2024. LegalBench-RAG: A Domain-Specific Benchmark for Evaluating Retrieval in Legal RAG Systems. *arXiv preprint* (2024). arXiv:2408.10343 https://arxiv.org/abs/2408.10343

[31] Yang Wang and Hassan A Karimi. 2024. Exploring Large Language Models for Climate Forecasting. *arXiv preprint arXiv:2411.13724* (2024).

[32] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600* (2018).

[33] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

[34] Siyu Yuan, Jiangjie Chen, Ziquan Fu, Xuyang Ge, Soham Shah, Charles Jankowski, Yanghua Xiao, and Deqing Yang. 2023. Distilling Script Knowledge from Large Language Models for Constrained Language Planning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 4303–4325.

[35] Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, et al. 2025. FutureX: An Advanced Live Benchmark for LLM Agents in Future Prediction. *arXiv preprint arXiv:2508.11987* (2025).

[36] Yuxuan Zhang, Zhiyuan Zhang, Yicheng Wang, Yuxuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, Yixuan Su, and Yixuan Su. 2023. PubHealth: A Benchmark for Public Health Question Answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Toronto, Canada, 9802–9822. doi:10.18653/v1/2023.acl-long.546

[37] Siyun Zhao, Yuqing Yang, Zilong Wang, Zhiyuan He, Luna K Qiu, and Lili Qiu. 2024. Retrieval augmented generation (rag) and beyond: A comprehensive survey on how to make your llms use external data more wisely. *arXiv preprint arXiv:2409.14924* (2024).

[38] Chang Zong, Yuchen Yan, Weiming Lu, Jian Shao, Eliot Huang, Heng Chang, and Yueting Zhuang. 2024. Triad: A Framework Leveraging a Multi-Role LLM-based Agent to Solve Knowledge Base Question Answering. arXiv:2402.14320 [cs.CL] https://arxiv.org/abs/2402.14320

Appendix

Case Studies

To evaluate the proposed supportiveness metric in RAG, we consider two representative cases shown in Figure 6. The first question asks whether the total crypto market cap will exceed $4 trillion on January 20, 2025. Among the retrieved articles, the blue-highlighted news discusses substantial rebounds in Bitcoin and Ether alongside ETF approvals—information directly indicative of a “Yes” outcome. Conversely, the white-highlighted news describes global market declines and cryptocurrency crashes, contradicting the supportive stance. Our metric assigns higher scores to the supportive article, reflecting its stronger evidential link to the answer.

The second question concerns whether Joe Rogan will attend the presidential inauguration. The supportive article details Rogan’s past political activities and his collaboration with Donald Trump during the campaign, which strengthens the likelihood of his attendance. In contrast, the non-supportive article mentions Rogan only in the context of broader shifts in media influence, without providing any concrete statement or event suggesting his presence at the inauguration. Our metric correctly ranks the supportive article higher, aligning with the ground truth answer “Yes”.

Across both cases, the metric effectively distinguishes articles offering direct, answer-relevant evidence from those that are tangential or context-only, thereby enhancing RAG answer accuracy.

Construction Details

During constructing, we use `gpt-4o-mini-2024-07-18` for all LLM callings. We set window size w to 3 which is enough large in our pilot study. For computing each probability in CIL, we call twice `gpt-4o-mini-2024-07-18` and get the average score. The constructing prompts we use are shown in prompts (a-f).

Prompts

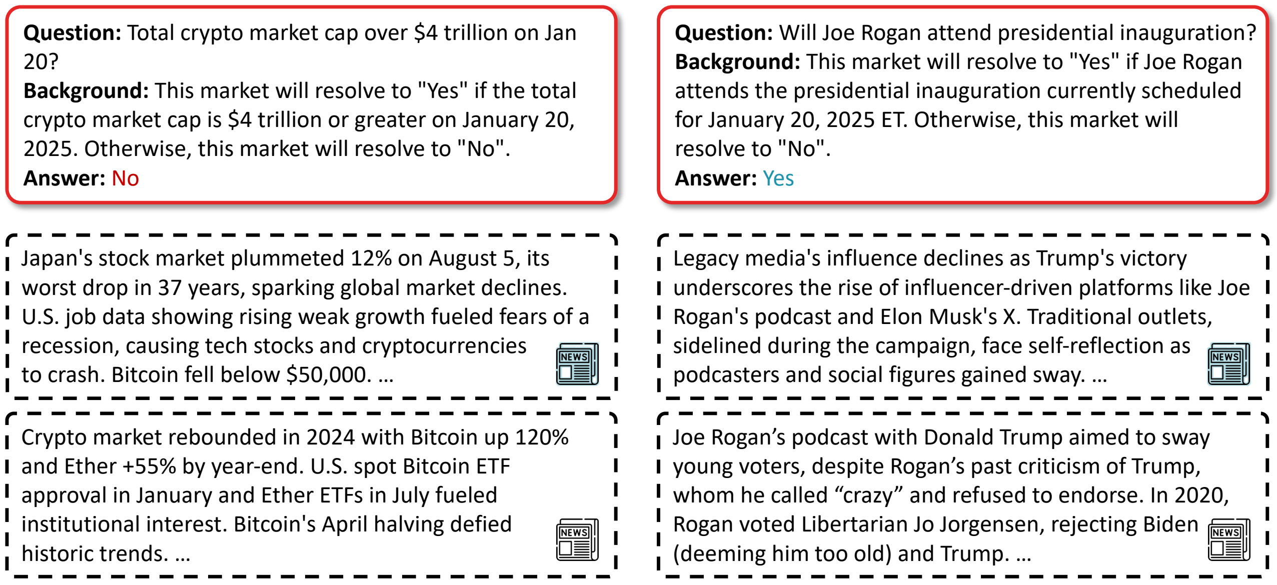
We list all prompts in the following Figures (a-h).

