

中国人民大学

专业学位硕士研究生学位论文

开题报告

论文题目：基于数据清洗与偏好对齐的医疗问答大模型的幻觉检测与缓解技术研究

姓 名	胡佩文
学 号	202410428
院（系）	智慧治理学院
专 业	电子信息
研究方向	人工智能与大模型应用
学术教师	祁琦
日 期	2025.12.15

说 明

1. 专业学位硕士学位论文的开题报告是保证论文质量的一个重要环节，为了加强对专业学位硕士研究生培养的过程管理，规范其学位论文的开题报告，特制此表。
2. 此表一式两份经导师和培养单位负责人签字后，交培养单位研究生教学管理办公室存档。
3. 专业学位硕士研究生在申请学位论文答辩时，必须提交该学位论文开题报告。
4. 标题应该包含论文的核心技术或新管理方法以及研究对象，注意要具体，不要粗泛。

一、 简况

论文类型	名称	中文	面向医疗问答大模型的幻觉检测与缓解技术研究——基于数据清洗与偏好对齐的系统性方案
		英文	Research on Hallucination Detection and Mitigation Technologies for Medical QA Large Models: A Systematic Scheme Based on Data Cleaning and Preference Alignment
	类别	1. 技术攻关研究✓ 2. 工程项目策划 3. 工程设计或技术改造 4. 新系统、新设备、新产品的研制与开发	
		形式	1、研究论文✓ 2、系统研制报告 3、工程设计
摘要	<p>随着大语言模型（Large Language Models, LLMs）在医疗健康领域的深入应用，其在临床辅助决策、医患沟通及健康咨询中的潜力已得到广泛验证。然而，当前主流模型普遍存在的“幻觉”（Hallucination）现象——即生成内容看似流畅合理但违背医学事实、逻辑自相矛盾或包含非既存知识——已成为制约其在严肃医疗场景落地的核心瓶颈。特别是在中文医疗语境下，由于医学术语的复杂性、中西医知识体系的异构性以及高质量中文医学对齐数据的稀缺性，幻觉问题呈现出高隐蔽性与高风险性的双重特征。</p> <p>本研究针对上述痛点，提出了一套从数据源头治理到模型末端对齐的系统性全流程解决方案，旨在构建“可信、可控、可溯源”的中文医疗问答大模型。研究工作主要包含三个层面：第一，构建基于医学知识图谱（Knowledge Graph, KG）的自动化数据清洗流水线。利用命名实体识别（NER）与关系抽取技术，将非结构化医学对话映射为结构化三元组，并通过与权威医学图谱（如 CMeKG、UMLS）进行实体一致性校验与逻辑冲突检测，从源头剔除训练语料中的噪声与事实性错误。第二，设计融合不确定性分析与检索增强（RAG）的多维度幻觉检测机制。结合白盒状态下的 Token 级熵值（Entropy）分析、特征值谱分析（EigenScore）与黑盒状态下的基于检索的事实核查（Fact-Checking），构建“双重防线”，有效识别并实时拦截包含错误用药建议、禁忌症冲突等高风险回答。第三，提出基于“对抗性实体替换”的直接偏好优化（DPO）对齐策略。通过构造语义高度相似但包含细微医学谬误的“困难负样本”（Hard Negatives），利用 SimPO 或 DPO 算法进行偏好对齐训练，迫使模型在参数更新中内化医学事实约束，从根本上降低幻觉生成概率。</p>		
主题词	主题词数量不多于三个，主题词之间空一格（英文用“/”分隔）		
	中文	医疗大模型 幻觉检测 偏好对齐	
	英文	Medical LLM / Hallucination Detection / Preference Alignment	

二、选题依据

1. 阐述该选题的研究意义，包括理论意义、实践意义或工程设计的价值和意义，国内外概况和发展趋势，选题的先进性和实用性，技术难度及工作量。（研究意义指本文研究的意义，不是项目的价值。研究意义要紧扣主题，说明为什么做此论文。注意问题不要罗列太多，偏重本文要解决的重要问题。然后说明标题提到的核心技术因某些优点可以解决此问题，最后解决此问题带来的价值即可。语言简明扼要，语句通顺。不要出现常识性和简单介绍性的内容。）

