导读：生成式人工智能的浪潮正席卷各行各业，在释放巨大生产力与创造力的同时，也带来了复杂多维的安全挑战。确保模型输出内容的安全可靠、保护核心参数资产免遭泄露、在全球应用中跨越文化与伦理的鸿沟，以及科学评估模型成效并系统治理相关风险，已成为当前亟待解决的关键议题。安全是生成式人工智能行稳致远的基石。为促进生成式人工智能安全领域的学术交流与前沿探讨，《网络安全与数据治理》期刊于2025年第11期特推出“生成式人工智能安全”主题专栏。本期专栏共遴选了5篇论文，内容涵盖“内容安全、数据安全、模型安全、文化适应与成效评估”5个维度，以期为相关领域的研究者提供参考借鉴。

专栏主编：

于静，博士，中央民族大学信息工程学院副教授，入选北京市科技新星人才项目，担任国际期刊TIFS编委、中国计算机学会区块链专委会执行委员、中国电子学会区块链分会委员。主要研究方向包括多模态信息内容安全、人工智能安全、跨模态人工智能等。在TIFS、TDSC、TIP、ICML、CVPR等国际期刊和会议发表论文70余篇。主持和参与国家自然科学基金、北京市科技计划项目等科研项目10余项。

赵悦，博士，中央民族大学教授，主要研究方向包括机器学习、少数民族语言语音处理技术、数据挖掘等。在藏语语音识别、语音合成、语音翻译方面取得一系列研究成果，先后主持和参与国家自然科学基金、国家社会科学基金等科研项目10余项，出版学术著作1部，教材3部。发表学术论文50余篇，获国家专利2项。

盖珂珂，博士，北京理工大学人工智能学院副院长，教授，博士生导师，入选国家级青年人才项目，担任IEEE技术与工程管理学会区块链专委会联合主席，中国计算机学会区块链专委会常务委员，新工科联盟区块链工委会秘书长，中国电子学会区块链分会常务委员。从事区块链、隐私计算、人工智能安全等方向的教学和科研工作，主持国家自然科学基金项目等科研项目10余项，发表论文200余篇。

领域大语言模型的内容安全控制研究

张欣欣[1]，李　涛[1]，赵龙彪[1]，贾真真[2]，周衡广[3]
（1. 中国人民解放军92981部队，北京　100161；2. 中国人民解放军91977部队，北京　100036；
3. 中国人民解放军91526部队，广东　湛江　524064）

摘　要：随着大语言模型在非通用领域中的广泛应用，其在知识管理、决策支持和安全信息交流等方面展现出巨大潜力。然而，这些领域具有高度的专业性和敏感性，在特定场景下确保输出内容的安全性与合规性是主要挑战。现有方法主要依赖模型的重新训练或微调，成本高且灵活性不足。提出了一种无需重新训练模型的精细化输出控制方法，将输出控制抽象为分类问题，利用分类算法对生成内容进行判断，决定是否输出。该机制结合数学建模与特征工程，力求在满足业务需求的同时，最大限度地减少潜在风险，提升输出的安全性与合规性。

关键词：大语言模型；安全控制；内容过滤；分类算法
中图分类号：TP309　　**文献标识码：**A　　**DOI：**10.19358/j.issn.2097－1788.2025.11.001
引用格式：张欣欣，李涛，赵龙彪，等. 领域大语言模型的内容安全控制研究［J］. 网络安全与数据治理，2025，44(11)：1－6.

Research on content safety control of domain-specific large language models

Zhang Xinxin[1], Li Tao[1], Zhao Longbiao[1], Jia Zhenzhen[2], Zhou Hengguang[3]
(1. Unit 92981 of the PLA, Beijing 100161, China; 2. Unit 91977 of the PLA, Beijing 100036, China;
3. Unit 91526 of the PLA, Zhanjiang 524064, China)

Abstract: With the increasing adoption of large language models in specialized domains, these models have demonstrated significant potential in areas such as knowledge management, decision support, and secure information exchange. However, given the high level of specialization and sensitivity in these domains, ensuring the safety and compliance of generated content in specific scenarios presents a major challenge. Current approaches predominantly rely on model retraining or fine-tuning, which are resource-intensive and lack flexibility. This study proposes a refined output control method that bypasses the need for model retraining. By framing output control as a classification problem, classification algorithms are employed to evaluate generated content and determine its appropriateness for release. This mechanism combines mathematical modeling and feature engineering to strike a balance between meeting business requirements and minimizing potential risks, thereby enhancing the safety and compliance of generated outputs.

