Benchmarking Large Language Models in Retrieval-Augmented Generation

Jiawei Chen[@ISCAS] , Hongyu Lin[@ISCAS] , et al. AAAI, 2024

徐梓航

华中科技大学计算机科学与技术学院

2024 年 12 月 09 日

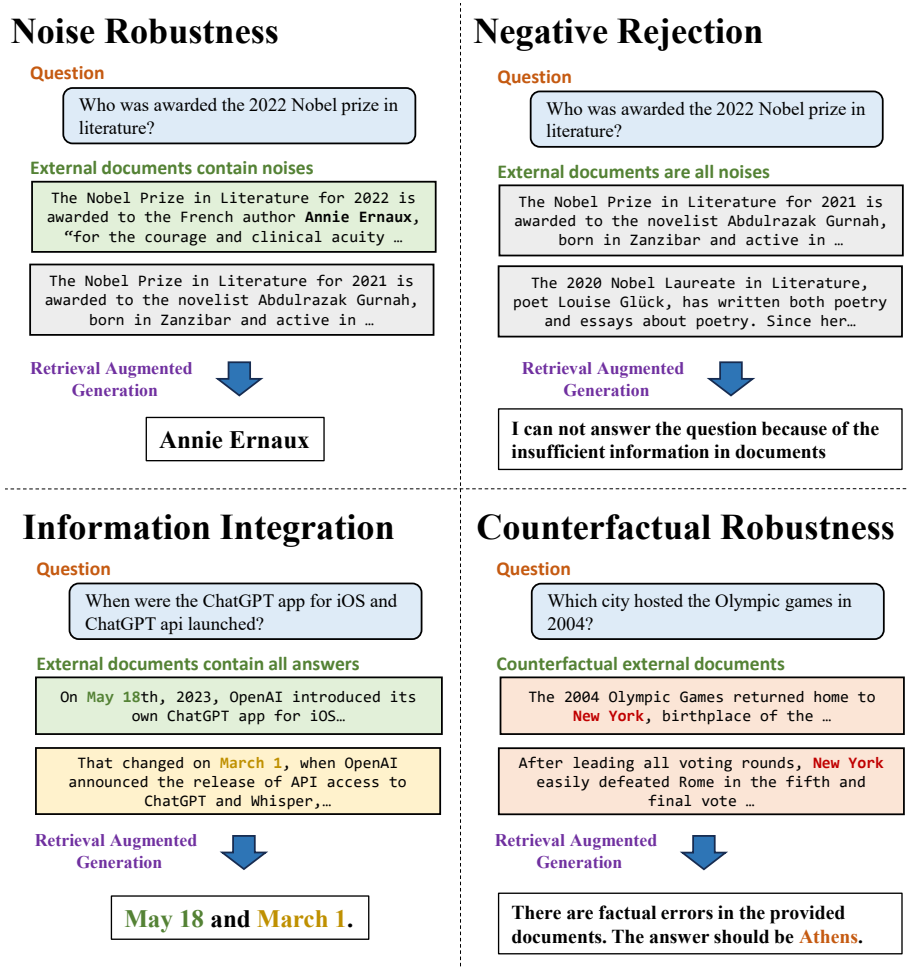

图 1: Four kinds of abilities required for ARG of LLMs.