1.1 阐述该选题的研究意义

理论意义：

本研究旨在深入探索垂直领域大模型知识边界的形成机制与事实对齐（Factuality Alignment）的内在规律。尽管通用大模型（如 GPT-4, Claude 3, DeepSeek-V3[5]）展现出强大的泛化能力，但在特定垂直领域（Domain-Specific），尤其是对准确性要求极高的医疗领域，模型往往因训练数据的长尾分布和预训练阶段的概率预测本质而产生“知识模糊”。本研究通过分析数据质量、训练目标（SFT vs. RLHF/DPO）与模型幻觉行为之间的因果关系，揭示医疗幻觉产生的深层机理（如数据源污染、解码策略偏差、知识回忆失败等）。同时，本研究将验证并拓展直接偏好优化（DPO）[9]、简单偏好优化（SimPO）[10]、Kahneman-Tversky 优化（KTO）[11] 等前沿算法在复杂逻辑推理与强事实约束任务中的适用性，丰富大模型可信性（Trustworthiness）与安全对齐（Safety Alignment）的理论框架，为解决“黑盒”模型的可解释性问题提供新的视角。此外，通过引入知识图谱（KG）作为逻辑约束，本研究探索了符号主义与联结主义在医疗 AI 中的融合路径[13]，为大模型的慢思考能力构建提供了理论依据。

实践意义与应用价值：

- 提升医疗安全性与合规性：医疗咨询具有“零容忍”的高风险特性。模型生成的错误剂量、禁忌症遗漏或误诊建议可能直接危害患者生命安全。本研究研发的多维度幻觉检测算法，能够作为“AI 守门员”，有效识别并拦截高危内容，显著降低医疗 AI 应用中的法律风险与伦理隐患，为 AI 医疗器械的审批与合规化应用提供技术保障。
- 优化智慧医疗服务效能：当前“AI 医生”在面对复杂、多轮的真实医患对话时，常出现前后矛盾或逻辑断裂。通过构建自动化数据清洗链路和基于对抗性样本的 DPO 微调，本研究旨在解决模型在长尾疾病、复杂并发症咨询中知识模糊的问题，提升辅助诊疗的准确率与用户信任度，推动智慧医疗从“玩具级”向“工具级”跨越，促进优质医疗资源的数字化下沉。
- 推动国产医疗大模型生态发展：针对中文医疗语境的特殊性（如中医辨证论治、特有药品名），本研究构建的高质量偏好数据集与评测基准，将填补现有开源资源的空白，为学术界和工业界提供可复用的数据资产与评测标准，助力国产医疗大模型（如 HuatuoGPT-II[3], BenTsao[14], Taiyi[31]）的迭代升级。

1.2 国内外概况和发展趋势：

医疗大模型发展现状：国际上，Google 的 Med-PaLM 2[2]通过指令微调在 USMLE

考试中达到专家水平。国内如香港中文大学的 HuatuoGPT 和哈工大的“本草”模型，分别利用混合训练策略和知识图谱注入提升了中文医疗问答能力。但现有研究多侧重于标准化考试，在真实医患对话中的事实准确性控制仍显不足。

幻觉检测技术进展：主流方法包括基于不确定性检测（如熵值分析）和基于检索的验证（Fact-Checking）。现有方法单一手段难以兼顾准确率与效率，缺乏针对医疗隐性幻觉的综合检测框架。

幻觉缓解技术：从 RLHF（基于人类反馈的强化学习）向 DPO（直接偏好优化）演进是当前趋势。DPO 训练更稳定，但在医疗垂直领域的应用尚处于起步阶段，特别是高质量医疗偏好对的构建策略仍有待探索。

2. 国内外研究现状分析。（代表性的方法、观点、技术、成果等的汇总、分析和对比。）

2.1 医疗大模型的发展现状

近年来，大语言模型在医疗领域的应用呈现爆发式增长，形成了“通用基座+领域微调”与“全流程预训练”并行的发展范式。

- **国际领先水平：**Google 发布的 Med-PaLM 2 [2]代表了当前医疗 LLM 的最高水平，通过在海量医学语料上进行指令微调（Instruction Tuning）和集成优化（Ensemble Refinement），其在 USMLE（美国执业医师资格考试）中达到了专家级水平（86.5%准确率），并展现出极强的长文本生成与推理能力。此外，BioMistral [4]基于 Mistral 架构，利用 PubMed Central 等生物医学文献进行持续预训练，显著提升了在生物医学任务上的性能，且保持了较小的参数规模（7B），便于本地部署和隐私保护。ChatDoctor [30]通过在 LLaMA 模型上微调约 10 万条真实医患对话，并引入 Wikipedia 和 MedlinePlus 作为外部知识库，增强了模型的问答能力。

