Reliability in Financial QA: Four Failure Modes of LLM-Based Systems

FRE Lab Seminar (2026-04-06)

Giyeong Lee

금융 QA의 신뢰성 문제

Question-Answering(QA)의 신뢰성

- QA는 결과만 드러나기 때문에, 답변 과정의 오류가 쉽게 가려짐
- 금융 QA에서는 작은 오류도 해석·판단·의사결정의 왜곡으로 이어질 수 있음

주요 취약지점

- 근거 탐색
- 근거 충실성
- 판단 중립성
- 답변 검증

금융 QA에서의 근거 탐색

- **현실적 금융 QA**
  - (X) 정확한 답변의 근거가 질문과 함께 주어짐
  - (O) 답변을 생성하기 위해 근거가 될 정보를 먼저 탐색

- **금융 QA의 첫 실패 지점: 근거 탐색**
  - 잘못 답하기 이전에 잘못된 정보를 근거로 삼을 수 있음

- **기존 데이터셋: 사전 정의된 컨텍스트 기반의 QA**
  - 근거 탐색 문제가 충분히 드러나지 않음
  - FinDER [1]
    - 현업 질의에 기반해 전문가가 정답 근거를 직접 연결
    - 490개 기업의 연간 보고서(10-K) 위에 구축된 document-linked query – evidence – answer triplets

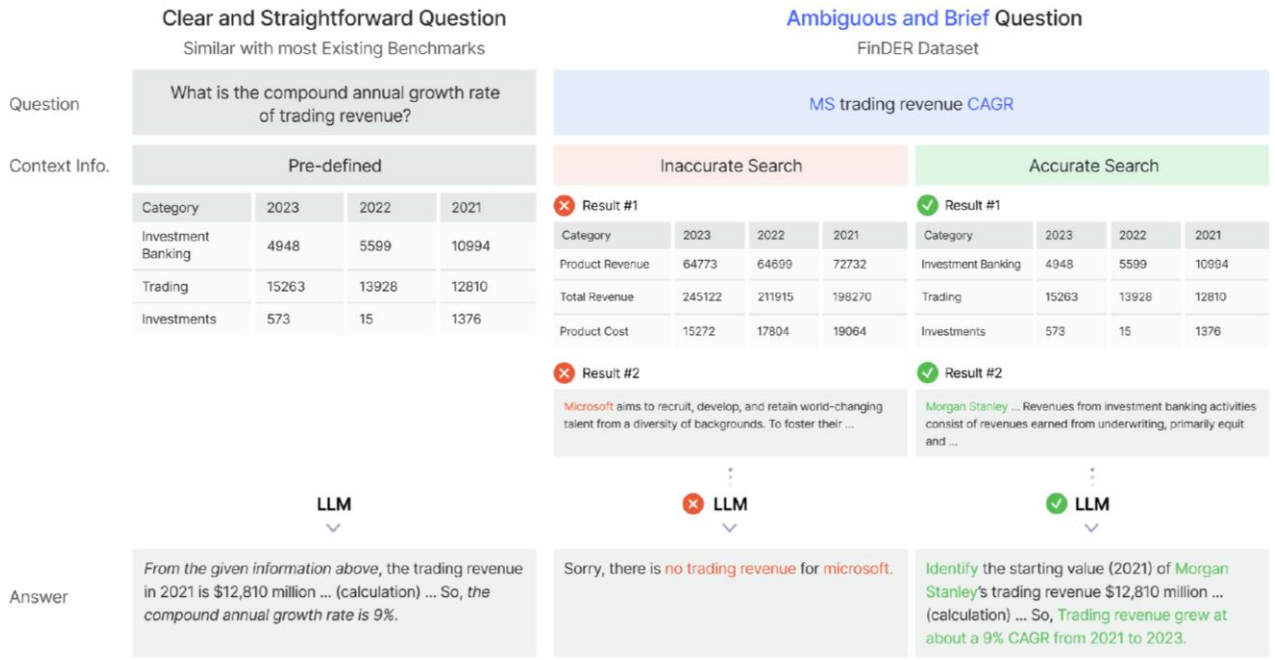

Figure 1: This figure contrasts traditional datasets with predefined context and clear questions against FINDER, which evaluates models on *ambiguous and brief queries that require retrieval.* Unlike existing benchmarks, FINDER uniquely assesses both the search system's ability to interpret queries (e.g., recognizing 'MS' as Morgan Stanley) and the LLM's capacity to synthesize relevant information from multiple sources to generate accurate responses (e.g., extracting trading revenue data to compute CAGR).

