Seminar on High-Performance Data Analytics

LLM Trustworthiness and Fact Validation

Ashutosh Kumar Jaiswal

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Introduction

What are Large-Language Models (LLMs)?

Massive neural networks trained on vast amount of text (web pages, books, code)
Learn statistical patterns of language to predict the next word (token)

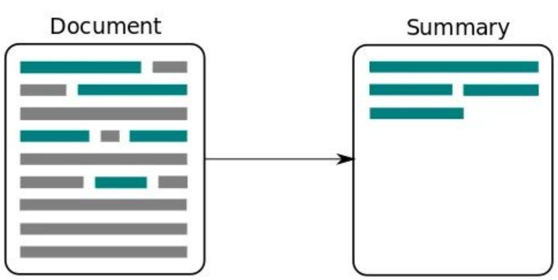
Source: https://productmindset.substack.com/p/prompt-engineering-explained

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Introduction

What can LLMs do?

Understand and Generate human-like Text, Code, etc.
Implied capabilities: Summarization, Translation, Answering and Reasoning



Source: https://medium.com/@thakermadhav/comparing-text-summarization-techniques-d1e2e465584e

Problem and Context

“Can We Trust the Machine?” - Why Trustworthiness Matters?

- LLMs are great but they can still have problems
  - Non-expert users can't tell if an answer is accurate or not (Closed-Book)
  - LLMs also make stuff up due to their token-by-token generative nature
  - Source of information is not available always, unlike Search Engines

LLMs need to be trustworthy because they are being employed in high-stakes environments like Law and Finance



Problem and Context

What "Trustworthiness" Means

The degree to which an LLM produces accurate, verifiable, and context-appropriate outputs while transparently signaling its own confidence and adhering to ethical constraints.

Factual Accuracy
Output aligns with authoritative sources

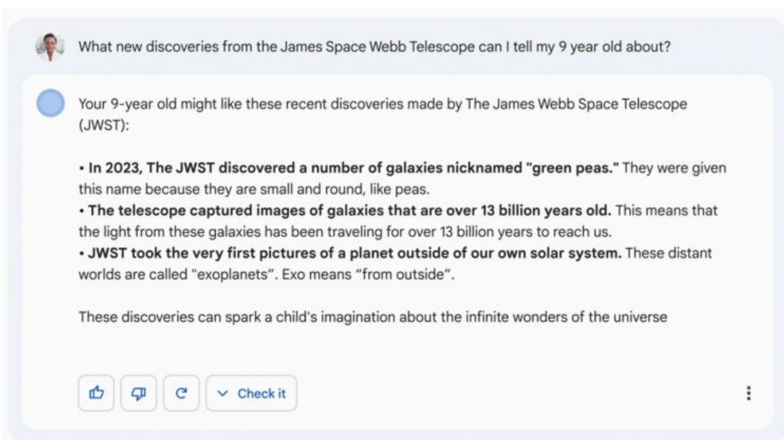
Evidence Transparency
Citations or snippets that back each claim

Consistency
Same answer across paraphrases; Robust

Problem and Context

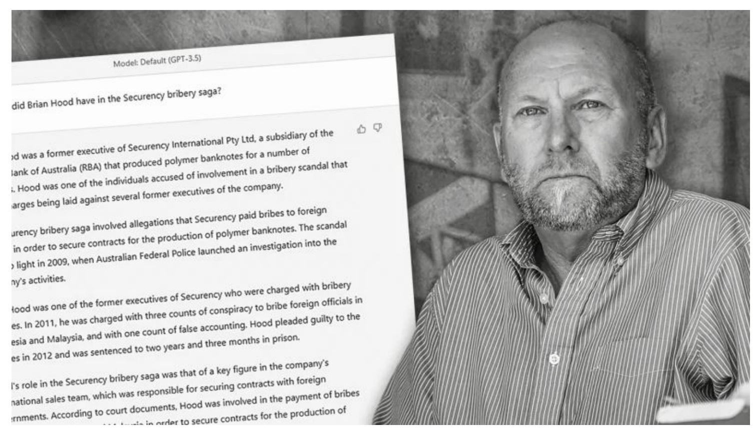
Real-life Examples of LLM Hallucination

Feb 8 2023: Google Bard's very first public demo wrongly claimed JWST took the first exoplanet photo, wiping ≈$100 B off Alphabet's market cap. (theverge.com, reuters.com)



Bard's very first answer contained a factual flub. Image: Google

Apr 2023: Australian mayor threatened the first defamation suit against OpenAI after ChatGPT falsely stated he had been jailed for bribery. (reuters.com)

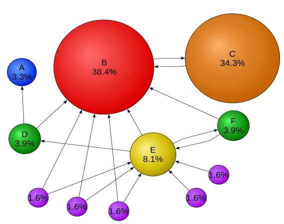


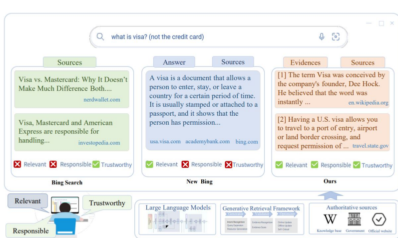
https://www.smh.com.au/technology/australian-whistleblower-to-test-whether-chatgpt-can-be-sued-for-lying-20230405-p5cy9b.html