- **噪声鲁棒性** (Noise Robustness)：能否有效处理和忽略无用信息。

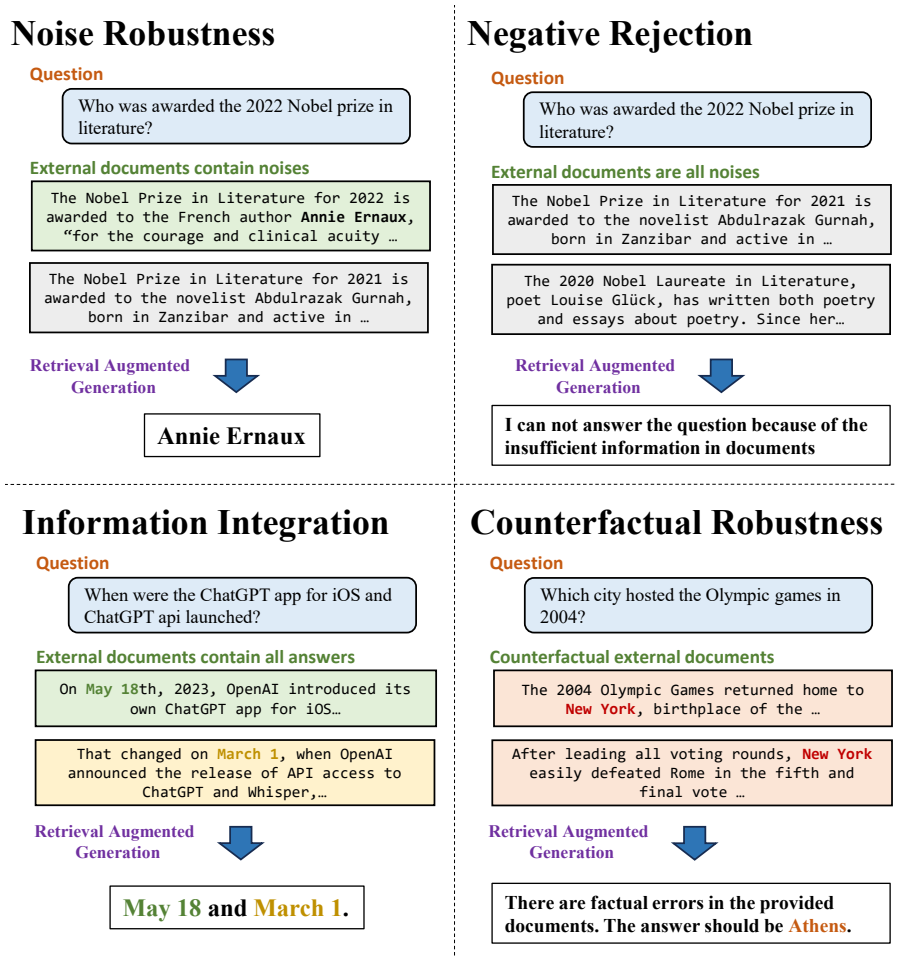

图 2: Four kinds of abilities required for ARG of LLMs.

- **负面拒绝** (Negative Rejection)：**确保在没有足够信息时不会生成不准确或虚假的答案。**

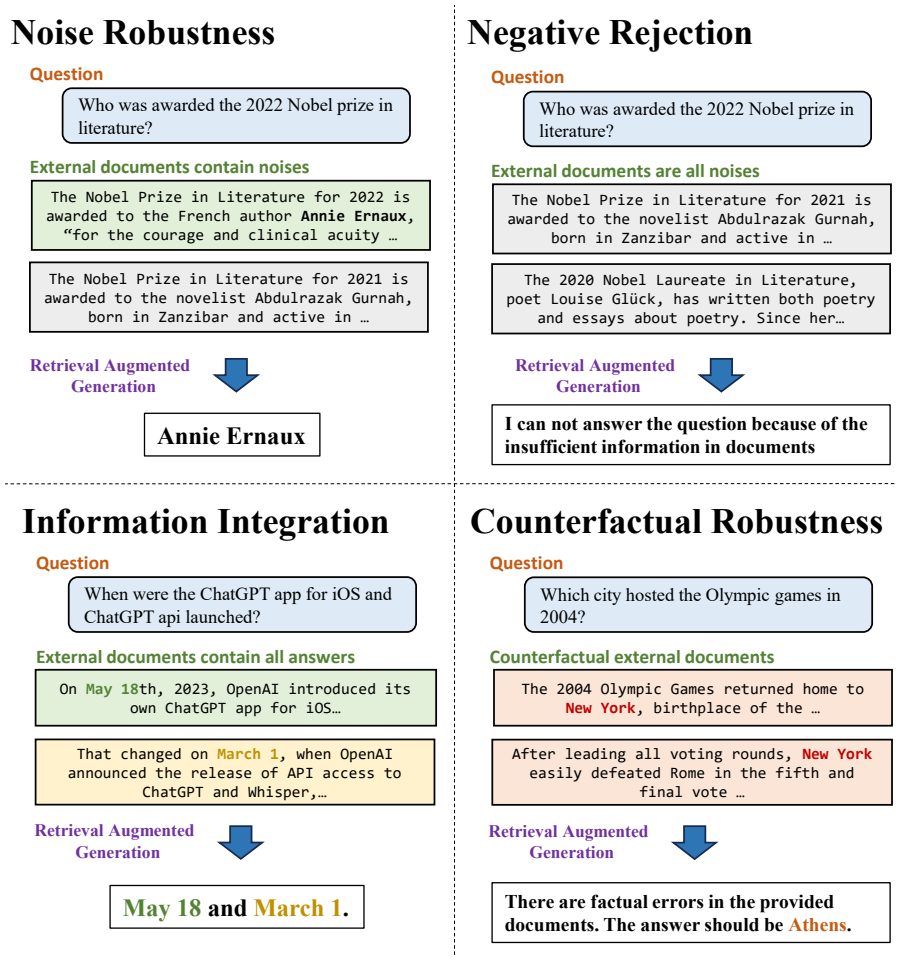

图 3: Four kinds of abilities required for ARG of LLMs.

