

FCS 哈工大团队提出MUBENCH基准，揭示机器遗忘在提升安全性的同时，面临严重的过度安全与通用能力受损问题

作者：writer 来源：科学网

本文原地址：<https://www.iikx.com/news/progress/42724.html>

本文仅供学习交流之用，版权归原作者所有，请勿用于商业用途！

FCS 哈工大团队提出MUBENCH基准，揭示机器遗忘在提升安全性的同时，面临严重的过度安全与通用能力受损问题。 论文标题：The gains do not make up for the losses: a comprehensive evaluation for safety alignment of large language models via machine unlearning

期刊：Frontiers of Computer Science

作者：Weixiang Zhao, Yulin Hu, Xingyu Sui, Zhuojun Li, Yang Deng, Yanyan Zhao, Bing Qin Wanxiang Che

发表时间：17 October 2025

DOI：10.1007/s11704-024-41099-x

微信链接：点击此处阅读微信文章

RESEARCH ARTICLE

The gains do not make up for the losses: a comprehensive evaluation for safety alignment of large language models via machine unlearning

Weixiang ZHAO¹, Yulin HU¹, Xingyu SUI¹, Zhuojun LI¹, Yang DENG², Yanyan ZHAO¹✉, Bing QIN¹, Wanxiang CHE¹

1. Department of Computing, Harbin Institute of Technology, Harbin 150006, China

2. School of Computing and Information Science, Singapore Management University, Singapore 178902, Singapore

Received October 11, 2024; accepted December 23, 2024

E-mail: yyzhao@ir.hit.edu.cn

© The Author(s) 2025. This article is published with open access at link.springer.com and journal.hep.com.cn

标题：The gains do not make up for the losses: a comprehensive evaluation for safety alignment of large language models via machine unlearning

作者：Weixiang ZHAO, Yulin HU, Xingyu SUI, Zhuojun LI, Yang DENG, Yanyan ZHAO, Bing QIN, Wanxiang CHE

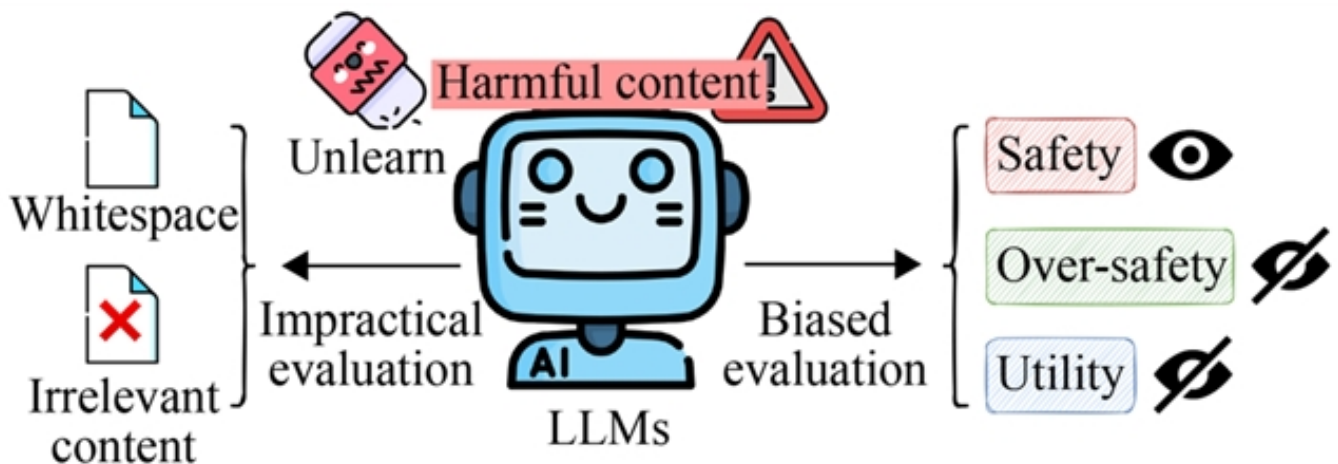
DOI：10.1007/s11704-024-41099-x



[扫码阅读原文↑](#)