Mitigation Strategies - An Overview

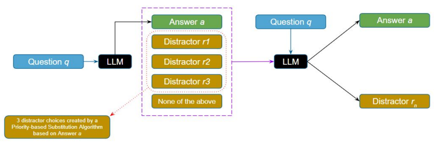
The following three concepts are explored in further detail:

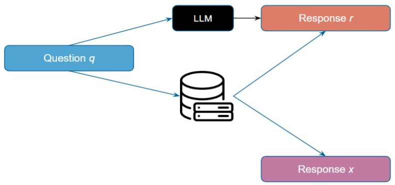
Fact-checking with Evidence



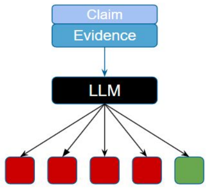


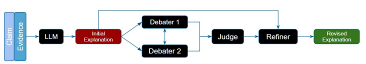
Self-Verification





Multi-step Refinement

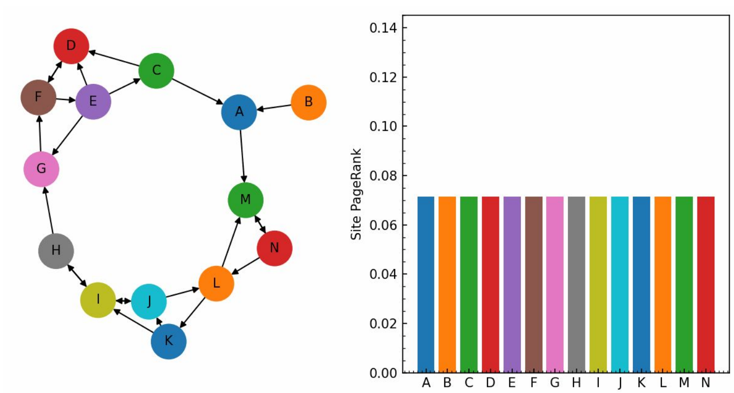




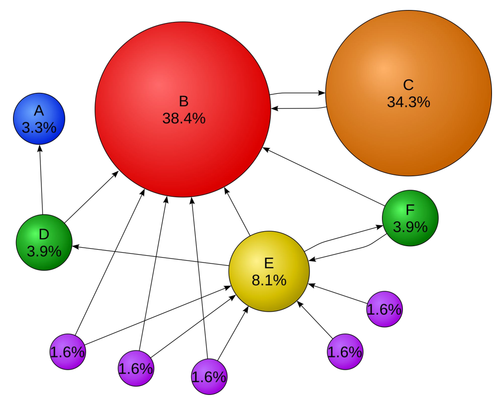
Fact-checking with Evidence

Know Where to Go: Making LLMs Relevant, Responsible & Trustworthy Searchers *by Xiang Shi · Jiawei Liu · Yinpeng Liu · Qikai Cheng · Wei Lu (2024)*

- A framework analogous to Google's PageRank algorithm for LLMs.
- The framework considers **Evidence Quality and Site Authority** to give sources a score, higher score implies that the source is reliable.



By Sage santo - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=153695322



By en:User:345Kai https://commons.wikimedia.org/w/index.php?curid=3470389

Fact-checking with Evidence

Know Where to Go: Motivation

- Complex/Vague queries may never find the right page
- LLMs hallucinate and/or quote the wrong / irrelevant source
- Our requirement: **Relevant + Reliable**

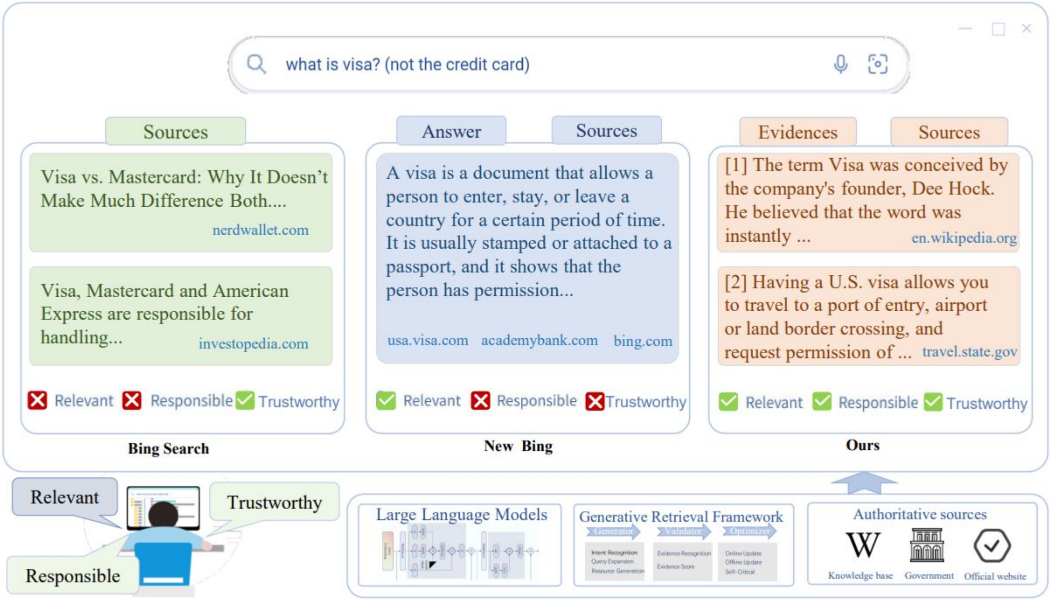


Image from [1]