- **信息整合**（Information Integration）：测试处理多源信息的能力。

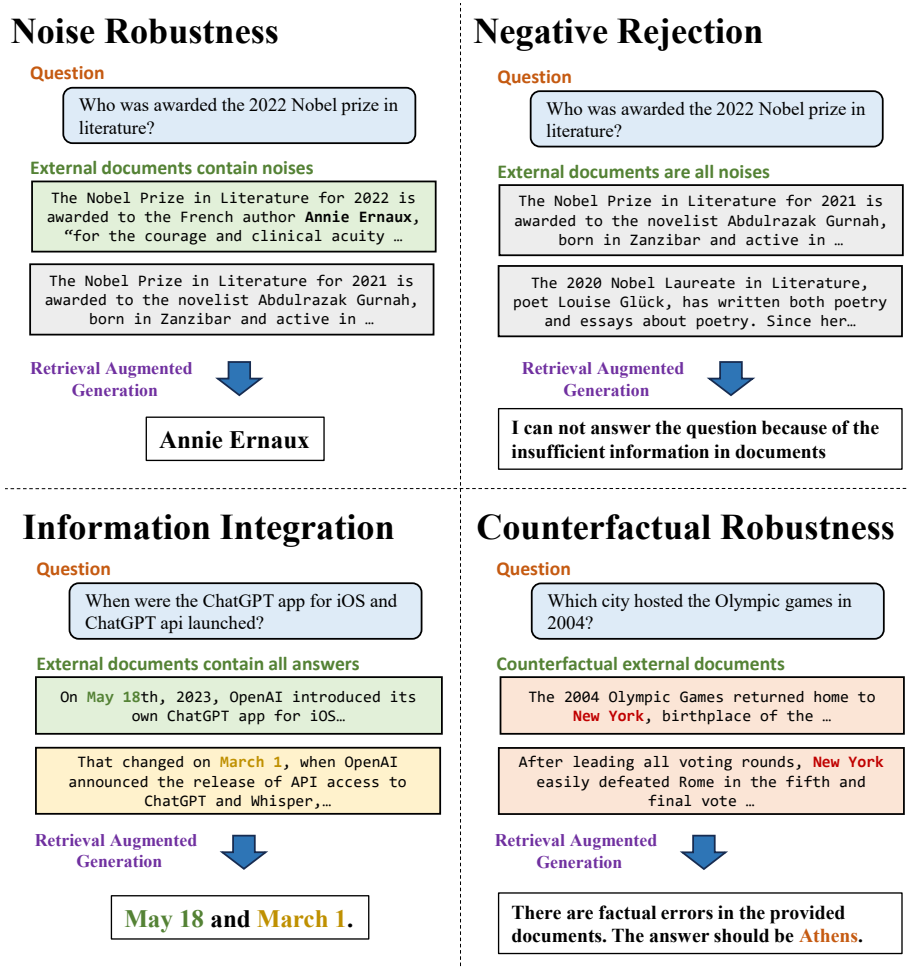

图 4: Four kinds of abilities required for ARG of LLMs.