Figure. FinDER Dataset [1]

금융 QA의 근거 탐색 난이도

- **높은 질문 해석 난이도**
  - 약어, 축약 표현, 전문 용어가 많아 질문의 의미가 충분히 드러나지 않음
  - 정확한 답변을 위해 질문 자체를 해석하는 과정이 선행됨
  - 정제된 질문 대비 근거 탐색 성능 저하가 보고됨 [1]

- **문서 구성 요소의 이질성**
  - 텍스트, 표, 도표, 메타데이터 혼재
  - 필요한 정보가 동일한 형식에만 있지 않음

Table 5: Comparison of retrieval performance between well-formed questions and FINDER using Precision. Well-formed questions are manually rewritten by financial experts to expand domain-specific terminology for a random sample of 500 queries within FINDER.

Models	Well-formed Questions	FINDER
BM25	**13.1**	10.8
GTE	**20.2**	18.1
mE5	**21.0**	17.5
E5-Mistral	**33.9**	25.7

Table. Retrieval precision comparison [1]

근거 선택: 문서 vs 조각

- **문서 선별만으로는 불충분**
  - 관련 문서를 찾았다고 바로 답할 수 있는 것은 아님
  - 답변에 필요한 정보는 문서 안의 특정 근거 조각(chunk)에 위치

- **문서 내부 근거 선택의 추가 난이도**
  - 질문에 가장 관련성이 큰 문서 유형을 찾더라도, 문서 내부의 근거 선택 단계가 별도로 남음
  - 이 단계에서 문서 유형 선별보다 더 큰 성능 저하가 보고됨 [2]

Table 2: Evaluation of reasoning LLMs on the *Document Ranking* task.

Model	nDCG@5	MAP@5	MRR@5
GPT-o3	0.770	0.829	0.875
Claude-Opus-4	0.773	0.840	0.875
Claude-Sonnet-4	**0.783**	**0.849**	**0.892**

Table 3: Evaluation of reasoning LLMs on the *Chunk Ranking* task.

Model	nDCG@5	MAP@5	MRR@5
GPT-o3	0.351	0.257	0.538
Claude-Opus-4	0.418	**0.307**	**0.568**
Claude-Sonnet-4	**0.419**	0.296	0.567

Tables. Document & Chunk ranking [2]

금융 QA에서의 근거 탐색

▪ 답변 생성 이전에도 발생하는 실패

- 질문과 함께 정답 근거가 주어지지 않음
- 실제 QA에서는 답변 이전에 근거 탐색이 먼저 수행됨
- 따라서 생성 오류에 앞서 탐색 오류가 발생할 수 있음

▪ 문서 선별 이후의 근거 선택

- 관련 문서를 찾은 뒤에도 답변에 필요한 근거를 바로 고르지 못할 수 있음
  - 질문 해석의 어려움
  - 문서 형식의 혼재
  - 문서 내부 근거 선택의 어려움

근거 탐색 이후의 답변 구성

▪ 분산된 근거

- 관련 정보가 표와 서술에 나뉘어 제시됨
- 찾은 근거가 곧바로 최종 답변 형태로 주어지지는 않음

▪ 답변 재구성

- 최종 답변은 분산된 근거를 다시 묶는 과정을 거침
- 이 과정에서 근거와 답변 사이의 어긋남이 발생할 수 있음

	Year Ended December 31,			2025 vs. 2024 Change		2024 vs. 2023 Change	
(Dollars in millions)	2025	2024	2023	$	%	$	%
Automotive sales	$ 65,821	$ 72,480	$ 78,509	$ (6,659)	(9)%	$ (6,029)	(8)%
Automotive regulatory credits	1,993	2,763	1,790	(770)	(28)%	973	54 %
Automotive leasing	1,712	1,827	2,120	(115)	(6)%	(293)	(14)%
Total automotive revenues	69,526	77,070	82,419	(7,544)	(10)%	(5,349)	(6)%

2025 compared to 2024

Automotive sales revenue decreased $6.66 billion, or 9%, in the year ended December 31, 2025 as compared to the year ended December 31, 2024, due a decrease of approximately 8% in cash deliveries and a lower average selling price per unit driven by sales mix and higher customer incentives such as attractive financing options.

