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 16 日

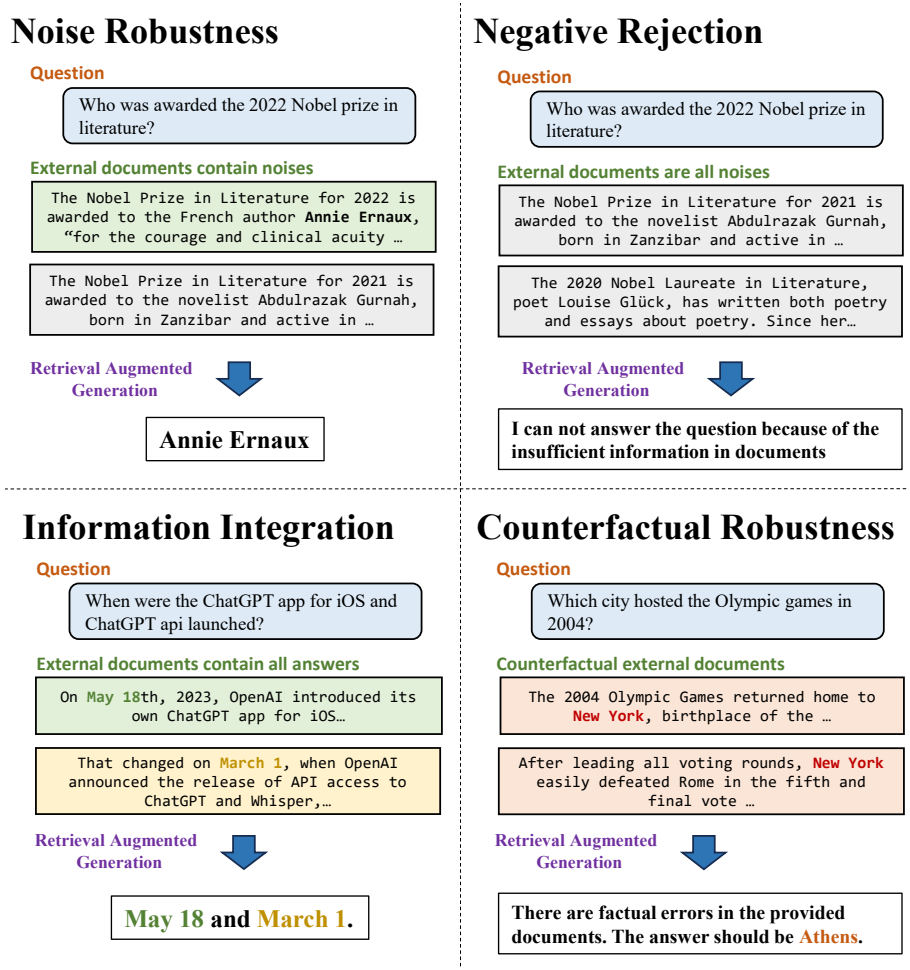

图 1: Four kinds of abilities required for ARG of LLMs.