Key words: large language model; safety control; content filtering; classification algorithm

0 引言

大型语言模型（Large Language Models，LLMs）近年来因其卓越的语言理解和生成能力而受到了广泛的关注。然而，这些模型也可能生成有害、侵犯隐私或者不安全的内容[1-2]，对用户和社会造成潜在的风险。而特定领域的大语言模型面向特定行业和特定需求，通常具有高度的专业性和敏感性，对安全要求更高。因此，对于非通用领域大模型来说，输出内容的安全性和合规性是主要的挑战之一。与现有方法不同，本研究提出的方法具有跨领域适用性，可以独立于 LLMs 的底层设计进行应用，并且通过干预模型输出来确保生成文本的安全性和合规性，从而为领域 LLMs 的安全控制提供了一种新颖且实用的解决方案。

为了有效控制大语言模型生成的内容，必须确保敏感信息的精准识别和安全过滤，同时满足特定场景的业务需求。为此，学者们提出了多种方法来增强模型的可靠性和内容质量，以应对这些问题。目前，主流的增强模型安全性和可靠性的方法是基于人类反馈的强化学习（Reinforcement Learning with Human Feedback，RLHF）[3]。通过人类反馈构建奖励模型，并利用该模型对 LLMs 进行训练，使其能够生成符合人类期望的内容。RLHF 架构的多个变体也相继提出，如 Safe-RLHF[4]、SENSEI[5] 和 f-DPG[6]，这些方法在不同方面进行了优化，如采用预训练的 LLMs 作为奖励模型，或者在信息检索领域中提升模型的表现[7]。然而，收集人类标注数据需要大量时间和成本。为了解决这一问题，一些研究提出了通过人工智能反馈代替人类反馈的强化学习[8]，从而降低对人类标注的依赖。还有研究致力于自动构建训练数据，以进一步降低成本和复杂性。为提高计算效率，差分偏好优化[9]是一种重要的尝试，该方法的核心思想是允许在不访问奖励模型的情况下使用相同的训练数据对 LLMs 进行训练。另一种常见的提高模型可靠性的方法是监督微调（Supervised Fine-Tuning，SFT）[10]，该方法通过大规模标注数据集对模型进行微调，以提升模型对用户需求的响应能力。RLHF 和 SFT 的共同点在于它们通过直接修改模型参数来提高模型的可靠性。

除了修改模型参数外，增强 LLMs 可靠性的另一种替代方法是直接干预输入提示或输出生成的过程。上下文学习（In-Context Learning，ICL）[11]是通过干预输入提示的一种主要方法。在 ICL 中，通过提供少量示例，可以引导 LLMs 完成特定任务，例如少样本学习[12]，从而减少生成不合规内容的风险。此外，一些研究集中于干预输出生成的方式。文献［13］提出了用于检索应用的输出格式化方法，避免 LLMs 在输出中重复相同词汇或短语。此外，Transformers 模块还提供了一些用于修正输出的函数，如 NoBadWordsLogitsProcessor 和 MinLengthLogitsProcessor。

现有的 LLMs 安全性控制方法主要依赖于预训练模型本身的优化或后处理技术。然而，这些方法通常存在局限性，例如依赖底层模型的设计或难以适用于不同领域的文本生成需求。为了解决上述方法灵活性不足的问题，有学者对 LLM 的输出过滤技术进行了一些研究，即在 LLM 生成文本后实施内容审查，无需修改模型参数[14]。针对输出内容的过滤技术，当前主要是通过预定义敏感

