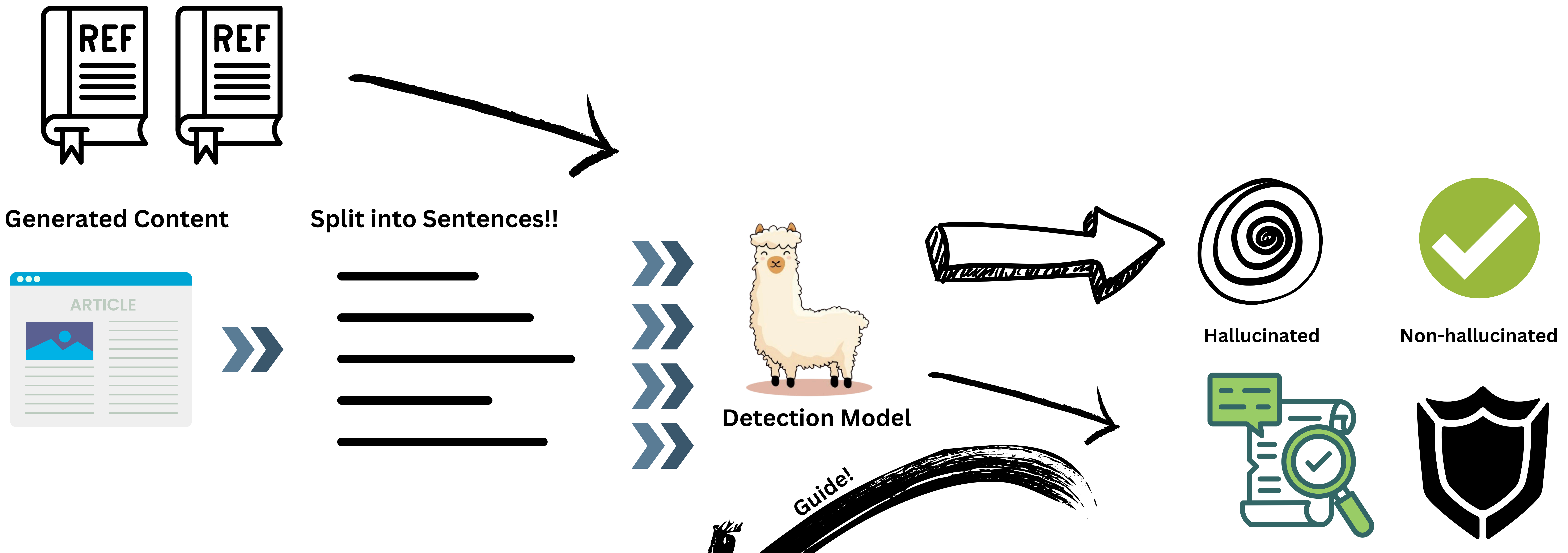
RAG-HAT: A Hallucination-Aware Tuning Pipeline for LLM in Retrieval-Augmented Generation

NEWSBREAK

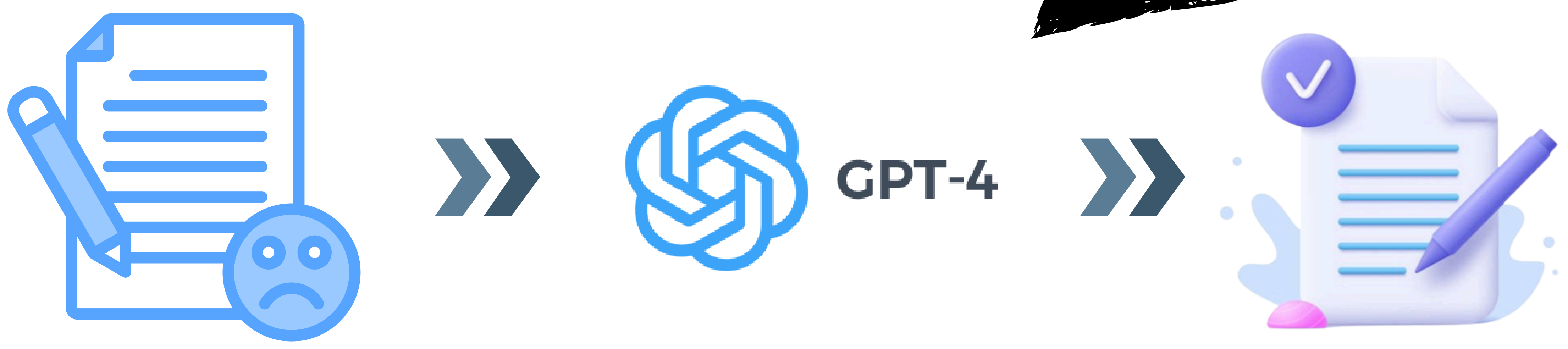
Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, Cheng Niu

