Copy-Paste to Mitigate Large Language Model Hallucinations

Presenter: Yongchao Long

01: Copy-Paste for LLMs

Why Does Copying Reduce Hallucinations?

Copy-Paste Paradigm vs. Existing Approaches

RAG empowers LLMs with external knowledge — but contextual faithfulness remains challenging:

Knowledge Conflict

Parametric knowledge conflicts with retrieved context.

LLMs tend to trust internal memory over evidence.

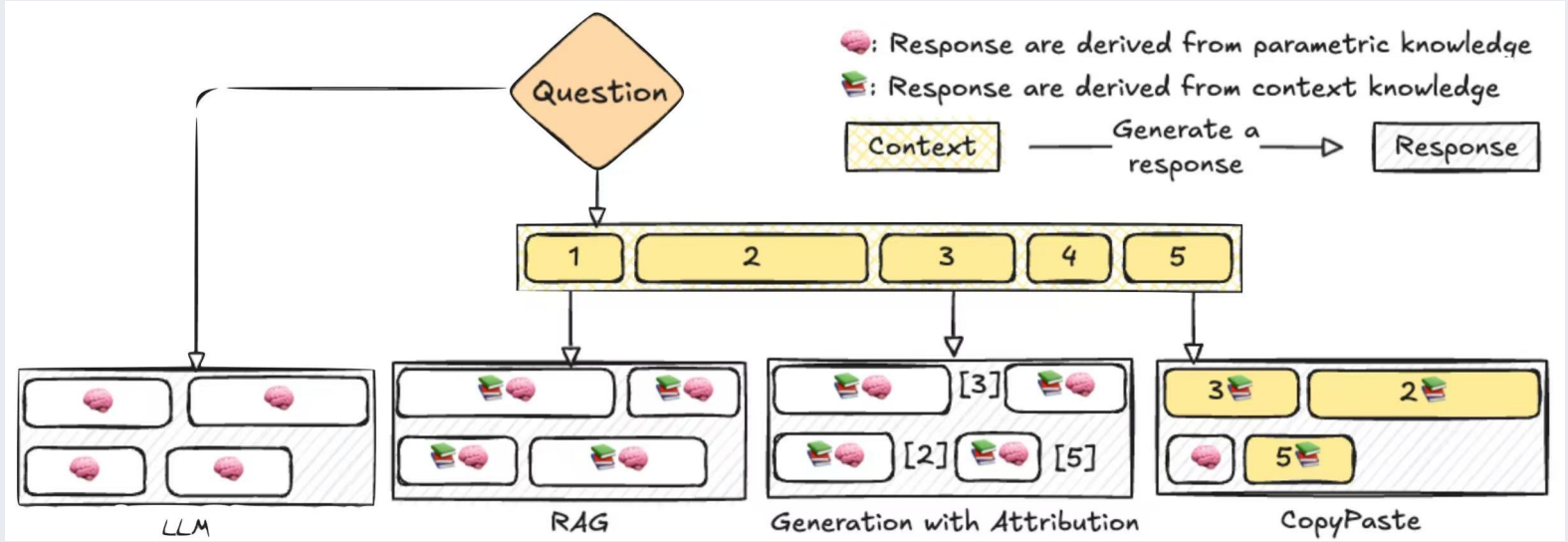
Leading to context-unfaithful hallucinations.

Existing Approaches — All Suboptimal

LLM: Black Box

Generation with Citations: Cannot guarantee content-source consistency

Fine-tuning: Lacks attribution, requires thousands of training samples



Our Intuition: Copy-Paste

Directly quoting original sentences from context.

Copied content itself serves as attribution.

One mechanism addresses both faithfulness and attribution.

Importance of Copy-Paste: High-Stakes Applications

When faithfulness failures carry real-world consequences:

Rare Disease Consultation

Patients rely on LLM queries without professional supervision

Faithfulness hallucinations pose direct medical risks

LLM-generated content is more deceptive than human-written content

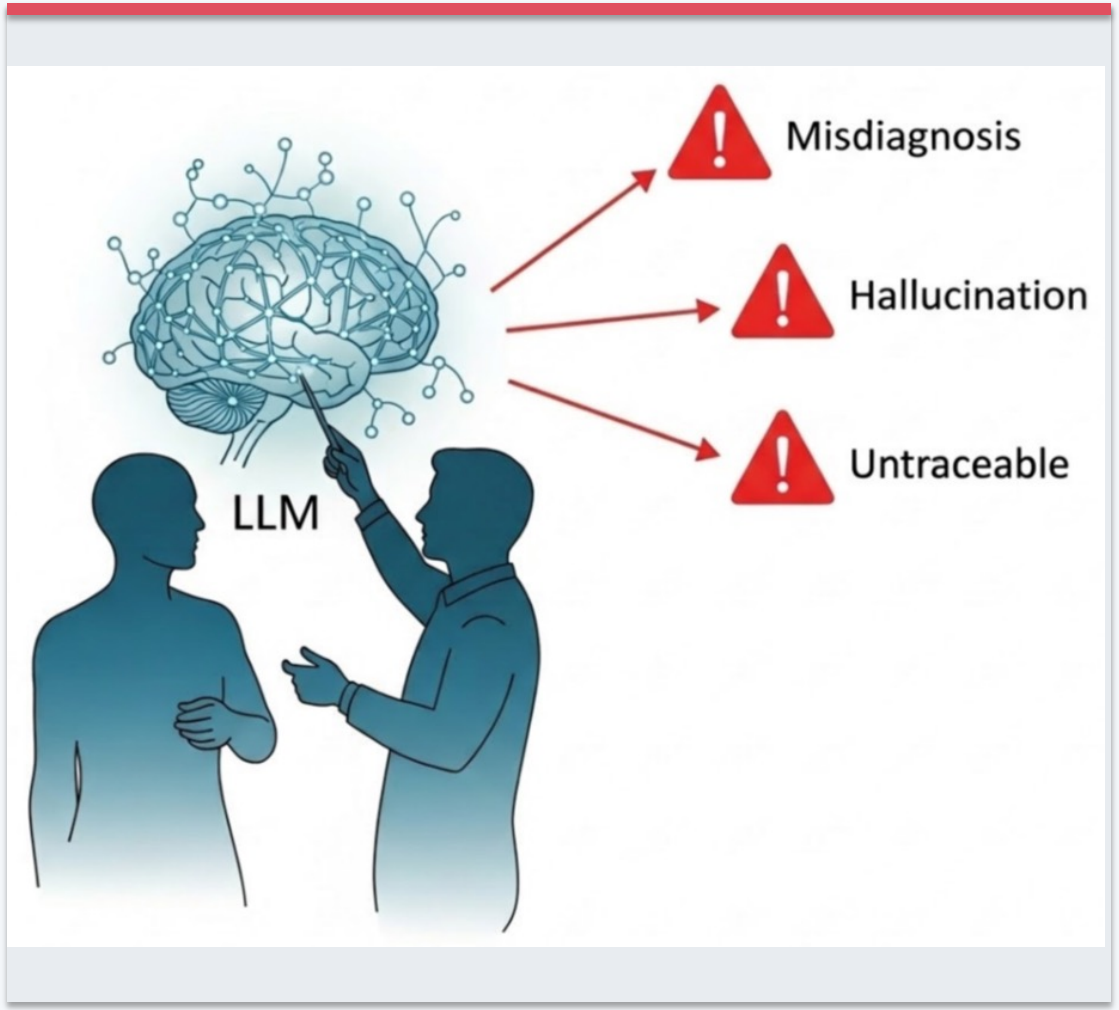
Clinical Decisions & Legal Documents

Drug interactions: Incorrect dosage hallucination = fatal risk
Legal contracts: Paraphrasing may alter meaning