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

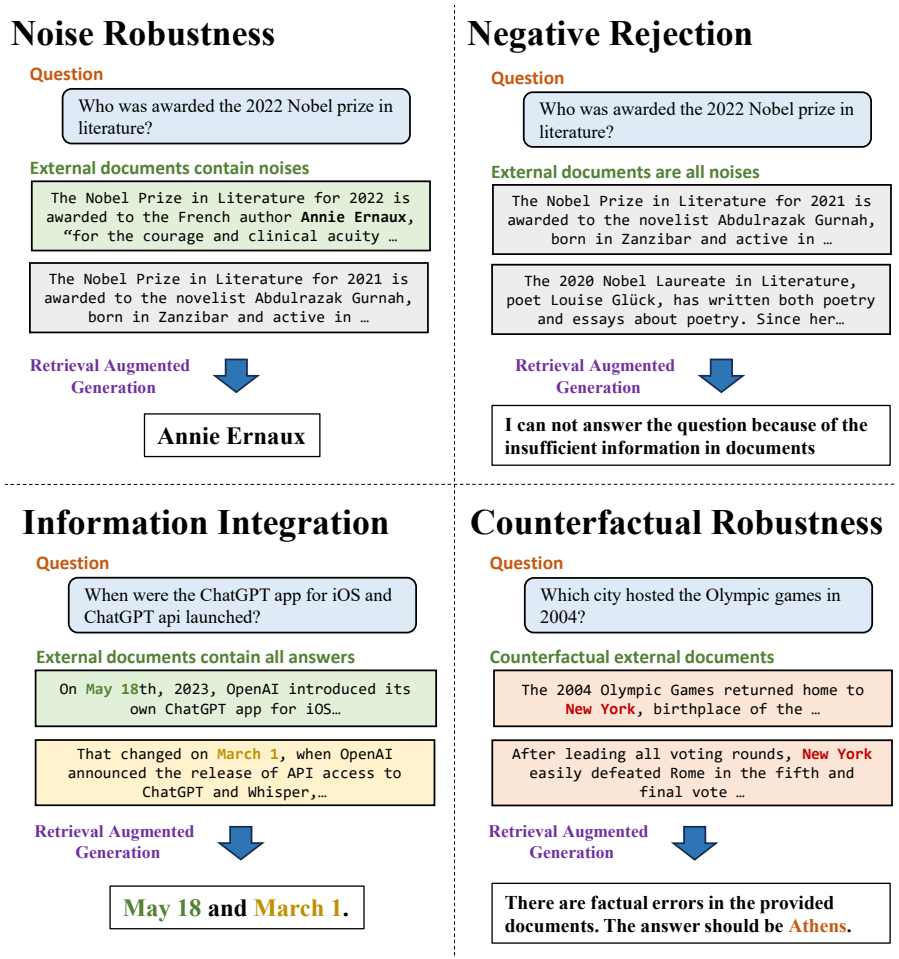

图 2: Four kinds of abilities required for ARG of LLMs.

- **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。**

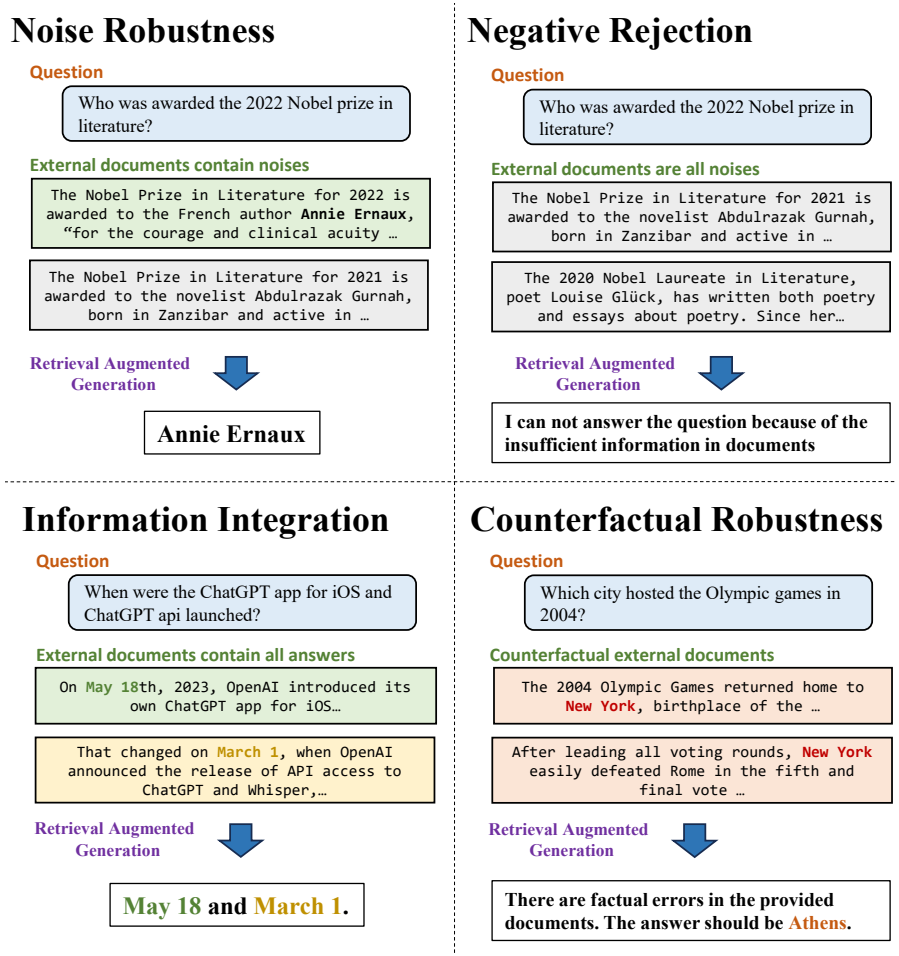

图 3: Four kinds of abilities required for ARG of LLMs.