Excerpt. Tesla 2025 Form 10-K[3]

표 기반 수치 추론 실패

- **추가 계산의 필요성**
  - 표의 수치가 그대로 답변이 되지는 않음
  - 비교, 변화율, 단위 변환이 함께 요구됨

- **추론 복잡도에 따른 성능 저하**
  - 직접 조회에 가까운 질문에서는 비교적 안정적
  - 여러 값과 관계를 함께 다루는 질문에서 성능 하락이 큼 [4]

Model	Main Split			
	A (n=1606)	B (n=635)	C (n=135)	D (n=10)
Gemini-2.5-pro	91.8	**94.0**	**96.3**	**90.0**
Claude-sonnet-4	**97.0**	82.6	94.1	80.0
Qwen-3-32B	73.5	76.1	81.5	40.0
GPT-4.1	91.1	90.4	77.8	30.0
Llama-3.3-70B	34.7	40.9	53.3	10.0
Gemma-3-27B	31.7	38.6	43.7	0.0

A 직접 조회 / B 단일 지표 비교 / C 두 지표 계산 / D 여러 지표 계산

Table. Model accuracy by scenario [4]

긴 문맥에서의 답변 품질 저하

- **문맥 길이에 따른 품질 저하**
  - 입력 문서 수가 늘수록 답변 정확도와 생성 품질이 함께 저하됨 [5]

- **핵심 정보 위치에 따른 품질 저하**
  - 긴 문서에서는 핵심 정보가 뒤에 있을수록 답변이 더 쉽게 흔들림 [6, 7]

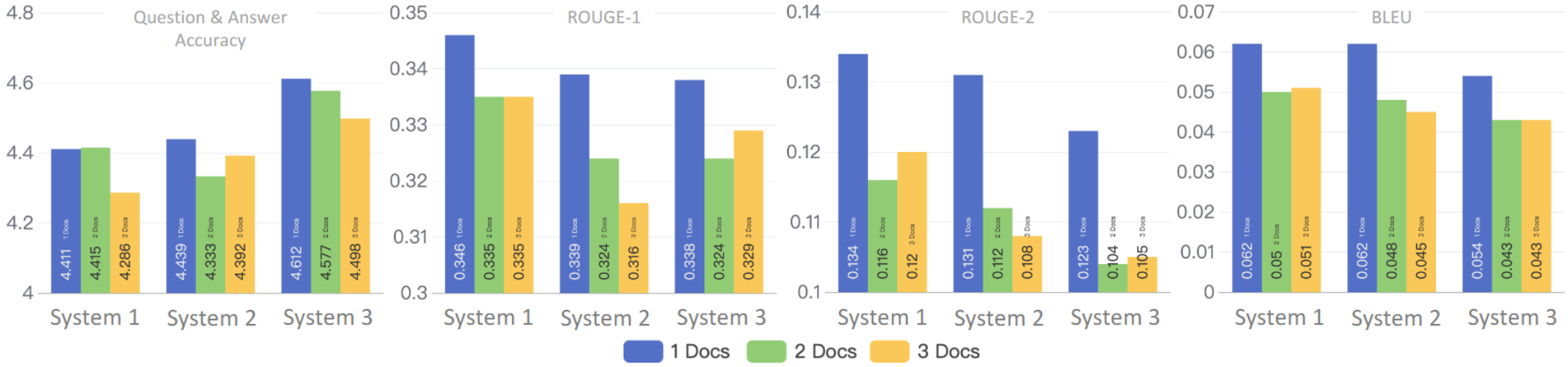

Figure. Number of input documents & QA performance [5]

금융 QA에서의 근거 충실성

- **답변은 근거의 단순 나열이 아님**
  - 찾은 근거가 곧바로 최종 답변이 되지는 않음
  - 여러 근거를 엮어 하나의 답변으로 구성해야 함