- **国内发展态势：**国内医疗大模型发展迅速，呈现百花齐放的态势。香港中文大学（深圳）团队发布的 HuatuoGPT-II[3] 采用“一阶段适应”（One-stage Adaptation）策略，将预训练与 SFT 数据统一格式，有效缓解了灾难性遗忘，在中文医疗问答中表现优异。哈尔滨工业大学发布的 本草（BenTsao）[14]模型则强调知识图谱的结构化注入，利用 CMeKG 增强模型的实体理解能力。Zhongjing（仲景）[15]模型则引入了完整的 RLHF 流程，并通过构建多轮对话数据集 CmtMedQA[16]提升了模型的主动问询能力。Taiyi（太一）[31]模型针对双语生物医学任务进行了微调，涵盖了丰富的中英文医学 NLP 数据集。此外，DeepSeek-V3[5]等通用模型凭借强大的推理能力（Reasoning）和超长上下文窗口，在医疗基准测试中也展现出惊人的潜力，甚至在部分中文医学推理任务上超越了专门微调的较小模型。

- **局限性：**现有研究大多侧重于提升模型在标准化考试（如 MedQA, MCMLE）中的单选题准确率，而在真实、复杂、多轮的医患对话场景中，对于长文本生成的逻辑一致性、循证医学依据（Evidence-Based Medicine）的引用准确性以及对“未知”问题的拒答能力仍显不足。特别是中西医结合场景下的幻觉问题尤为突出[1,24]，且缺乏统一的、细粒度的幻觉评测标准。

2.2 幻觉检测技术的研究进展

幻觉检测是治理幻觉的第一道防线，目前学界主要分为三类技术路线：

1. **基于不确定性的白盒检测（Uncertainty-based Detection）：**该类方法依赖

模型内部状态。研究表明，模型生成幻觉时，其 Token 的对数概率（Logits）分布往往表现出较高的熵值（Entropy）或较低的置信度。SelfCheckGPT [6]提出了一种零资源（Zero-resource）检测方法，通过对同一提示进行多次采样并计算样本间的一致性（Consistency）来判定幻觉，若多次生成内容大相径庭，则判定为幻觉。INSIDE[20]方法进一步探索了利用模型内部隐藏状态（Hidden States）的特征值（EigenScore）来检测幻觉，该方法在语义空间中衡量生成内容的自洽性，比单纯依赖输出概率更为鲁棒。

2. **基于检索的黑盒验证（Retrieval-based Fact-Checking）**：利用外部知识库（如 PubMed, Wikipedia, 临床指南）作为“真理源”。FacTool[8] 框架通过调用搜索引擎或数据库检索证据，再利用 LLM 验证生成内容与证据的蕴含关系（Entailment），实现了多任务的真实性检测。FActScore [7]提出将长文本拆解为原子事实（Atomic Facts），逐一验证每个原子事实的准确性，提供了细粒度的量化指标。RAGTruth [17]构建了一个基于 RAG 场景的幻觉语料库，证明了基于高质量数据微调的小模型在幻觉检测上可以媲美 GPT-4。

3. **基于大模型自查与交互（LLM-as-a-Judge & Interactive）**：利用 GPT-4/Gemini-2.5 等强模型作为裁判，通过思维链（CoT）推理来评估生成内容的真实性。G-Eval[26] 框架证明了使用 CoT 的大模型评分与人类评估具有高度一致性。MetaQA [25]引入了蜕变测试（Metamorphic Testing）的思想，通过对输入问题进行语义不变的变换（如重述），观察模型输出是否保持一致，从而在不依赖外部知识库的情况下检测幻觉。

分析：单一方法难以应对医疗幻觉的隐蔽性。白盒方法在面对“过度自信”的模型时容易失效；黑盒检索方法成本高昂且依赖检索质量。本研究拟结合不确定性（低成本初筛）与 RAG 验证（高精度复核），构建混合检测框架，并引入知识图谱作为结构化验证源。

