NEWSBREAK

RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, Cheng Niu

ACL 2024

Bangkok, Thailand

RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models

Cheng Niu[1], **Yuanhao Wu**[1], **Juno Zhu**[1], **Siliang Xu**[1], **Kashun Shum**[1], **Randy Zhong**[1], **Juntong Song**[1], and **Tong Zhang**[2]

[1]NewsBreak
[2]University of Illinois Urbana-Champaign
cheng.niu@newsbreak.com

Figure 1: Data gathering pipeline. Taking a data-to-text writing task as an example, our data gathering pipeline includes 2 steps: 1) response generation. We generated responses with multiple LLMs and natural prompts. 2) human annotation. Human labeler annotated hallucinated spans in LLM responses.

ACL 2024
Bangkok, Thailand

RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models

Cheng Niu[1], **Yuanhao Wu**[1], **Juno Zhu**[1], **Siliang Xu**[1], **Kashun Shum**[1], **Randy Zhong**[1], **Juntong Song**[1], and **Tong Zhang**[2]

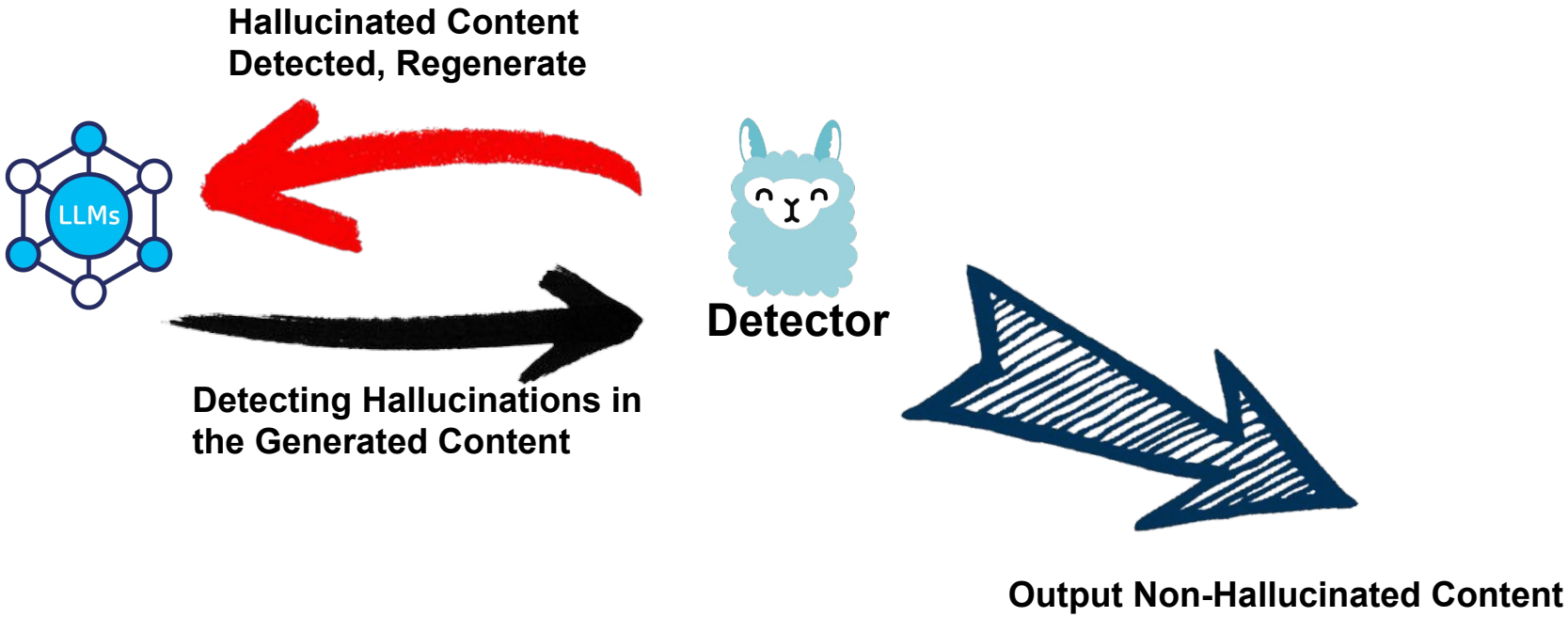
[1]NewsBreak
[2]University of Illinois Urbana-Champaign
cheng.niu@newsbreak.com