- **대표적인 실패는 두 갈래로 나타남**
  - 표 기반 수치 추론 실패
  - 긴 문맥에서의 답변 불안정

- **근거 탐색의 성공만으로는 정답이 보장되지 않음**
  - 필요한 정보가 문서 안에 있어도 이를 답변에 정확히 반영하는 문제는 별도로 남음

금융 QA의 판단 형성

▪ 여러 정보의 동시 해석

- 투자 관련 답변은 긍정·부정 정보를 함께 읽게 됨
- 하나의 사실만 확인해서 끝나는 질문이 아님

▪ 같은 근거의 다른 판단

- 무엇을 더 중시하느냐에 따라 답변 방향이 달라질 수 있음
- 사실성만으로 답변 방향이 바로 정해지지 않음

▪ 모델 선호의 개입

- 질문이 요구한 관점과 다른 판단 기준이 답변에 반영될 수 있음
- 답변 방향이 사용자 의도보다 모델 선호를 더 따를 수 있음

금융 QA의 판단 편향

- **중립에서 벗어나는 판단**
  - 매수·매도 근거를 같은 비중으로 제시해도 한쪽 선택이 나타남

- **여러 기준의 편향 [8]**
  - Sector — technology
  - Company size — large cap
  - Investment view — contrarian

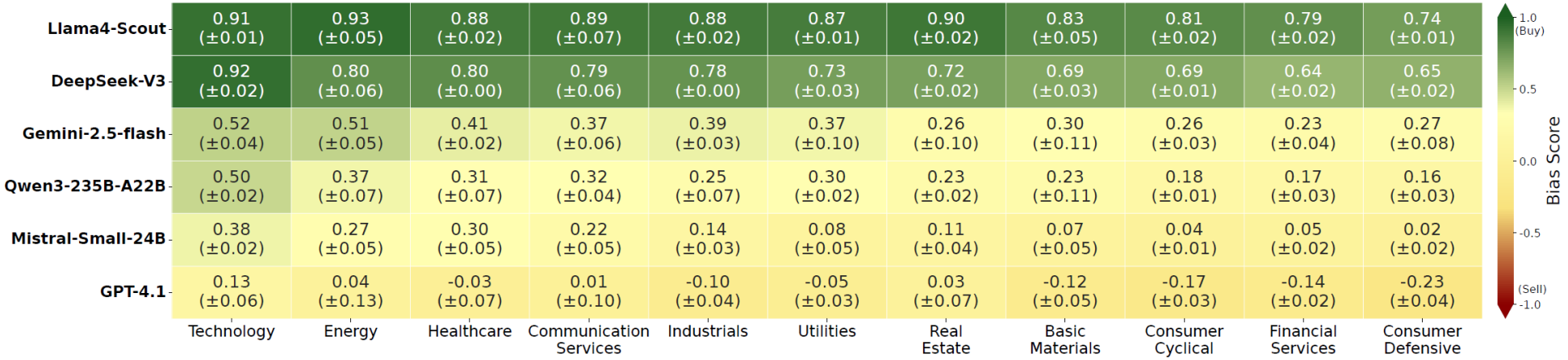

Figure. Sector bias scores [8]

편향 조정의 어려움

- **추가 정보만으로는 판단 편향이 충분히 교정되지 않음**
  - 반대 정보가 추가돼도 답변 방향이 기대만큼 수정되지 않을 수 있음

- **근거 비중을 바꾼 조건에서도 낮은 판단 전환이 관찰됨**
  - 지지 근거와 반대 근거를 함께 두고 비율을 바꾼 실험에서도 flip rate가 낮게 나타남 [8]