2.3 幻觉缓解与对齐训练技术

在检测之外，如何从根本上减少幻觉是研究重点，主要分为数据中心与模型中心两类方法。

1. **数据中心方法（Data-Centric）**：强调“Garbage In, Garbage Out”。研究指出，训练数据中的事实错误是幻觉的主要来源。Woodpecker[12] 框架提出了一种无需训练的幻觉修正方法，通过提取关键概念并检索视觉/文本证据进行后处理修正。在医疗领域，利用知识图谱（Knowledge Graph）进行数据增强和清洗已成为趋势，如利用 SNOMED CT 或 UMLS 对训练语料进行实体对齐和矛盾剔除[23]，构建高质量的指令微调数据集。此外，自动化数据清洗流水线（Automated Pipeline）结合 LLM 与启发式规则，已被证明能显著提升临床数据的质量[29]。

2. 模型中心方法（Model-Centric）：

- **Prompt Engineering & Inference Strategy:** Chain-of-Verification (CoVe)[18] 通过让模型生成验证问题并自我回答来减少幻觉。Thinking（如树搜索、慢思考）被引入推理过程，通过多步推理和自我反思来抑制即时生成的错误。

- **Alignment Training:** 传统的 RLHF 虽然有效，但 PPO 算法训练不稳定。DPO（直接偏好优化）[9] 作为 RLHF 的替代方案，直接基于偏好数据优化策略，训练更加稳定。近期提出的 SimPO（简单偏好优化）[10] 进一步简化了 DPO，去除了参考模型（Reference Model），通过长度归一化和目标边距（Target Margin）在减少幻觉的同时避免了生成长

度的冗余，且计算效率更高。KTO (Kahneman-Tversky 优化) [11] 基于前景理论，利用二元信号 (好/坏) 进行对齐，更适合利用大量非配对的医疗反馈数据。针对多模态医疗模型，MMedPO [19] 和 SymMPO [28] 等方法通过构建视觉-文本的一致性偏好对，有效缓解了多模态幻觉。OPA-DPO [22] 则进一步通过在线策略校准来减少视觉语言模型中的幻觉。

分析： DPO 及其变体 (SimPO, KTO) 在通用领域已证明了抑制幻觉的有效性，但在医疗垂直领域的应用尚处于起步阶段。特别是如何构建高质量的、针对医疗事实错误的“对抗性负样本” (Hard Negatives, 例如：将“阿司匹林”替换为“布洛芬”以模拟易混淆药物)，而非简单的随机负样本，是提升对齐效果的关键难点[21]。本研究将重点探索基于知识图谱构建对抗性偏好对的策略。

3. 主要参考文献 (列出作者、论文名称、期刊名称、出版年月)。

- [1] Alansari, A., & Luqman, H. (2025). "Large Language Models Hallucination: A Comprehensive Survey." *arXiv preprint arXiv:2510.06265*.
- [2] Singhal, K., et al. (2023). "Towards Expert-Level Medical Question Answering with Large Language Models." *Nature*, 620(7972), 172-180. (Med-PaLM 2)
- [3] Chen, J., et al. (2023). "HuatuogPT-II, towards taming language model to be a doctor." *arXiv preprint arXiv:2311.09774*.
- [4] Labrak, Y., et al. (2024). "BioMistral: A Collection of Open-Source Pretrained Large Language Models for Medical Domains." *arXiv preprint arXiv:2402.10373*.
- [5] DeepSeek-AI. (2024). "DeepSeek-V3 Technical Report." *arXiv preprint arXiv:2412.19437*.
- [6] Manakul, P., et al. (2023). "SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models." *EMNLP 2023*.
- [7] Min, S., et al. (2023). "FactScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation." *EMNLP 2023*.
- [8] Chern, I., et al. (2023). "FacTool: Factuality Detection in Generative AI -- A Tool Augmented Framework for Multi-Task and Multi-Domain Scenarios." *arXiv preprint arXiv:2307.13528*.
- [9] Rafailov, R., et al. (2024). "Direct Preference Optimization: Your Language Model is Secretly a Reward Model." *NeurIPS 2023*.
- [10] Meng, Y., et al. (2024). "SimPO: Simple Preference Optimization