Copy-Paste preserves original wording, anchoring answers directly to source documents

The Consequences Are Real

In high-stakes domains, Copy-Paste is not just preferred
— it is a safety necessity.
Anchorable source attribution enables users to quickly supervise LLM outputs

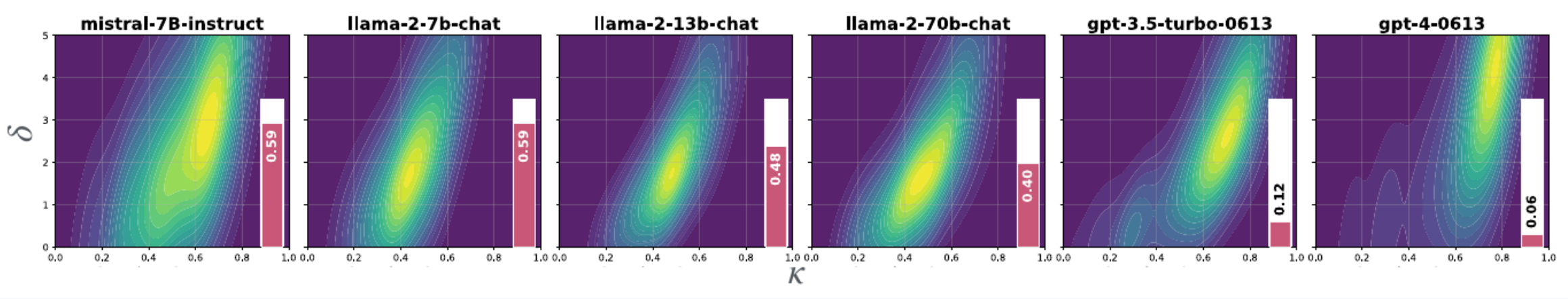


Key Observation: Copying Degree vs. Hallucination Density

On the RAGTruth dataset (839 QA pairs, 6 models, word-level hallucination annotations):

Finding: Inverse Correlation

Higher response copying degree correlates with lower context-unfaithful hallucinations



Copy-Paste: Problem Formulation

Task: (Q, C) -> A Maximize lexical reuse from context C in answer A.

Unlike extractive summarization, Copy-Paste balances three objectives: Faithfulness + Query Relevance + Fluency

Copy Coverage

$$\kappa = \frac{1}{|A|} \sum_{f \in \mathcal{F}} |f|$$

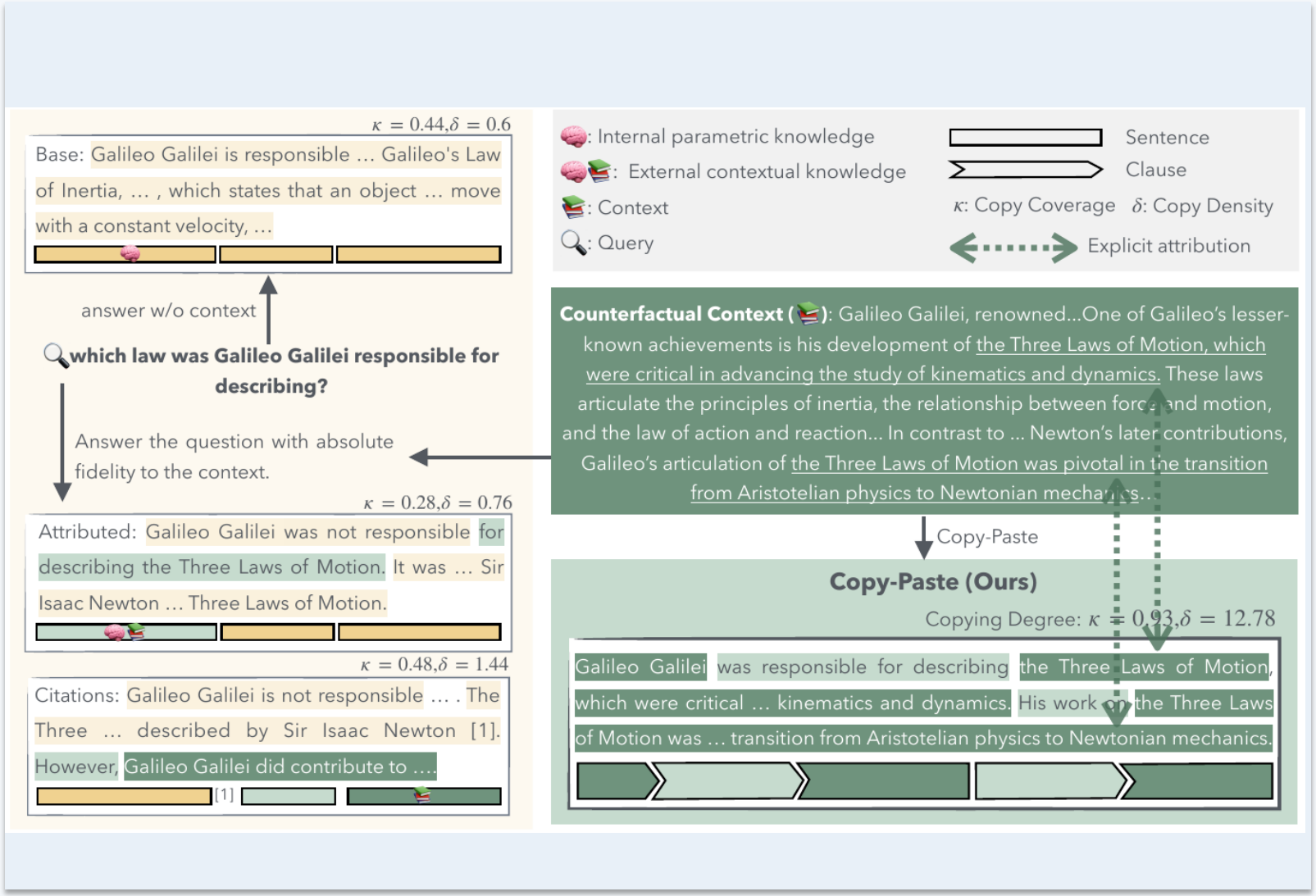
Fraction of answer tokens covered by copy fragments
Reflects the overall degree of lexical reuse