- **反事实鲁棒性** (Counterfactual Robustness)：在面对误导性信息时能够作出正确的判断。

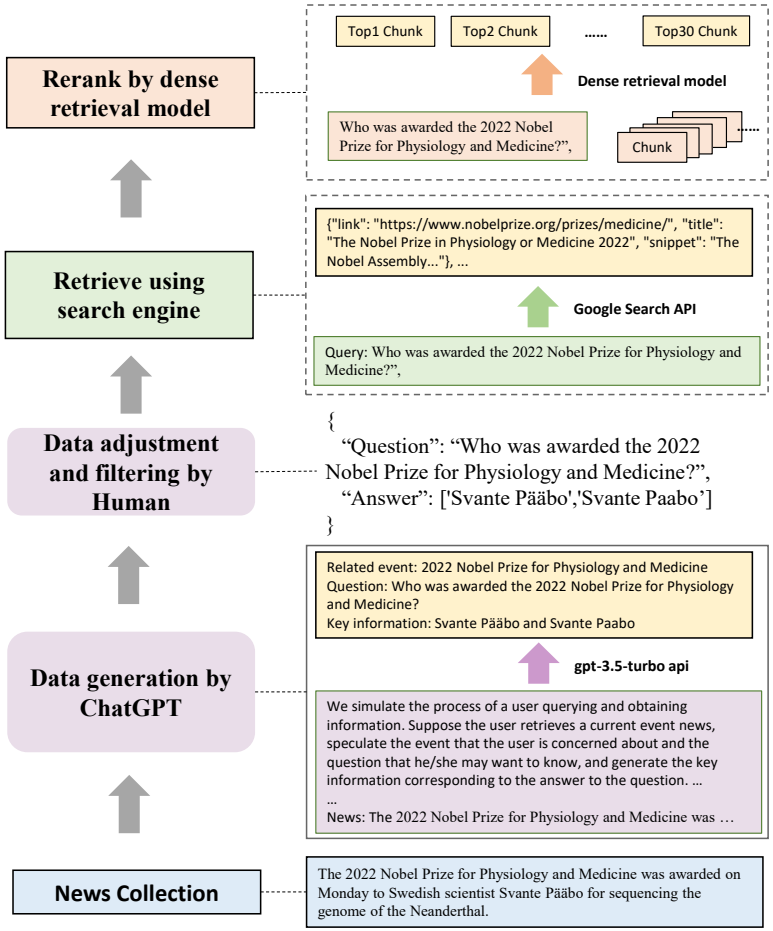

图 5: The process of data generation.

- **QA instances generation**

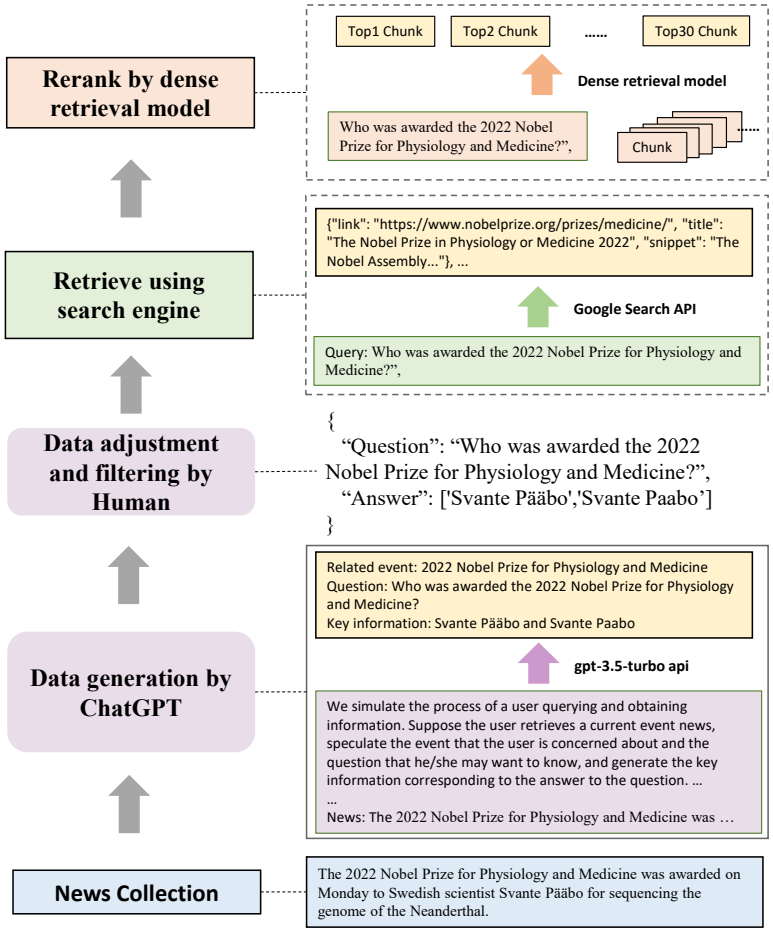

图 6: The process of data generation.