Methods	Question Answering			Data-to-text Writing			Summarization			Overall		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
$\text{Prompt}_{\text{gpt-3.5-turbo}}$	18.8	84.4	30.8	65.1	95.5	77.4	23.4	89.2	37.1	37.1	92.3	52.9
$\text{Prompt}_{\text{gpt-4-turbo}}$	33.2	90.6	45.6	64.3	**100.0**	78.3	31.5	97.6	47.6	46.9	97.9	63.4
$\text{SelfCheckGPT}_{\text{gpt-3.5-turbo}}$	35.0	58.0	43.7	68.2	82.8	74.8	31.1	56.5	40.1	49.7	71.9	58.8
$\text{LMvLM}_{\text{gpt-4-turbo}}$	18.7	76.9	30.1	68.0	76.7	72.1	23.3	81.9	36.2	36.2	77.8	49.4
Finetuned Llama-2-13B	**61.6**	76.3	**68.2**	**85.4**	91.0	**88.1**	**64.0**	54.9	**59.1**	**76.9**	80.7	**78.7**

Table 5: The response-level hallucination detection performance for each baseline method across different tasks and different models.

Methods	Question Answering			Data-to-text Writing			Summarization			Overall		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
$\text{Prompt Baseline}_{\text{gpt-3.5-turbo}}$	7.9	25.1	12.1	8.7	45.1	14.6	6.1	33.7	10.3	7.8	35.3	12.8
$\text{Prompt Baseline}_{\text{gpt-4-turbo}}$	23.7	52.0	32.6	17.9	**66.4**	28.2	14.7	**65.4**	24.1	18.4	**60.9**	28.3
Finetuned Llama-2-13B	**55.8**	**60.8**	**58.2**	**56.5**	50.7	**53.5**	**52.4**	30.8	**38.8**	**55.6**	50.2	**52.7**

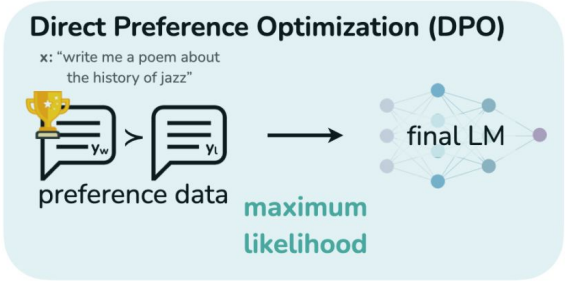
Table 6: The span-level detection performance for each baseline method across different tasks and different models.


Hallucinated Content
Detected, Regenerate
LLMs
Detector
Detecting Hallucinations in
the Generated Content
Output Non-Hallucinated Content

Hallucinations still persist, even after the model has undergone delicate instruction tuning …

Model	Question Answering			Data-to-text Writing			Summarization			Overall	
	# Resp.	# Span	Density	# Resp.	# Span	Density	# Resp.	# Span	Density	# Resp.	# Span
GPT-3.5-turbo-0613	75	89	0.12	272	384	0.18	54	60	0.05	401	533
GPT-4-0613	48	51	0.06	290	354	0.27	74	80	0.08	406	485
Llama-2-7B-chat	510	1010	0.59	888	1775	1.27	434	517	0.58	1832	3302
Llama-2-13B-chat	399	654	0.48	983	2803	1.53	295	342	0.41	1677	3799
Llama-2-70B-chat†	320	529	0.40	863	1834	1.15	212	245	0.26	1395	2608
Mistral-7B-Instruct	378	594	0.59	958	2140	1.51	617	828	0.86	1953	3562

Table 3: Hallucination counts and density of models. †: We used 4-bit quantized version of Llama-2-70B-chat.



Reinforcement learning becomes a clear choice when hallucinations still persist, even after the model has undergone delicate instruction tuning.