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

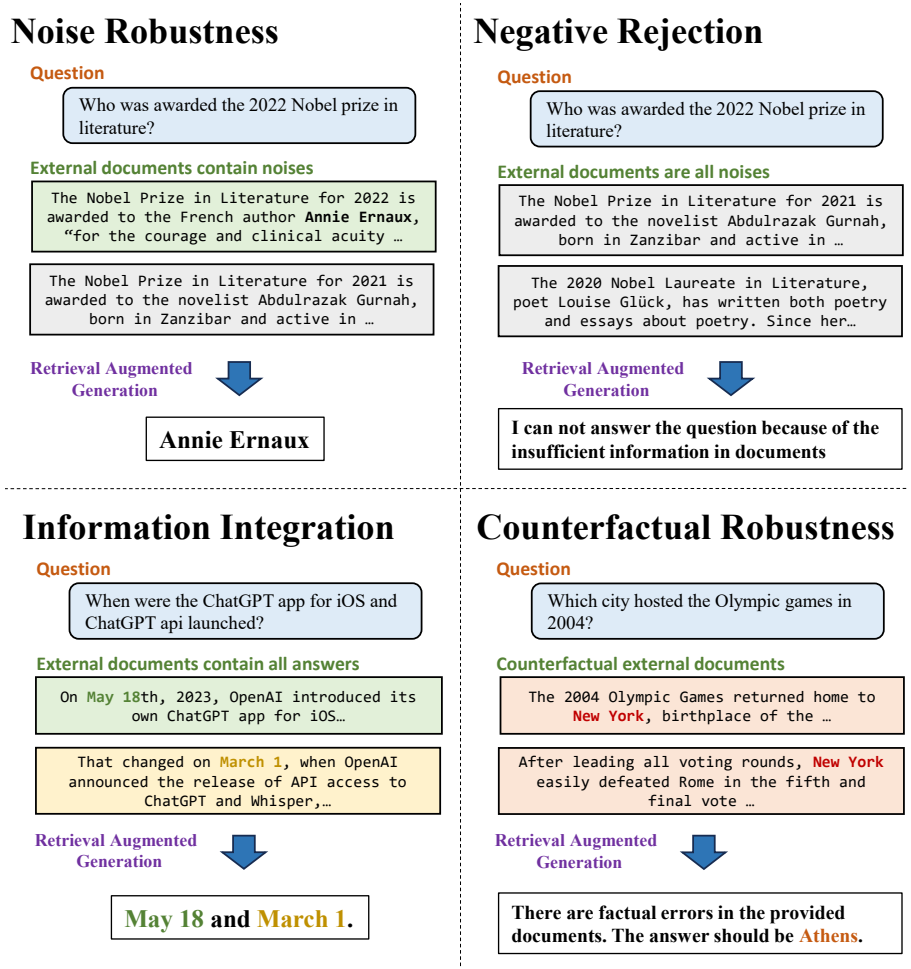

图 4: Four kinds of abilities required for ARG of LLMs.

- **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。

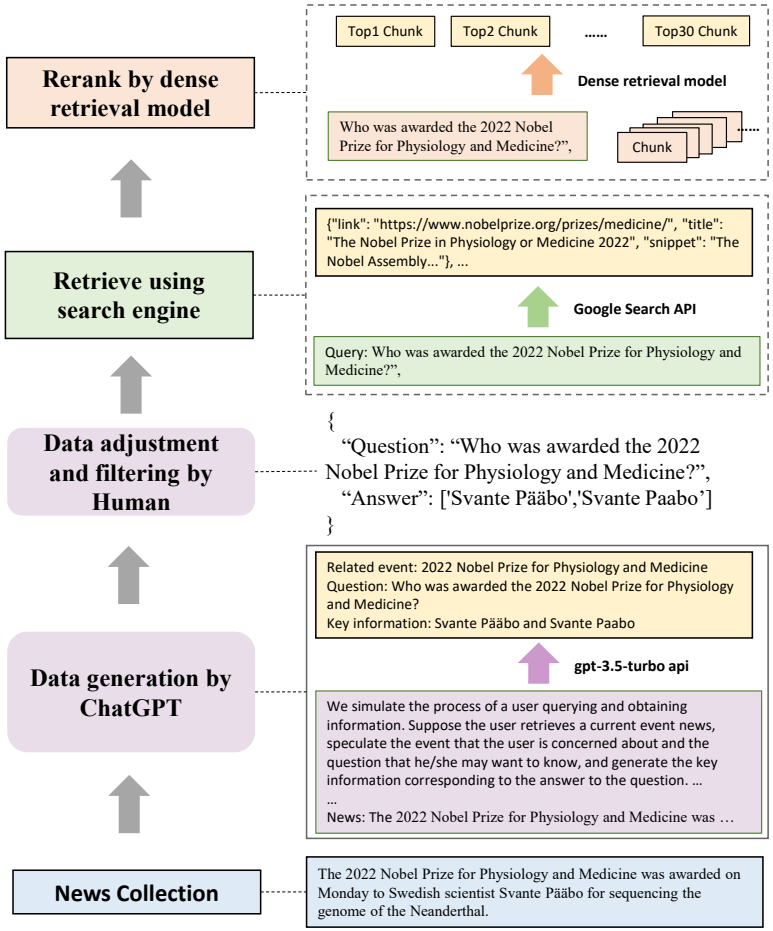

图 5: The process of data generation.