- with a Reference-Free Reward." *arXiv preprint arXiv:2405.14734*.
- [11] Ethayarajh, K., et al. (2024). "KT0: Model Alignment as Prospect Theoretic Optimization." *ICML 2024*.
- [12] Yin, S., et al. (2024). "Woodpecker: Hallucination Correction for Multimodal Large Language Models." *arXiv preprint arXiv:2310.16045*.
- [13] Zhang, Y., et al. (2025). "DR.KNOWS: Integrating Knowledge Graphs for Diagnostic Reasoning in Healthcare." *JMIR Medical Informatics*.
- [14] Wang, H., et al. (2023). "Huatuo: Tuning LLaMA Model with Chinese Medical Knowledge." *arXiv preprint arXiv:2304.06975*. (BenTsao)
- [15] Yang, R., et al. (2024). "Zhongjing: Enhancing the Chinese Medical Capabilities of Large Language Model through Expert Feedback." *AAAI 2024*.
- [16] Zhu, N., et al. (2024). "CMtMedQA: A Chinese Multi-turn Medical Question Answering Benchmark." *ACL BioNLP 2024*.
- [17] Cheng, S., et al. (2024). "RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models." *ACL 2024*.
- [18] Dhuliawala, S., et al. (2023). "Chain-of-Verification Reduces Hallucination in Large Language Models." *arXiv preprint arXiv:2309.11495*.
- [19] Zhu, K., et al. (2024). "MMedPO: Multimodal Medical Preference Optimization for Hallucination Mitigation." *ICML 2024*.
- [20] Su, W., et al. (2024). "Unsupervised Real-Time Hallucination Detection based on the Internal States of Large Language Models." *Findings of ACL 2024*.
- [21] Liu, X., et al. (2025). "Entity-based Hard Negative Mining for Medical Preference Alignment." *Electronics*.
- [22] Microsoft Research. (2025). "OPA-DPO: Efficiently Minimizing Hallucinations in Large Vision-Language Models." *Microsoft Research Blog*.
- [23] Li, P., et al. (2025). "Automated Knowledge Graph Construction for Medical LLMs using SNOMED CT." *PMC*.

- [24] Gu, Y., et al. (2025). "Med-VH: A Survey of Visual Hallucinations in Medical Large Multimodal Models." *medRxiv*.
- [25] Zhang, Y., et al. (2025). "MetaQA: Hallucination Detection in Large Language Models with Metamorphic Relations." *FSE 2025*.
- [26] Liu, Y., et al. (2023). "G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment." *EMNLP 2023*.
- [27] Wang, Z., et al. (2025). "Sequential Preference Optimization: Aligning LLMs with Multiple Dimensions of Human Preferences." *AAAI 2025*.
- [28] Liu, W., et al. (2025). "SymMPO: Symmetric Multimodal Preference Optimization for Hallucination Mitigation." *arXiv preprint arXiv:2506.11712*.
- [29] Chen, X., et al. (2025). "Automated Data Cleaning Pipeline for Clinical Trial Data Using LLMs." *arXiv preprint arXiv:2508.05519*.
- [30] Li, Y., et al. (2023). "ChatDoctor: A Medical Chat Model Fine-Tuned on LLaMA Model using Medical Domain Knowledge." *arXiv preprint arXiv:2303.14070*.
- [31] Luo, L., et al. (2024). "Taiyi: A Bilingual Fine-Tuned Large Language Model for Diverse Biomedical Tasks." *JAMIA*.
- [32] Zhang, J., et al. (2024). "SafetyBench: A Comprehensive Benchmark for Safety of Large Language Models." *arXiv preprint*.