In RAG-HAT, We conduct Direct Performance Alignment on the selected LLM. It's a widely used technology that can do preference alignment without explicitly training a reward model.

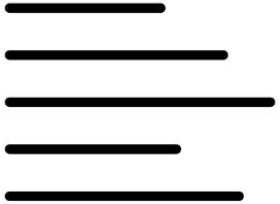
Our choice is based on the following pain points we identified during our investigations:

1. First, organizing hallucination mitigation tasks into a preference dataset for reward modeling is challenging, as annotators often struggle with determining which type of hallucination is more severe. In our previous work on RAGTruth, our annotators spent a lot of time struggling to achieve consensus.
2. Second, treating "not hallucinating" as a simple preference is problematic because it essentially requires LLMs to be "always correct," which is an overly rigid expectation. To address this pain point, we will introduce some adjustments later to make the alignment process more natural.

Detect → Rewrite → Mitigate

Generated Content

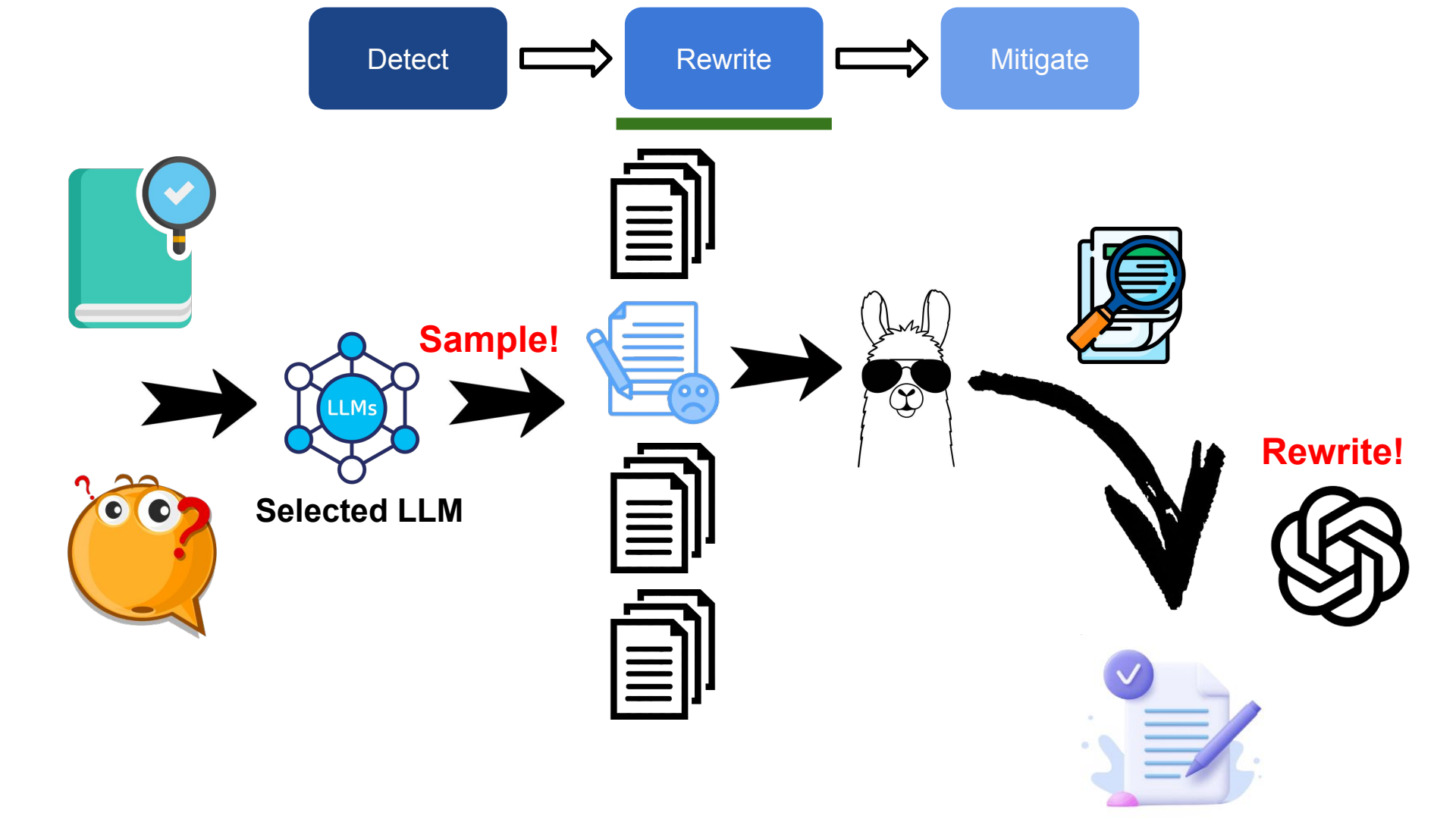
Split into Sentences!!