Fact-checking with Evidence

Know Where to Go: Proposed Solution

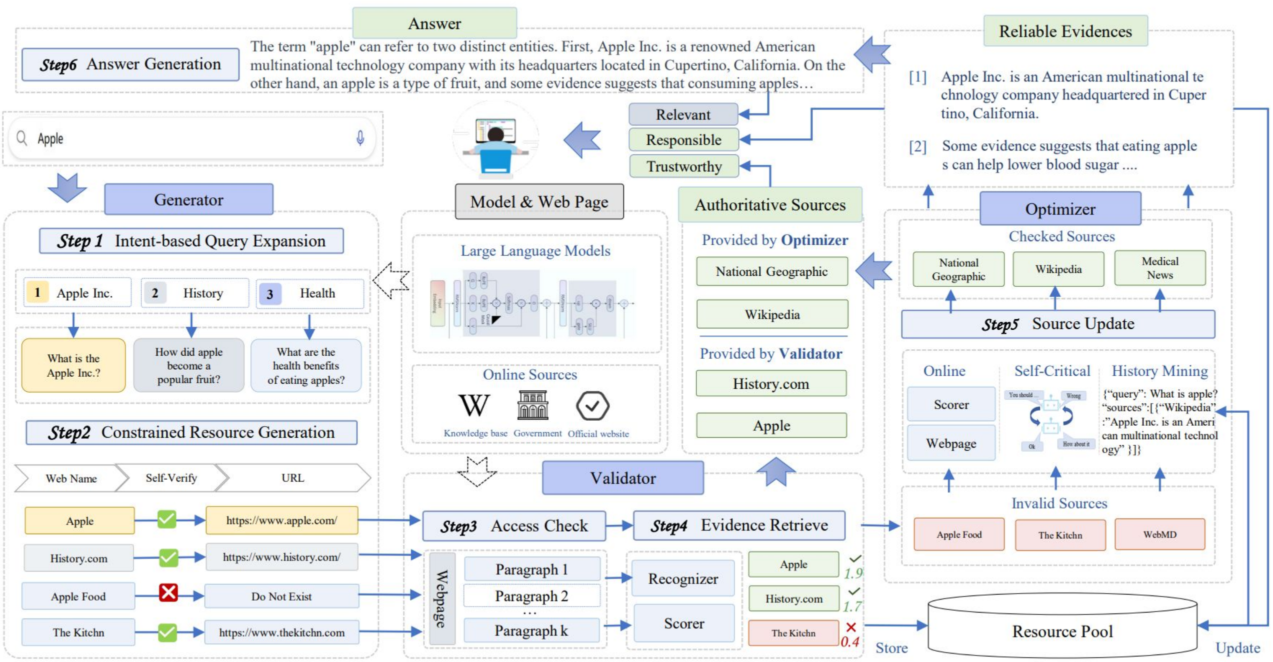


Image from [1]

Fact-checking with Evidence

Know Where to Go: Proposed Solution (Simplified)

Generator (The Writer)

Understands intent:
Expands the query into several clearer sub-queries (e.g., “Apple Inc.” vs “apple nutrition”).

Two-step source guess:
(1) List likely site names → (2) Map each name to its root URL.

Self-check loop:
Drops any site/URL that doesn’t actually load → boosts live-link rate from **≈ 39 % to ≈ 71 %**.

Validator (Fact-Checker)

***Open & ping*:**
Confirms every URL is accessible and up-to-date.

Find the proof:
Slides a window over the page text and has the LLM: (1) **Recognize** candidate answer sentences → (2) **Score** them for relevance/confidence.

Keeps only high-score sentences as explicit evidence.

Optimizer (The Fixer)

Self-critical*:* if Validator flags a dud link, swap it out instantly.

Online*:* run a real search engine with the refined queries to harvest fresh sources.

History mining*:* reuse proven links from past, similar questions in the *Source Pool*.

Iterates until every claim is backed by solid evidence.

Fact-checking with Evidence

Know Where to Go: Walkthrough

User question: “What are the health benefits of eating apples?”

1. **Generator**
    - Intents: *nutrition*, *apple products*, *agriculture*.
    - Expanded query for nutrition: “health benefits of eating apples”.
    - Site suggestions: *Healthline*, *WebMD*.
      Self-check confirms the domains healthline.com, webmd.com.
2. **Validator**
    - Opens both sites, finds articles.
    - Extracts sentences like “Apples are high in fiber, which supports gut health.”
    - Scores those sentences as good evidence.
3. **Optimizer**
    - Notices no issues, maybe still adds *Mayo Clinic* found via Bing search as an extra source.
    - Stores (query, sources, evidence) in the Source Pool for next time.

Outcome: The final answer the user sees cites real, accessible websites with specific sentences, and the system has already learned something useful for future “apple health” questions.

Self-Verification

TrustScore: Reference-Free Evaluation of LLM Response Trustworthiness

by Danna Zheng · Danyang Liu · Mirella Lapata · Jeff Z. Pan (2024)

- A method that checks an LLM's confidence in it's response to a certain query **(Behavioural Consistency)**
- If a model chooses the same response in the presence of incorrect response choices, it is likely that the response aligns with the model's parametric knowledge.