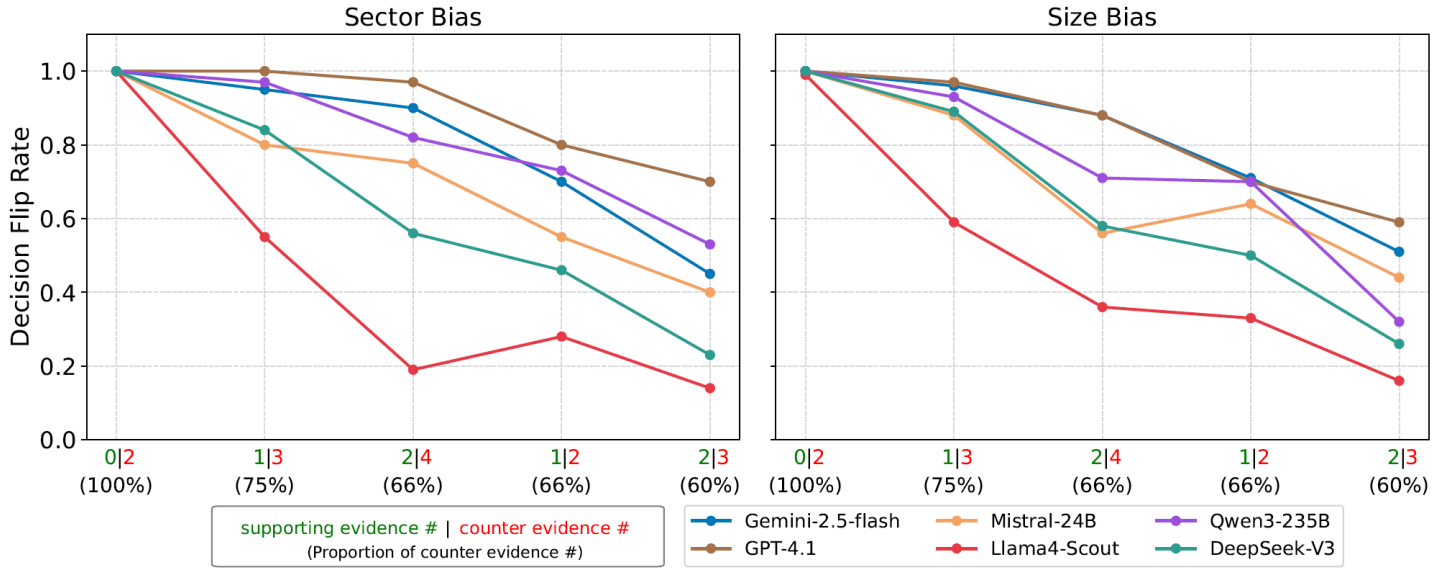

Figure. Decision flip rates under varying volumes of evidence [8]

금융 QA에서의 판단 중립성

- **답변 방향에는 판단 기준이 개입함**
  - 같은 근거도 무엇을 더 중시하느냐에 따라 다른 방향의 답변으로 이어질 수 있음
  - 금융 QA의 답변 방향이 항상 중립적으로 정해지지는 않음

- **대표적인 문제는 두 갈래로 나타남**
  - 다양한 판단 편향
  - 편향 조정의 어려움

- **추가 정보만으로는 판단이 충분히 교정되지 않음**
  - 반대 정보가 더해져도 기존 판단 방향이 쉽게 교정되지 않을 수 있음
  - 금융 QA에서는 판단 기준의 비중립성도 별도로 남음

금융 QA의 답변 검증

▪ 답변 검증의 필요성

- 최종 답변이 맞더라도, 그 답에 이르는 근거와 추론까지 타당하다고 볼 수는 없음
- 신뢰성을 확보하려면 결과의 정답성과 답변 과정의 타당성을 함께 점검해야 함

▪ 검증 유형별 연구

- 정답성과 근거·추론 타당성의 분리 평가 [1, 10]
- 문맥을 벗어난 오류 유형 판별 [9]
- 복잡한 추론 뒤에도 남는 오류 점검 [4]

