基于大语言模型微调的出院小结生成“幻觉”抑制方法*

姜胜耀[1] 袁 铖[2] 朱立峰[1] 李寅驰[1] 范亚蔚[2] 张维彦[2] 阮 彤[2] 邵 炜[1]

([1] 上海交通大学医学院附属瑞金医院 上海 200025 [2] 华东理工大学 上海 200237)

〔摘要〕 **目的/意义** 解决大语言模型在出院小结生成过程中存在的“幻觉”问题，提升大语言模型的生成能力与上下文一致性。**方法/过程** 构建高质量、多层次的医疗指令数据集，采用基于分阶段训练的指令微调策略，引导大语言模型从简单到复杂任务逐步学习。在微调过程中引入数据回放与混合训练机制，确保大语言模型在新任务中保留和利用已有知识。**结果/结论** 该方法显著降低了大语言模型生成“幻觉”的发生率，提高了医疗文本生成准确性和可靠性。将课程学习理论与回放机制有效结合，不仅提升了模型对复杂任务的适应性，还确保了生成内容的专业性，同时展现出较高的实用性和可靠性。

〔关键词〕 大语言模型；出院小结生成；“幻觉”抑制；微调

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2025.02.003

Hallucination Suppression Methods for Discharge Summary Generation Based on Large Language Model Fine-tuning

JIANG Shengyao[1], YUAN Cheng[2], ZHU Lifeng[1], LI Yinchi[1], FAN Yawei[2], ZHANG Weiyan[2], RUAN Tong[2], SHAO Wei[1]

[1]Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China; [2]East China University of Science and Technology, Shanghai 200237, China

〔**Abstract**〕 **Purpose/Significance** To address the hallucination problem in discharge summary generation by using large language models, and to enhance the generative ability and contextual consistency of large language models. **Method/Process** The study constructs a high-quality, multi-level medical instruction dataset and employs a staged training-based instruction fine-tuning strategy to guide the model in learning tasks from simple to complex. A data replay and mixed training mechanism is introduced during the fine-tuning process to ensure that the large language model retains and utilizes existing knowledge when tackling new tasks. **Result/Conclusion** Experimental results show that the method proposed in this paper significantly reduces the occurrence of hallucinations in large language model generation and improves the accuracy and reliability of medical text generation. The superiority of this method lies in the effective integration of curriculum learning theory and the replay mechanism, which not only enhances the model's adaptability to complex tasks, but also ensures the professionalism of the generated content, while showing high practicality and reliability.

〔**Keywords**〕 large language models (LLM); discharge summary generation; hallucination suppression; fine-tuning

〔修回日期〕 2025-01-29

〔作者简介〕 姜胜耀，工程师，发表论文 8 篇；通信作者：李寅驰。

〔基金项目〕 上海市卫生和健康发展研究中心横向课题医学人工智能场景应用案例研究与社会实验调研。

1 引言

随着医疗信息化迅速发展，出院小结已成为现代医疗实践的重要组成部分，显著提高了医疗服务的效率和数据共享的便捷性。然而，当前大语言模型（large language models，LLM）在生成出院小结时常面临“幻觉”问题，即生成的内容与实际情况不符，严重影响了医疗记录的准确性，也可能对患者安全构成威胁。例如，对于检验结果这类数值密集型数据，LLM 可能难以正确地将数值与对应项进行匹配。此外，由于文书记录较长，模型可能会给出错误的诊断结论或推测，这些结果并非最终诊断，进而可能误导医生的后续治疗决策。因此，亟待开发有效的解决方案，以提高模型处理复杂医疗文档时的生成能力和上下文一致性。

本研究提出一种基于分阶段训练的指令微调策略。通过构建高质量、多层次的医疗指令数据集，逐步引导模型从基础任务过渡到复杂任务。在数据准备阶段，收集多种类型的医疗信息，确保训练数据的全面性和准确性。接着，任务被划分为初级、中级和高级 3 个难度阶段，使模型逐步适应医疗文本的复杂性，提升其对专业术语和结构化信息的理解能力。此外，引入数据回放与混合训练机制，有效巩固模型对已学知识的掌握。本研究的创新点在于结合课程学习理论与数据回放机制，显著提升了 LLM 在医疗文本生成中的准确性和专业性，并大幅抑制了“幻觉”，FactKB、UniEval 两个指标分别平均提高 17.60%、19.40%，为未来出院小结自动化生成提供了有效的解决方案。

