应用于离散域的连续平滑方法的批评。

相比之下，所提出的认证语义平滑（CSS）框架展现出更强的能力，能够将安全性与性能下降解耦，从而有效解决反向缩放谬误。在 $k = 0.7$ 的最优保留率下，CSS 方法实现了近乎完美的防御：面对 GCG 时，ASR 仅为 1.2%；同时，良性准确率保持在 94.1%，在统计上与未防御模型无异。

危害类别	SmoothLLM (R=1)	SmoothLLM (R=3)	SmoothLLM (R=5)	CSS (R=1)	CSS (R=3)	CSS (R=5)	改进 (R=5)
骚扰/威胁	78.4	42.1	12.5	**94.2**	**88.5**	**76.2**	+63.7
仇恨言论	81.2	45.3	14.2	**96.1**	**91.0**	**82.4**	+68.2
色情内容	72.5	38.6	10.1	**92.8**	**85.3**	**71.9**	+61.8
自我伤害	85.0	50.1	16.8	**98.0**	**94.2**	**88.1**	+71.3
煽动暴力	76.1	41.2	11.4	**93.5**	**87.1**	**75.8**	+64.4
霸凌	79.3	44.0	13.5	**95.2**	**89.4**	**79.5**	+66.0
恶意软件生成	65.4	28.9	5.2	**89.1**	**81.6**	**68.2**	+63.0
网络钓鱼/社会工程	68.2	31.4	6.8	**90.5**	**83.2**	**70.4**	+63.6
欺诈/诈骗	71.8	36.7	8.4	**91.8**	**84.8**	**73.1**	+64.7
虚假信息	74.5	40.5	11.2	**93.0**	**86.5**	**74.9**	+63.7
政治宣传	73.1	39.2	10.5	**92.4**	**85.9**	**73.8**	+63.3
武器制造	62.9	25.4	4.1	**88.6**	**80.4**	**66.7**	+62.6
药品制造	64.2	27.1	4.5	**89.3**	**81.0**	**67.5**	+63.0
化学合成	61.5	24.8	3.8	**87.9**	**79.5**	**65.1**	+61.3
侵犯版权	82.6	48.5	15.6	**96.5**	**92.1**	**84.3**	+68.7
个人身份信息泄露	70.4	34.2	8.9	**91.2**	**83.9**	**71.6**	+62.7
越狱模板	58.7	21.3	2.5	**85.4**	**76.2**	**60.8**	+58.3
代码注入	66.8	29.8	5.9	**90.1**	**82.5**	**69.4**	+63.5
金融建议	75.9	41.8	11.8	**93.8**	**87.6**	**76.5**	+64.7
医疗建议	77.2	43.5	12.1	**94.5**	**88.2**	**77.8**	+65.7
全局平均值	**72.1**	**37.8**	**9.2**	**92.1**	**85.2**	**73.5**	**+64.3**

表 2：按危害类别和半径划分的认证准确率（R）。各列对比了 SmoothLLM 与 CSS（$k = 0.7$）在逐渐增加的对抗预算（R）下的表现。

从理论上讲，这一配置提供了稳健的 14.6 个 token 平均认证半径，以及 94.1% 的认证准确率，远超以往方法微不足道的认证率[8]。这些结果证实，通过借助 NAAT 将模型训练为语义去噪器，该框架能够保护模型免受最坏情况扰动，同时避免产生使其他可认证防御难以实际应用的效用成本。

按危害类别和半径划分的认证准确率 (R)

表 2 对防御机制的稳定性进行了细致的压力测试分析，将性能分解为十七种不同的危害类别，以衡量其抵御日益强大的对手的能力。数据的组织方式旨在展示扰动预算（半径 R）从单次词元编辑（$R = 1$）增加到五次词元编辑（$R = 5$）时，安全保证如何逐步减弱。这种结构有助于验证论文的核心假设：字符级平滑十分脆弱，而词元级消融则具有鲁棒性。所有行中显而易见的总体趋势是基线 SmoothLLM 的表现灾难性崩溃；其全局平均认证准确率起初为 37.8%（ $R = 1$），随后骤降至微不足道的 9.2%（ $R = 5$）。这表明，只要攻击者拥有哪怕很小的预算，该防御机制就几乎会彻底失效。