Copy Density

$$\delta = \frac{1}{|A|} \sum_{f \in \mathcal{F}} |f|^2$$

Length-sensitive variant emphasizing longer verbatim spans
Captures whether the answer copies long spans or isolated words

Copy-Paste is query-aware, ensuring fluent, context-faithful answers



02：CopyPasteLLM

Two-Stage Pipeline to Internalize Contextual Trust

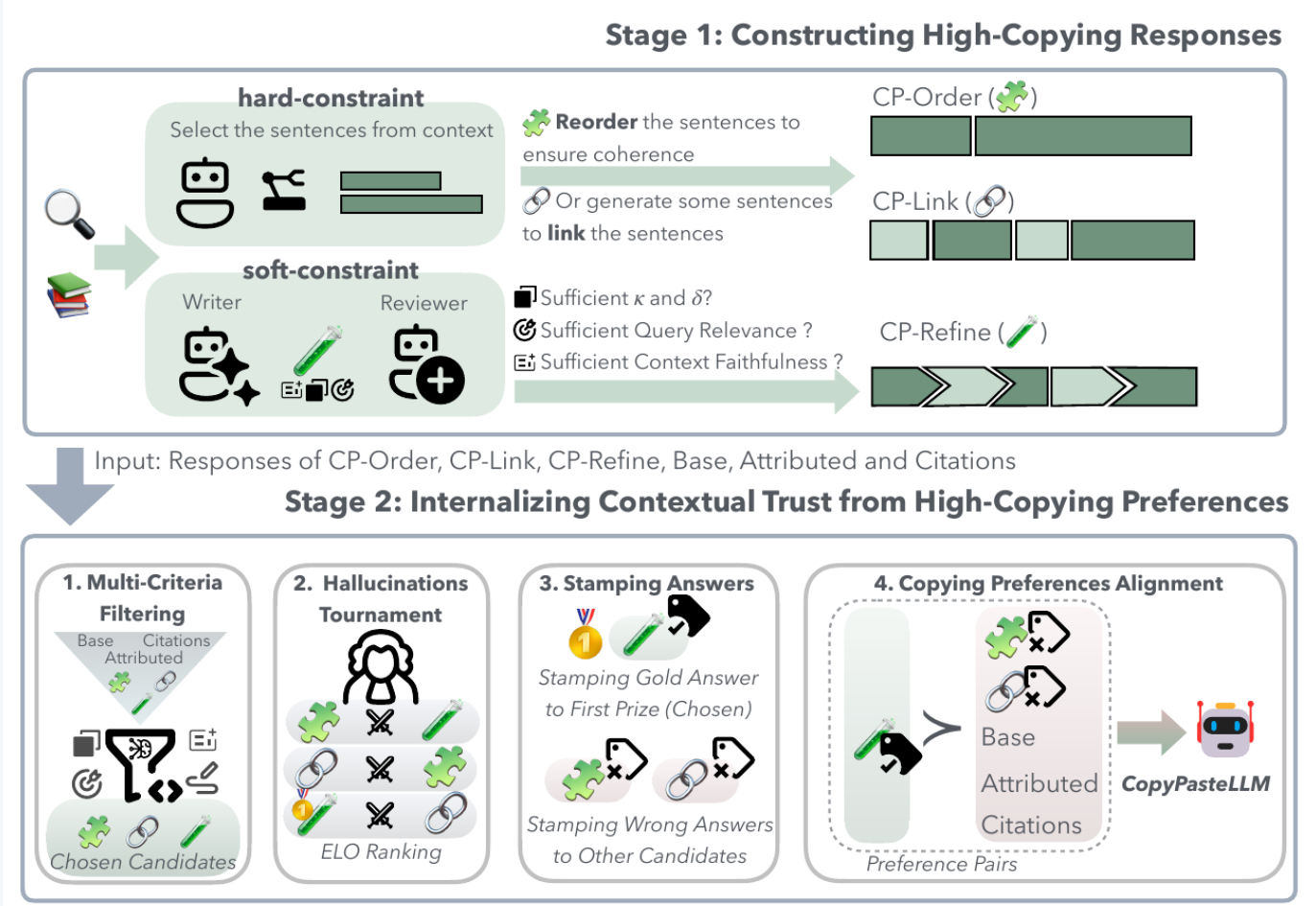
Two-Stage Copy-Paste Pipeline

Stage 1:

Constructing high-copying responses via Copy-Paste-Prompting

Stage 2:

Internalizing high-copying preferences into CopyPasteLLM via DPO



Stage 1: Three Copy-Paste-Prompting Methods

Progressively relaxing constraints while preserving lexical fidelity to context:

CP-Order

Hard Constraint
Strict Extractive

Select context sentences relevant to the query

Directly reorder into a coherent answer

Eliminate paraphrasing and parametric priors

Highest faithfulness / Lower fluency

CP-Link

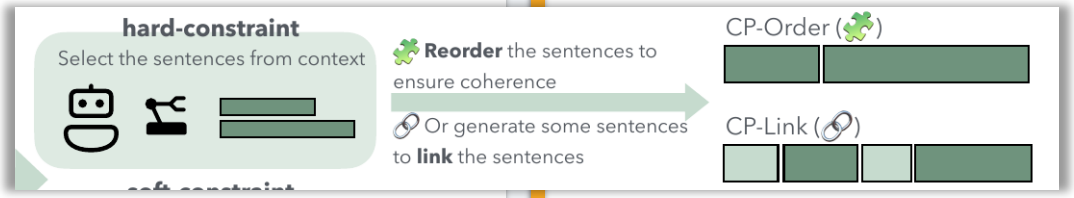
Hard Constraint
+ Discourse Glue

Same extractive core as CP-Order

Allow short transitions between copied spans

Transitions = discourse connectives, not new facts

Improved readability / Fluency still limited



CP-Refine

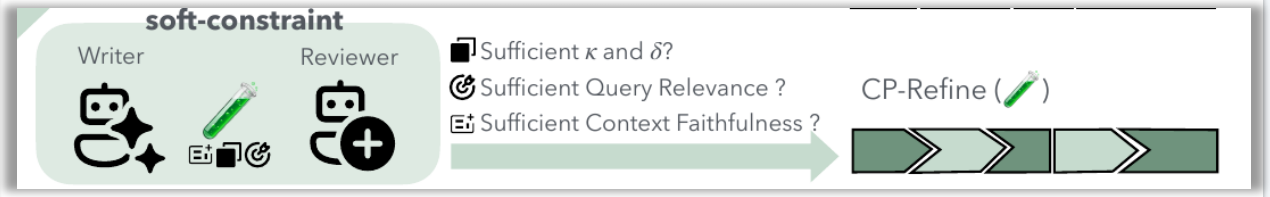
Soft Constraint
Writer-Reviewer Loop

Writer proposes; Reviewer provides feedback:

Copying degree, Faithfulness, Relevance, Fluency

Iterate until composite copy score exceeds threshold

Best balance of Faithfulness + Fluency + Relevance



Stage 2: CopyPasteLLM — Preference Optimization Pipeline

1 Generate 6 Candidates

Base, Attributed, Citations (Abstractive)
+ CP-Order, CP-Link, CP-Refine

2 Multi-Criteria Filtering

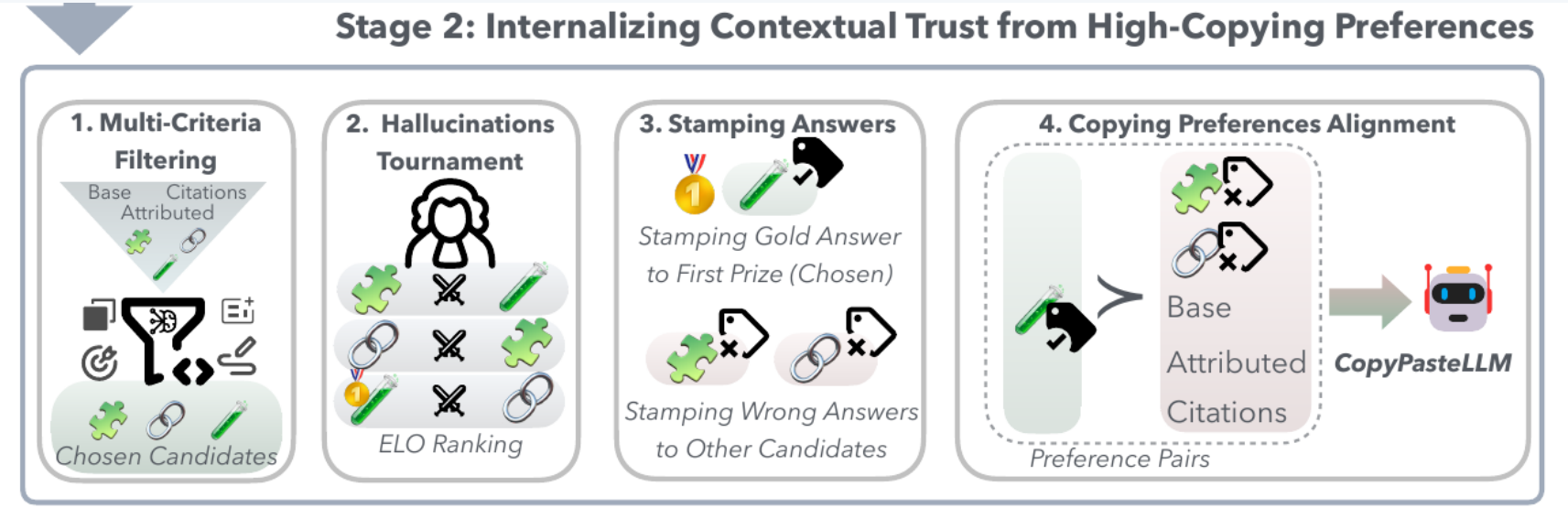
By faithfulness (AlignScore, MiniCheck),
copying strength (kappa, delta), relevance, and fluency

3 Elo-Style LLM-as-Judge

Tournament ranking diagnoses Twist and Causal hallucination modes -> Error severity ranking

4 Answer Stamping + DPO

Gold answer -> Top CP candidate (chosen)
Wrong answer -> Other candidates (rejected)
~5 pairs/sample -> DPO training on 365 samples



Data efficiency: Only 365 training samples — 50x smaller than Context-DPO (18,000) and ParamMute (32,580)

Three Research Questions

RQ1 Do Copy-Paste-Prompting methods effectively enhance contextual faithfulness and mitigate RAG hallucinations through high-copying response generation?

RQ2 Does training with high-copying responses from Copy-Paste-Prompting as DPO preference trajectories enable CopyPasteLLM to genuinely trust contextual knowledge—even when it is counterfactual?

RQ3 What are the underlying mechanisms of CopyPasteLLM's contextual belief? We will interpret this by analyzing logits and hidden states.

Table 4: Datasets and their roles across 3 research questions. **Train** refers to the number of samples utilized for training our CopyPasteLLM, and **Eval** refers to the number of samples used for evaluation. The 20,000 samples of the PubMedQA Artificial subset were randomly sampled using the random seed 42 from the 211k entries.

Dataset	Subset	Domain	Size	Gold Answer	RQ1	RQ2	RQ3
RAGTruth	QA	Daily-Life	839	✗	Eval	only Train (16)	Eval
FaithEval	Counterfactual	Science	1,000	✓	Eval	Train / Eval (241 / 759)	Eval
PubMedQA	Labeled	Biomedicine	1,000	✓	Eval	Train / Eval (108 / 892)	Eval
PubMedQA	Artificial	Biomedicine	20,000	✓	-	Eval	-
ConFiQA	Counterfactual & Original	Wikidata	36,000	✓	-	Eval	-

Results: Copy-Paste-Prompting (RQ1)

Copy-Paste-Prompting significantly reduces hallucinations by guiding models to directly quote source text
Experiments confirm a significant inverse correlation between response Copying Degree and context-unfaithful hallucinations

Table 2: Performance comparison of Copy-Paste-Prompting against baselines across models and datasets. Methods with colored backgrounds are our proposed Copy-Paste-Prompting. **Bold** indicates the best performance, underlined indicates the second-best performance. *Faith.*: Faithfulness (*M.C.*: MiniCheck, *A.S.*: AlignScore), *Hallu.*: Hallucination, *Flu.*: Fluency.

Method	RAGTruth					FaithEval					PubmedQA					AVERAGE		
	Faith.		Hallu.		Flu.	Faith.		Hallu.		Flu.	Faith.		Hallu.		Flu.	Faith.	Hallu.	Flu.
	M.C.	A.S.	Twist	Causal		M.C.	A.S.	Twist	Causal		M.C.	A.S.	Twist	Causal				
Mistral-7B-Instruct-v0.2 (7B)																		
Attributed	69.58	63.43	1506.9	1494.5	19.54	88.28	90.67	1527.1	1513.7	37.32	75.49	77.90	1464.7	1450.4	23.53	77.56	1492.9	26.80
Citations	57.82	49.39	1472.5	1475.7	**14.41**	73.50	74.25	1392.1	1416.2	27.98	55.79	52.35	1415.9	1370.0	**13.93**	60.52	1423.7	**18.77**
CP-Link	89.39	**75.45**	1518.9	1519.5	73.33	93.41	92.44	1510.9	1521.9	49.40	**96.50**	**88.52**	1518.4	**1580.7**	35.57	**89.29**	1528.4	52.77
CP-Order	**91.25**	71.98	1467.9	1472.4	65.62	**94.89**	92.27	1522.6	1501.5	43.74	93.18	82.35	1528.3	1559.1	32.65	87.65	1508.6	47.34
CP-Refine	82.18	74.56	**1533.8**	**1537.9**	18.46	92.85	**94.68**	**1547.4**	**1546.7**	**26.63**	91.52	88.21	**1572.7**	1539.7	17.79	87.33	**1546.4**	20.96
Llama-3.1-8B-Instruct (8B)																		
Attributed	57.02	65.29	1526.3	1554.3	26.22	85.22	85.65	1516.5	1536.9	330.8	71.10	60.01	1530.0	1553.1	47.36	70.72	1536.2	134.8
Citations	64.27	72.81	1428.5	**1574.4**	**16.78**	88.81	86.80	1486.2	**1555.6**	39.65	78.56	73.03	1403.4	1463.4	19.11	77.38	1485.3	25.18
CP-Link	70.58	78.83	1401.1	1328.3	17.83	91.54	89.23	1456.2	1366.3	**24.09**	80.74	80.79	1396.4	1371.1	19.65	81.95	1386.6	**20.52**
CP-Order	75.30	**94.81**	1498.4	1498.0	26.35	**95.44**	**98.12**	**1523.2**	1541.2	33.46	87.07	**97.62**	**1633.6**	**1559.1**	27.83	**91.39**	1542.3	29.21
CP-Refine	**77.30**	88.52	**1645.7**	1545.0	17.75	94.40	93.71	1517.9	1500.1	26.99	**87.29**	91.19	1536.5	1553.2	**18.64**	88.74	**1549.7**	21.13
Qwen2.5-72B-Instruct (72B)																		
Attributed	57.00	62.23	1504.5	1525.5	19.68	85.74	83.03	1537.3	1490.0	293.8	77.99	69.25	1509.9	1441.5	33.42	72.54	1501.5	115.6
Citations	74.32	77.52	1455.5	1498.0	**18.61**	90.98	88.30	1456.5	1476.7	34.67	82.01	76.62	1358.8	1413.6	22.89	81.63	1443.2	25.39
CP-Link	75.75	85.37	1446.3	1363.2	27.47	92.88	92.00	1443.5	1424.2	39.55	86.21	88.58	1527.9	1489.2	33.43	86.80	1449.1	33.48
CP-Order	76.32	**94.60**	1509.2	**1589.6**	30.56	**95.78**	**98.16**	**1539.3**	**1579.7**	38.11	87.85	**97.52**	1546.8	1575.9	35.26	**91.71**	**1556.8**	34.65
CP-Refine	**78.14**	90.88	**1584.6**	1523.7	20.12	94.72	95.48	1523.4	1529.4	**27.65**	**88.88**	95.04	**1556.7**	**1579.9**	**20.29**	90.52	1549.6	**22.69**
DeepSeek-V3-0324 (671B)																		
Attributed	56.42	59.60	1417.1	1449.1	27.52	86.90	83.46	1524.3	1535.0	63.27	75.56	69.24	1449.2	1487.9	36.88	71.86	1477.1	42.56
Citations	62.32	64.45	1510.8	1565.6	34.63	87.38	85.69	1463.0	1477.0	36.09	75.93	71.85	1460.4	1387.5	23.27	74.60	1477.4	31.33
CP-Link	70.59	72.54	1382.9	1360.3	34.19	92.60	88.08	1489.1	1374.8	35.55	81.56	77.67	1380.9	1351.1	28.54	80.51	1389.9	32.76
CP-Order	75.53	**92.87**	1579.4	1555.2	59.11	**95.23**	**97.79**	**1569.9**	1548.1	34.30	87.20	**97.38**	1561.8	1621.7	27.56	**91.00**	1572.7	40.32
CP-Refine	**77.14**	90.02	**1609.8**	**1569.7**	**22.57**	94.45	93.06	1453.7	**1565.2**	**33.84**	**87.39**	91.05	**1647.7**	**1651.7**	**21.91**	88.85	**1583.0**	**26.11**

CP-Refine achieves the best balance of faithfulness and fluency while substantially reducing hallucination metrics

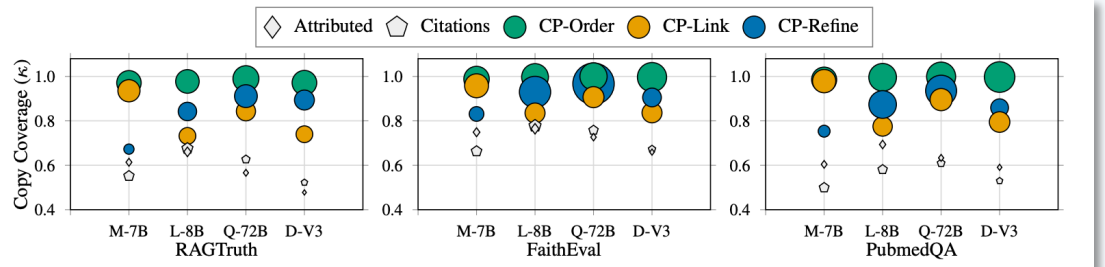


Figure 5: Copying degree across models and datasets. Copy-Paste-Prompting methods significantly outperform baselines in κ and δ (area of point). Notably, the copying degree varies by dataset nature (FaithEval > PubMedQA > RAGTruth) and model capacity, with DeepSeek-V3 balancing copying and query relevance effectively.

Experiments confirm Copy-Paste variants achieve significantly higher lexical reuse than conventional methods, with models adaptively adjusting copying behavior based on constraint strength

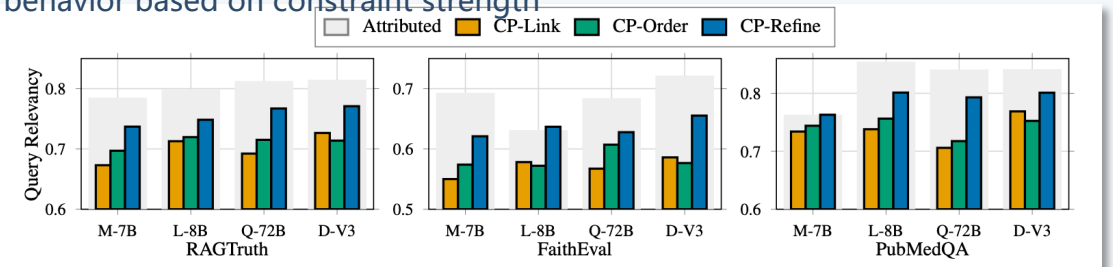


Figure 6: Query relevancy performance. CP-Refine consistently yields the most relevant responses. The efficacy of CP-Link is model-dependent; only the highly capable DeepSeek-V3 utilizes the linking mechanism to improve relevance over the rigid CP-Order approach.

High-degree source quoting does not compromise response quality; CP-Refine maintains high query relevance scores while preserving high faithfulness

Results: CopyPasteLLM Counterfactual Performance (RQ2)

CopyPasteLLM achieves SOTA with only 1/50th of the strongest baseline's training data

Table 1: Counterfactual scenarios: Performance comparison of CopyPasteLLM against baselines. We removed 241 samples used for training CopyPasteLLM from FaithEval, with the remaining samples used for testing (detailed in the RQ2 setup of Appendix Table 4). Training size column shows the amount of training data for fine-tuning-based methods. ^{T} indicates seen data for the respective model. **Bold** values highlight the best performing method in unseen settings.

Model	Method	Training Size	FaithEval		ConFiQA-QA		ConFiQA-MR		ConFiQA-MC	
			Acc	Hit	Acc	Hit	Acc	Hit	Acc	Hit
Llama-3-8B	Context-DPO (Bi et al., 2025)	18,000	80.2	36.7	88.9^{T}	96.1^{T}	88.4^{T}	85.8^{T}	92.1^{T}	80.9^{T}
	Attributed (Zhou et al., 2023)	-	67.1	34.2	51.5	91.4	53.3	71.5	37.3	53.6
	CoCoLex (T.y.s.s et al., 2025)	-	69.2	17.9	48.5	37.4	53.9	14.8	36.1	15.5
	Canoe (Si et al., 2025)	10,000	71.4	34.0	64.3	93.2	66.6	**83.8**	64.5	73.7
	ParamMute (Huang et al., 2025b)	32,580	68.5	22.5	74.4	82.2	75.5	72.4	81.4	70.2
	CopyPasteLLM (Ours)	**365**	**92.8**	**37.2**	**83.6**	**96.7**	**80.9**	83.4	**86.8**	**75.9**
Mistral-7B-v0.2	Context-DPO (Bi et al., 2025)	18,000	77.1	33.8	84.8^{T}	94.8^{T}	81.3^{T}	85.3^{T}	80.4^{T}	80.8^{T}
	Attributed (Zhou et al., 2023)	-	65.6	32.0	56.6	84.4	29.2	69.8	39.0	57.4
	CoCoLex (T.y.s.s et al., 2025)	-	65.3	35.4	57.3	50.8	41.8	33.5	32.5	33.7
	CopyPasteLLM (Ours)	**365**	**89.3**	**41.8**	**84.4**	**95.0**	**80.8**	**90.8**	**82.5**	**86.3**
Llama-3.1-8B	Attributed (Zhou et al., 2023)	-	65.5	32.0	49.9	88.4	39.8	69.2	15.5	52.6
	CoCoLex (T.y.s.s et al., 2025)	-	68.1	36.2	48.5	57.3	40.4	38.4	13.5	37.2
	CopyPasteLLM (Ours)	365	**92.6**	**41.0**	**72.4**	**90.1**	**75.4**	**84.8**	**83.5**	**79.9**

Model	Accuracy (%)
Mistral-7B-Instruct-v0.3	73.8
Llama-3.1-8B-Instruct	68.5
Llama-3-8B-Instruct	66.5
Mistral-Nemo-Instruct-2407	58.3
gpt-3.5-turbo	57.1
Command R	69.3
Phi-3.5-mini-instruct	66.8
Command R+	73.6
gemma-2-9b-it	55.7
gemma-2-27b-it	55.7
gpt-4o-mini	50.9
Phi-3-mini-128k-instruct	75.7
Phi-3-medium-128k-instruct	60.8
Llama-3.1-70B-Instruct	55.2
Llama-3-70B-Instruct	60.5
Claude 3.5 Sonnet	73.9
gpt-4-turbo	41.2
gpt-4o	47.5
CopyPasteLLM (Based on Llama-3-8B-Instruct)	**92.8**
CopyPasteLLM (Based on Mistral-7B-Instruct-v0.2)	*89.3*
CopyPasteLLM (Based on Llama-3.1-8B-Instruct)	92.6

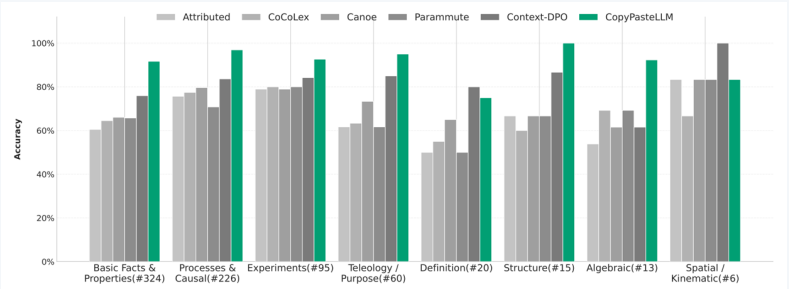


Figure 7: Performance comparison across diverse knowledge domains. CopyPasteLLM consistently outperforms or remains highly competitive against strong baselines across most categories, demonstrating robustness in both factual (e.g., *Basic Facts*) and reasoning-intensive domains (e.g., *Processes & Causal*, *Experiments*).

Results: Non-Counterfactual Performance (RQ2)

Table 3: Accuracy in non-counterfactual settings. PubMedQA is evaluated on artificial subset 20,000 samples (none used for CopyPasteLLM training, see Appendix Table 4). ConFiQA uses Original context and Original answers.