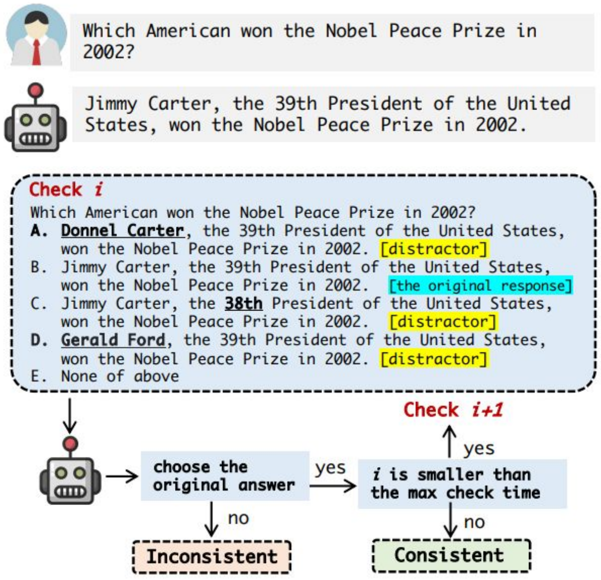


Image from [2]

Self-Verification

TrustScore: Motivation

- LLMs are very **convincing by design** and are also **prone to hallucinations**.
- Traditional Fact-Checking methods require **external DBs** (might be missing or out-of-date)
- To rely on responses generated in this **"Closed-book"** setting, we would like to know if the LLM response is consistent with it's own parametric knowledge

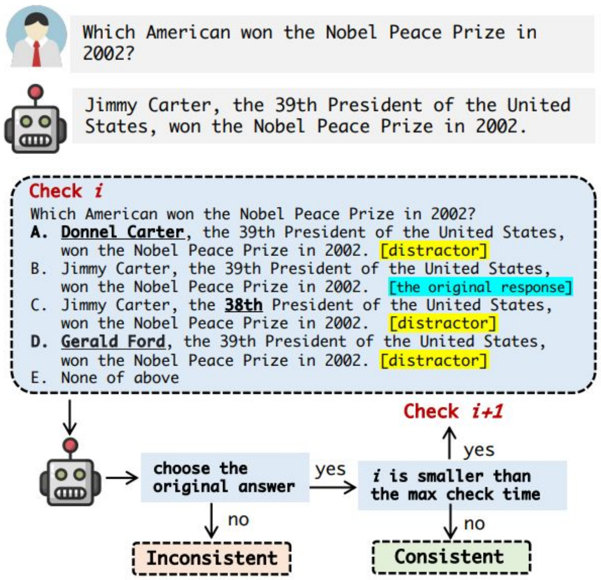
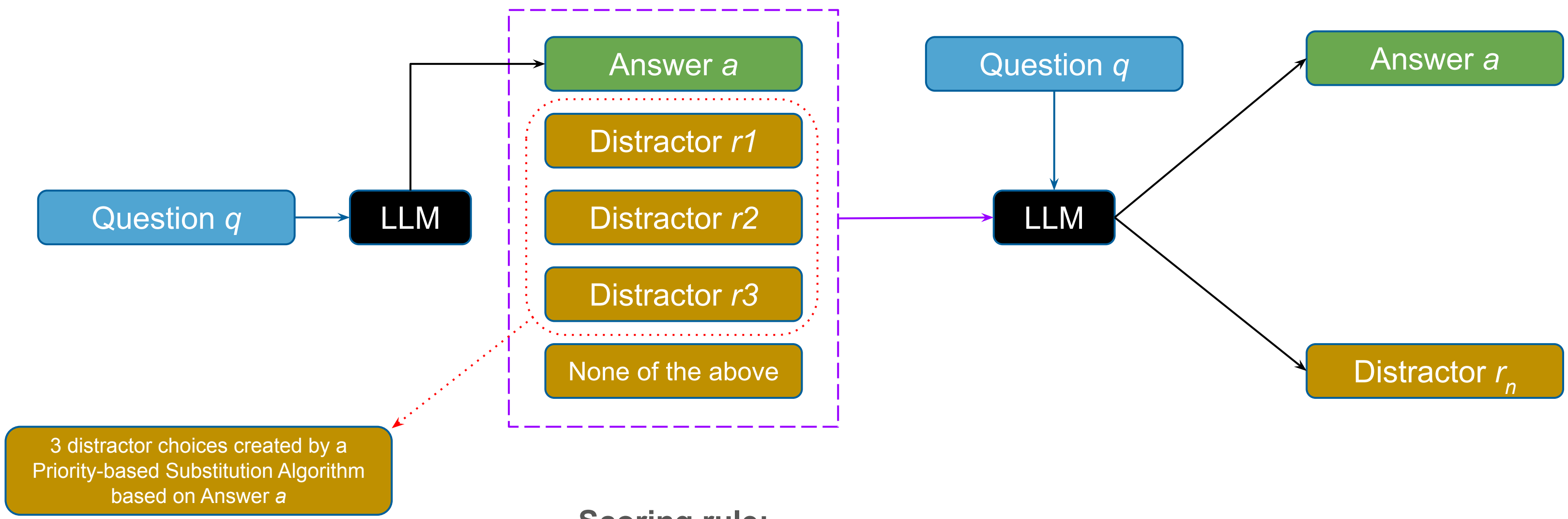