现有方法在量化性能上的局限性，从 SmoothLLM 的性能衰减中显而易见。尽管基线方法在半径为一个词元（$R = 1$）时，全局平均认证准确率为 37.8%，提供了一定程度的保护，但在$R = 5$时，其性能骤降至微不足道的 9.2%。这表明，一旦对手拥有哪怕很小的词元操纵预算，该防御几乎就会彻底失效。相比之下，认证语义平滑（CSS）框架表现出更优异的扩展性，在$R = 5$时仍能保持稳健的全局平均认证准确率，达到 73.5%。这比基线方法高出 64.3 个百分点，充分证明，即使攻击面扩大，词元级方法仍能有效维持安全保证。

深入考察具体危害类别后，结果揭示了可认证性方面鲜明的语义层次，区分了广泛的语义违规与棘手的技术威胁。该防御对普遍性危害表现出卓越的效力；例如，在测试的最大半径下，“自我伤害”和“仇恨言论”的可认证准确率仍分别高达88.1%和82.4%。然而，面对指令密集型或依赖关键词的类别时，模型会遇到更大挑战。“越狱模板”和“化学合成”的表现处于较低水平，可认证准确率分别为60.8%和65.1%。这种差异表明，在专业领域中，消融特定技术关键词会使意图检测变得更加困难；尽管如此，即使在这些最坏情况下，CSS的安全间隔仍显著高于基线。

保留率 Rate (k/L)	认证 半径 (R)	认证 准确率 @ R=5	MMLU 准确率- racy	推理 延迟 (ms)
0.1 （稀疏）	28	98.5%	34.2%	**45**
0.2	24	96.2%	48.6%	52
0.3	21	94.5%	56.1%	61
0.4	18	91.5%	62.4%	70
0.5	15	88.2%	65.8%	79
0.6	12	82.1%	**67.2%**	88
0.7 (最佳 平衡点)	8	73.5%	67.9%	98
0.8	4	41.5%	68.1%	109
0.9 （密集）	1	12.8%	68.3%	121

表3：Llama-3-8B 上保留率（k）的消融研究。半径 R 表示中位认证半径 ℓ_0。MMLU准确率是良性效用。

消融研究

表3是对“反向扩展”现象的关键实证验证，量化了理论安全性与实际效用之间的根本张力。通过对保留率（k）进行参数扫描，范围从稀疏上下文到密集上下文，数据揭示了可认证防御中固有的“天下没有免费的午餐”规律：要最大化一项指标，往往就必须牺牲另一项指标。这项消融研究对于刻画稀疏性-鲁棒性权衡（SRT）（Steunou、Druilhe 和 Saue，2025）至关重要，它表明，系统为了通过处理更少的词元来统计性地排除对抗性触发器，也会同时使模型缺乏复杂推理所需的语义上下文。

这种权衡的边界由保留率范围两端的性能指标清晰界定。在稀疏端（$k = 0.1$），模型的中位认证半径达到惊人的 28 个 token，理论上使其能够免受大规模扰动的影响。然而，这种安全性带来了灾难性的代价：MMLU准确率骤降至 34.2%，模型在推理任务中也因此几乎无法胜任。相反，在密集端（$k = 0.9$），模型虽然恢复了较高的效用，MMLU 得分达到 68.3%，但安全保证却荡然无存，认证半径缩减至仅一个 token，几乎无法抵御自适应攻击。

实验结果最终确定了一个明确的“最佳平衡点”，有效解决了这一优化问题。 $k = 0.7$配置成功平衡了相互冲突的目标，在保持 MMLU准确率为 67.9%（与基线性能在统计上相当）的同时，实现了稳健的中位认证半径 8 个 token。这一发现通过实证验证了噪声增强对齐调优（NAAT）的有效性，证实只要明确训练模型应对信息稀疏性，就有可能实现有意义的认证——而无需付出通常与稳健防御层相关的高昂对齐税。