词库或正则表达式匹配拦截的基于规则的过滤，这种方法实现简单但泛化能力有限，难以识别语义变体以及进行细粒度权限控制[15]。

为了有效控制非通用领域大语言模型生成的内容，本文提出了一种基于数学建模、特征工程和分类算法的安全过滤控制方法，通过应用一个安全过滤器来干预LLMs的输出（即干预大语言模型生成序列的轨迹），进而确保生成内容符合安全和合规标准，以生成用户期望的结果。该方法不仅独立于LLMs的设计，还能够灵活地应用于不同领域的文本生成场景，具有广泛的适用性和较强的实用价值。

本文主要贡献如下：

本文提出了一种面向特定领域大语言模型的内容安全控制机制，设计了一个添加于LLMs输出层的外部过滤器，从而实现无需访问其模型参数即可控制输出内容。这是一个新颖的“无需学习”的LLMs安全控制策略，它不依赖LLMs的底层设计，可以应用于多种特定领域的LLMs，具有良好的通用性和适应性。

此外，本文针对特定领域的行业特点和安全隐私特性，抽取了一些特征因素，并结合分类算法和特征工程，在大语言模型内容安全控制领域做出了一些新的尝试。与现有基于规则或词典的安全过滤方法不同，特征工程技术结合分类算法能够更精确地识别和过滤潜在的风险文本，极大提升了检测精度和适用范围。

1　大语言模型文本生成

大型语言模型在文本生成过程中特别关注其结构和生成机制。在本文中，文本是指由一系列词汇单元（即标记）组成的序列，而所有可能的文本集合构成了模型的输入空间。自然语言中的某个特定表达可以被视为一段特定的文本，例如“我真的很棒”。

文本生成过程可以看作是一个从初始文本逐步扩展的迭代系统。其核心机制是基于当前已生成的文本内容来预测下一个最有可能的标记，以逐步生成完整的句子或段落。具体来说，文本生成的主要步骤如下：

（1）初始文本设定：首先，从一个初始文本开始，该初始文本作为生成过程的起点。比如，可以设定初始文本为“你好”，这一初始内容为模型的生成提供了上下文基础。

（2）标记预测：在每一步迭代中，模型会基于当前文本内容预测下一个最有可能的标记。这一预测过程通过计算每个可能标记出现在当前文本后面的概率来实现。这些概率反映了不同词汇单元在特定上下文中的适宜程度。

（3）标记选择：接下来，模型根据预测的概率分布，选择出最有可能的标记作为下一个词汇单元，并将其添加到当前文本的末尾。这样，当前文本得到扩展。

（4）文本扩展与迭代：这一过程会不断重复，即利用扩展后的文本作为新的上下文，再次进行标记预测和选择，直至生成出符合预期长度或语义要求的完整文本。

例如，假设初始文本为“我”，模型可能预测“真的”是接下来最合适的标记，生成的文本更新为“我真的”；随后，模型继续预测，认为“很棒”是下一个最有可能的标记，最终生成“我真的很棒”。

在这个生成过程中，还使用了一个称为“温度”的超参数，用于控制生成结果的多样性和稳定性。温度参数的设置影响模型对标记的选择方式：较高的温度值会增加生成文本的随机性，使得输出更加多样化；而较低的温度值则会使模型更倾向于选择概率最高的标记，输出更为保守和精确。因此，温度的调节对文本生成的灵活性和质量有重要影响。

综上所述，大型语言模型的文本生成是一个逐步扩展文本的过程，从初始文本开始，通过迭代地预测和选择新的词汇单元，逐步构建出语义连贯的句子或段落。这一生成过程可以被视为一个基于动态系统的递归式建模，其中标记预测和选择构成了生成的核心逻辑。

2　设计方法及步骤

本文提出了一种基于特征工程和分类算法的安全过滤器框架，重点是干预LLMs生成的文本序列。通过应用安全过滤器干预LLMs输出以生成安全内容，来提高LLMs输出的安全性和可控性。

安全过滤器核心由3个模块组成：首先，输入文本的预处理模块负责对原始文本进行分词、去噪等处理；其次，特征提取模块从处理后的文本中提取关键词、情感特征等信息；最后，分类算法模块根据提取的特征对文本进行分类，并决定是否允许输出该文本或进行修改。该过滤器通过干预生成文本的方式，确保输出符合安全性和合规性要求。具体流程如图1所示。