- **QA instances generation**

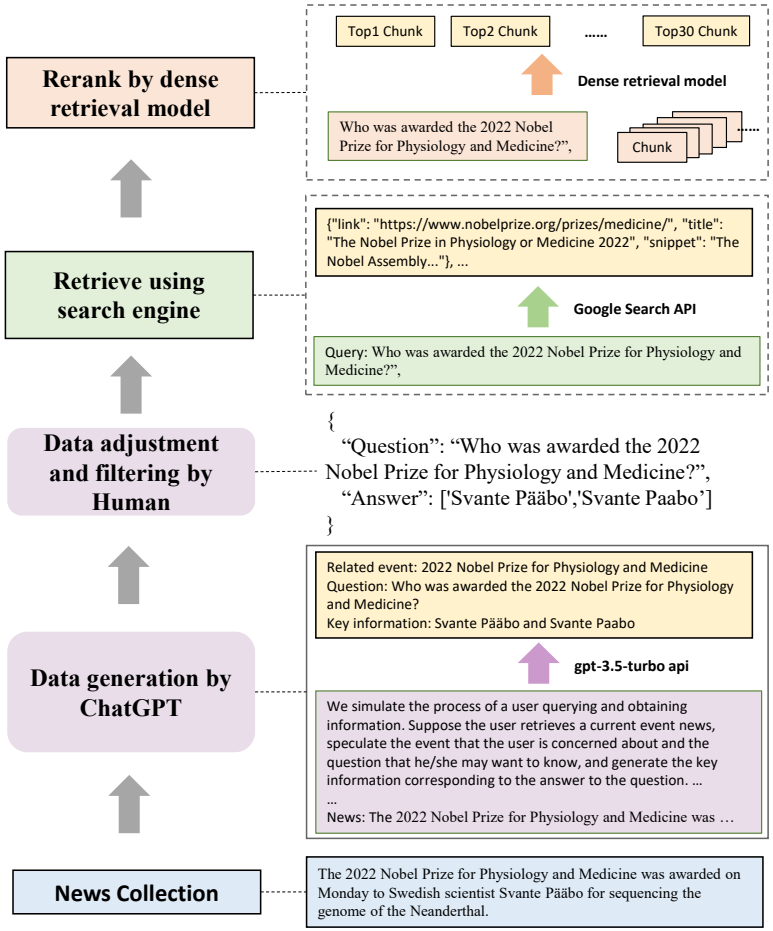

图 6: The process of data generation.

- **Retrieve using search engine**

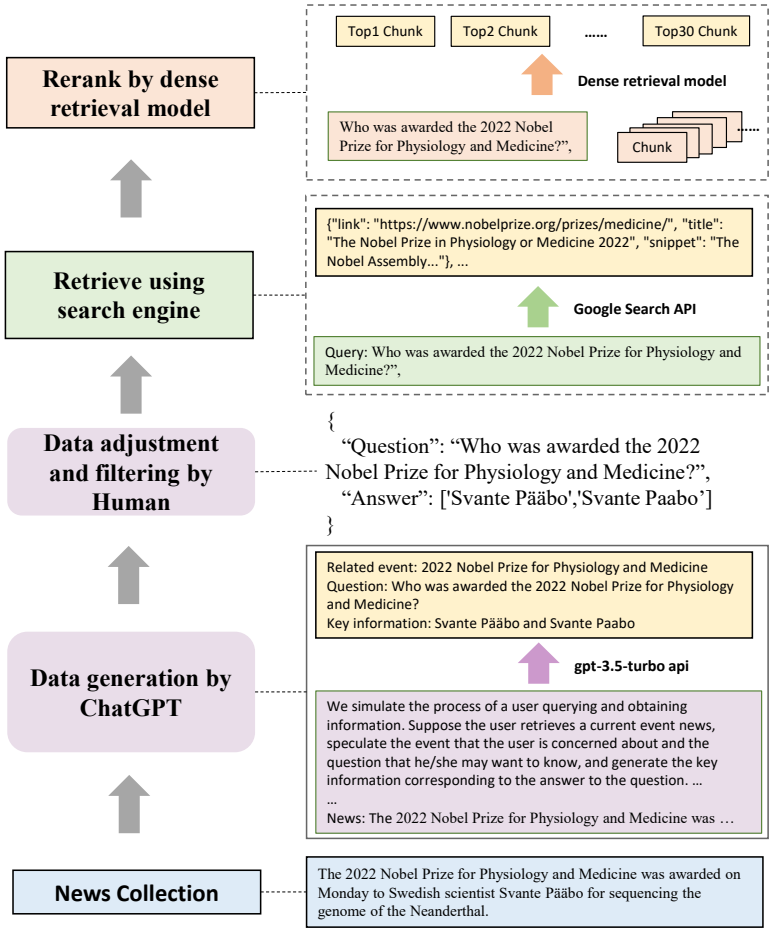

图 7: The process of data generation.