Image from [2]

Self-Verification

TrustScore: Proposed Solution (Behavioral Consistency)



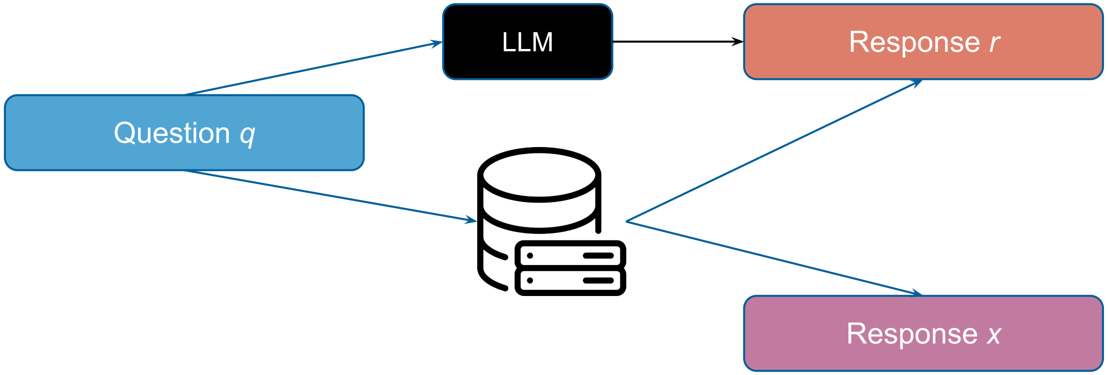
Scoring rule:

LLM chooses "Answer *a*" n times → Trust_{BC} = 1 (consistent)

LLM fails to pick "Answer *a*" even once → Trust_{BC} = 0 (inconsistent)

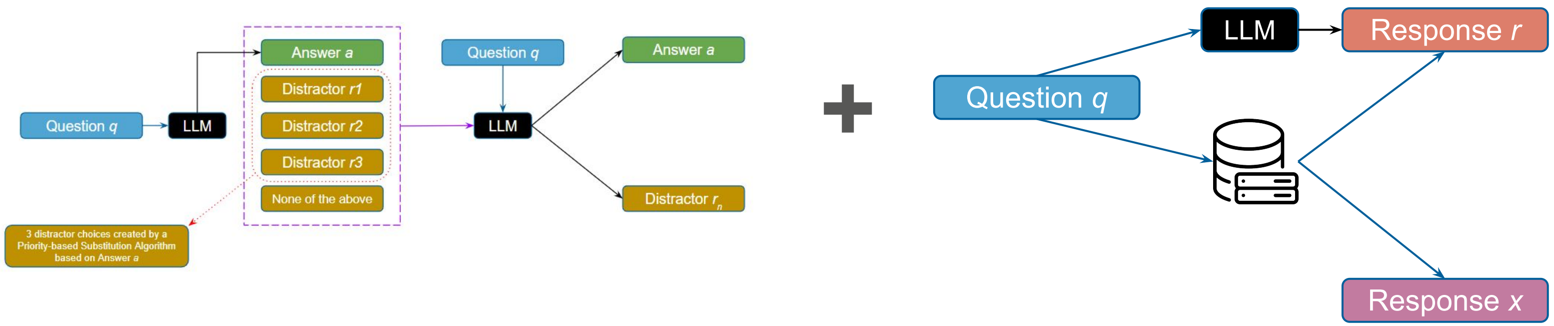
Self-Verification

TrustScore: Proposed Solution (Fact-Checking)



Self-Verification

TrustScore: Proposed Solution (Trust$_{ov}$)



Trust$_{ov}$ Score:

1. If Knowledge Base contradicts answer → low overall score, no matter BC
2. If Knowledge Base supports answer → high score, BC can boost further
3. If no Knowledge Base evidence → rely on BC alone

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

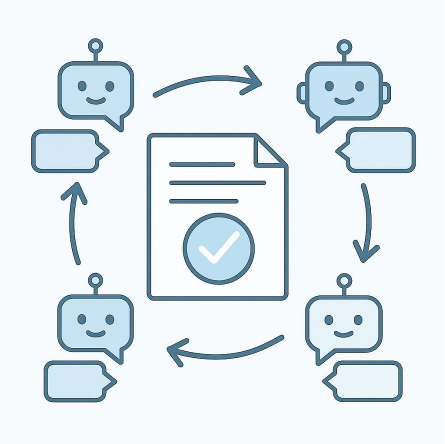
Multi-step Refinement

Can LLMs Produce Faithful Explanations For Fact-checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate

by Kyungha Kim · Sangyun Lee · Kung-Hsiang Huang · Hou Pong Chan · Heng Ji (2024)

- LLMs are good at checking facts but they struggle with **explaining** their verdicts.

- The paper proposes a framework that leverages **multiple AI-agents that debate and refine** each other's responses which reduces hallucinations and keeps the explanations **closely linked with the evidence** present.