2.1　问题抽象

2.1.1　定义分类问题

输入：大模型生成的输出文本数据 D。

目标：对于每个文本片段 d（即一系列词汇单元），判断是否可以向用户 u 输出，即进行二分类：

类别1：允许输出。

类别0：不允许输出（需要替换为代号或屏蔽）。

2.1.2　分类依据

特征：综合考虑用户权限、数据敏感性、项目归属、

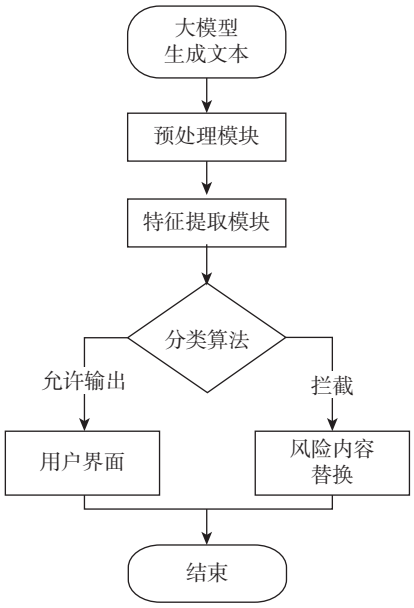

图 1　安全过滤器流程图

IP 地址等因素，提取用于分类的特征。

标签：根据权限策略，确定每个文本片段的真实标签（输出/不输出）。

2.2　特征工程

2.2.1　特征提取

对于每个待分类的文本片段 d，提取以下特征：

（1）用户特征

权限等级 $L(u)$：数值表示，数值越高权限等级越大。如普通员工为 1，项目成员为 2，高层管理为 3。

参与项目集合 Proj(u)：编码为多热向量。

IP 地址 IP(u)：编码为网络段所属类别。

（2）数据特征

敏感度等级 $S(d)$：数值表示，如低为 1，中为 2，高为 3。

所属项目集合 Proj(d)：编码为多热向量。

（3）交互特征

用户与项目交集大小为 $|\text{Proj}(u)\cap\text{Proj}(d)|$。

用户 IP 地址是否在允许范围内：二值特征，1 表示是，0 表示否。

（4）其他环境特征

访问时间：可以编码为工作时间（1）或非工作时间（0）。

地理位置：根据 IP 地址映射，编码为类别特征。

2.2.2　特征向量表示

将上述特征编码后，形成特征向量 $\boldsymbol{X}$：

$$\boldsymbol{X}=\begin{bmatrix} L(u),\ S(d),\ |\text{Proj}(u)\cap\text{Proj}(d)|, \\ \text{IP 允许性，其他环境特征} \end{bmatrix} \tag{1}$$

通过对特征的提取与编码，本研究能够有效地构建用于分类的特征向量，确保特征能够充分反映用户、文本及环境的各类信息。这种特征工程的方法使得分类模型能够更为准确地判断文本的输出安全性。

2.3　分类算法

根据场景需求选择分类算法，如逻辑回归（Logistic Regression）、决策树（Decision Tree）、随机森林（Random Forest）、梯度提升树（如 XGBoost、LightGBM）、支持向量机（SVM）和神经网络（Neural Networks）。

不同的场景对模型性能的要求各不相同：逻辑回归具有良好的可解释性，适合于对透明性要求较高的场景；决策树和随机森林在处理复杂特征交互方面表现优异；而梯度提升树在很多竞赛和实际应用中都显示出极强的性能表现。对于需要处理大量非线性特征的场景，支持向量机和神经网络是优选方案。

为客观比较不同分类算法在安全过滤任务中的性能并支撑选型决策，本文在模拟军工数据集上进行了对比实验。关键性能指标对比如表 1 所示。

表 1　分类算法性能对比

算法	准确率/%	召回率/%	可解释性
逻辑回归	91.5	88.2	极高
随机森林	94.2	92.5	中等
XGBoost	93.8	92.8	中等
支持向量机	92.6	90.8	低