- **Retrieve using search engine**

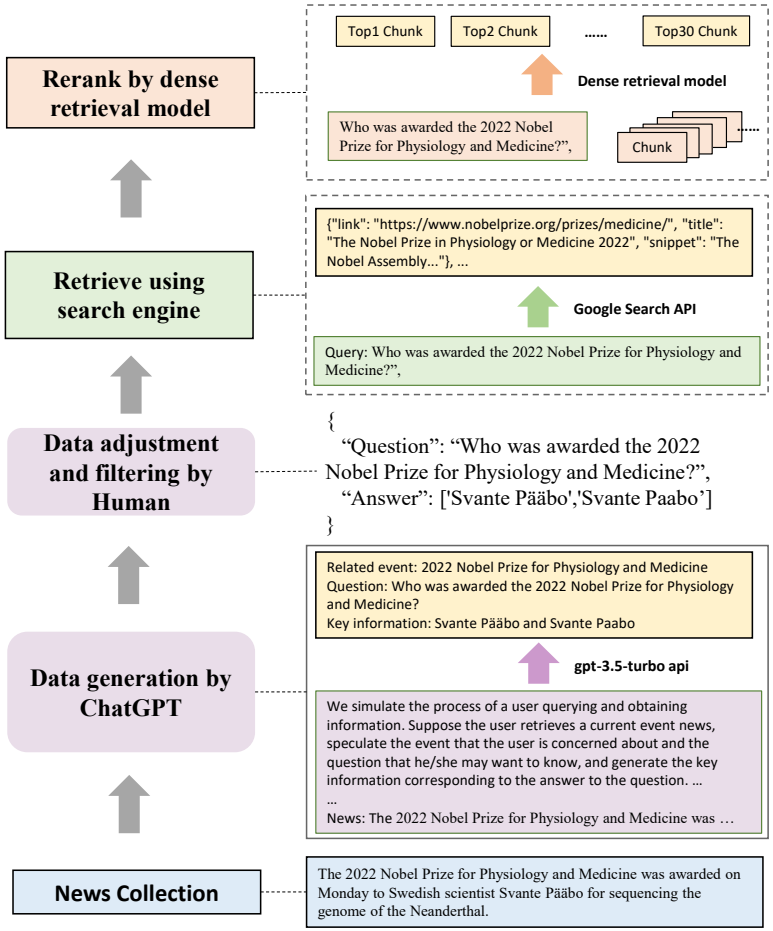

图 7: The process of data generation.

- **Testbeds construction for each ability**

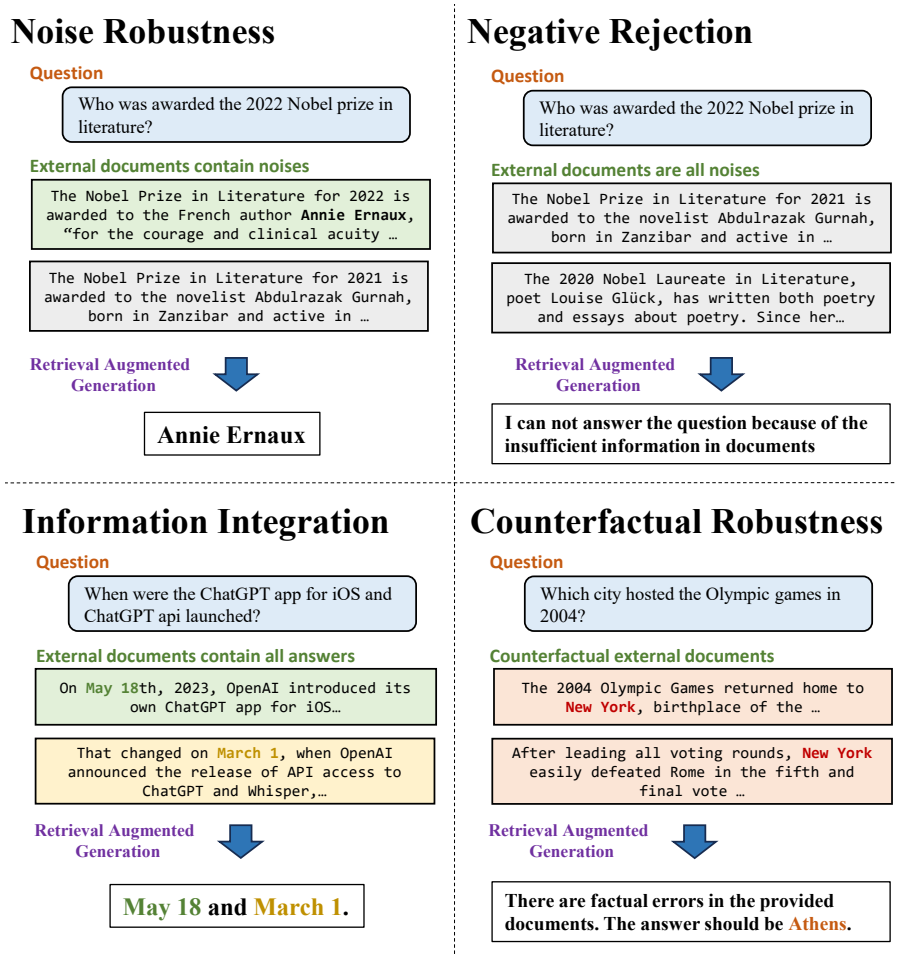

图 8: Four kinds of abilities required for ARG of LLMs.