2 相关研究

2.1 大语言模型

当预训练语言模型的参数达到一定规模时，会出现“涌现”现象。2022 年 11 月 ChatGPT[1] 以其卓越的自然语言交互与多场景内容生成能力迅速引起广泛关注。其基于 Transformer 架构，采用自回归结构，仅保留解码器，每次生成一个标记（Token）时依赖于先前生成的标记。然而，由于缺少整体语义的全局约束，在长文本生成过程中可能逐渐偏离原始语义或逻辑，产生“幻觉”现象。随后，业界推出诸多新模型，如 Palm[2]、LLaMA[3]、ChatGLM[4]、百川[5] 等。这些模型在语言生成任务中表现出良好的泛化能力，成为推动出院小结生成的重要工具。研究[6] 表明，结合医学知识库或专用数据，LLM 可以生成复杂的医疗文书，如临床记录等。然而，受医学内容专业性和敏感性影响，模型生成准确性仍不足，尤其是“幻觉”现象影响病历质量。

2.2 “幻觉”抑制

“幻觉”问题在出院小结生成中尤为严重，可能影响医疗服务的准确性和患者安全。目前常见的“幻觉”抑制手段包括提示工程和模型优化[7]。提示工程通过优化上下文或推理逻辑，引导模型生成正确回答[8]。例如，Vu T 等[9] 提出 FreshPrompt，通过搜索相关上下文样例增强提示。模型优化则利用模型输出分布对比和知识差异识别抑制“幻觉”。Shi W 等[10] 提出上下文感知解码策略，通过放大有无上下文条件下的输出差异，提升模型准确性。在出院小结生成中，“幻觉”抑制技术对提升医疗文本生成准确性和可靠性具有重要意义。

2.3 指令微调

出院小结生成作为专业领域任务，仅依赖预训练模型难以规避“幻觉”问题，因而指令微调成为常用策略。指令微调[11] 通过构造（指令，输入，输出）三元组引导模型学习特定领域知识，调整参数以约束输出符合预期。现有指令微调方法可分为 3 类。一是基于现有任务数据集，通过添加任务描述信息指导模型完成任务，例如 Super - Natural Instructions[12] 基于 Natural Instructions 从 61 个任务扩充到 1 600 多个任务。二是基于真实对话数据，利用人类或人机对话构建指令数据集，例如 ShareGPT[13] 通过记录 ChatGPT 与用户对话构建指令数据集。三是基于种子指令，通过提示生成新指令，例如 Self - Instruct[14] 从 100 多条种子指令扩展生成新指令。

3 模型训练方法

3.1 训练流程

为减少 LLM 在出院小结生成任务中的“幻觉”现象，以出院小结为生成对象，提出一种基于分阶段训练的指令微调策略，见图 1。首先，基于出院小结数据设计一系列辅助 LLM 理解的指令任务，并构造相应的微调数据集；其次，根据各指令任务对模型能力的要求，划分训练阶段，从易到难分阶段进行微调训练，以逐步提高模型能力；最后，在中、高级阶段各回放前一阶段部分数据，进行混合训练。该方法显著降低了“幻觉”发生率，提升了生成内容的可靠性。