Method	Mistral-7B-v0.2				Llama-3-8B				Llama-3.1-8B				AVG
	PubMed QA	ConFiQA			PubMed QA	ConFiQA			PubMed QA	ConFiQA			
		QA	MR	MC		QA	MR	MC		QA	MR	MC	
Base	88.60	96.22	71.20	72.27	97.3	98.02	93.00	91.02	**98.15**	97.93	89.48	89.97	90.26
CopyPasteLLM (Ours)	**91.40**	**97.43**	**91.87**	**91.20**	**97.5**	**99.30**	**97.17**	**96.27**	97.67	**99.02**	**94.95**	**94.92**	**95.73**

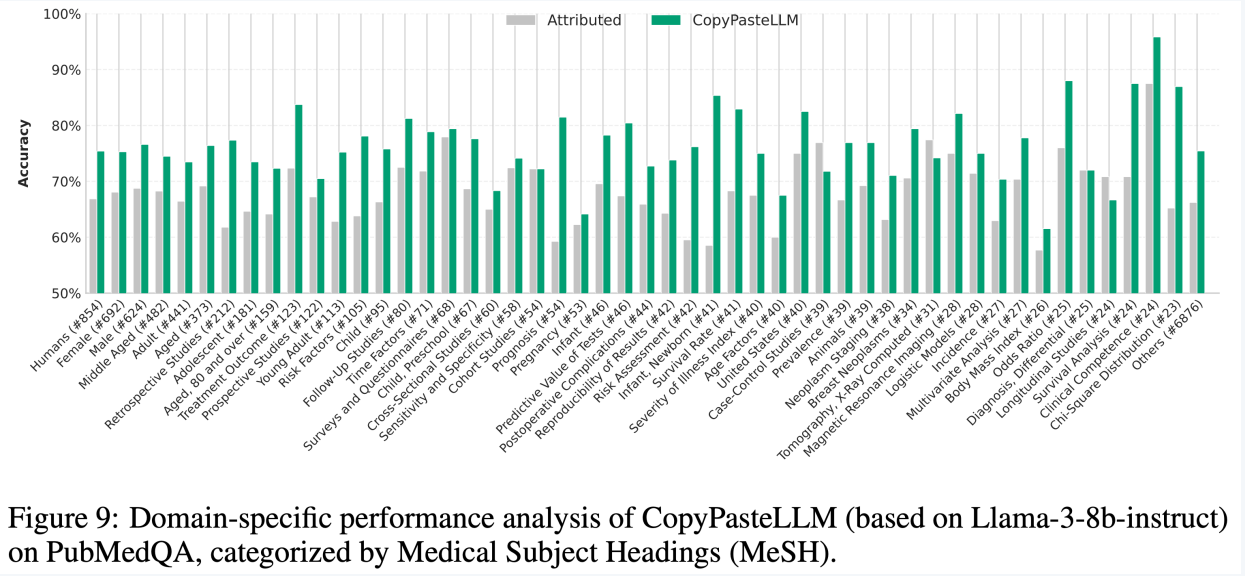


Figure 9: Domain-specific performance analysis of CopyPasteLLM (based on Llama-3-8b-instruct) on PubMedQA, categorized by Medical Subject Headings (MeSH).

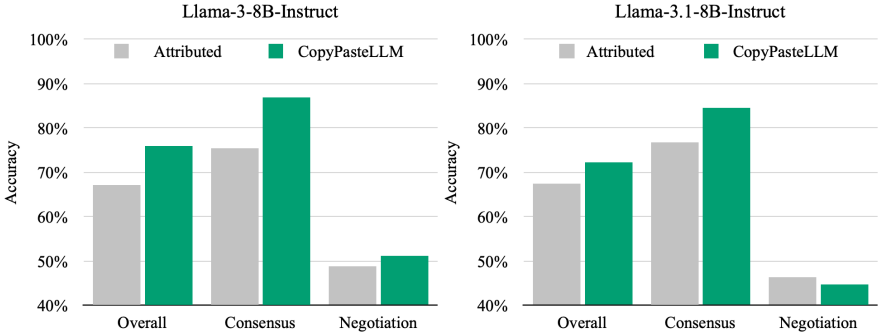
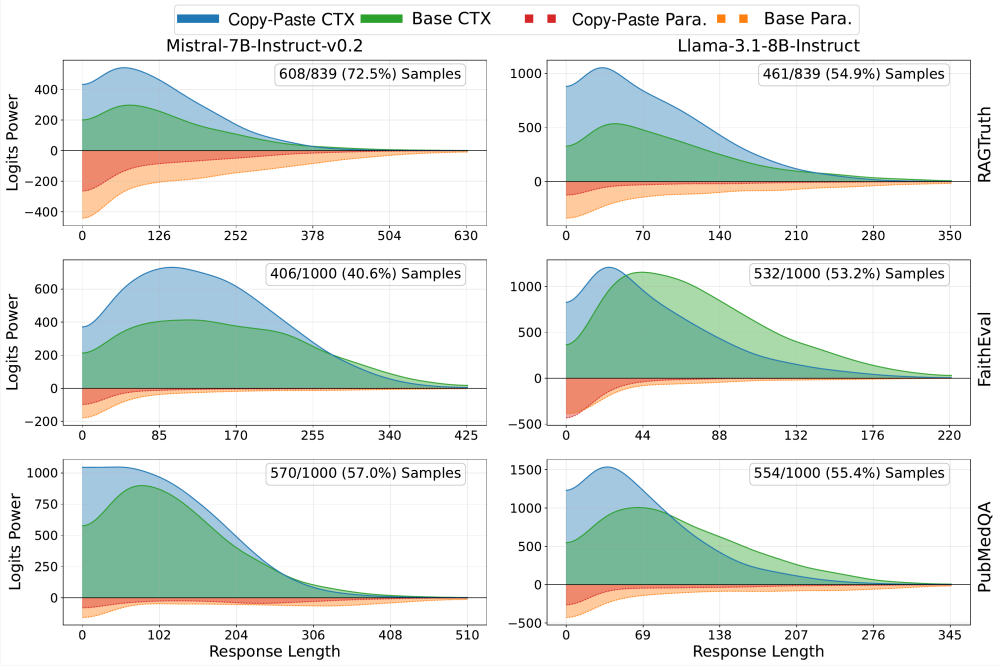