摘要：

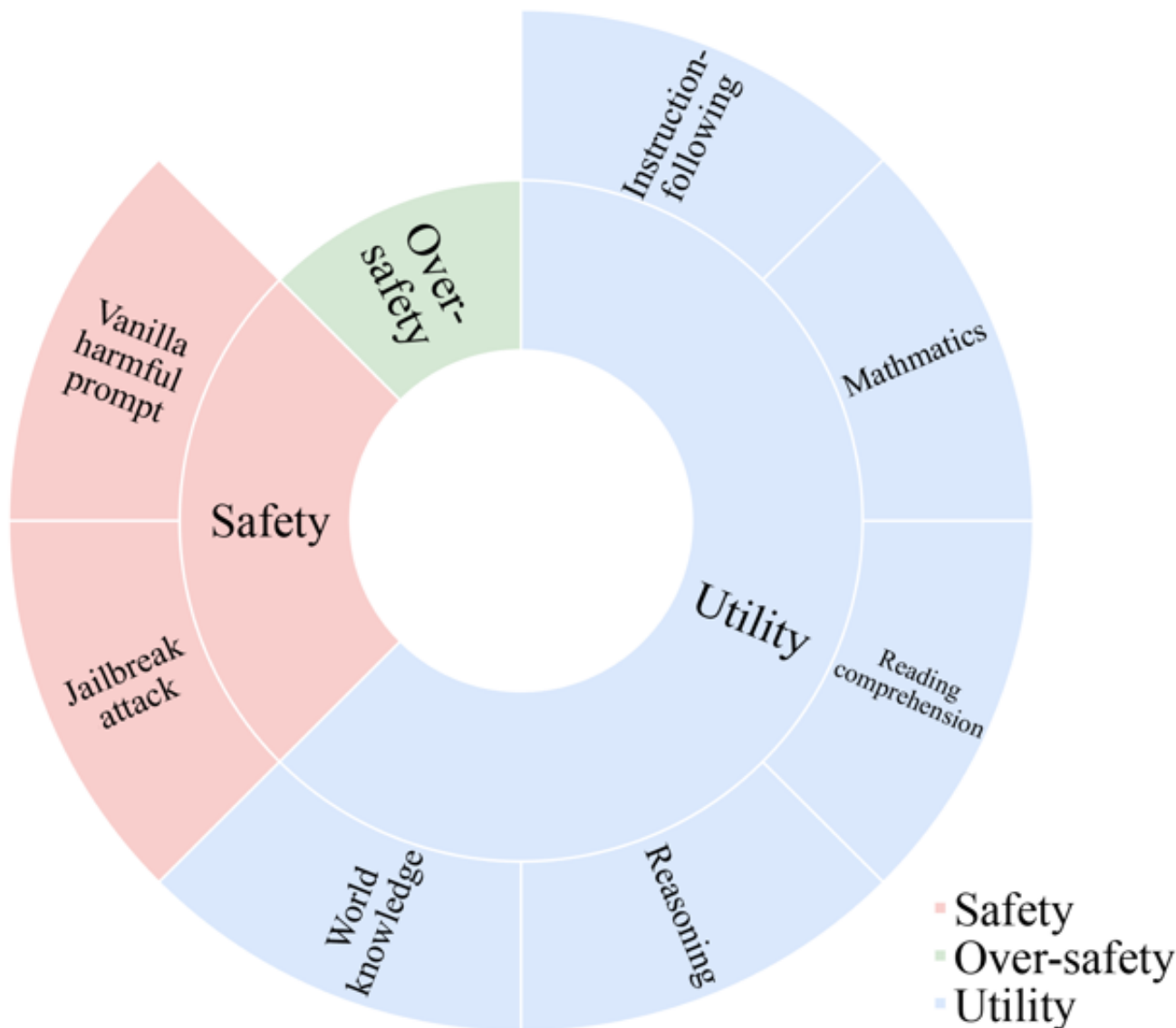
机器遗忘（Machine Unlearning, MU）已被视为一种让大语言模型（LLM）遗忘有害内容并实现安全对齐的新兴技术。然而，哈尔滨工业大学研究团队指出，现有评估标准过于单一，忽视了遗忘过程对模型通用性能的影响。为此，团队构建了包含18个数据集的全面评估基准MUBENCH，从安全性、过度安全性和通用效用三个维度对主流MU技术进行了深入测试。研究发现，虽然MU能有效抵御简单有害输入，但在面对越狱攻击时表现脆弱，且往往会导致模型产生严重的过度防御（Over-safety）和通用知识损失。



内容摘要

随着大语言模型（LLM）的广泛应用，确保模型不生成有害、偏见或非法的内容至关重要。传统的对齐方法（如RLHF）成本极高，因此机器遗忘（MU）作为一种更高效的替代方案受到了关注——它旨在精确地从已训练模型中删除特定负面知识。然而，当前的研究存在一个关键技术痛点：评估过程往往陷入安全盲区，即只关注模型是否变安全了，却忽略了它是否变得太安全以至于无法正常回答无害问题，以及其基础逻辑和常识能力是否遭到破坏。

针对这一问题，哈工大计算学部的研究团队提出了MUBENCH框架，旨在通过多维度的压力测试来审视MU的真实效果。实验涵盖了五种主流的遗忘方法，并引入了10种不同类型的越狱攻击（Jailbreak Attacks）。研究揭示了一个严峻的现实：现有的MU技术在增强安全性方面表现出极大的不稳定性，它们虽然能擦除表层的有害关联，但模型在面对精心设计的越狱提示词时依然会崩溃。



更重要的是，研究发现得不偿失现象普遍存在：为了实现安全目标，模型往往会产生严重的遗忘副作用，导致在数学推理、代码编写和常识问答等通用任务上的表现大幅下滑，并表现出过度的拒绝行为（即使是安全请求也会被模型拒绝）。该研究强调，未来的安全对齐工作必须在安全性与模型可用性之间寻求更精细的平衡，而非简单地进行知识擦除。

期刊简介

Frontiers of Computer Science (FCS) 是由教育部主管、高等教育出版社和北京航空航天大学共同主办，南京大学支持，SpringerNature 公司海外发行的英文学术期刊。本刊于 2007 年创刊，月刊，全球发行。主要刊登计算机科学领域具有创新性的综述论文、研究论文等。本刊主编为周志华院士，执行主编为周昆教授。编委会及青年 AE 团队由国内外知名学者及优秀青年学者组成。本刊被 SCI、Ei、DBLP、INSPEC、SCOPUS 和中国科学引文数据库 (CSCD) 核心库等收录，为 CCF 推荐 B 类期刊；两次入选中国科技期刊国际影响力提升计划；入选第 4 届中国国际化精品科技期刊；两次入选中国科技期刊卓越行动计划（一期梯队、二期领军）。

来源：Frontiers of Computer Science

更多 科学进展 请访问 <https://www.iikx.com/news/progress/>

本文版权归作者所有，请勿用于商业用途，[爱科学iikx.com](https://www.iikx.com)转发