Hallucinated

Non-hallucinated

Interpretations


Detect
Rewrite
Mitigate
LLMs
Sample!
Selected LLM
Rewrite!

Detect → Rewrite → Mitigate

Defensive Advice

… According to their structured data, they offer a variety of beers and a Honey Avocado Ale that is **popular among customers ...**

... {
"review_stars": 5.0,
"review_date": "2021-12-17 02:46:28",
"review_text": "Beautiful views, great beer, and family friendly. **Highly recommend the Honey Avocado Ale** ...

Some might argue that the claim of the Honey Avocado Ale being "popular among customers" could be considered a hallucination since the JSON data does not explicitly state its popularity …

Figure 1: An Example of Defensive Advice: The LLM made a minor extension partially based on the provided references. Defensive advice highlights that the statement is not well supported.

Overly Cautious Penalization

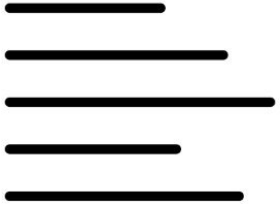
Teaching yourself can be done by learning something new with each task you complete, as mentioned in passage 1. However, taking notes is not always necessary. To learn a language quickly, studying or working abroad can be helpful, as mentioned in passage 2. Wikipedia ...

Teaching yourself can be done by learning something new with each task you complete, as mentioned in passage 1. ~~However, taking notes is not always necessary.~~ To learn a language quickly, studying or working abroad can be helpful, as mentioned in passage 2. Wikipedia ...

Detect → Rewrite → Mitigate

Generated Content

Split into Sentences!!