- **Accuracy: Measure noise robustness and information integration.**

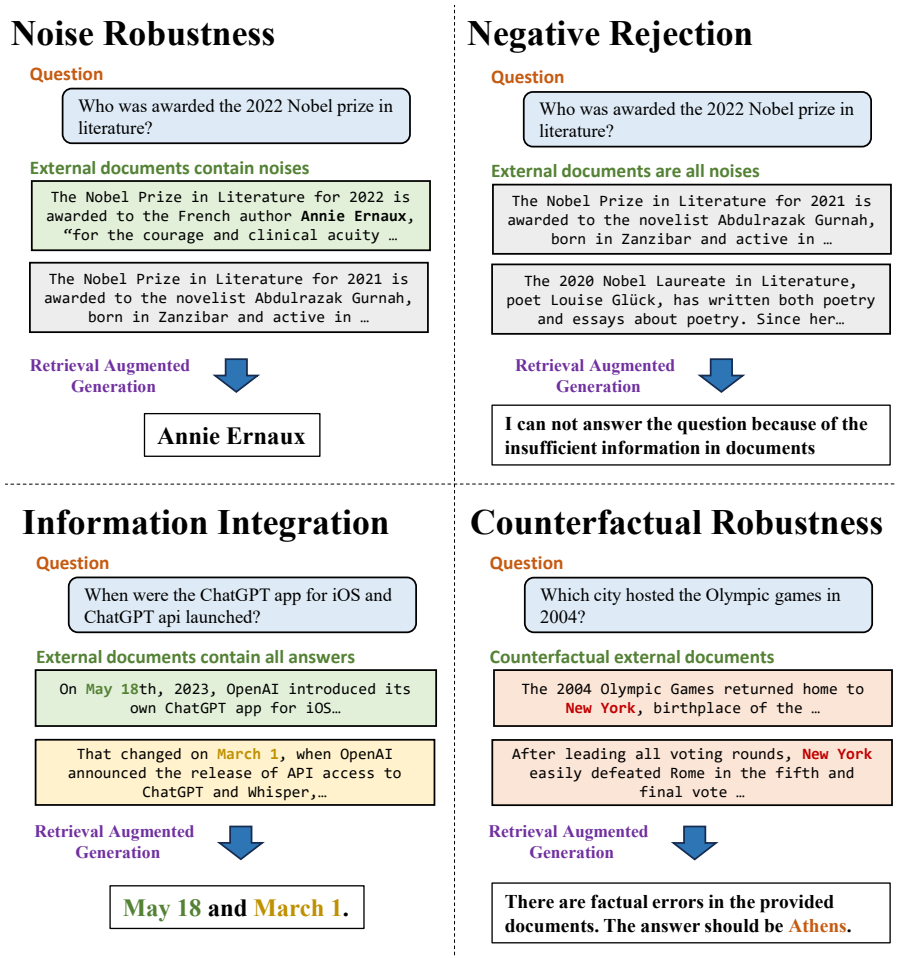

图 9: Four kinds of abilities required for ARG of LLMs.