Generated using ChatGPT

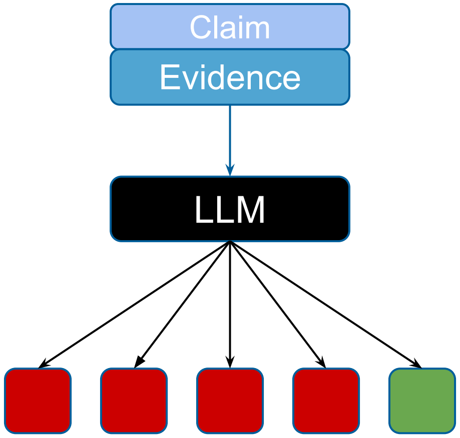
Multi-step Refinement

Multi-Agent Debate Refinement (MADR): Motivation

- In the age of misinformation, people accept a fact-check only when they understand why a claim is true or false.

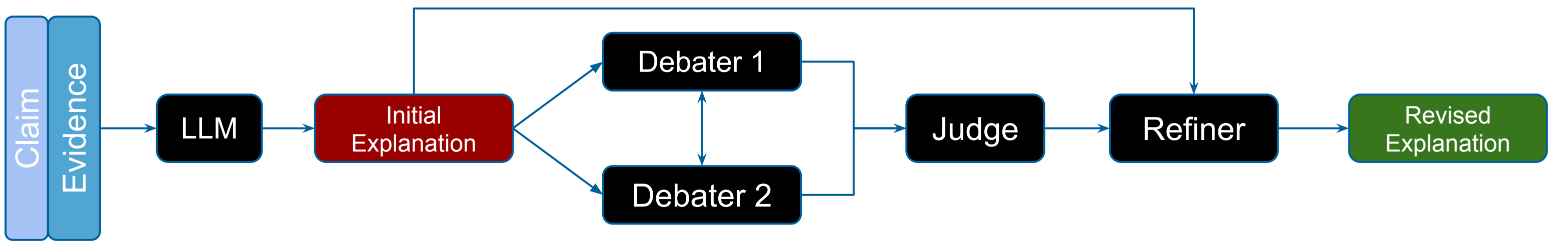
- Zero-Shot Prompting an LLM produces hallucinated responses 80% of the time.

- Multi-hop claims are hard. Explaining a verdict often means connecting several pieces of evidence; one slip can flip the story.



Multi-step Refinement

Multi-Agent Debate Refinement (MADR): Proposed Solution



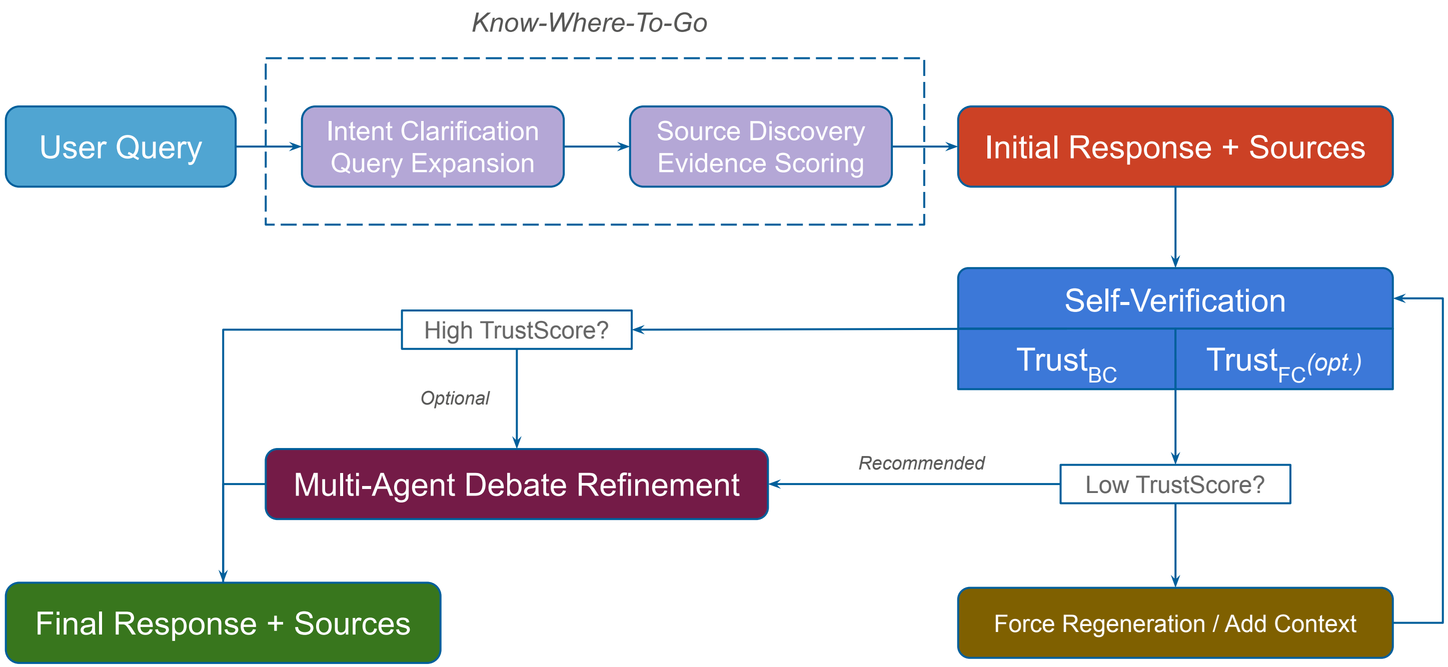
Debater 1: Finds errors guided by a predefined error typology

Debater 2: Finds errors freely

Judge: Checks whether both Debaters now agree

Refiner: Rewrites the Initial Explanation with the agreed feedback

Conceptual Pipeline



Conceptual Pipeline

- **Components' Strengths**
  - Retrieval section prevents external hallucinations (wrong facts).
  - TrustScore flags parametric inconsistencies.
  - MADR converts a bare verdict into a human-readable chain-of-thought.