领域安全基线要求召回率 > 85%，四类算法均满足安全标准，在拦截高风险内容上均具备基础可靠性，满足领域安全管控的入门门槛，其中随机森林和 XGBoost 在召回率上优于支持向量机和逻辑回归；随机森林与 XGBoost 在准确率上小幅领先，差距处于可接受范围；逻辑回归凭借线性权重可解释性，天然适配需透明决策的强监管场景，而随机森林和 XGBoost 通过特征组合捕捉复杂规则，适合对精度要求极高的非敏感场景，支持向量机在高维空间构建分隔边界，适用于特征维度较高的任务。

3　数学模型

3.1　分类函数

为了根据特征向量 $\boldsymbol{X}$ 判断文本片段是否可以输出，定义分类函数 $f(\boldsymbol{X})$ 输出为预测标签 y，即 y 为文本片段的安全标签，$y=1$ 表示允许输出（安全或风险可接受），$y=0$ 表示应当拦截（存在敏感或高风险内容）：

$$y=f(\boldsymbol{X})=\begin{cases}1, & P(y=1\mid\boldsymbol{X})\geqslant\tau \\ 0, & P(y=1\mid\boldsymbol{X})<\tau\end{cases} \tag{2}$$

其中，τ 为（0，1）区间内的实数判决阈值，当预测概率

$P(y=1\mid \boldsymbol{X})$不小于 τ 时，系统判定该文本片段可以安全输出，否则予以拦截或进行降级处理。

3.2 逻辑回归算法

以逻辑回归为例，其模型形式为：

$$P(y=1\mid \boldsymbol{X})=\sigma(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}+b) \tag{3}$$

其中：$\sigma(z)=\frac{1}{1+\mathrm{e}^{-z}}$为 sigmoid 函数，$\boldsymbol{w}$ 为权重向量，b 为偏置项。

当 $P(y=1\mid \boldsymbol{X})\geqslant 0.5$ 时，预测 $y=1$（允许输出）；否则预测 $y=0$。

4 示例演示

假设在某军工企业内部的装备智能问答大语言模型中有两个用户：普通员工（用户 A）和高层管理（用户 B）。智能问答系统在生成答案时需要利用逻辑回归模型决定是否输出文本片段，最终形成每个用户在提出同一个问题后生成不一样的回答的结果。

4.1 示例特征

假设智能问答系统在生成回答的初始阶段过程中，此时文本片段编号为 1；敏感度等级 $S(d)=2$；所属项目集合 $\mathrm{Proj}(d)=\{$项目 $X\}$。

对于用户 A，涉密等级为一般，权限等级 $L(u)=1$；用户 A 负责文印工作，并未参与任何项目，参与项目集合 $\mathrm{Proj}(u)=\{\}$；在工位访问智能问答大语言模型，IP 允许性 =1；访问时间为工作时间，时间特征 =1。则用户 A 特征向量表示为：$\boldsymbol{X}_{\mathrm{A}}=[1,2,0,1,1]$。

对于用户 B，涉密等级为重要，权限等级 $L(u)=3$；具有任何项目权限，参与项目集合 $\mathrm{Proj}(u)=\{$全部项目$\}$；在办公室访问智能问答大语言模型，IP 允许性 =1；访问时间为工作日午休时间，时间特征 =1。则用户 B 特征向量表示为：$\boldsymbol{X}_{\mathrm{B}}=[3,2,1,1,1]$。

4.2 分类结果计算

假设逻辑回归模型的权重向量为 $\boldsymbol{w}=[0.5,-0.8,0.3,0.7,0.2]$，偏置项为 $b=-0.1$。

计算 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{X}_{\mathrm{A}}+b=-0.3$；计算 sigmoid 函数值 $\approx 0.425\,6$。

由于 $P(y=1\mid \boldsymbol{X}_{\mathrm{A}})<0.5$，因此预测结果为 $y=0$，即不允许用户 A 访问该文本片段 1，也就是说，LLMs 为用户 A 生成的回答中不包含编号为“1”的本文内容；同理，$P(y=1\mid \boldsymbol{X}_{\mathrm{B}})>0.5$，预测结果为 $y=1$，即允许用户 B 访问该文本片段 1。