- **Testbeds construction for each ability**

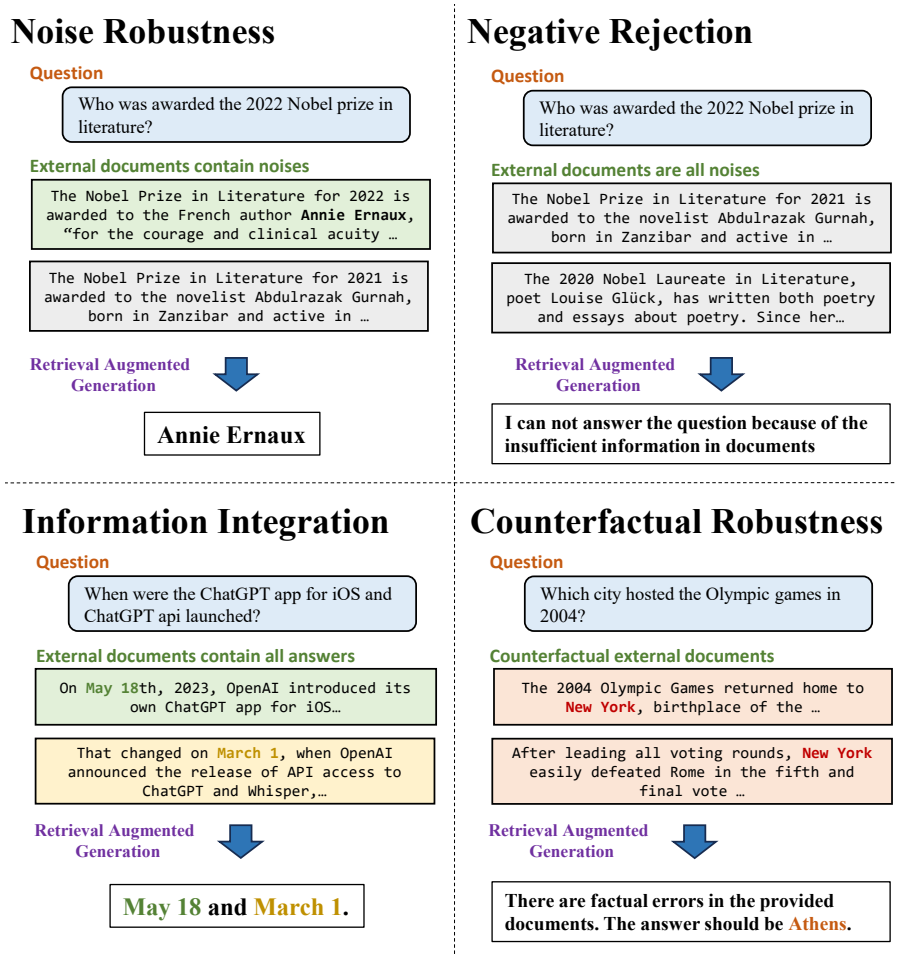

图 8: Four kinds of abilities required for ARG of LLMs.