- **Efficiency knobs**
  - Cheap "ping & page-rank" retrieval runs first (cache high-authority sources)
  - Costlier debate step activates only if TrustScore < threshold.
  - Debaters can be parallelized.

Key Takeaways

Layered Defences > Single-Shot Answers

Evidence retrieval (Know-Where-To-Go), Self-Verification (TrustScore), and Multi-Agent Debate (MADR) tackle different failure modes, so together they reduce hallucinations far more than any one method alone.

Be efficient, light checks first, heavy checks when needed

Run Fast retrieval + TrustScore by default and trigger the costlier MADR loop only when confidence is low. Result is reliability and efficiency.

Trust demands visible reasoning, not just a verdict.

MADR turns raw LLM outputs into transparent, multi-hop explanations users can check for themselves which is important for high-stakes domains like law, finance, medicine, etc.

Optional Slides

Fact-checking with Evidence

Know Where to Go: Proposed Solution

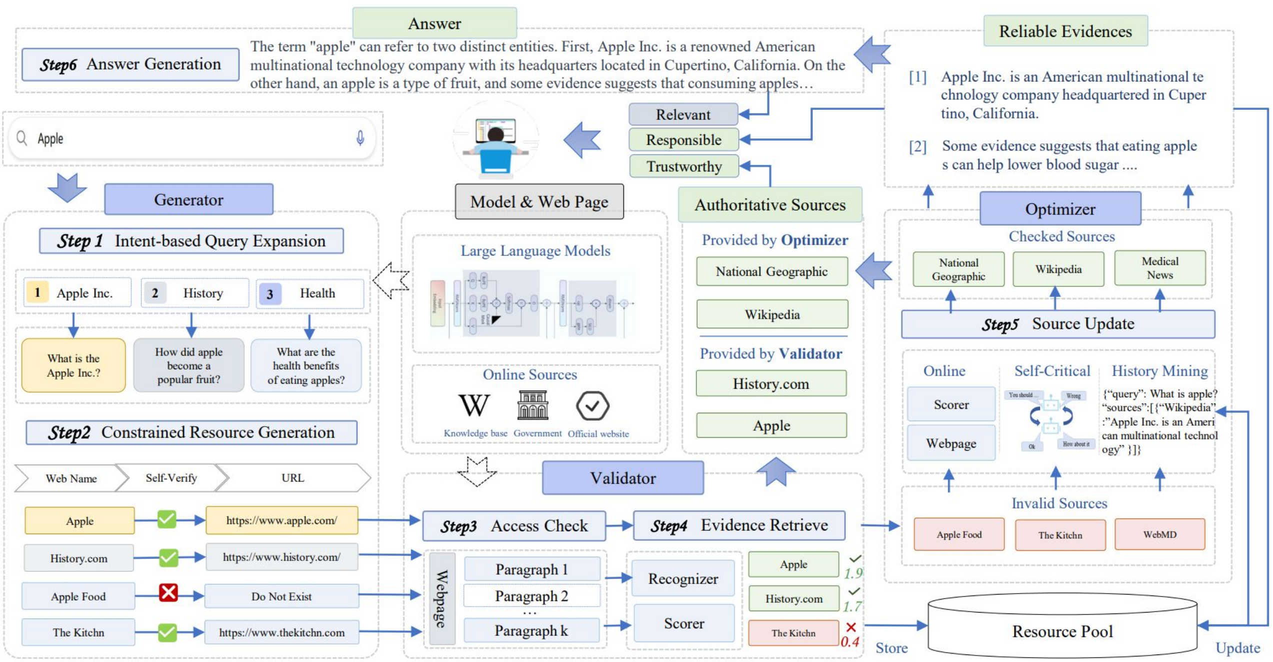


Image from [1]

Fact-checking with Evidence

Know Where to Go: Performance Evaluation

System	Statistical Metrics				Performance Metrics				
	S_{count}	E_{count}	$Q_{correct}$	T_{avg}	Timeliness ↑	Access ↑	Consistency ↑	Validity ↑	Precision ↑
New Bing	903 (460)	427	28	1.59	97.56 (99.57)	97.56 (99.57)	97.56 (99.57)	73.53	72.83
Perplexity.ai	**1595 (1038)**	1107	42	**2.42**	99.37 (99.33)	99.24 (99.23)	99.31 (99.23)	73.35	67.57
WebGPT (175B)	950 (505)	950	154	1.68	97.15 (96.83)	96.94 (96.44)	96.73 (96.23)	84.63	77.36
WebGLM (10B)	1355 (513)	**1355**	152	1.76	99.55 (99.81)	99.40 **(99.81)**	**99.40 (99.81)**	85.38	74.83
Our Method (7B)	295 (173)	565	**178***	1.29	**100.00**	**99.81**(99.66)	97.21(95.62)	**87.92***	**78.41***

Scount: How many distinct web sources were returned.
Ecount: How many evidence sentences were extracted.
Qcorrect: For how many questions did they get any correct answer.
Tavg: Avg. number of topics hit per query (diversity).