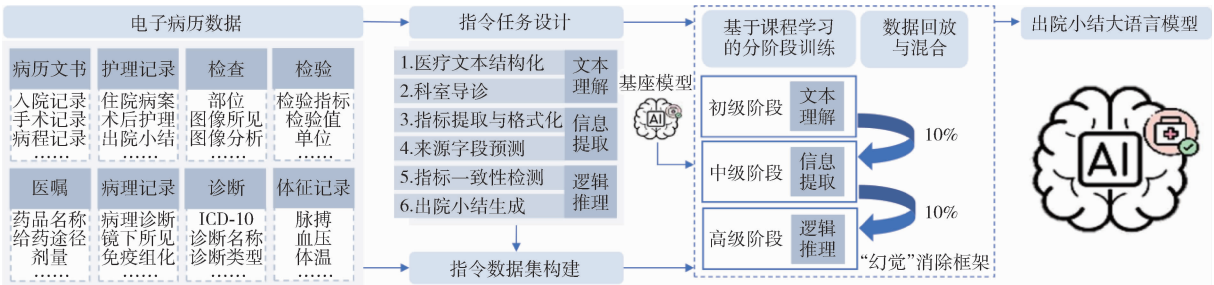

图 1 基于分阶段训练的指令微调流程

3.2 设计指令数据

为抑制“幻觉”产生，设计多个指令任务以辅助出院小结生成任务。由于该任务涉及复杂的文本生成与医疗文本理解，直接微调模型往往效果不佳。为增强模型的任务适应性，避免复杂任务理解不足引发的“幻觉”，设计 5 个辅助任务。指令任务共包含 6 种，见表 1。这些任务在文本理解、信息提取和逻辑推理方面紧密相连，层层递进，以提升模型能力。首先，医疗文本结构化任务和科室导诊任务提升医疗术语识别和疾病 - 科室映射能力，帮助模型识别并提取医疗信息，为异常检验检查或疾病史给出科室随访建议。其次，指标提取与格式化任务和来源字段预测任务增强了从病历中提取数值和关键信息的能力，确保模型快速准确地定位和提取医疗数据，提升出院小结中的数据完整性和准确性。最后，指标一致性检测任务和出院小结生成任务提升了逻辑推理能力，确保生成内容一致性，并通过信息整合生成高质量出院小结。针对每项任务，编写一系列基础提示阐述任务要求，提供处理策略或明确所需的输出格式。为了增强提示的表现力，帮助模型理解任务提示，引入 ChatGPT 进行提示扩充，在不改变原始意图的前提下，通过生成多样化的表达方式丰富提示内容。

表 1 出院小结指令任务

编号	任务名	任务介绍	衡量模型能力	指标
任务 1	医疗文本结构化	根据所需字段，将文本数据转为 json 数据	对医疗文本的术语识别与理解	ACC
任务 2	科室导诊	根据患者情况（主诉），推荐其应前往的科室	对症状文本和科室的理解	ACC
任务 3	指标提取与格式化	提取文本数据中的检验指标信息	对医疗文本（尤其是检验文本）的数值提取	ACC
任务 4	来源字段预测	根据给定医疗文本，判断其可能的来源字段	对医疗文本和文书字段的关键信息提取和语义理解	ACC
任务 5	指标一致性检测	判断文本中的检验信息与患者的检验数据是否完全匹配	对检验数值的敏感性，对医疗文本的比较和推理能力	ACC
任务 6	出院小结生成	根据给定病历信息，输出指定的出院小结某字段内容	对医疗文本的总体理解和推理生成	ACC、BertScore、Segment - ACC

3.3 训练策略

受课程学习[15-16]和数据回放混合[17]等方法的启示，提出两项训练措施。一是基于课程学习的分阶段训练方法，通过逐步增加任务复杂性，帮助模型适应更高挑战。二是结合数据回放与混合训练策略，通过回放历史数据并混合不同阶段的数据，增强模型泛化能力和稳定性。

3.3.1 基于课程学习的分阶段训练 对 6 项指令任务难度进行评级排序，划分为初级、中级、高级 3 个训练阶段，以便引导模型逐层学习。（1）初级。聚焦于医疗文本结构化和科室导诊两项任务，旨在为模型构建一个稳固且全面的基础医疗知识框架。输入为医疗文本的原始内容，输出为结构化的数据或推荐的就诊科室。在结构化任务中，模型需识别并转换医疗文本中的专业术语与关键信息，形成规范的数据格式。在科室导诊任务中，模型根据患者主诉和初步诊断信息，判断其应就诊的科室。（2）中级。涵盖指标提取与格式化、来源字段预测两项任务，旨在深化模型的指标理解与信息提取能力。输入为病历文本，输出为提取并格式化的健康指标或推测的来源字段。在指标提取与格式化任务中，模型从文本中识别关键健康指标并规范化输出，确保数据一致性。在来源字段预测任务中，模型根据文本内容推测信息来源，以构建完整的医疗信息链。（3）高级。该阶段是训练过程中最重要的一环，包含指标一致性检测和出院小结生成两项任务，旨在锤炼模型的信息对比和整合推理能力。输入为多来源医疗文本，输出为一致性检测报告或完整的出院小结。在一致性检测任务中，模型比对医疗记录中的指标数据，识别可能存在的不一致情况。在出院小结生成任务中，模型整合多种医疗信息，生成符合规范的出院小结报告。

3.3.2 数据回放与混合训练机制 为了提升模型的记忆巩固能力和任务适应灵活性，在各训练阶段融入数据回放与混合训练机制。输入包括当前阶段的训练数据以及前一阶段的部分数据，输出为更新后的模型参数，使其在学习新任务的同时保持对先前知识的记忆。通过对输入数据的联合优化，模型参数在不同阶段持续更新，从而提升对不同任务的适应性和稳定性。该机制使模型能够充分利用先前阶段积累的基础知识，为复杂任务处理提供坚实支撑。例如，在生成出院小结的高级阶段，模型能够高效调用其在中级阶段习得的关键技能，如精准高效的指标提取能力、敏锐准确的来源字段预测技巧。这些技能的协同作用，使模型能够更加精确无误、全面详尽地完成总结任务，避免信息遗漏或理解偏差导致的“幻觉”现象。

3.4 微调方法

遵循与传统指令微调相同的自回归训练方法，允许模型在生成每个输出时考虑先前所有的输入信息，有效捕捉序列中的长距离依赖。考虑到实际应用场景中模型面临的输入是多样化的，且指令微调的核心目的是依据给定的任务描述和输入数据生成相应的输出，因此在损失计算环节，对任务描述和输入部分实施遮蔽处理，以更精确地对模型进行优化。此外，采用 LoRA[18] 方法提升指令微调效率。在保留原有模型大部分参数的前提下，高效微调模型的少量参数，从而大幅提高模型在特定任务中的表现，计算方式如下。其中，θ_0 是基座模型的初始参数，$\Delta\theta$ 是通过指令微调得到的增量参数。LoRA 通过引入低秩矩阵适配技术，使模型参数更新既高效又灵活，从而在确保训练效果的同时，大幅提高训练速度和灵活性。

$$\theta^* = \theta_0 + \Delta\theta \tag{1}$$

4 实验

4.1 评估方案

4.1.1 数据集 收集来自 15 个科室的病历数据，每科室包含 5 000 名患者信息，涵盖病历文书、护理记录、检查、检验、医嘱、病理报告、诊断、体征记录 8 种医疗信息类型，见表 2。为确保数据质量，遵循严格的筛选标准，纳入完整记录的病历，排除缺失严重的数据。在数据清洗过程中，去除重复记录、统一日期等数据格式，删除出院小结和其

他病历文书结论不一致的样本，并对患者姓名、医生姓名等敏感信息进行匿名化脱敏处理，以最大限度地保护患者隐私。最终，数据按 8∶1∶1 的比例划分为训练集、验证集和测试集，分别用于模型的训练、参数调优与性能评估。

表 2 医疗信息类型

序号	信息类型	包含内容	结构
1	病历文书	入院记录、手术记录、查房记录等	带 html 标签的非结构化数据
2	护理记录	出院小结等	xml 数据
3	检查	检查信息	结构化数据
4	检验	检验信息	结构化数据
5	医嘱	化验、配药、文字性提醒事项等	结构化数据
6	病理报告	所做病理检查信息及其报告	结构化数据
7	诊断	患者从入院到出院期间各阶段医生给出的诊断信息	结构化数据
8	体征记录	患者从入院到出院期间，所做的体征检测信息	结构化数据

4.1.2 指标 使用准确度（accuracy，ACC）评估前 5 个指令任务效果，见公式（2）。根据出院小结不同部分的内容特性，分别选用 ACC、块准确度 Segment－ACC、语义相似度 BertScore 评估出院小结生成任务。一份完整的出院小结分为 6 部分，见表 3。对于格式固定、内容明确、范围统一的部分（Sec1、Sec2）采用 ACC 评估，计算方法与前文一致，其中$Text_{gold}$为临床医生书写的参考文本。对于格式固定、内容明确，但范围没有明确要求的部分（Sec3）采用 Segment－ACC 评估，见公式（3）。其中，$Segment_{model}$指模型生成文本的分词结果，$Segment_{material}$指患者住院期间的检验检查分词结果。对于需要压缩和综合分析大量信息进行摘要的部分（Sec4、Sec5、Sec6）选用 BertScore 作为评估指标。

$$ACC = \frac{Text_{model} \cap Text_{gold}}{Text_{gold}} \tag{2}$$

$$Segment - ACC = \frac{Segment_{model} \cap Segment_{material}}{Segment_{material}} \tag{3}$$

表 3 出院小结组成部分

编号	部分	描述内容	指标
Sec1	患者基本信息	个人详细信息、入院/出院时间、诊断相关信息	ACC
Sec2	出院诊断	患者出院时的最终诊断结果	ACC
Sec3	住院期间医疗情况	住院期间主要检验检查结果（例如：心电图）	Segment－ACC
Sec4	病程与治疗情况	住院期间的疾病发展、治疗、手术及术后恢复情况	BertScore
Sec5	出院时情况	患者健康状况和出院时恢复情况	BertScore
Sec6	出院后用药及建议	出院后药物（例如：服药频率与剂量）和随访计划	BertScore

4.2 实验设置

选用 ChatGLM3－6B 作为基座模型，在本地部署相应的实验环境，对其进行指令调优。实验环境配置，见表 4。超参数设置，见表 5。

表 4 实验环境配置

配置	规格
CPU	Intel（R）Xeon（R）Gold 6348 CPU @ 2.60 GHz
GPU	A100－40G＊8
Memory	504 G
深度学习框架	Torch 2.1.2

表 5 超参数配置

超参数	参数值
LoRA Rank	32
LORA alpha	64
优化器	AdamW
批处理大小	128
学习率	1e－4

4.3 实验结果与分析

4.3.1 主实验 为研究训练策略对出院小结生成

任务（任务 6）的影响，以 ChatGLM3 – 6B 模型作为基线，比较本文方法训练的 ChatGLM3 – 6B 模型与其他开源医疗大模型的生成效果，见表 6。所有模型均以零样本方式生成。本文模型在出院小结各部分的生成效果显著优于其他开源大模型，其对出院小结具有深度理解和有效推理能力。在 Sec2 生成方面，提升最为显著。这可能是由于该部分要求模型具备较强的语义理解和概括能力，而本文方法有效增强了模型在该方面的表现。相比之下，在 Sec4 生成方面提升相对有限，可能是因为该部分信息简洁集中，模型生成准确性本已较高，进一步提升空间较小。

表 6 各模型出院小结 6 部分生成性能（%）

模型	Sec1	Sec 2	Sec 3	Sec 4	Sec 5	Sec 6	平均
ChatGLM3 – 6B	46.85	65.24	32.14	58.74	51.79	55.42	51.70
BenTaso – 7B	44.35	36.85	5.23	50.01	38.74	48.85	37.34
HuatuoGPT – Ⅱ – 7B	45.83	52.07	21.03	65.17	36.73	52.04	45.48
Alpacare – 13B	35.08	25.80	5.07	28.99	14.41	23.79	22.19
本文模型（训练后的 ChatGLM3 – 6B）	94.48	98.48	93.81	88.31	83.99	86.42	90.92

4.3.2 *消融实验* 通过移除特定模块评估整体性能，若性能下降，证明所消除模块的有效性，反之亦然。共设计 6 组对比实验。（1）不训练。直接采用 ChatGLM3 – 6B 模型在各项任务上进行推理。（2）直接训练。将所有指令任务混合在一起，无阶段划分地进行微调训练。（3）随机分阶段训练。将指令任务随机分为 3 个阶段，每阶段 2 个任务，逐阶段对模型进行微调训练。（4）按能力分阶段训练。依据任务所需能力强弱，按从弱到强划分阶段，逐阶段进行微调训练。（5）回放 20% 训练。基于数据回放与混合训练策略，在按能力分阶段训练基础上，向中、高级阶段各混入 20% 的上一阶段指令数据进行混合微调训练。（6）回放 10% 训练（本文方法）。基于数据回放与混合训练策略，在按能力分阶段训练基础上，向中、高级阶段各混入 10% 的上一阶段指令数据进行混合微调训练。实验结果，见表 7。

表 7 不同方法在 6 种指令任务上的性能（%）

序号	方法	任务 1	任务 2	任务 3	任务 4	任务 5	任务 6	平均
1	不训练	11.82	48.29	62.87	16.04	6.39	51.70	32.85
2	直接训练	84.05	86.84	89.74	75.15	84.79	88.66	84.87
3	随机分阶段训练	78.56	89.08	89.94	76.80	86.73	89.30	85.07
4	按能力分阶段训练（回放 0%）	82.95	88.88	90.41	77.56	86.39	89.79	86.00
5	回放 20% 训练	84.86	89.33	90.45	79.09	85.54	89.70	86.50
6	回放 10% 训练（本文方法）	85.86	89.55	91.05	80.02	89.49	90.92	87.81

直接训练、随机分阶段训练、按能力分阶段训练 3 种情况下的实验结果表明，分阶段训练能够有效提升多数指令任务的性能。且按任务能力需求从弱到强逐步分阶段训练，能够进一步优化模型表现。这表明在指令微调训练期间，根据任务复杂度调整训练阶段具有重要意义。按能力分阶段训练、回放 20% 训练、回放 10% 训练（本文方法）3 种情况下的实验结果表明，混合前一阶段的指令数据可以有效提高 6 个指令任务的平均性能。然而，过多回放数据（如回放比例 20%）会导致性能下降，可能是由于模型未能充分适应当前阶段数据。