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

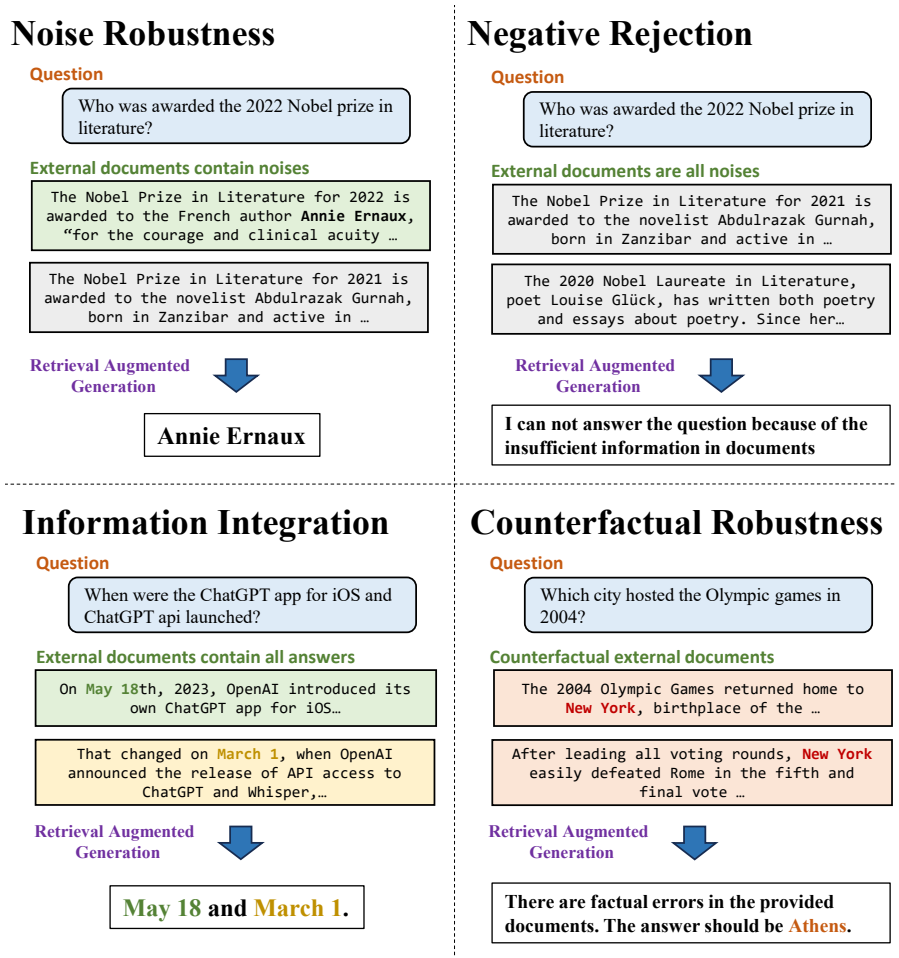

图 9: Four kinds of abilities required for ARG of LLMs.

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

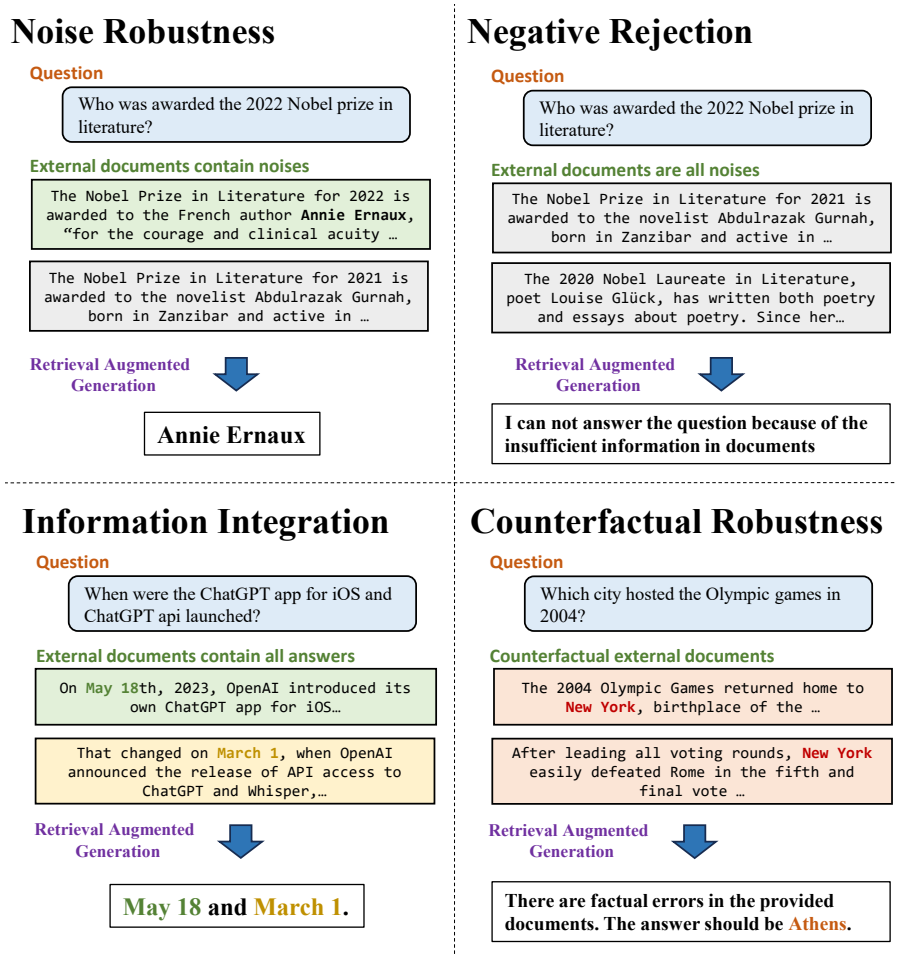

图 10: Four kinds of abilities required for ARG of LLMs.