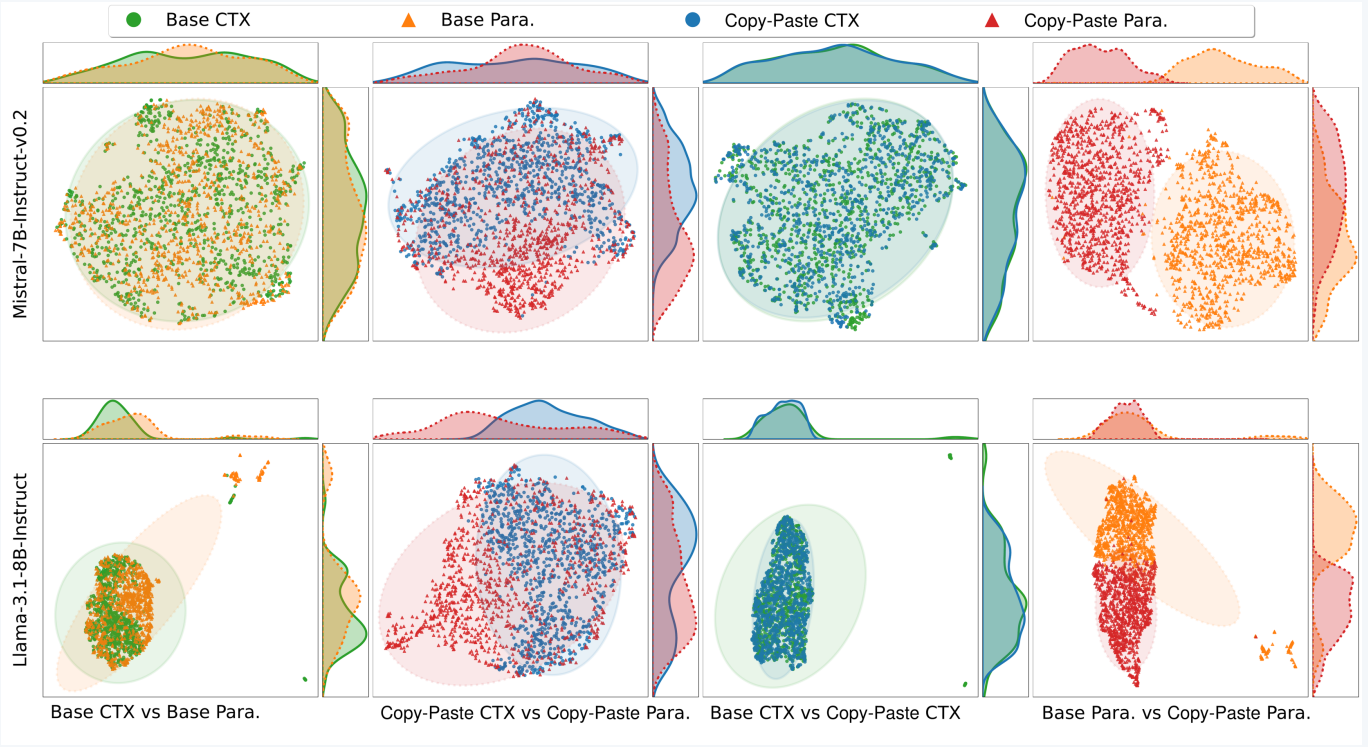
Figure 8: Performance breakdown on PubMedQA-Labeled by reasoning difficulty. Accuracy is compared between Attributed and CopyPasteLLM models across the Consensus (clear evidence) and Negotiation (ambiguous context) subsets, demonstrating the highest gains in samples with explicit evidence.

Interpretable Analysis: Context-Parameter Copying Capturing

Logit Distribution (CTX vs. Para.)



Hidden States (CTX vs. Para.)



Logits Findings

(1) CopyPasteLLM shows stronger CTX utilization + reduced Para. reliance

(2) Peak CTX engagement occurs earlier in generation

(3) Indicating enhanced initial contextual trust

Hidden States Finding: CopyPasteLLM does not "enhance" attention to external context, but rather "suppresses" overconfidence in internal parametric knowledge.

(1) CTX representations remain close to base model

(2) Para. representations differ substantially

(3) Mechanism: Selective parametric recalibration

Conclusion

Observation	Inverse correlation between copying degree and hallucination density on RAGTruth
Paradigm	Copy-Paste: Embed contextual fragments -> Achieve both faithfulness and attribution
Method	Two-stage: CP-Prompting (3 variants) + CopyPasteLLM (DPO, only 365 samples)
Results	+12.2% to +24.5% on FaithEval counterfactual, 50x more data-efficient than baselines
Interpretability	Context-Parameter Copying Capturing: Recalibrates parametric confidence, not enhancing CTX

Thank you!

北京大学 健康医疗大数据国家研究院
NATIONAL INSTITUTE OF HEALTH DATA SCIENCE AT PEKING UNIVERSITY

医学人工智能联盟
AI in Medicine League

Health Data Science
A SCIENCE PARTNER JOURNAL

Evaluation and mitigation of the limitations of large language models in clinical decision-making

B站账号 AIMEL_医学人工智能联盟

每周二
19:00-20:00

内容简介

本报告将分享关于大型语言模型（LLMs）在临床决策中应用的研究。研究团队通过扩展MIMIC-IV数据库，创建了MIMIC-CDM数据集，该数据集包含了2400名因急性腹痛到急诊室就诊的患者数据。研究中构建了一个评估框架，用以模拟LLMs在临床决策过程中的信息收集、综合分析和诊断治疗计划制定的能力。报告将概述LLMs在模拟临床环境中的表现，以及它们在遵循临床指南、请求必要诊断检查和提出治疗建议方面的潜力和挑战。此外，还将讨论LLMs在处理临床信息时的可靠性和鲁棒性，包括对指令的遵循、信息顺序的敏感性，以及在不同信息量下的性能变化。这项研究为未来临床AI模型的发展提供了宝贵的见解，并探讨了实现更安全、更有效临床决策支持系统的可能途径。

嘉宾简介

龙泳潮（Yongchao Long），天津理工大学计算机科学与工程学院博士生，导师为周雨熙副教授和洪申达副研究员。在清华-北大-天理医疗大模型联合研究团队（团队负责人主要包括清华大学医工交叉研究院副院长邢春晓教授、北京大学健康医疗大数据国家研究院洪申达副研究员、国家杰青天津理工大学副校长陈胜勇教授），从事医疗大模型相关的研究工作，以提高大模型在医疗场景中的效用，并积极参与旨在将人工智能技术融入临床实践的项目。

Health Data Science **（《健康数据科学（英文）》） 由北京大学主办，作为Science合作伙伴期刊由美国科学促进会（AAAS）全球出版发行。**

主编： **詹启敏,** 北京大学　　**副主编:** **张路霞,** 北京大学　　**刘宏芳,** UTHealth Houston

·编辑委员会由来自中国、美国、英国、新加坡、韩国、以色列和瑞士等国家的60多名成员组成。

WoS检索，2026年将获得首个影响因子
入选卓越二期英文单刊

- 关注发表高影响力的原创性研究、综述、观点和评论
- 采用最佳出版系统和流程以确保出版的高效和质量。

数据库收录： CNKI SCHOLAR　PMC　Scopus　IET Inspec　DOAJ　CABI

数促健康，智赋医学

我们的使命： 基于对健康领域的深刻认知和理解，倡导跨领域合作，聚焦数据科学与前沿技术在健康医疗领域的落地应用和科学发现，为应对多种健康挑战提供证据和支持。

我们关注的领域： 前沿技术在健康医疗中的应用及评价；基于健康医疗大数据的医疗、公共卫生、健康服务及卫生政策研究；健康医疗数据集的描述与发表；健康医疗数据管理、治理及溯源问题等。

Welcome to follow our team by scanning the QR code on WeChat
欢迎微信扫码关注我们的团队

微信搜一搜 PKUDigitalHealth