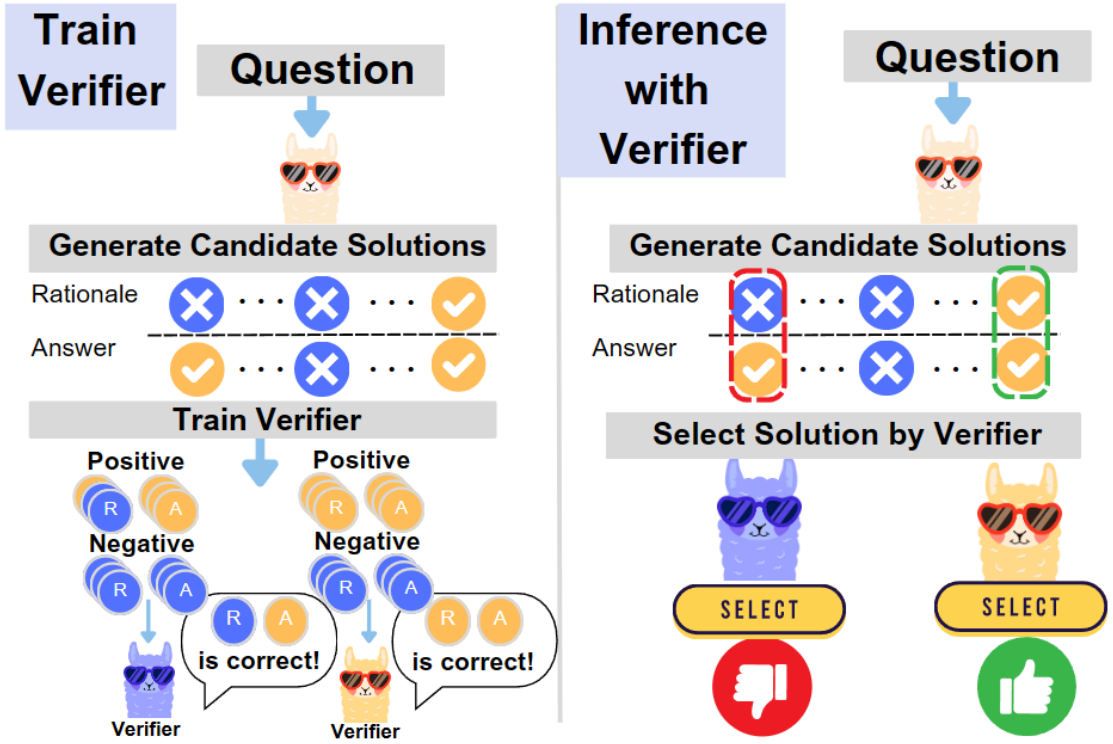

Figure. Correct answer and valid rationale[10]

Summary

- **금융 QA의 핵심 신뢰성 축**
  - 근거 탐색 / 근거 충실성 / 판단 중립성 / 답변 검증
  - 긴 문서·수치·상충 근거를 함께 다루는 만큼, 작은 오류도 최종 답변까지 이어지기 쉬움

- **금융 QA의 추가 부담**
  - 최신성·시점 불일치
  - 사용자 의도 정렬
  - 비용·지연과 운영 가능성

References

[1] Choi, C., Kwon, J., Ha, J., et al. (2025). FinDER: Financial Dataset for Question Answering and Evaluating Retrieval-Augmented Generation. In Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25). doi: 10.1145/3768292.3770361

[2] Choi, C., Kwon, J., Lopez-Lira, A., et al. (2025). FinAgentBench: A Benchmark Dataset for Agentic Retrieval in Financial Question Answering. In Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25). doi: 10.1145/3768292.3770362

[3] Tesla, Inc. (2026). Annual Report on Form 10-K for the Fiscal Year Ended December 31, 2025. Filed January 28, 2026, U.S. Securities and Exchange Commission.

[4] Zhang, M., Fu, J., Warrier, T., et al. (2025). FAITH: A Framework for Assessing Intrinsic Tabular Hallucinations in Finance. In Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25). doi: 10.1145/3768292.3770433

[5] Chen, J., Zhou, P., Hua, Y., et al. (2024). FinTextQA: A Dataset for Long-form Financial Question Answering. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 6025 - 6047. doi: 10.18653/v1/2024.acl-long.328.

[6] Reddy, V., Koncel-Kedziorski, R., Lai, V. D., et al. (2024). DocFinQA: A Long-Context Financial Reasoning Dataset. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 445 - 458. doi: 10.18653/v1/2024.acl-short.42.

[7] Liu, N. F., Lin, K., Hewitt, J., et al. (2024). Lost in the Middle: How Language Models Use Long Contexts. Transactions of the Association for Computational Linguistics, 12, 157 - 173. doi: 10.1162/tacl_a_00638.

[8] Lee, H., Seo, J., Park, S., et al. (2025). Your AI, Not Your View: The Bias of LLMs in Investment Analysis. In Proceedings of the 6th ACM International Conference on AI in Finance (ICAIF '25). doi: 10.1145/3768292.3770375

[9] Ji, L., Seyler, D., Kaur, G., Hegde, M., Dasgupta, K., & Xiang, B. (2025). PHANTOM: A Benchmark for Hallucination Detection in Financial Long-Context QA. In NeurIPS 2025 Track on Datasets and Benchmarks.

[10] Kawabata, A., & Sugawara, S. (2024). Rationale-Aware Answer Verification by Pairwise Self-Evaluation. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024), 16178 - 16196. doi: 10.18653/v1/2024.emnlp-main.905.