三、课题技术路线及研究方案

<p>1、主要研究内容（研究内容注意充实、细致和具体，不能泛泛介绍。注意思路清晰、流畅，突出重点，要明确指出重点研究内容。）</p> <p>本研究将围绕“数据治理（事前）—实时检测（事中）—模型对齐（事后）”的三层闭环架构，构建面向中文医疗问答大模型的系统性幻觉治理方案。主要研究内容包含以下三个方面：</p> <p>1.1 构建基于知识图谱的医疗指令数据清洗与增强流水线 针对开源医疗数据（如 MedDialog, Huatuo-26M）中存在的噪声、非结构化及逻辑谬误问题，设计一套自动化的数据清洗方案。</p> <ul style="list-style-type: none">• 医学实体抽取与链接（NER+EL）：利用通用大模型（如 Qwen-2.5-72B）或专门的医学 BERT 模型，从非结构化医患对话中提取疾病、药物、症状、检查项目等关键实体，并将其链接到标准医学术语集（如 ICD-10, SNOMED CT, MeSH）。• 基于 KG 的实体一致性校验算法：将提取的实体映射为(h, r, t)三元组，并与成熟医学知识图谱（CmeKG[14]或 UMLS[23]）进行比对。重点研发逻辑冲突检测算法，自动识别违反常识的组合，对于检测到逻辑冲突的样本，采用分级处理策略：高置信度冲突（如禁忌症）直接剔除，以保证安全性；低置信度或模糊冲突则利用 GPT-4o 基于 KG 三元组进行重写修正，以保留语料的多样性。例如，若数据中出现“孕妇服用利巴韦林”，通过检索 KG 发现（利巴韦林，禁忌人群，孕妇），则判定该数据存在逻辑冲突并予以清洗或修正。• 数据增强与重写：利用 GPT-4o 或 DeepSeek-V3 根据清洗后的高质量三元组生成多样化的指令微调（SFT）数据，确保数据的纯净度与多样性，解决长尾疾病数据稀缺问题。 <p>1.2 研发多维度的医疗幻觉检测机制</p> <p>针对医疗幻觉高隐蔽性（如剂量错误、伪造引用）的特点，构建“白盒+黑盒”融合的实时检测框架。</p> <ul style="list-style-type: none">• 白盒检测（内部状态）：研究基于不确定性估算（Uncertainty Estimation）的方法。监测生成 Token 的熵值（Entropy）、自一致性（Self-Consistency[6]）以及特征值（EigenScore[20]）。当模型在关键医学实体（如药品名、数值）上的生成熵值超过动态阈值时，标记为高风险幻觉候选。阈值将根据实体在医学语料中的词频（TF-IDF）或其知识图谱中的度（Degree）进行自适应调整，对于长尾罕见病实体，设置更为严格的低熵值阈值。• 黑盒检测（外部验证）：研究基于检索增强（RAG）的一致性验证方法。利用 FacTool [8]和 FactScore[7]的思想，将生成的长文本回答拆解为原子事实（Atomic Facts），调用搜索引擎或检索本地权威医疗指南库（PubMed, 默克诊疗手册），利用自然语言推理（NLI）模型判断生成内容与检索证据的“蕴含/冲突/中立”关系，实现对事实性错误的精准拦截。 <p>1.3 研究基于直接偏好优化（DPO/SimPO）的幻觉缓解技术</p> <p>探索无需显式奖励模型的对齐方法，直接利用偏好数据优化模型策略，使其内化医学事实约束。</p>

- **医疗偏好数据对(Medical Preference Pairs)构造:** 重点研究对抗性负样本(Hard Negatives)的生成策略。不同于传统的随机负样本, 本研究将通过实体替换(Entity Replacement)技术, 在正确回答的基础上, 利用知识图谱寻找相似但错误的实体(如将“头孢拉定”替换为“阿莫西林”用于青霉素过敏患者), 或颠倒因果关系, 构造出语义通顺但事实错误的负样本 y_{bad} , 形成 (x, y_{good}, y_{bad}) 数据对。为构建高质量的‘困难负样本’, 将引入语义相似度约束(Semantic Similarity Constraint), 确保替换后的实体在向量空间中与原实体相近(如均为‘头孢类抗生素’), 但在临床知识图谱中存在属性差异(如‘肾毒性’不同), 从而迫使模型学习细粒度的医学特征而非简单的语义模式。

- **模型对齐训练:** 对比分析 DPO、SimPO(Simple Preference Optimization)与 KTO(Kahneman-Tversky Optimization)在医疗场景下的表现。SimPO 算法通过引入长度归一化和目标边距(Target Margin), 理论上能更好避免模型生成冗长且空洞的“安全回答”(Safe but Useless), 本研究将重点验证其在医疗事实对齐中的有效性。

2、需要突破和解决的难题或关键问题

1. **高隐蔽性医疗幻觉的精准识别问题:** 医疗领域的幻觉往往表现为细微的数值错误(如“5mg”误写为“50mg”)或实体混淆(如适应症相似但禁忌不同的药物), 在语法上高度流畅, 传统基于语义相似度的评估指标(如 BLEU, ROUGE)无法有效识别。如何突破单一检测手段的局限, 建立对细粒度医学知识错误敏感的检测机制(例如基于 KG 的路径推理), 是本研究需解决的关键问题。