- **Rejection rate: Measure negative rejection.**

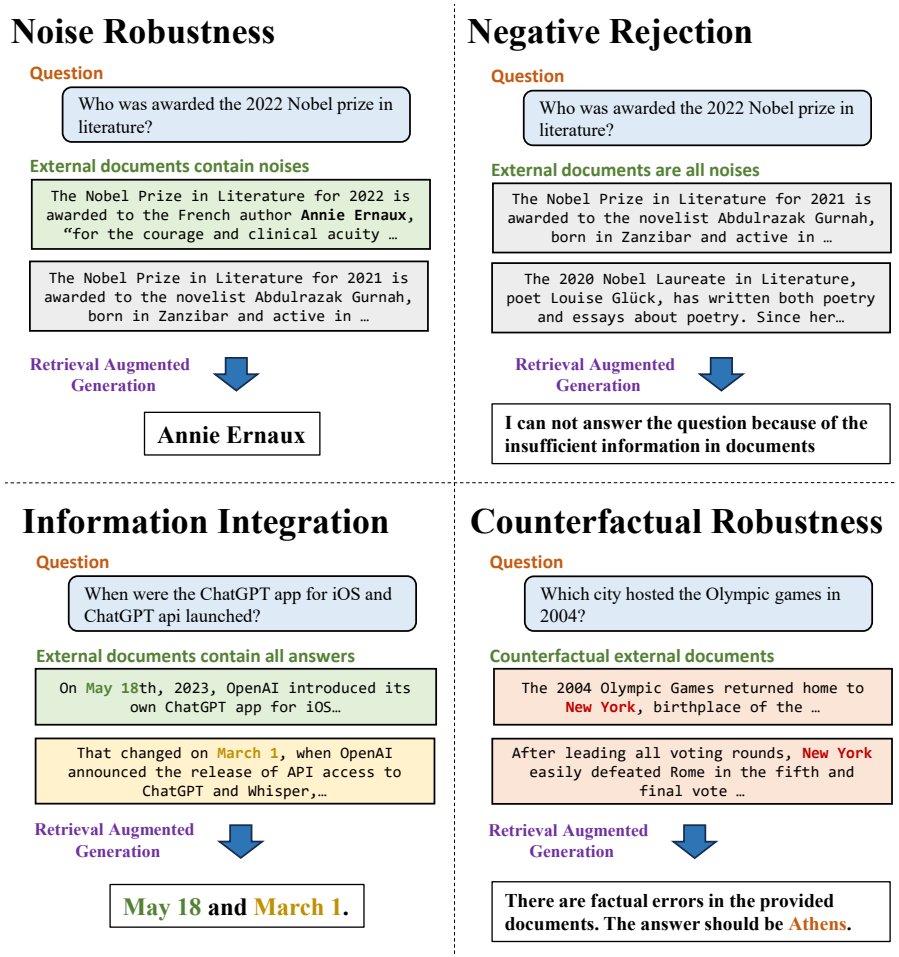

图 10: Four kinds of abilities required for ARG of LLMs.

- **Error detection rate: Measure counterfactual robustness.**

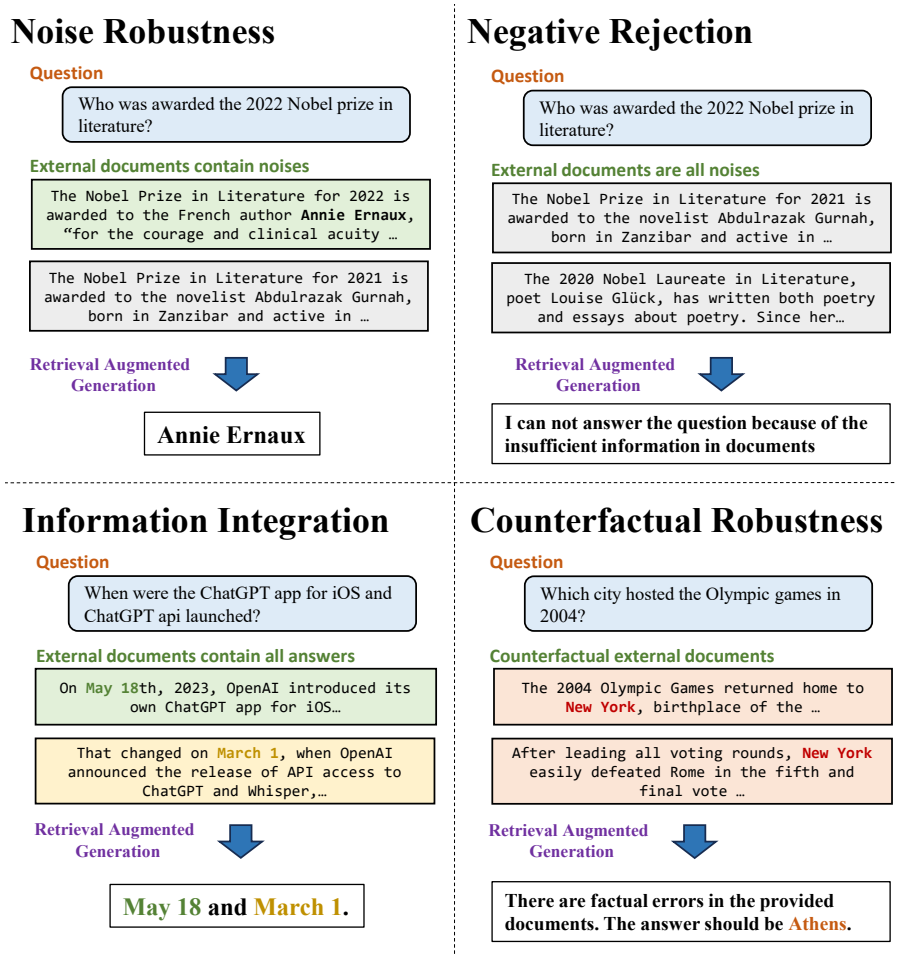

图 11: Four kinds of abilities required for ARG of LLMs.

- **Error correction rate: Measure counterfactual robustness.**

English

System instruction

You are an accurate and reliable AI assistant that can answer questions with the help of external documents. Please note that external documents may contain noisy or factually incorrect information. If the information in the document contains the correct answer, you will give an accurate answer. If the information in the document does not contain the answer, you will generate 'I can not answer the question because of the insufficient information in documents.' If there are inconsistencies with the facts in some of the documents, please generate the response 'There are factual errors in the provided documents.' and provide the correct answer.