4.3.3 *“幻觉”评估* “幻觉”抑制是目前 LLMs 文本生成任务中的热门话题，在事关生命安全的医疗文本生成领域，显得尤为重要。为了探究

本文策略对“幻觉”的抑制效果，以出院小结生成任务为对象，从机器指标和人工评估两方面进行深入分析。（1）机器指标。选用专为“幻觉”检测所设计的 FactKB[19]、UniEval[20]指标评估出院小结生成的真实性，得分越高，与标准答案越一致，“幻觉”可能性越小。实验结果，见图 2—图 3。本文方法增强了出院小结生成的真实性，且广泛适用于不同科室，在 FactKB 上平均提升 17.60%，在 UniEval 上平均提升 19.40%，对“幻觉”抑制有显著效果。（2）人工评估。随机抽取 20 份病历，邀请 3 名临床专家（副主任医师及以上）对本文模型生成、ChatGLM3－6B 基座模型生成、临床医生书写的出院小结进行评估，共分为 3 个等级，出现数值或文本等语义错误视为严重“幻觉”，出现冗余描述但语义正确视为轻度“幻觉”，无冗余描述且语义正确视为没有“幻觉”，评估结果，见图 4。

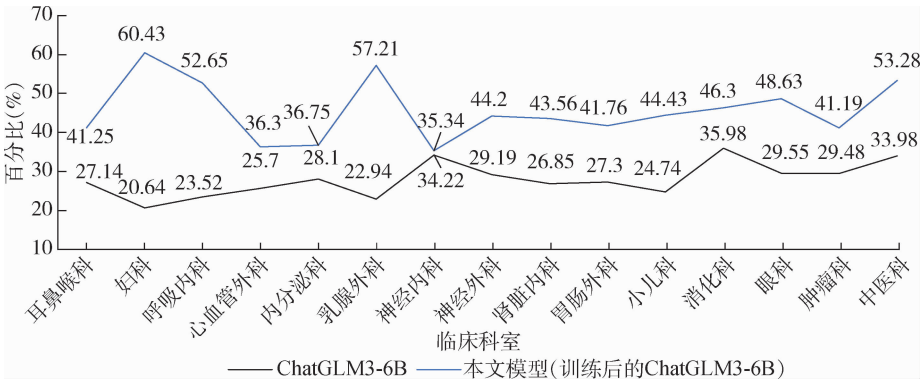

图 2　不同科室 FactKB 得分对比

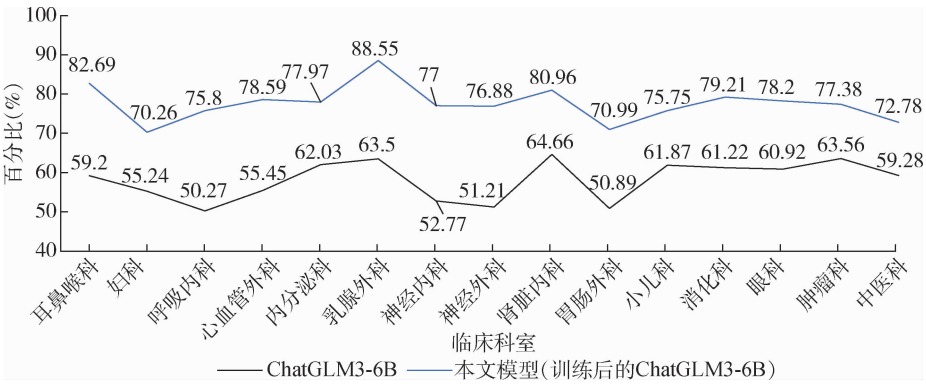

图 3　不同科室 UniEval 得分对比

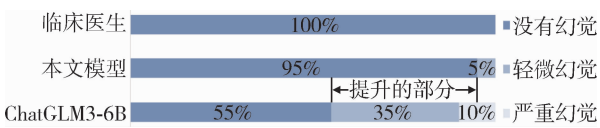

图 4　出院小结“幻觉”人工评分对比

本文方法显著提升了出院小结的生成质量，不会出现严重“幻觉”，近乎达到临床医生的文本质量。

“幻觉”抑制实例，见图 5。ChatGLM3－6B 模型对数值信息不敏感，且未正确生成诊断结论等文本信息，产生了“幻觉”。而基于本文方法训练的 ChatGLM3－6B 模型能较好地理解数值信息，准确概括诊断结论，有效抑制了“幻觉”发生，基本达到医生水准的书写质量。