2. **高质量医疗对抗性负样本的自动化构建问题:** DPO/SimPO 算法的效果高度依赖于偏好数据的质量。简单的随机负样本(易区分样本)无法提供足够的梯度信息, 导致训练效率低下。如何设计算法自动生成语义通顺但包含特定医学逻辑错误(如因果倒置、禁忌症替换)的“困难负样本”, 以有效驱动模型学习医学事实边界, 是本研究的技术难点[21]。

3. **复杂推理下的事实一致性保持:** 在多轮对话和长文本生成(如病历分析)中, 模型容易出现前后矛盾(例如前文说“无过敏史”, 后文建议“避免过敏原”)。如何在长窗口(Long Context)下保持知识调用的稳定性, 避免“遗忘”或“篡改”前文信息, 是模型训练需要解决的问题。

3、特色与创新之处(在系统功能、核心流程、架构或其他方面与同类系统比较, 突出本文的亮点(这是论文优劣的主要指标)。)

1. **引入知识图谱(KG)的深层逻辑校验机制:** 不同于传统仅依靠关键词或规则的数据清洗方法, 本研究采用基于 CmeKG[14]的实体一致性校验算法。通过将非结构化文本映射为结构化三元组并进行矛盾检测, 能够识别出深层的医学逻辑冲突(如违反性别、年龄、病史的禁忌症), 显著提升数据治理的深度与准确性。

2. **构建“不确定性+RAG”的双重混合检测防线:** 本研究整合白盒检测(计算成本低、响应快, 基于 SelfCheckGPT/EigenScore)与黑盒 RAG 验证(准确性高、可解释性强, 基于 FacTool/FActScore)的优势, 采用级联式混合检测架构。这种多维度的检测机制既克服了单一模型“过度自信”的缺陷, 又兼顾了系统在实际落地中的时效性与安全性。

3. **提出基于“对抗性实体替换”的 SimPO/DPO 负样本构造策略:** 在模型对齐阶段,

本研究创新性地引入医学本体知识，设计了“对抗性实体替换”方法构造负样本。通过针对性地修改关键医学实体（如将“抗生素”替换为“抗病毒药”），使模型在 SimPO/DPO 训练中必须“学会”区分细微的医学事实差异，从而更本质地抑制事实性幻觉。这在医疗垂直领域的模型对齐研究中具有显著的创新性，区别于通用的 RLHF[15]方法。

4、拟采取的研究方法设计方法，技术路线

总体技术路线：

数据工程(清洗/增强)→SFT 模型训练→幻觉检测系统构建(不确定性+RAG) →DPO 偏好对齐优化→综合评估

具体方法：

1. 数据工程方法：

- **采集：**整合 MedQA[2], MedMCQA, PubMedQA, Huatuo-26M[14], CmtMedQA[16]等公开数据集。
- **Pipeline 设计：**
 - 1) **预处理：**去除 PII（个人隐私信息）、特殊字符过滤。
 - 2) **KG 校验：**利用医疗 NER 模型提取实体，映射至 CMeKG/UMLS 图谱，进行三元组矛盾检测（如：Drug A --treats--> Disease B? Check KG）。剔除逻辑错误样本。
 - 3) **构造偏好对：**对于一条指令 x ，保留原始高质量回答作为 y_{win} ；利用规则（实体替换、数值扰动）或弱模型生成 y_{lose} ，构建 (x, y_{win}, y_{lose}) 三元组用于 DPO/SimPO 训练。

2. 检测系统设计方法：

- **模块一（白盒）：**计算生成序列的 Token 级熵值 $H(t)$ 和困惑度（PPL）。对高熵值实体进行高亮标记，利用 INSIDE 方法计算特征值谱，评估内部状态的混乱度。
- **模块二（黑盒）：**提取回答中的原子事实（Atomic Facts），调用 Google Search API 或检索本地 PubMed/指南库，获取 Top-K 相关文档。利用微调后的 DeBERTa 或 Qwen-7B 作为 NLI 模型，判断“蕴含/冲突”。
- **决策逻辑：**综合熵值评分与 NLI 冲突率，输出最终的幻觉风险等级（Low/Medium/High）。