4.3 解释与处理

根据分类结果，大语言模型在生成回答时，会对用户 A 隐藏文本片段 1 或者用其他文本替换；而对于具有更高权限的用户 B，大语言模型的回答中是包含文本片段 1 的。

通过这种方式，逻辑回归模型能够结合用户特征和数据特征，实现对大语言模型输出的安全控制，确保敏感信息仅对具备相应权限的用户可见。

为了进一步增强示例的效果，本研究还进行了多场景的模拟实验，展示了该方法在不同类型用户和不同敏感度数据下的适用性。实验结果表明，分类算法能够灵活调整输出决策，有效降低敏感数据泄露的风险。

本部分展示了安全过滤器在文本生成过程中对内容质量和合规性的控制效果，安全过滤器通过对生成过程中的每个文本片段进行筛选，确保生成的内容安全合规，避免产生泄密信息。

图 2 展示了在使用安全过滤器时，生成文本过程中 $P(y=1\mid \boldsymbol{X})$值的轨迹。说明如下：安全过滤器的主要目标是防止生成内容的 $P(y=1\mid \boldsymbol{X})$ 值下降到 0.5 以下，并保持内容的质量与目标的一致性。

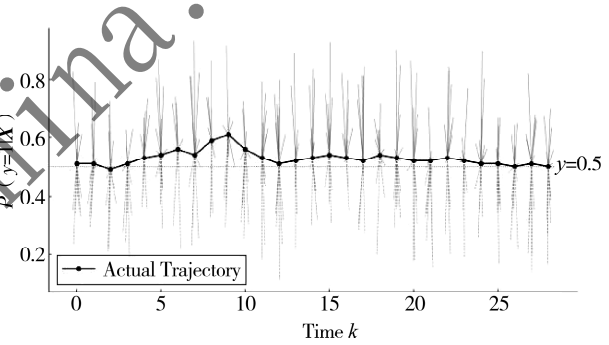

图 2 安全过滤器预测轨迹图

（1）$P(y=1\mid \boldsymbol{X})$ 值（纵轴）：表示生成过程中生成文本的 $P(y=1\mid \boldsymbol{X})$ 值。大于 0.5 表示内容比较安全，小于 0.5 表示内容存在风险。

（2）时间（横轴）：表示生成过程中每个文本片段的时间步，即生成的文本顺序。

（3）实线条：代表在每一步中，符合安全过滤器约束的文本片段，这些文本符合生成目标，因此被允许生成。

（4）虚线条：代表在每一步中，被安全过滤器排除的文本片段，即不符合安全标准的文本片段。通过这种筛选，确保生成内容不会偏离预期。

（5）圆点实线（Actual Trajectory）：表示最终实际生成的 $P(y=1\mid \boldsymbol{X})$ 值的轨迹。圆点实线显示了在每个时间步中，实际生成的文本片段的选择结果。可以看到，生成过程中，圆点实线大部分保持在大于 0.5 的 $P(y=1\mid \boldsymbol{X})$ 值区域，表明生成的文本相对安全。

安全过滤器通过以下几种方式有效地控制文本生成的内容：

文本片段筛选：在每个时间步，安全过滤器会将候选文本片段分为符合约束和不符合约束的两部分。实线条代表符合标准的文本片段，虚线条代表不符合标准的文本片段。

防止低值出现：通过对每一步的文本片段进行筛选，安全过滤器确保 $P(y=1 \mid \boldsymbol{X})$ 值不会降至 0.5，防止生成内容偏离目标。由此，生成的文本能够保持正向性，符合期望。

实际生成效果：圆点实线代表的实际生成路径表明，安全过滤器能够引导生成内容趋向于符合预期目标，并避免不合规的内容出现。

综上所述，安全过滤器能够有效地控制生成过程中的文本片段选择，确保生成的内容安全可靠。通过对每个时间步的候选文本片段进行筛选，安全过滤器保持让 $P(y=1 \mid \boldsymbol{X})$ 值大于等于 0.5，避免生成不符合预期的内容。这一方法不仅提高了文本的质量，也确保了内容的安全性和合规性。