Timeliness: Does that web page still exist under the same name?
Access: Is the URL alive and reachable?
Consistency: Does the URL actually belong to the claimed site?
Validity: Does the page genuinely answer the question?
Precision: Of the evidence sentences pulled out, how many are truly on-point?

Image from [1]

Fact-checking with Evidence

Know Where to Go: Module-level Ablation Study

	Timeliness ↑	Access ↑	Consistency ↑	Validity ↑	Precision ↑
Full	**100.00**	**99.81(99.66)**	97.21 **(95.62)**	87.92	**78.41**
w/o opt.	**100.00**	99.51 (98.77)	**98.05** (95.06)	**88.76**	67.56
w/o val.	96.94 (94.24)	89.44 (82.98)	85.83 (76.44)	58.89	–

Image from [1]

Scount: How many distinct web sources were returned.
Ecount: How many evidence sentences were extracted.
Qcorrect: For how many questions did they get any correct answer.
Tavg: Avg. number of topics hit per query (diversity).

Timeliness: Does that web page still exist under the same name?
Access: Is the URL alive and reachable?
Consistency: Does the URL actually belong to the claimed site?
Validity: Does the page genuinely answer the question?
Precision: Of the evidence sentences pulled out, how many are truly on-point?

GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Self-Verification

TrustScore: Generating High-Quality Distractors

Three distractors built by a **Priority-based Substitution Algorithm:**

- Swap the most informative tokens first (entities > nouns/numbers > others)
- Pull replacements from:
  - DBpedia entities (preferred)
  - Semantically close words (embedding)
  - Random words with matching Part-Of-Speech

When did all night long come out lionel richie?
A) All night long came out in 1975. [distractor]
B) All night long came out in 1986. [distractor]
C) All night long came out in 1983. [original response]
D) All night long came out in 1999. [distractor]
E) None of the above.

Example taken from [2]

Multi-step Refinement

Multi-Agent Debate Refinement (MADR): Overview

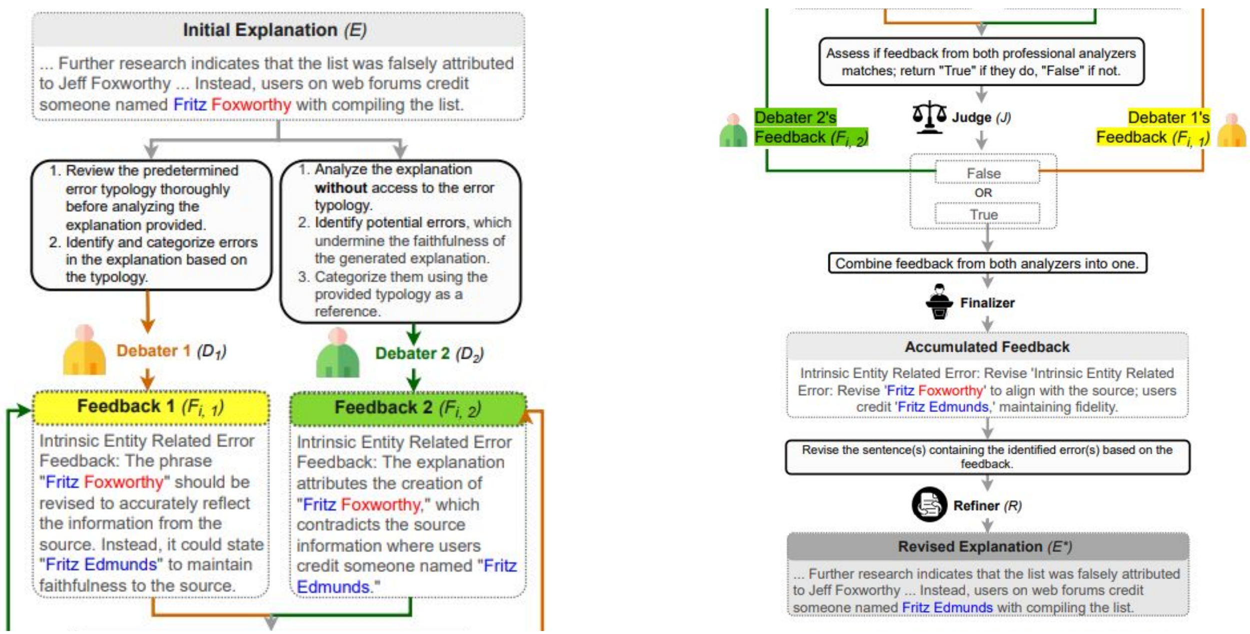


Image from [3]

References

[1] Shi, Xiang, et al. "Know where to go: Make LLM a relevant, responsible, and trustworthy searchers." Decision Support Systems 188 (2025): 114354.

[2] Zheng, Danna, et al. "Trustscore: Reference-free evaluation of llm response trustworthiness." arXiv preprint arXiv:2402.12545 (2024).

[3] Kim, Kyungha, et al. "Can llms produce faithful explanations for fact-checking? towards faithful explainable fact-checking via multi-agent debate." arXiv preprint arXiv:2402.07401 (2024).