User input Instruction

Document:\n{DOCS} \n\nQuestion:\n{QUERY}

Chinese

System instruction

你是一个准确和可靠的人工智能助手，能够借助外部文档回答问题，请注意外部文档可能存在噪声事实性错误。如果文档中的信息包含了正确答案，你将进行准确的回答。如果文档中的信息不包含答案，你将生成“文档信息不足，因此我无法基于提供的文档回答该问题。”如果部分文档中存在与事实不一致的错误，请先生成“提供文档的文档存在事实性错误。”，并生成正确答案。

User input Instruction

文档：\n{DOCS} \n\n问题：\n{QUERY}

图 12: Provide 5 external documents for each question.

	English					Chinese				
Noise Ratio	0	0.2	0.4	0.6	0.8	0	0.2	0.4	0.6	0.8
ChatGPT [Ope22]	**96.33**	**94.67**	**94.00**	**90.00**	**76.00**	**95.67**	**94.67**	**91.00**	**87.67**	**70.67**
ChatGLM-6B [THU23a]	93.67	90.67	89.33	84.67	70.67	94.33	90.67	89.00	82.33	69.00
ChatGLM2-6B [THU23b]	91.33	89.67	83.00	77.33	57.33	86.67	82.33	76.67	72.33	54.00
Vicuna-7B-v1.3 [CLL+23]	87.67	83.33	86.00	82.33	60.33	85.67	82.67	77.00	69.33	49.67
Qwen-7B-Chat [Qwe23]	94.33	91.67	91.00	87.67	73.67	94.00	92.33	88.00	84.33	68.67
BELLE-7B-2M [JDG+24]	83.33	81.00	79.00	71.33	64.67	92.00	88.67	85.33	78.33	67.68

表 1: The experimental result of noise robustness measured by accuracy (%) under different noise ratios. We can see that the increasing noise rate poses a challenge for RAG in LLMs.

- complex instructions following ability of LLMs
- CELLO Benchmark
- Conduct extensive experiments to compare the performance of representative models.

思考

- **数据集的多样性和代表性**：论文中提出的 CELLO 基准测试涵盖了多种复杂指令的特征，并从真实世界场景中构建了评估数据集。然而，数据集的多样性和代表性始终是一个可以进一步探讨的话题。未来的工作可以探索如何确保数据集覆盖更广泛的语言、地区和文化背景，以及如何平衡不同领域和任务类型的样本。
- **评估标准的细化**：尽管论文提出了四个评估标准（Count limit、Answer format、Task-prescribed phrases、Input-dependent query），但这些标准是否可以进一步细化，以便更精确地捕捉模型在处理复杂指令时的细微差异，是一个值得考虑的问题。
- **模型的可解释性**：论文主要关注模型的性能评估，但对于模型的决策过程和内部机制的可解释性讨论不多。未来的研究可以探索如何提高模型在处理复杂指令时的透明度和可解释性，以便更好地理解模型的行为。

思考

- **模型的适应性和泛化能力**：论文中的实验主要关注模型在特定数据集上的表现。未来的研究可以探讨模型在面对新的、未见过的复杂指令时的适应性和泛化能力，以及如何通过持续学习或迁移学习来提高这些能力。
- **多模态和跨领域指令的理解**：随着多模态学习和跨领域应用的兴起，未来的研究可以考虑如何评估和提高模型在处理包含图像、声音等多种模态信息的复杂指令时的性能。
- **实时性能和资源消耗**：论文中的评估主要关注模型的准确性和完成度。在实际应用中，模型的实时性能和资源消耗（如计算时间、内存使用等）也是非常重要的考量因素。未来的工作可以探讨如何在保证性能的同时优化模型的效率。

[CLHS23] Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. Benchmarking Large Language Models in Retrieval-Augmented Generation. *arXiv e-prints*, page arXiv:2309.01431, September 2023.

[CLL+23] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[JDG+24] Yunjie Ji, Yong Deng, Yan Gong, Yiping Peng, Qiang Niu, Baochang Ma, and Xiangang Li. Your paper title. *Journal Name*, 123:456–789, 2024.

[lat23] latexstudio. Hust-beamer-theme, 2023.

[Ope22] OpenAI. Chatgpt: Optimizing language models for dialogue. https://openai.com/blog/chatgpt, 2022.

[Qwe23] QwenLM. Qwen-7b. https://github.com/QwenLM/Qwen-7B, 2023.

[THU23a] THUDM. Chatglm-6b. https://github.com/THUDM/ChatGLM-6B, 2023.

[THU23b] THUDM. Chatglm2-6b. https://github.com/THUDM/ChatGLM2-6B, 2023.

Thanks!