ChatGLM3-6B	本文模型	临床医生
入院诊断：乳腺恶性肿瘤…	入院诊断：乳房肿块(右乳，癌疑)…	入院诊断：乳房肿块(右乳，癌疑)…
血压(BP)：null	血压(BP)：112/78mmHg	血压(BP)：112/78mmHg
脉搏(P)：76次/分	脉搏(P)：78次/分	脉搏(P)：78次/分
呼吸(R)：20次/分	呼吸(R)：20次/分	呼吸(R)：20次/分
体温(T)：约37℃	体温(T)：36.8℃	体温(T)：36.8℃
病程与治疗情况：	病程与治疗情况：	病程与治疗情况：
2019年3月20日，患者行右乳癌单纯乳房切除术+前哨淋巴结活检术备腋窝淋巴结清扫术+左乳象限切除术…	2019-03-25全麻下行左乳象限切除术+右乳癌单纯乳房切除术+前哨淋巴结活检术+低位组淋巴结清扫术…	2019-03-25全麻下行左乳象限切除术+右乳癌单纯乳房切除术+前哨淋巴结活检术+低位组淋巴结清扫术…
住院期间医疗情况：	住院期间医疗情况：	住院期间医疗情况：
血常规： 白细胞计数 4.20×10^9/L 中性粒细胞% 62.1% 淋巴细胞% 28.4% 嗜酸性粒细胞% 5.8↑% 红细胞计数 3.81×10^12/L 血红蛋白 114g/L 血小板计数 205×10^9/L …	【血常规】： 白细胞计数 3.80×10^9/L 中性粒细胞% 53.7% 淋巴细胞% 32.4% 嗜酸性粒细胞% 5.3↑% 红细胞计数 3.91×10^12/L 血红蛋白 120g/L 血小板计数 207×10^9/L…	血常规： 白细胞计数 3.80×10^9/L 中性粒细胞% 53.7% 淋巴细胞% 32.4% 嗜酸性粒细胞% 5.3↑% 红细胞计数 3.91×10^12/L 血红蛋白 120g/L 血小板计数 207×10^9/L…

图 5 “幻觉”抑制效果对比

5 结语

本研究针对 LLM 在出院小结生成任务中出现的“幻觉”现象，提出一种基于分阶段训练和数据回放的指令微调策略（代码开源地址：https://github.com/ycycyc02/Hallucination-Suppression-Framework）。通过构建多层次医疗指令数据集，并结合课程学习理论从易到难进行训练，提升了医疗文本生成的准确性和专业性。实验结果表明，该方法有效降低了“幻觉”发生率，增强了模型对复杂医疗文档的理解与生成能力，并巩固已学知识的应用，为医疗领域实际应用提供支持。

作者贡献：姜胜耀负责数据收集、提供临床专业指导、论文审核；袁铖负责实验设计、论文撰写；朱立峰、李寅驰负责提供指导、论文审核；范亚蔚负责数据分析；张维彦负责论文修订；阮彤、邵炜负责论文审核。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 AN J, DING W, LIN C. ChatGPT: tackle the growing carbon footprint of generative AI [J]. Nature, 2023, 615 (7953): 586.

2 CHOWDHERY A, NARANG S, DEVLIN J, et al. Palm: scaling language modeling with pathways [J]. Journal of machine learning research, 2023, 24 (240): 1-113.

3 TOUVRON H, LAVRIL T, IZACARD G, et al. Llama: open and efficient foundation language models [EB/OL]. [2024-12-20]. https://arxiv.org/abs/2302.13971.

4 GLM T, ZENG A, XU B, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 all tools [EB/OL]. [2024-12-20]. https://arxiv.org/abs/2406.12793.

5 YANG A, XIAO B, WANG B, et al. Baichuan 2: open large-scale language models [EB/OL]. [2024-12-20]. https://arxiv.org/abs/2309.10305.

6 KUMICHEV G, BLINOV P, KUZKINA Y, et al. MedSyn: LLM-based synthetic medical text generation framework [C]. Vilnius: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, 2024.

7 TONMOY S M, ZAMAN S M, JAIN V, et al. A comprehensive survey of hallucination mitigation techniques in large language models [EB/OL]. [2024-12-20]. https://arxiv.org/abs/2401.01313.

8 FELDMAN P, FOULDS J R, PAN S. Trapping llm hallucinations using tagged context prompts [EB/OL]. [2024-12-20]. https://arxiv.org/abs/2306.06085.

9 VU T, IYYER M, WANG X, et al. Freshllms: refreshing large language models with search engine augmentation [EB/OL]. [2024-12-20]. https://arxiv.org/abs/2310.03214.

10 SHI W, HAN X, LEWIS M, et al. Trusting your evidence: hallucinate less with context-aware decoding [C]. Mexico: The 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2024.

（下转第 35 页）

Phenominal: an EHR - integrated web application for structured deep phenotyping at the point of care [J]. BMC medical informatics and decision making, 2022, (22): 198.

12 MATTHIAS B, BRITTA B. Extraction of umls® concepts using Apache cTAKES™ for German language [J]. Studies in health technology and informatics, 2016 (223): 71 - 76.