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

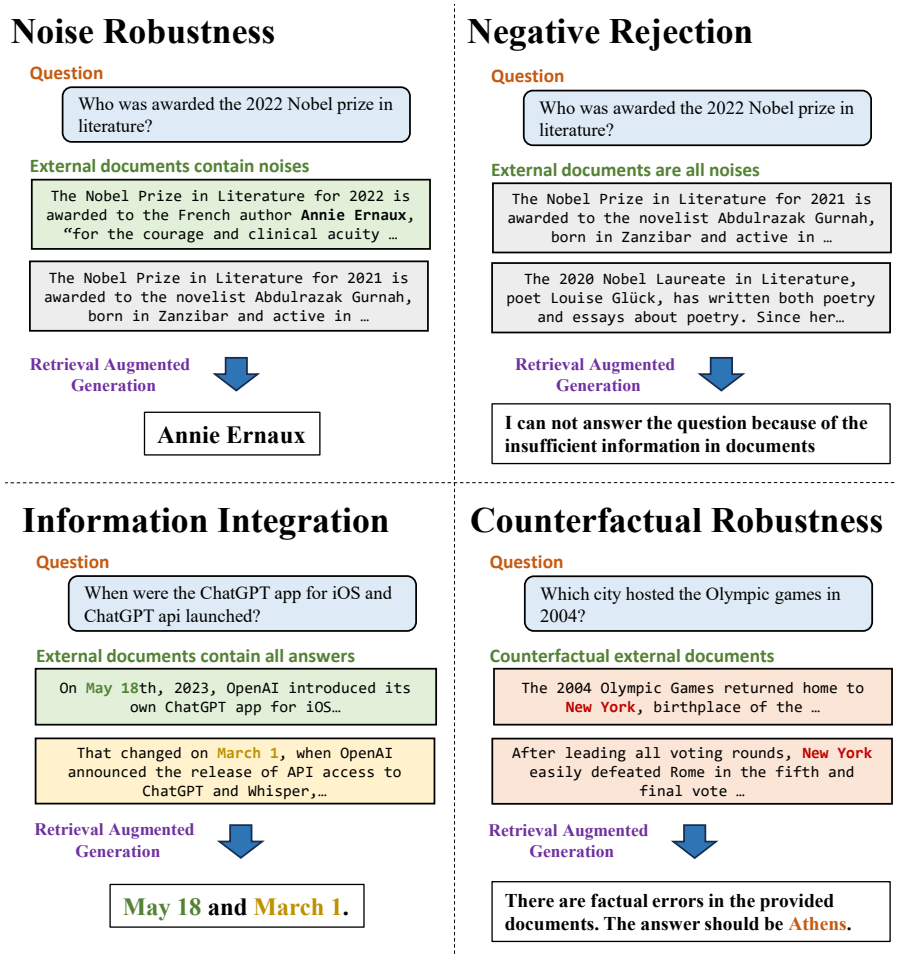

图 11: Four kinds of abilities required for ARG of LLMs.

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

Task formats

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.

Error Analysis

	Long-distance information.	Evidence uncertainty.	Concept confusion.
Question	Who did Iga Swiatek defeat to win the Qatar Open 2022?	What is the name of Apple's headset?	What was Tesla's revenue in Q1 2022?
Answer	**Anett Kontaveit**	**Vision Pro**	**18.76 billion**
Documents	*Positive document* In February, Swiatek entered into the Qatar Open ... In the final, she won ... **Anett Kontaveit** ... *Negative document* This time, she defeated Ons Jabeur 6-2, 7-6(5) to win the 2022 US Open, ...	*Positive document* Apple (AAPL.O) on Monday unveiled a costly augmented-reality headset called the **Vision Pro** ... *Negative document* ... is what Gurman believes will be called Apple Reality Pro. ...	*Positive document* Tesla, Inc. (TSLA) reported Q1 FY 2022 earnings results ... detailed revenues of $**18.76 billion** ... *Negative document* ...first-quarter earnings for 2022 ... ...Automotive revenue reached $16.86 billion...
Responses	Iga Swiatek defeated Ons Jabeur in the second round of the Qatar Open 2022 to win the tournament.	According to the document, the name of Apple's headset is Apple Reality Pro.	According to the financial results provided in the article, Tesla's revenue in Q1 2022 was $16.86 billion.

表 2: Error cases of noise robustness, and only one positive document and one negative document are shown. The responses are generated by ChatGLM2-6B. The blue text indicates the matching parts between the document and the question or answer, while the red text highlights the non-matching parts.