3. 模型训练与优化方法：

- **基座模型：**选择 Qwen-2.5-7B/14B-Instruct 和 Llama-3-8B-Instruct 作为主要实验对象，HuatuoGPT-II 和 BioMistral 作为对比基座。
- **SFT 阶段：**使用清洗后的数据进行全量微调或 LoRA 微调，让模型适应医学指令格式。
- **对齐阶段（Alignment）：**应用 SimPO (Simple Preference Optimization) 算法。

相比 DPO，SimPO 在损失函数中引入了长度归一化项 $\frac{\beta}{|y|} \log \pi(y|x)$ 和目标边距 γ ，公式如下：

$$\mathcal{L}_{SimPO} = -\log \sigma \left(\frac{\beta}{|y_w|} \log \pi_{\theta}(y_w|x) - \frac{\beta}{|y_l|} \log \pi_{\theta}(y_l|x) - \gamma \right)$$

这种方法能更有效地利用对抗性负样本，提升模型对事实错误的敏感度。

5、实验方案的可行性分析

- **数据资源：**

CMeKG（Chinese Medical Knowledge Graph）：由北京大学、鹏城实验室等构建的中文医学知识图谱，包含数百万节点，是本研究“数据治理”和“因果图构建”的基础设施。

Huatuo-26M / MedQA：公开的大规模中文医疗问答数据集，包含真实医患对话和考试题目，可用于 SFT 基座训练。

CMtMedQA：包含中文医疗问答的评测集，可作为测试基准。

- **算力支持：**

实验室/实习单位提供 A100/A800 GPU 集群，或平台租赁 GPU 集群，满足大模型微调与推理的算力需求。

- **技术基础：**

本人熟悉 Transformer 架构、PyTorch 框架、HuggingFace 生态，且在实习期间已有 RLHF 及知识图谱相关项目的实践经验。

四、工作进度安排

应包括文献调研，工程或系统设计，新设备、新产品的研制和调试，实验操作，实验数据的分析处理，撰写论文等。

2025.11 - 2025.12：阅读国内外关于医疗 LLM（Med-PaLM 2, HuatuoGPT）、幻觉检测（SelfCheckGPT, FacTool）、偏好对齐（DPO, SimPO）的最新文献（30+篇）；收集 MedQA, CMtMedQA 等数据集，搭建基于 CMeKG 的实体校验脚本，完成初步数据清洗。

2026.01 - 2026.02：幻觉检测算法研发（实现基于不确定性的白盒检测模块、搭建基于 RAG 的原子事实验证链路（检索+NLI））；构建包含正负样本的“医疗幻觉评测集”（Benchmark），涵盖药物、疾病、检查等多个维度。

2026.03：模型训练与 DPO 优化（完成基座模型的 SFT 训练、构造医疗偏好数据对（含对抗性负样本），进行 DPO 对齐实验）

2026.04：实验评估与系统调优（使用 FactScore、Win Rate（GPT-4 Judge）、Rouge-L 等多维度指标进行评估、对比消融实验（SFT vs DPO vs SimPO；有无 KG 清洗），分析不同策略效果）；与 HuatuoGPT-II, BioMistral 等 SOTA 模型进行对比

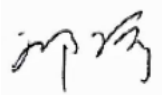


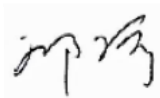

2026.05 - 2026.06：撰写论文与答辩准备整理实验数据，撰写硕士学位论文，准备答辩

演示。

五、预期成果

1. 工程系统：构建一套完整的医疗指令数据自动化清洗脚本库
2. 发布一个经过 SimPO/DPO 对齐、具备低幻觉率的中文医疗大模型 Demo，支持实时幻觉风险提示。
3. 评测基准：建立一套包含正负样本的医疗幻觉评测数据集。
4. 学术论文：完成高质量硕士学位论文一篇。

六、 审核意见

<p>导师意见</p> <p>审核通过</p> <div><div>导师签名:</div><div></div><div>2025 年12 月15 日</div></div>	
<p>开题报告专家委员会意见</p> <p>审核通过</p> <div><div>专家委员会主席签字:</div><div></div><div>专家委员会委员签字:</div><div></div><div>2025 年12 月15 日</div></div>	
<p>培养单位负责人意见</p> <div><div>培养单位负责人签名:</div><div>年 月 日</div></div>	