5 结论

本文提出了一种面向特定领域的大语言模型内容安全控制机制，利用特征工程和分类算法构造安全过滤器实现对模型输出的精细化控制，确保文本生成系统能够生成安全合规的内容。安全过滤器作为基础 LLMs 的附加模块，它干预基础 LLMs 的输出，而无需对 LLMs 进行额外的训练。该机制显著提升敏感内容的识别能力的同时，避免了对底层模型的改动，增强了方法的适用性与扩展性。

尽管所提出的安全过滤器机制在静态文本生成场景中表现出良好性能，但在多轮对话、上下文嵌套以及语言风格漂移等复杂应用环境中仍面临一定挑战：在多轮对话中，模型需保持前后语义一致性，现有分类机制难以捕捉长距离上下文依赖；面对用户语言风格漂移和语义隐喻表达，分类器可能难以准确判断风险等级；此外，将该机制扩展至语音生成、图像生成等多模态输出仍需设计新的特征提取方式与判别策略。

未来将进一步探索通过集成多种 AI 模型和引入风格不变性编码器等技术手段优化安全过滤器，并尽可能应用于更多行业场景，不断提升其适用性和智能化水平。

参考文献

[1] SHEN T, JIN R, HUANG Y, et al. Large language model alignment: a survey [J]. arXiv preprint arXiv: 2309.15025, 2023.

[2] 王耀祖，李擎，戴张杰，等. 大语言模型研究现状与趋势 [J]. 工程科学学报，2024，46（5）：1201－1215.

[3] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback [J]. Advances in Neural Information Processing Systems, 2022, 35: 27730－27744.

[4] DAI J, PAN X, SUN R, et al. Safe RLHF: safe reinforcement learning from human feedback [J]. arXiv preprint arXiv: 2310.12773, 2023.

[5] LIU R, ZHANG G, FENG X, et al. Aligning generative language models with human values [C]//Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Seattle: Association for Computational Linguistics, 2022: 241－252.

[6] GO D, KORBAK T, KRUSZEWSKI G, et al. Aligning language models with preferences through f-divergence minimization [J]. arXiv preprint arXiv: 2302.08215, 2023.

[7] 刘昆麟，屈新纪，谭芳，等. 大语言模型对齐研究综述 [J]. 电信科学，2024，40（6）：173－194.

[8] GLAESE A, MCALEESE N, TREBACZ M, et al. Improving alignment of dialogue agents via targeted human judgements [J]. arXiv preprint arXiv: 2209.14375, 2022.

[9] RAFAILOV R, SHARMA A, MITCHELL E, et al. Direct preference optimization: your language model is secretly a reward model [J]. Advances in Neural Information Processing Systems, 2024, 36: 14528－14541.

[10] 张钦彤，王昱超，王鹤羲，等. 大语言模型微调技术的研究综述 [J]. 计算机工程与应用，2024，60（17）：1－15.

[11] DONG Q, LI L, DAI D, et al. A survey on in-context learning [J]. arXiv preprint arXiv: 2301.00234, 2022.

[12] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners [J]. arXiv preprint arXiv: 2005.14165, 2020.

[13] CAO N D, IZACARD G, RIEDEL S, et al. Autoregressive entity retrieval [J]. arXiv preprint arXiv: 2010.00904, 2020.

[14] LIU H, HUANG H, GU X, et al. On calibration of LLM-based guard models for reliable content moderation [J]. arXiv preprint arXiv: 2410.10414, 2024.

[15] HAN S, AVESTIMEHR S, HE C. Bridging the safety gap: a guardrail pipeline for trustworthy LLM inferences [C] //Proceedings of the 2025 IEEE International Conference on Machine Learning and Applications. San Francisco: IEEE, 2025: 1108－1126.

（收稿日期：2025－05－01）

作者简介：

张欣欣（1992－），女，硕士，助理工程师，主要研究方向：人工智能、大数据。

李涛（1982－），男，博士，副研究员，主要研究方向：人工智能、大数据。

赵龙彪（1986－），男，硕士，工程师，主要研究方向：人工智能、大数据。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部