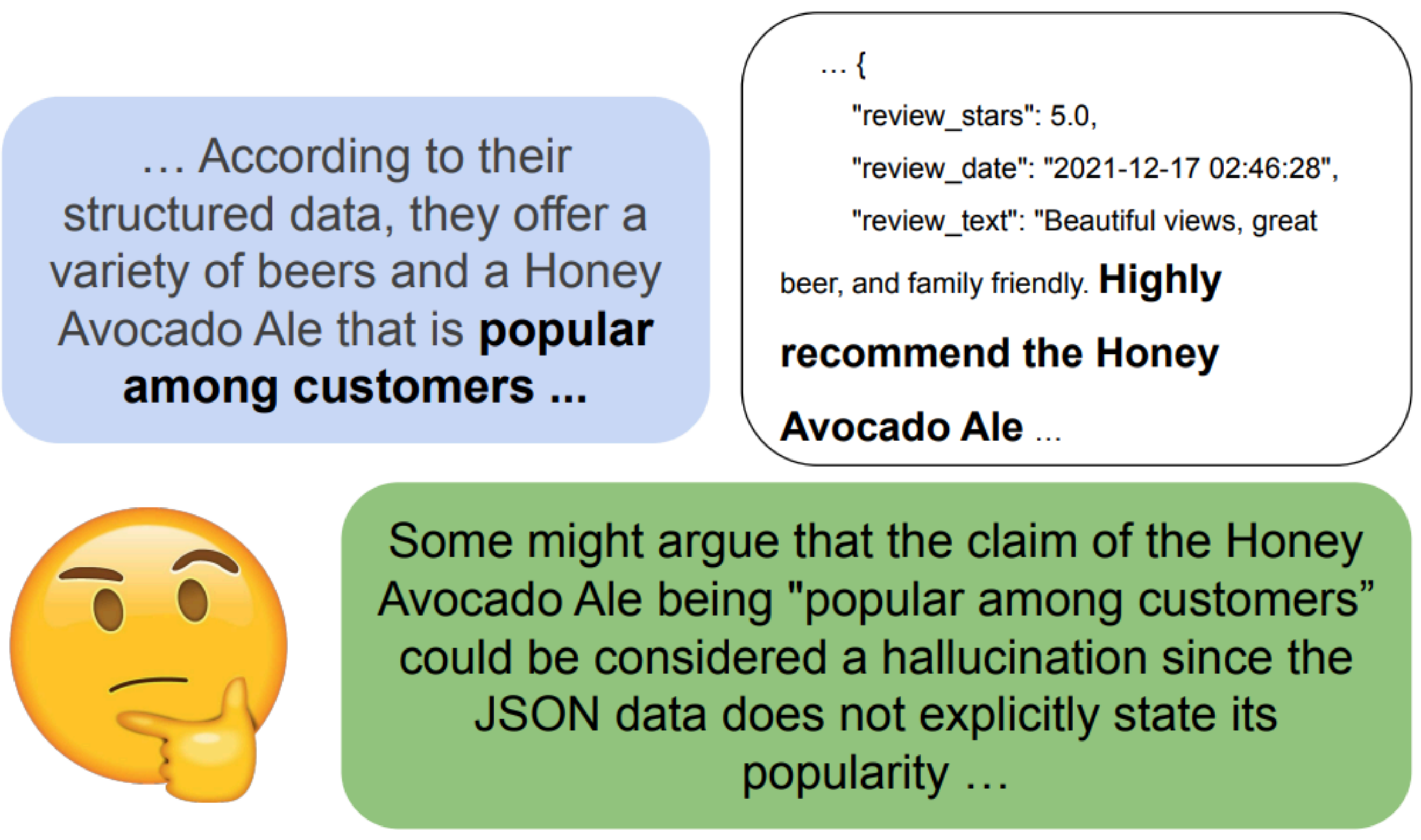
• Detection: Sentence-Level Interpretable Hallucination Detector



• Rewriting: Utilize GPT to Do Answer Rewrite

Detected Hallucinated Content





… 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 …

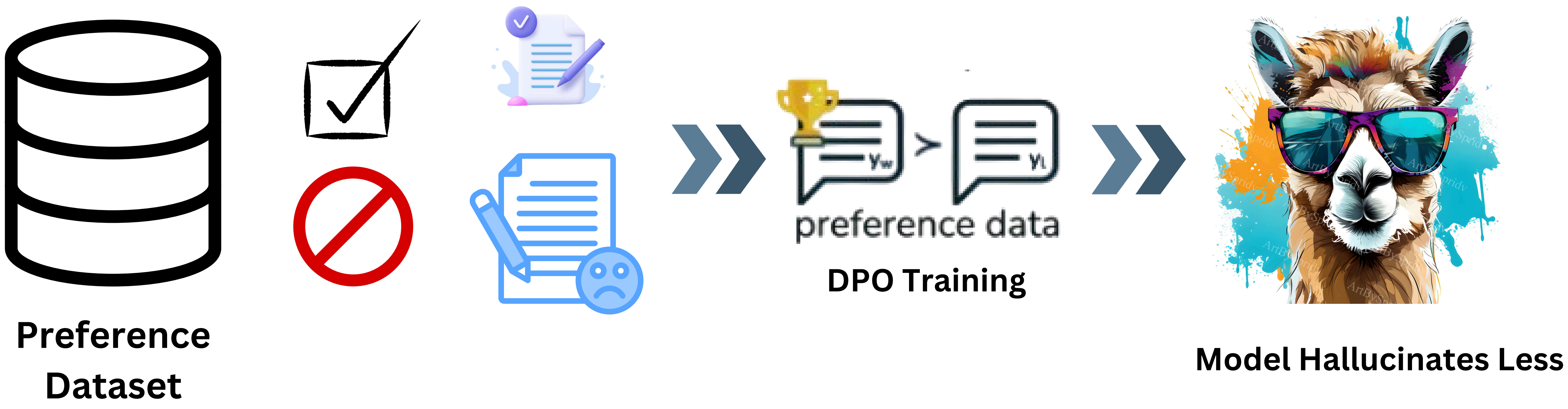
Figure 1: An Example of Defensive Advice: The LLM made a minor extension partially based on the provided references. Defensive advice highlights that the statement is not well supported.

• Mitigation: 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

• Mitigation: Conduct DPO Training



Evaluations

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.

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

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