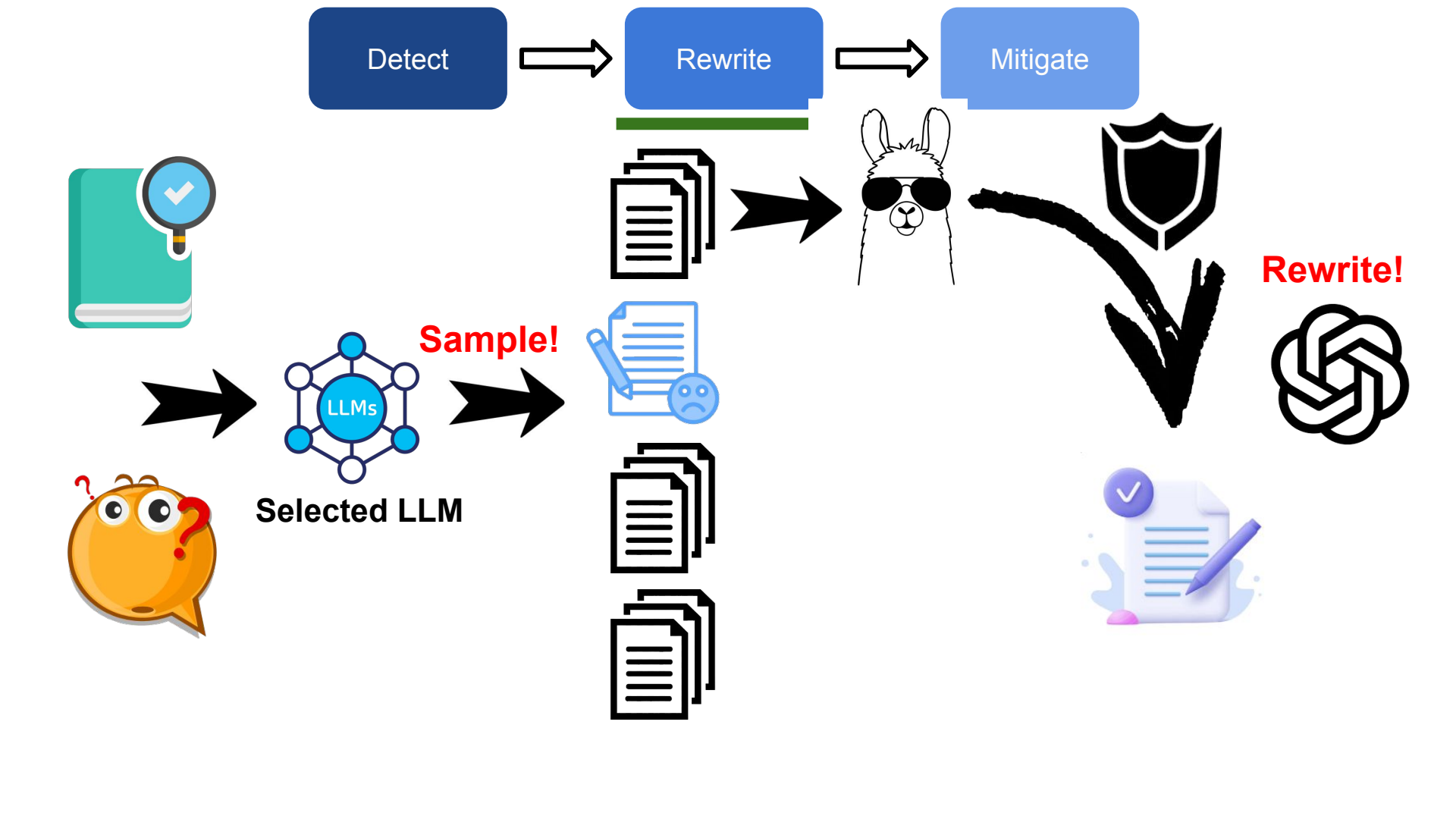


Hallucinated

Non-hallucinated

Interpretations

Defensive Advice


Detect
Rewrite
Mitigate
LLMs
Sample!
Selected LLM
Rewrite!

DATASET	METHOD	Detector	GPT-4 Turbo	Human	Average
RAGTruth Test Set	Qwen	36.9(-)	51.3(-)	34.4(-)	40.9(-)
	Qwen(Regenerate)	-	44.2(↓13.8%)	-	44.2(↓13.8%)
	RAG-HAT	22.7(↓**38.5%**)	41.3(↓**19.5%**)	25.7(↓**25.3%**)	29.9(↓**26.9%**)
WebGLM 1000	Qwen	21.3(-)	46.7(-)	-	34(-)
	Qwen(Regenerate)	-	38.8(↓17.0%)	-	38.8(↓17.0%)
	RAG-HAT	12.0(↓**43.7%**)	37.9(↓**19.0%**)	-	24.9(↓**26.8%**)

Table 3: Hallucination Rate: 1,000-Example WebGLM Set and RAGTruth Test Set (Total 450 Examples): Our detection model cannot fairly benchmark the hallucination rate of the regeneration approach since it serves as the trigger for regeneration.

PAIRED METHOD	WIN RATE (GPT-4 Turbo)
RAG-HAT (full) :: (w/o defensive, w/ OCP)	**51.5**
RAG-HAT (full) :: (w/o defensive, w/o OCP)	**54.1**

Table 6: Impact of Training Dataset Composition on Answer Quality: Pairwise Comparison

DATASET	METHOD	GPT-4 Turbo	Human
RAGTruth Dataset	Qwen	41.1	33.2
	RAG-HAT	**57.3**	**40.8**
WebGLM 1000	Qwen	39.5	-
	RAG-HAT	**58.5**	-

Table 5: Answer Quality Win Rates: 1,000-Example WebGLM Set and RAGTruth Test Set

Thank You!