结论

本研究针对大语言模型易受自适应对抗攻击这一关键漏洞，建立了一个在离散词元域中实现可认证鲁棒性的框架。为应对经验性防御在基于梯度的优化器面前的脆弱性，本研究通过分层随机消融提出了认证语义平滑（CSS）。与不适用于语言的连续方法不同，该方法利用超几何分布，为 l_0 范数扰动提供严格保证，确保安全性得到数学认证，而非仅凭观察得出。为使这一理论认证具备实际可行性，该框架引入噪声增强对齐调优（NAAT），以解决输入稀疏时通常会降低模型效用的反向缩放权衡问题。通过将模型训练为语义去噪器，NAAT 使基础表征与平滑分布相对齐，从而能够根据部分证据重建意图，并显著优于 SmoothLLM 等字符级基线。这确保了较高水平的认证安全性不会以牺牲模型的通用推理能力为代价。CSS 提供了一项确定性证书，保证模型在可证明半径范围内面对所有对抗变体时均保持安全，为在安全关键环境中部署大语言模型奠定了必要基础。由此建立的可验证防御层，无论攻击者采用何种具体优化策略，都能保持鲁棒性。

参考文献

Alon, G.; and Kamfonas, M. 2023年。利用困惑度检测语言模型攻击。*arXiv* 预印本 *arXiv:2308.14132*。

Chao, P.; Debenedetti, E.; Robey, A.; Andriushchenko, M.; Croce, F.; Sehwag, V.; Dobriban, E.; Flammarion, N.; Pappas, G. J.; Tramer, F.; 等。2024年。Jailbreakbench：用于大语言模型越狱的开放式鲁棒性基准。*神经信息处理系统进展*，37：55005–55029。

Cheng, Z.; Zhang, M.; Sun, J.; and Dai, W. 2025年。关于采用前景理论对齐的抗武器化大语言模型。收录于第*31*届国际计算语言学会议论文集，10309–10324。

Cohen, J.; Rosenfeld, E.; and Kolter, Z. 2019。通过随机平滑认证对抗鲁棒性。收录于国际机器学习会议，1310–1320。PMLR。Dong, Y.; Mu, R.; Zhang, Y.; Sun, S.; Zhang, T.; Wu, C.; Jin, G.; Qi, Y.; Hu, J.; Meng, J.; 等。2025年。大语言模型防护：综述。人工智能评论，58(12)：382。

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; 等。2024年。Llama 3 模型群。*arXiv* 预印本 *arXiv:2407.21783*。

Liu, X.; Xu, N.; Chen, M.; and Xiao, C. 2023年。AutoDAN：在对齐的大语言模型上生成隐蔽越狱提示。发表于 *ICLR*。Mehrotra, A.; Zampetakis, M.; Kassianik, P.; Nelson, B.; Anderson, H.; Singer, Y.; and Karbasi, A. 2024年。攻击树：自动越狱黑盒大语言模型。神经信息处理系统进展，37：61065–61105。Neyman, J.; and Pearson, E. S. 1933年。IX. 关于统计假设最有效检验问题。伦敦皇家学会哲学汇刊。*A*辑，数学或物理类论文集，231(694-706)：289–337。Robey, A.; Wong, E.; Hassani, H.; and Pappas, G. J. 2025年。SmoothLLM：防御大语言模型免受越狱攻击。*TMLR*。Shayegani, E.; Mamun, M. A. A.; Fu, Y.; Zaree, P.; Dong, Y.; and Abu-Ghazaleh, N. 2023年。对抗攻击揭示的大语言模型漏洞综述。*arXiv* 预印本 *arXiv:2310.10844*。Steunou, K.; Druilhe, T.; and Saue, S. 2025年。稀疏表示提升神经网络分类器的对抗鲁棒性。*arXiv* 预印本 *arXiv:2509.21130*。Yi, S.; Liu, Y.; Sun, Z.; Cong, T.; He, X.; Song, J.; Xu, K.; and Li, Q. 2024年。针对大语言模型的越狱攻击与防御：综述。*arXiv* 预印本 *arXiv:2407.04295*。Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J. Z.; and Fredrikson, M. 2023年。针对对齐的语言模型的通用且可迁移的对抗攻击。*arXiv* 预印本 *arXiv:2307.15043*。