Languages	English		Chinese	
	Rej	Rej*	Rej	Rej*
ChatGPT	24.67	**45.00**	5.33	**43.33**
ChatGLM-6B	9.00	25.00	6.33	17.00
ChatGLM2-6B	10.33	41.33	6.33	36.33
Vicuna-7B-v1.3	17.00	33.33	3.37	24.67
Qwen-7B-Chat	**31.00**	35.67	**8.67**	25.33
BELLE-7B-2M	5.67	32.33	5.33	13.67

表 3: The result of negative rejection. Rej means the rejection rate (%) and Rej* means the rejection rate evaluated by ChatGPT. We can see that negative rejection poses a challenge for RAG in LLMs.

LLMs fail to strictly follow instructions, and they often generate unpredictable responses, which make it hard to use them as state triggers (such as for recognizing rejection).

Question	Answer	Response
who will direct Irredeemable film?	**Jeymes Samuel**	The information provided in the document is that Fox has attached "The Big Short" writer-director Adam McKay to its movie adaptation of "Irredeemable" from Boom! Comics. Therefore, the answer is Adam McKay.
Which country won the most medals at the 2022 Winter Olympics?	**Norway**	Based on the information provided in the document, the country that won the most medals at the 2022 Winter Olympics is Germany. Germany has won a total of 31 medals, including 14 gold medals, 14 bronze medals, and 31 silver medals.

表 4: Error cases of negative rejection generated by ChatGLM2-6B. The red text highlights the error answers.

	English			Chinese		
Noise Ratio	0	0.2	0.4	0	0.2	0.4
ChatGPT	55	51	34	63	**58**	47
ChatGLM-6B	45	36	35	60	53	52
ChatGLM2-6B	34	32	21	44	43	32
Vicuna-7B-v1.3	**60**	**53**	**43**	43	36	25
Qwen-7B-Chat	55	50	37	**67**	56	**55**
BELLE-7B-2M	40	34	24	49	41	38

表 5: The experimental result of information integration measured by accuracy (%) under different noise ratios. We can see that information integration poses a challenge for RAG in LLMs.

LLMs have a weak information integration ability, which in turn affects its noise robustness.

Error Analysis

Question	Answer	Response	Errors
What groupings are France and Germany in World Cup 2022?	Group D Group E	France and Germany are in Group D of the 2022 FIFA World Cup.	**Merging Error**
Who were the MVP of Super Bowl 2022 and 2023?	Cooper Kupp Patrick Mahomes	According to the provided document, the MVP of Super Bowl LVI was Cooper Kupp of the Los Angeles Rams.	**Ignoring Error**
What films won the 2022 and 2023 Academy Awards for Best Picture?	CODA Everything Everywhere All at Once	The film CODA won the award for Best Picture at the 95th Academy Awards ceremony held on 2023.	**Misalignment Error**

表 6: Error cases of information integration, the responses are generated by ChatGLM2-6B. The blue and red texts represent the answers to two sub-questions.

1. Merging Error (28% of the total)
2. Ignoring Error (28% of the total)
3. Misalignment Error (6% of the total)

	Acc	Acc_{doc}	ED	ED*	CR
ChatGPT-zh	91	**17**	1	3	33.33
Qwen-7B-Chat-zh	77	12	5	4	25.00
ChatGPT-en	89	9	**8**	**7**	**57.14**

表 7: The result of counterfactual robustness. ACC is the accuracy (%) of LLMs without external documents. ACC_{doc} is the accuracy (%) of LLMs with counterfactual documents. ED and ED* are error detection rates evaluated by exact matching and ChatGPT, respectively. CR is the error correction rate.

Evaluated four abilities of retrievalaugmented generation in LLMs:

- 噪声鲁棒性（Noise Robustness）
- 负面拒绝（Negative Rejection）
- 信息整合（Information Integration）
- 反事实鲁棒性（Counterfactual Robustness）

RetrievalAugmented Generation Benchmark (RGB):

- The instances of RGB are generated from latest news articles and the external documents obtained from search engines.
- Current LLMs have limitations in the 4 abilities. This indicates that there is still a significant amount of work needed to effectively apply RAG to LLMs.
- To ensure accurate and reliable responses from LLMs, it is crucial to exercise caution and carefully design for RAG.

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

- **开发专门模块**：设计专门用于检测和过滤噪声及虚假信息的模块，帮助 LLMs 更好地筛选有用信息，减少因误信不良信息而导致的错误
- **优化提示工程**：利用更精细的提示策略指导 LLMs 如何处理特定类型的任务，例如明确指示其在遇到不确定证据时应采取谨慎态度。

[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!