13 LUO L, YAN S, LAI P T, et al. Phenotagger: a hybrid method for phenotype concept recognition using human phenotype ontology [J]. Bioinformatics, 2021, 37 (13): 1884 - 1890.

14 杨晨柳, 胡佳慧, 方安, 等. 临床文本自然语言处理系统构建研究——以 cTAKES 为例 [J]. 医学信息学杂志, 2018, 39 (12): 52 - 57.

15 DURANGO M C, TORRES - SILVA E A, OROZCO - DUQUE A. Named entity recognition in electronic health records: a methodological review [J]. Healthcare informatics research, 2023, 29 (4): 286 - 300.

16 张睿, 郑周荣, 陈薇. 罕见病辅助诊断临床决策支持系统综述 [J]. 中国数字医学, 2021, 16 (5): 86 - 90, 120.

17 BARBOSA - GOUVEIA S, VAZQUER - MOSQUERA M E, GONZALEZ - VIOQUE E, et al. Rapid molecular diagnosis of genetically inherited neuromuscular disorders using next - generation sequencing technologies [J]. Journal of clinical medicine, 2022, 11 (10): 2750.

18 LIEVIN V, JONAS M H, ALLAN L, et al. Findzebra online search delving into rare disease case reports using natural language processing [J]. Plos digital health, 2023, 2 (6), 269.

19 NICOLAS G, ANTOINE N, SALOMON R, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse [J]. Orphanet journal of rare diseases, 2018, 13 (1): 85 - 96.

20 JIA J, AN Z, MING Y, et al. ERAM: encyclopedia of rare disease annotations for precision medicine [J]. Nucleic acids research, 2018, 46 (1): 937 - 943.

21 WEBER G M. Federated queries of clinical data repositories: scaling to a national network [J]. Journal of biomedical informatics, 2015 (55): 231 - 236.

22 QIU L, SRUTHI G, VAIBHAV R, et al. Multi - disease predictive analytics: a clinical knowledge - aware approach [J]. ACM transactions on management information systems, 2021, 12 (3): 1 - 34.

23 PENG J, XUE H, HUI W, et al. An online tool for measuring and visualizing phenotype similarities using HPO [J]. BMC genomics, 2018, 19 (6): 89 - 97.

24 GAO J, ZHANG X, TIAN L, et al. mTGNN: multi - task graph neural network based few - shot learning for disease similarity measurement [J]. Methods, 2022 (198): 88 - 95.

25 SCHAAF J, SEDLMAYR M, SEDLMAYR B, et al. Evaluation of a clinical decision support system for rare diseases: a qualitative study [J]. BMC medical informatics and decision making, 2021 (21): 1 - 11.

(上接第 21 页)

11 WEI J, BOSMA M, ZHAO V Y, et al. Finetuned language models are zero - shot learners [EB/OL]. [2024 - 12 - 20]. https://arxiv.org/abs/2109.01652.

12 WANG Y, MISHRA S, ALIPOORMOLABASHI P, et al. Super - NaturalInstructions: generalization via declarative instructions on 1600 + NLP Tasks [C]. Abu Dhabi: The 2022 Conference on Empirical Methods in Natural Language Processing, 2022.

13 ShareGPT Dataset [EB/OL]. [2024 - 12 - 20]. https://sharegpt.com/.

14 WANG Y, KORDI Y, MISHRA S, et al. Self - Instruct: aligning language models with self - generated instructions [C]. Toronto: The 61st Annual Meeting of the Association for Computational Linguistics, 2023.

15 BENGIO Y, LOURADOUR J, COLLOBERT R, et al. Curriculum learning [C]. Montreal: The 26th Annual International Conference on Machine Learning, 2009.

16 PLATANIOS E A, STRETCU O, NEUBIG G, et al. Competence - based curriculum learning for neural machine translation [C]. Minneapolis: The 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019.

17 BUZZEGA P, BOSCHINI M, PORRELLO A, et al. Dark experience for general continual learning: a strong, simple baseline [C]. New York: The 34th International Conference on Neural Information Processing Systems, 2020.

18 HU E, SHEN Y, WALLIS P, et al. Low - rank adaptation of large language models [EB/OL]. [2024 - 12 - 20]. https://arxiv.org/abs/2106.09685.

19 FENG S, BALACHANDRAN V, BAI Y, et al. FactKB: generalizable factuality evaluation using language models enhanced with factual knowledge [C]. Singapore: The 2023 Conference on Empirical Methods in Natural Language Processing, 2023.

20 ZHONG M, LIU Y, YIN D, et al. Towards a unified multi - dimensional evaluator for text generation [C]. Abu Dhabi: The 2022 Conference on Empirical Methods in Natural Language Processing, 2022.