Agentic RAG Tool

We show the article retrieval tool for all agentic RAG methods:

```python
class AgenticRAGTool:
    name = "query_rag"
    description = (
        "To search in a previously built RAG index
             to find the most relevant chunks of
           text."
    )
    parameters = [
        {
            'name': 'query',
            'type': 'string',
            'description': 'the query text to
                search for relevant text chunks.',
            'required': True
        },
        {
            'name': 'top_k',
            'type': 'integer',
            'description': 'the number of top
                relevant chunks to retrieve (
                default is 3).',
            'required': False
        }
    ]
```

Listing 1: Python Class Definition for Agentic RAG Tool



Question: Total crypto market cap over $4 trillion on Jan 20?
Background: This market will resolve to "Yes" if the total crypto market cap is $4 trillion or greater on January 20, 2025. Otherwise, this market will resolve to "No".
Answer: No

Japan's stock market plummeted 12% on August 5, its worst drop in 37 years, sparking global market declines. U.S. job data showing rising weak growth fueled fears of a recession, causing tech stocks and cryptocurrencies to crash. Bitcoin fell below $50,000. ...

Crypto market rebounded in 2024 with Bitcoin up 120% and Ether +55% by year-end. U.S. spot Bitcoin ETF approval in January and Ether ETFs in July fueled institutional interest. Bitcoin's April halving defied historic trends. ...

Question: Will Joe Rogan attend presidential inauguration?
Background: This market will resolve to "Yes" if Joe Rogan attends the presidential inauguration currently scheduled for January 20, 2025 ET. Otherwise, this market will resolve to "No".
Answer: Yes

Legacy media's influence declines as Trump's victory underscores the rise of influencer-driven platforms like Joe Rogan's podcast and Elon Musk's X. Traditional outlets, sidelined during the campaign, face self-reflection as podcasters and social figures gained sway. ...

Joe Rogan's podcast with Donald Trump aimed to sway young voters, despite Rogan's past criticism of Trump, whom he called "crazy" and refused to endorse. In 2020, Rogan voted Libertarian Jo Jorgensen, rejecting Biden (deeming him too old) and Trump. ...

Figure 6: Case studies. The supportive and non-supportive articles are selected by `CIL` scores.

`(a) Entity Query Generation`

I will provide you with a forecasting question and the background information for the question. Extract the named entities, events of the question. Each entity and event are up to 5 words. The named entities can only be people, organizations, countries, locations while can not be date or time. Put all result items in a list that I can parse by JSON as ["entity 1", "entity 2", "event 1", "event 2", ...].
Question: *Q*
Question Background: *B*
Question Date: *D*
Output:

`(b) Resolving Steps Query Generation`

I will provide you with a forecasting question and the background information for the question. I will then ask you to generate short search queries (up to max words words each) that I'll use to find articles on Google News to help answer the question. The articles should be mainly about event arguments such as subjects, objects, locations, organizations of the events in question and background information. You must generate this exact amount of queries: num keywords. Put all result items in a list that I can parse by JSON as ["step 1", "step 2", "step 3", ...].
Question: *Q*
Question Background: *B*
Question Date: *D*
Output:

`(c) Similar Events Query Generation`

I will provide you with a forecasting question and the background information for the question. I will then ask you to generate short search queries (up to max words words each) that I'll use to find articles of similar events on Google News to help answer the question. The similar events are events happended on other similar entities in the history. Or events happended on question entities but on other date. You must generate this exact amount of queries: num keywords. Put all result items in a list that I can parse by JSON as ["event 1", "event 2", "event 3", ...].
Question: *Q*
Question Background: *B*
Question Date: *D*
Output: