GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

HPS

Seminar Report

LLM Trustworthiness and Fact Validation

Ashutosh Kumar Jaiswal

MatrNr: 22540934

Supervisor: Sadegh Keshtkar

Georg-August-Universität Göttingen
Institute of Computer Science

October 3, 2025

Abstract

Large Language Models (LLMs) are increasingly deployed in knowledge-intensive and high-stakes domains such as law, finance, and healthcare. However, they often generate confident yet incorrect statements, and unlike web search, the origin of these responses/outputs is unknown, making it difficult for non-experts to assess correctness. Ensuring the *trustworthiness* of such models thus requires mechanisms to evaluate factual accuracy, verify evidence grounding, and calibrate confidence while maintaining practical latency and cost.

The report surveys and integrates three complementary research directions aimed at enhancing LLM reliability: (i) *Evidence-centric retrieval and validation* (KNOW WHERE TO GO) that scores web sources by evidence quality and domain authority [Shi+25]; (ii) *Reference-free self-verification* (TRUSTSCORE) that measures behavioral consistency to determine whether a model stands by its own answers in the presence of adversarial distractors [Zhe+24]; and (iii) *Multi-Agent Debate Refinement* (MADR), a cooperative critique framework that transforms bare factual verdicts into faithful, evidence-grounded explanations [Kim+24].

The report also proposes a conceptual layered pipeline combining these paradigms: lightweight retrieval and self-verification by default, escalating to debate-based refinement when uncertainty or inconsistency is detected. Together, they target complementary failure modes, specifically, external hallucinations (incorrect or low-quality sources), parametric inconsistencies (self-contradictions), and unfaithful explanations (unsupported reasoning). The studies in consideration demonstrate stronger evidence precision in retrieval, moderate correlation of TRUSTSCORE with human judgments without gold references, and substantial reductions in hallucinated explanations/content through debate-driven refinement.

Keywords: LLMs, Trustworthiness, Fact Validation, Evidence Retrieval, Self-Verification, Multi-Agent Debate, Explainability, Evaluation

Declaration on the use of ChatGPT and comparable tools in the context of examinations

In this work I have used ChatGPT or another AI as follows:

- ☐ Not at all
- ☑ During brainstorming
- ☐ When creating the outline
- ☑ To write individual passages, altogether to the extent of 0% of the entire text
- ☐ For the development of software source texts
- ☐ For optimizing or restructuring software source texts
- ☐ For proofreading or optimizing
- ☑ Further, namely: - Make coherent paragraphs from notes taken during research

I hereby declare that I have stated all uses completely.
Missing or incorrect information will be considered as an attempt to cheat.

Contents

List of Tables

List of Figures

List of Abbreviations

LLM Large Language Model

GPT Generative Pretrained Transformer

RAG Retrieval-Augmented Generation

RLHF Reinforcement Learning from Human Feedback

KWTG Know Where to Go

TrustBC Behavioral Consistency component of TrustScore

TrustFC Fact-Checking component of TrustScore

TrustOV Overall combined trust score (TrustBC + TrustFC)

MADR Multi-Agent Debate Refinement

QA Question Answering

CoT Chain-of-Thought

URL Uniform Resource Locator

UX User Experience

MSMARCO MS MARCO (Microsoft Machine Reading Comprehension)

1 Introduction

Large Language Models (LLMs) have evolved from general-purpose text generators into powerful reasoning and decision-support systems across domains such as finance, healthcare, law, and public administration. Their broad competence, emergent reasoning abilities, and capacity to summarize or synthesize information at scale make them attractive for real-world deployment. However, the same generative flexibility introduces uncertainty: LLMs generate outputs token by token, optimizing for plausibility rather than truth. As a result, they may produce statements that sound confident yet are factually incorrect or unverifiable—a phenomenon widely known as *hallucination*.

Unlike web search, where retrieved pages visibly indicate provenance, the source of an LLM's knowledge is largely implicit. This opacity makes it difficult for users—especially non-experts—to assess correctness. The issue becomes acute in high-stakes contexts such as legal reasoning, clinical decision support, or financial forecasting, where unverified claims can mislead users or incur reputational and ethical risks. Real-world incidents illustrate the stakes: in February 2023, Google's Bard demo misstated an astronomy fact [Ver23; Reu23b]; in April 2023, an Australian mayor prepared a defamation suit after ChatGPT fabricated details about him [Reu23a]. These examples underline that linguistic fluency alone is not a guarantee of factual reliability. So, how can we evaluate and improve the trustworthiness of LLM outputs—that is, ensure that generated statements are both *factually correct* and *faithful* to underlying evidence?

In human–AI collaboration, trustworthiness is foundational. Users must not only receive correct answers but also understand *why* an answer is credible. Current evaluation methods often focus on accuracy against reference answers, yet many open-ended queries lack gold labels. Similarly, system-level metrics like perplexity or log-probabilities tell us how surprised the model is at each step, but they don't reliably reflect whether the final answer is correct or how certain the model should really be. What is needed are mechanisms that can both *assess internal confidence* and *validate external evidence* in a scalable way.

This report builds on three complementary research directions:

1. **Evidence-centric retrieval and validation**—"Know Where to Go" integrates LLMs into the search loop, ranking sources by authority and evidence quality rather than trivial methods such as keyword overlap [Shi+25].
2. **Reference-free self-verification**—*TrustScore* evaluates a model's behavioral consistency to test whether it truly "believes" its own answer without needing ground-truth labels [Zhe+24].
3. **Faithful multi-agent explanation**—*MADR* employs debate among agents to iteratively refine fact-check explanations, reducing unsupported claims and hallucinated responses [Kim+24].

This report dives into studies that design practical mechanisms to (i) estimate whether a generated answer is likely correct, (ii) attach verifiable evidence supporting each claim, and (iii) expose transparent reasoning that users can audit—all while balancing latency, cost, and interpretability.

Limitations of current practices: Traditional retrieval-augmented generation (RAG) mitigates some hallucinations but struggles when retrieved documents are outdated, incomplete, or low-quality. Single-pass generation often introduces claims absent from evidence, while confidence estimates derived from token probabilities remain poorly calibrated. Moreover, reference-based metrics are limited to benchmark datasets and fail to generalize to open-ended, real-world queries.

This report makes two contributions:

1. It synthesizes three studies based on retrieval, self-verification, and explanation refinement into a unified conceptual pipeline for trustworthy answering.

2. It provides practical recommendations for integrating trust signals and fact validation into LLM-based systems.

Outline of the Report: §2 reviews key background knowledge and terminology, §3 presents Fact-checking with Evidence, §4 goes into Self-Verification via Behavioral Consistency, §5 talks about Multi-Agent Debate Refinement, §6 goes over a conceptual pipeline combining the previously discussed approaches, along with some key takeaways and §7 concludes the entire report.

2 Background and Terminology

In this section we introduce key concepts and definitions that underlie our report.

2.1 Large Language Models (LLMs)

LLMs are neural networks based on the *Transformer* architecture, which uses self-attention to read and relate different parts of text efficiently. This design replaced earlier recurrent/convolutional approaches and enabled training on massive datasets. [Vas+17] LLMs are first *pretrained* by next-token prediction on large text corpora (web pages, books, code). This teaches broad language and world knowledge directly in the parameters (*parametric knowledge*). After pretraining, models are typically *instruction-tuned* to follow user prompts better (e.g., supervised fine-tuning on demonstrations), and then *aligned* with human preferences via *Reinforcement Learning from Human Feedback (RLHF)* or related methods. These steps improve helpfulness and safety but do not fully solve factual errors. [Ouy+22]

To reduce reliance on memory alone, systems often use *Retrieval-Augmented Generation (RAG)*: the model looks up documents and conditions on them while generating an answer. This blends parametric knowledge with *non-parametric* evidence (retrieved text), improving transparency and up-to-dateness. [Lew+20]

This is relevant in context of LLM Trustworthiness because decoding is probabilistic and optimized for plausibility, LLMs can produce fluent but incorrect or unverifiable claims. Instruction tuning and RAG help, yet errors still occur, especially when evidence is missing, low-quality, or misused.

2.2 Trustworthiness

The report uses *trustworthiness* to mean the degree to which an LLM's outputs are (i) *accurate*, (ii) *verifiable*, and (iii) *context-appropriate*, while also appropriately representing uncertainty. A trustworthy model should not only provide correct statements, but also signal when it is unsure and avoid misleading or harmful claims.

In recent surveys, trustworthiness is often decomposed into multiple dimensions such as reliability, safety, explainability, resistance to misuse, and robustness [Liu+23]. While such taxonomies are useful, our discussions in this report mostly revolve around factual correctness, evidential grounding, and uncertainty transparency.

2.3 Factuality vs. Faithfulness

These two terms are sometimes used interchangeably, but they refer to distinct phenomena:

- **Factuality** refers to how well a generated statement matches external, real-world facts. If the model says "Paris is the capital of France," that is factually correct; if it says "Paris is the capital of Spain," that is factually incorrect.

- **Faithfulness** pertains to explanations or claims being strictly grounded in the evidence provided (or retrieved). A faithful explanation does *not* introduce unsupported claims or deviate from what the evidence supports. In other words, every claim in the explanation should be traceable to actual evidence or premises.

Thus, even a correct answer can be accompanied by an unfaithful explanation (if the explanation fabricates or overreaches). The literature sometimes frames this as *narrow truthfulness* (factuality) versus *broad truthfulness* (which includes faithfulness and avoidance of misleading framing) [Ma25].

2.4 Closed-book vs. Open-book

- In a **closed-book** setting, the model answers purely from its internal (parametric) knowledge, without access to external documents or evidence. - In an **open-book** (or retrieval-augmented) setting, the model is provided with external sources (e.g. retrieved passages, databases) and may base its answer on those sources. Open-book methods help reduce "external hallucinations" (fabrication of external facts), but they do not guarantee faithfulness or internal consistency.

A related concept is *situated faithfulness*, introduced in recent work: when external sources conflict with internal knowledge, the model should dynamically assess which to trust based on confidence in each [Hua+24].

2.5 Hallucination

A *hallucination* is any piece of generated content that is ungrounded, fabricated, or contradicts known evidence or reality. Hallucinations may be: - *Factual hallucinations*, where the statement is factually false; - *Faithfulness violations*, where the explanation or claim goes beyond or distorts the evidence.

Reducing hallucinations is a central goal in deploying trustworthy LLMs. Even the most advanced models still hallucinate, especially in edge domains or when training data is scarce. [Hua+23].

2.6 Behavioral Consistency

A measure of "internal confidence" in context of LLMs is *behavioral consistency*. In approaches like *TrustScore*, after an LLM provides an answer, one generates distractor alternatives and asks the model to choose among them. If the LLM reliably re-selects its original answer (across variations), that suggests internal consistency and "belief" in its response [Zhe+24]. This method does not rely on external labels and helps estimate whether the model "stands by" its own answer.

2.7 Source Authority and Evidence Reliability

When retrieving external sources, not all documents are equally trustworthy. A system must assess:
- *Source authority*: how credible or expert is the domain or publisher?
- *Evidence quality*: how well does the snippet directly support the claim?

In the "Know Where to Go" framework, the model is integrated into the retrieval loop: generation, validation, and optimization guide which sources to trust [Shi+25]. Some newer work also studies how to transparently integrate internal (parametric) knowledge and external sources into citation generation [Kha+24].

2.8 Zero-Shot Prompting

In zero-shot prompting, the model is given only an instruction and the query—no examples and, in the closed-book case, no external documents. The model relies on its *parametric knowledge* to produce an answer. Zero-shot is fast and simple (low latency/cost), so it is a common baseline and often the first step in pipelines (e.g., to expand intents or draft candidate answers). But because decoding is optimized for plausibility, zero-shot answers—especially in closed-book settings—can be *confident yet ungrounded*.

—

With these definitions and background, we now have a shared vocabulary for the rest of the report.

3 Fact-checking with Evidence (Know Where To Go)

The proposed system puts evidence first: it uses retrieval and source vetting so the model only answers from pages that look reliable and contain quote-level support. Concretely, a Generator–Validator–Optimizer (G–V–O) loop (i) expands and clarifies the query, (ii) checks that candidate pages are reachable and actually answer the question, and (iii) learns over time which domains to prefer. This shifts effort from post-hoc fact-checking to *pre-commitment* on credible sources with explicit snippets. [Shi+25]

3.1 Motivation

Two gaps in "LLM + web" search motivate the proposed system. First, traditional search often misses the right page for vague or complex queries (on MS MARCO, about ∼37.76%

of Bing queries receive no effective result). Second, chat-style LLMs can hallucinate or cite weak sources, so fluent answers are not reliably grounded. [Shi+25]

3.2 Core Mechanism

The paper designs an evidence-first retrieval pipeline that makes the model *commit upfront* to credible sources and quote-level evidence. The system has three main modules—**Generator**, **Validator**, and **Optimizer**—plus a **Resource Pool** that stores (query, source, evidence) mappings for reuse. Figure 1 in the paper gives a high-level view of this flow. [Shi+25]

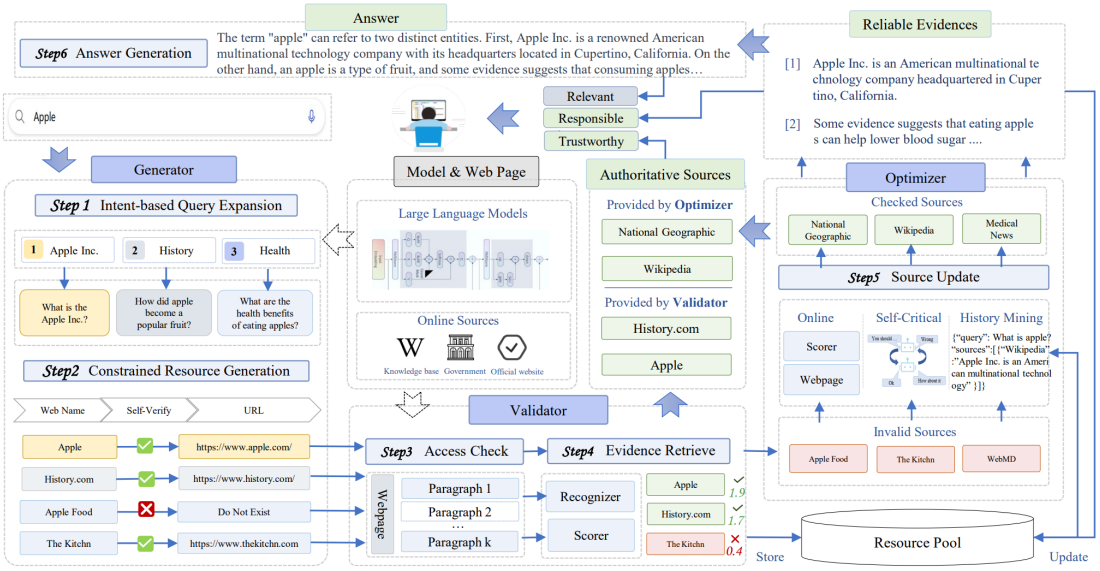

Figure 1: Know Where to Go: Proposed Solution Overview [Shi+25]

3.2.1 Generator

This module links a user query directly to trustworthy *online sources* (domains/root URLs), not just produce an answer.

(a) Intent-based Query Expansion: Many queries are ambiguous (e.g., "Apple" could mean the company or the fruit). The generator expands the query along likely intents and variants so later steps can search and verify against the right families of pages. The paper constructs multi-level intent prompts (10 broad themes → 100 sub-themes) to make this expansion robust across topics.

(b) Constrained Online Source Generation: Instead of hallucinating long, deep URLs, the model first proposes *webpage names* (e.g., "History.com", "WebMD"), then maps each name to a *root URL* (e.g., `https://www.history.com/`). This reduces the "illusion" problem (invented pages/paths) and makes later access checks realistic. In their Alpaca example on MS MARCO queries, adding these constraints increases URL *accessibility* from about 39.25% to 71.07%.

(c) Self-verification of names/URLs: The generator asks itself to re-check that a proposed site actually exists and that the name matches the domain (catching cases like a valid-sounding site name with no real URL). Outputs are tagged into: (1) nonexistent

site name; (2) name exists but URL invalid/mismatched; (3) name and URL exist but may still be off-topic (the last case is handled by the Validator).

3.2.2 Validator

This module, for the Generator's candidates, verifies that pages are live and contain sentences that actually answer the query, and score how reliable that evidence is.

(a) Access Check and Retrieval: The validator concatenates the expanded queries with the candidate sources to form search expressions and queries Bing for the top-K results. It then checks *timeliness* and *accessibility* (are URLs live now?) and parses accessible pages to plain text.

(b) Evidence Recognition + Scoring: The page text is split into fixed windows; an LLM-based *Evidence Recognizer* pulls out candidate sentences that address the query, and an *Evidence Scorer* assigns a score by asking the model a constrained Yes/No question ("Does this text answer the query?"), turning that probability into a numeric score. Two strategies are used:

- *Score-only*: higher recall, coarser spans.
- *Hybrid*: recognizer narrows candidates, scorer re-checks for precision (supported by data augmentation to avoid overfitting to phrasing).

Validated sentences become *quote-level evidence* attached to the source.

3.2.3 Optimizer

This module improves sources over time by replacing weak ones, discovering better ones, and reusing proven links from prior queries.

(a) Self-Critical Strategy: If validation shows a site is invalid (e.g., a hallucinated domain), the system feeds that finding back to the model as a natural-language instruction ("Replace site X with a valid alternative"), regenerates candidates, and re-validates.

(b) Online Strategy.: The Validator's evidence module also helps *discover* additional pages close to the user's intent (e.g., by following search results or related pages), then runs the same access/evidence checks before accepting them.

(c) History Mining Strategy: A *Resource Pool* stores past (query, sources, evidence). For a new query, the system measures similarity to past queries; if similarity exceeds a threshold δ, their sources are proposed as candidates and re-validated. This gives a warm start on recurring topics and improves stability when the web changes.

3.3 Complete Pipeline

1. **Generator:** expands the query by intent, proposes site names, and maps them to root URLs under constraints; it self-verifies site existence.
2. **Validator:** runs AccessCheck, retrieves top-K results, extracts sentence-level evidence, and scores reliability (score-only or hybrid).
3. **Optimizer:** replaces weak sources (self-critical), discovers new ones online, and mines the Source Pool to reuse reliable domains; all updates are re-validated before use.

3.4 Evaluation

To judge whether the proposed system improves both *relevance* and *reliability*, the paper reports two groups of signals: (A) Statistical Signals that describe what was retrieved or generated, and (B) Performance Signals that measure trust and evidence quality. They are summarized in Table 1

Table 1: Signals used in "Know Where To Go" Paper. Higher is better (↑)

Signal	**What it measures**
S_{count}, E_{count}	Source/evidence volume recalled
Q_{correct}, T_{avg}	Correct answers; topic breadth per query
Timeliness↑, Access↑	Source freshness; URL reachability
Consistency↑	Name↔URL match (right site)
Validity↑	Fraction of sources that truly answer the query
Precision↑	Sufficiency of quoted snippets for the answer

Against strong baselines models (New Bing, Perplexity.ai, WebGPT, WebGLM), the proposed method reports *higher* validity and precision (e.g., Validity ≈ 89.7%, Precision ≈ 78.4% for their 7B parameter model), and overall improvements of +2.54% (validity) and +1.05% (precision) versus advanced generative methods. [Shi+25]

4 Self-Verification via Behavioral Consistency (TrustScore)

4.1 Motivation

Closed-book answers are generated from a model's internal (parametric) knowledge, so users cannot easily verify correctness. External fact-checking can help, but evidence may be missing, hard to retrieve, or inconclusive. The key idea in TrustScore is to ask: *does the model itself behave as if its answer is correct?* If an answer reflects internal knowledge, the model should consistently prefer it over strong alternatives. This offers a reference-free trust signal that complements evidence-based checks when references exist [Zhe+24].

4.2 Core Mechanism

At a high level, TrustScore turns a free-form answer into a series of *self-checks*. Given a question q and the model's answer a, it creates strong but incorrect alternatives (*distractors*) and repeatedly asks the model to choose among them. The design is reference-free and model-agnostic, so it works even when gold answers or reliable evidence are unavailable.

4.2.1 Behavioral Consistency (TrustBC)

Given a question q and the model's free-form answer a, TrustBC runs a series of multiple-choice checks. Each check presents a alongside several *distractors* plus a "None of the

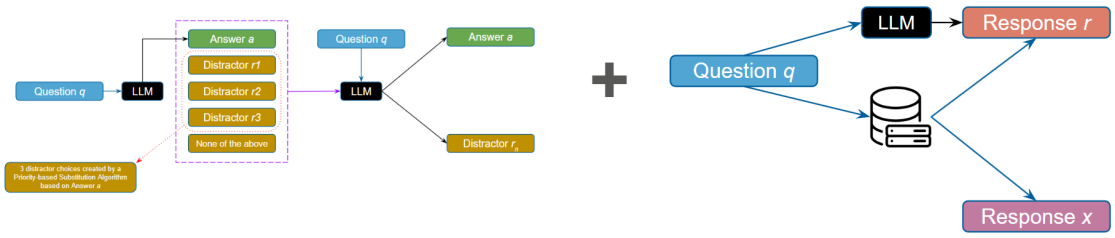

Figure 2: TrustScore (TrustOV): Proposed Solution Overview (TrustBC + TrustFC)

above" option; the model must pick the best choice. If the model *always* re-selects a across up to n checks, TrustBC returns 1 (consistent); otherwise 0 (inconsistent). High behavioral consistency suggests the model "stands by" its original answer [Zhe+24]. This is represented on the left side in Figure 2

4.2.2 Distractors' Generation

Distractors are built to be *plausible* but wrong using a priority-based substitution algorithm: (i) substitute words unique to a before words shared with q; (ii) prefer entities, then nouns/numbers, then other POS; (iii) source substitutes from (1) filtered DBpedia entities, (2) embedding-nearest words, and (3) POS-filtered random words. This yields strong foils that meaningfully probe the model's internal belief [Zhe+24].

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.

Figure 3: Distractor Generation: Model's free-form answer is Option C, other options are distractors

4.2.3 Optional Fact-Checking (TrustFC)

When external knowledge is available, TrustFC retrieves evidence and uses an entailment model to decide whether r is *supported*, *contradicted*, or *unknown*. Right side of figure 2 shows a simple diagram of the same.

4.2.4 Combined score (TrustOV)

TrustOV merges TrustBC and TrustFC with simple rules: prioritize factual consistency over behavioral consistency; if evidence contradicts a, mark it untrustworthy; otherwise increase the weight of answers the model consistently endorses [Zhe+24].

4.3 Evaluation

The authors built *MixedQA* (1,000 questions sampled from NQ, WebQuestions, TriviaQA, HotpotQA, PopQA) and collect human binary labels (correct/incorrect) for 3,000 answers

from FLAN-T5-XXL, LLaMA-7B, and GPT-3.5. They reported Pearson correlations between metrics and human judgments [Zhe+24]:

- *Reference-free Trust (Internal):* TrustBC correlation with human correctness (closed-book, no references).
- *Reference-based Trust (External):* TrustFC correlation when evidence is available.
- *Combined Trust:* TrustOV correlation, reflecting both internal consistency and external support.

TrustBC outperforms other reference-free baselines and, for some models, gets close to reference-based metrics. TrustFC correlates well when evidence exists; and TrustOV is the most reliable overall, consistently matching or exceeding alternatives across models [Zhe+24].

5 Reliable Explanations via Multi-Agent Debate Refinement (MADR)

5.1 Motivation

Users trust fact-checks more when they see *why* a verdict is correct. However, zero-shot LLM explanations often drift from the evidence: in the authors' study, about 80% of such explanations contained hallucinated or unreliable details [Kim+24]. This is especially problematic in multi-hop settings, where a claim must be checked against multiple pieces of evidence. The goal is therefore to generate explanations that stay strictly aligned with the provided evidence.

5.2 Core Mechanism

MADR pipeline generates reliable explanations by the following *"Critique-Revise"* loop among specialized LLM agents [Kim+24]:

1. **Initial Writer:** A base LLM produces an initial explanation E for a given claim and evidence.
2. **Two Debaters ($\mathbf{D_1}$, $\mathbf{D_2}$):** D_1 reviews E using a predefined *error typology* (intrinsic/extrinsic entity, event, and noun-phrase errors; reasoning coherence; overgeneralization; irrelevant evidence), while D_2 critiques E freely, i.e. without seeing the typology to detect different issues. Each produces structured feedback $F_{i,1}$ and $F_{i,2}$ at iteration i.
3. **Judge:** A judge agent compares the two feedback sets; if they align, the debate round stops, otherwise the debaters refine and reconcile their feedback until agreement (or a fixed number of iterations).
4. **Refiner:** The refiner rewrites only the flagged parts of E using the *accumulated* feedback, producing E^*. The loop then continues for a small, fixed number of rounds, yielding a final explanation that better matches the evidence.

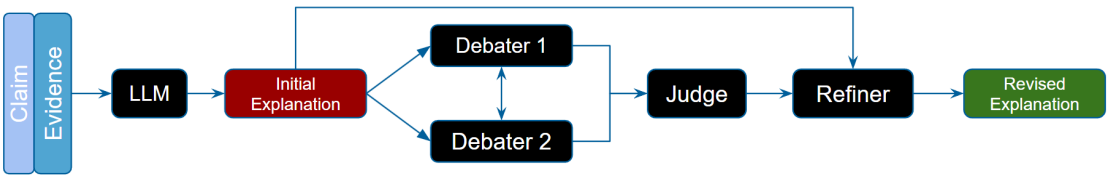

Figure 4: Multi-Agent Debate Refinement (MADR): Proposed Solution Overview (Simplified)

5.3 Evaluation

The authors evaluate on *PolitiHop* dataset with automatic G-Eval (GPT-4-Turbo) at sentence- and document-level, with/without an explicit error typology, plus human judgments for faithfulness and error counts.

1. MADR produces the most faithful explanations with the fewest errors among baselines (zero-shot, Chain-of-Thought, self-refine) — e.g., the highest share of faithful outputs (approx. 30%) and the lowest total errors on PolitiHop.

2. The *sentence-level + typology* protocol aligns best with human judgments (highest Kendall's τ), suggesting that fine-grained scoring with an explicit error schema is the most reliable automatic signal [Kim+24].

6 Discussion

This section goes over a hypothetical pipeline that attempts to conceptually unify the three prior approaches discussed in the report, highlights strengths and limitations and end with some miscellaneous practical advice.

6.1 A Layered Pipeline

We combine the three paradigms into a practical workflow (Figure 5): run fast retrieval and evidence scoring by default (§3); compute TrustScore to decide whether to *accept*, *re-prompt*, or *escalate* (§4); and activate MADR only when confidence is low or a faithful explanation is explicitly required (§5). This "defense in depth" maps failure modes to targeted checks: retrieval reduces *external* hallucinations [Shi+25]; behavioral consistency probes *parametric* uncertainty [Zhe+24]; debate reduces unfaithful explanations [Kim+24].

6.2 Strengths and Limitations

Evidence-centric retrieval curbs external hallucinations and improves auditability, but can under-weight new or low-link sources; mitigate with curated whitelists and periodic refreshes [Shi+25]. **Self-verification** is cheap and reference-free, but depends on strong distractors and can be biased toward the initial answer [Zhe+24]. **Debate refinement** improves faithfulness and user trust, yet adds latency and cost; returns diminish after a few rounds [Kim+24].

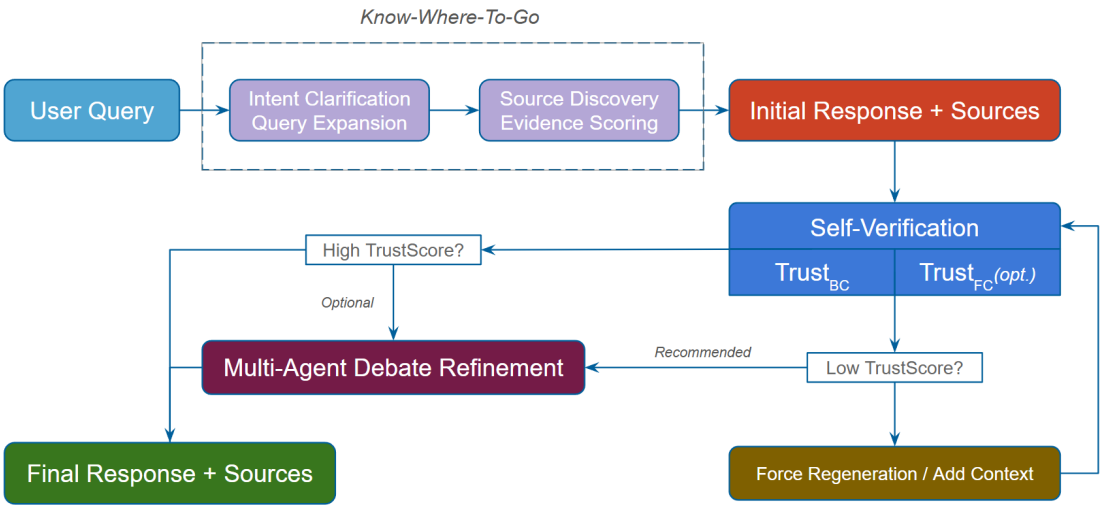

Figure 5: (Conceptual) Unified pipeline for Trustworthy Answering and Fact Validation

6.3 Operational Guidance

- **Caching & Reuse:** Cache high-authority domains and validated quote-level snippets; prefer them when semantically similar queries recur.
- **Thresholds by Risk:** Tune τ and evidence sufficiency rules per domain (e.g., require two independent snippets for high stakes domain such as medical or legal).
- **Selective Escalation:** Trigger MADR only for low TrustOV, multi-hop claims, or explanations shown to non-experts.
- **Abstention & UX:** Prefer an explicit "cannot verify" state over speculative answers; output confidence scores and citations by default.

7 Conclusion

LLM outputs remain persuasive yet fallible because decoding optimizes plausibility, not truth. This report argued for a layered, defense-in-depth approach to trust: (i) *pre-commit* to reputable sources and quote-level evidence (Know Where To Go); (ii) use *reference-free* behavioral consistency (TrustScore) as an early, cheap signal of internal belief; and (iii) apply *debate-based refinement* (MADR) when confidence is low or faithful explanations are required. Combined in an adaptive pipeline, these methods substantially reduce hallucinations without incurring constant heavy costs (assuming the pipeline being tuned for efficiency). The pipeline is practical but not without limits, such as, source-authority bias, distractor quality, and added latency at refinement time. It is advised to tune thresholds by domain risk, prefer abstention over speculation, and present citations by default. Future work may include learning authority priors from feedback, strengthening adversarial defenses (prompt injection, retrieval poisoning), and improving human-facing calibration and UX for uncertainty.

References

[Hua+23] Longke Huang et al. "A Survey on Hallucination in Large Language Models". In: *arXiv preprint arXiv:2311.05232* (2023). URL: https://arxiv.org/abs/2311.05232.

[Hua+24] Yukun Huang et al. "Enhancing Large Language Models' Situated Faithfulness to External Contexts". In: *arXiv preprint arXiv:2410.14675* (2024). URL: https://arxiv.org/abs/2410.14675.

[Kha+24] Muhammad Khalifa et al. "Source-Aware Training Enables Knowledge Attribution in Language Models". In: *arXiv preprint arXiv:2404.01019* (2024). URL: https://arxiv.org/abs/2404.01019.

[Kim+24] Kyungha Kim et al. "Can LLMs Produce Faithful Explanations For Fact-checking? Towards Faithful Explainable Fact-Checking via Multi-Agent Debate". In: *arXiv preprint arXiv:2402.07401* (2024). URL: https://arxiv.org/abs/2402.07401.

[Lew+20] Patrick Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 33. 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

[Liu+23] Yang Liu et al. "Trustworthy LLMs: A Survey and Guideline for Evaluating Large Language Model Trustworthiness". In: *arXiv preprint arXiv:2308.05374* (2023). URL: https://arxiv.org/abs/2308.05374.

[Ma25] Uwe Maes and et al. "Mitigating Misleadingness in LLM-Generated Natural Language Explanations". In: *CEUR Workshop Proceedings* 3957 (2025). URL: https://ceur-ws.org/Vol-3957/AXAI-paper11.pdf.

[Ouy+22] Long Ouyang et al. "Training language models to follow instructions with human feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

[Reu23a] Reuters. *Australian mayor readies world's first defamation lawsuit over ChatGPT content*. 2023. URL: https://www.reuters.com/technology/australian-mayor-readies-worlds-first-defamation-lawsuit-over-chatgpt-content-2023-04-05/ (visited on 09/28/2025).

[Reu23b] Reuters. *Google AI chatbot Bard offers inaccurate information in company ad*. 2023. URL: https://www.reuters.com/technology/google-ai-chatbot-bard-offers-inaccurate-information-company-ad-2023-02-08/ (visited on 09/28/2025).

[Shi+25] Xiang Shi et al. "Know Where to Go: Make LLM a Relevant, Responsible, and Trustworthy Searcher". In: *Decision Support Systems* 188 (2025), p. 114354.

[Vas+17] Ashish Vaswani et al. "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017. URL: https://papers.neurips.cc/paper/7181-attention-is-all-you-need.pdf.

[Ver23] The Verge. *Google's Bard made an error in its first demo*. 2023. URL: https://www.theverge.com/2023/2/8/23590864/google-ai-chatbot-bard-mistake-error-exoplanet-demo (visited on 09/28/2025).

[Zhe+24] Danna Zheng et al. "TrustScore: Reference-Free Evaluation of LLM Response Trustworthiness". In: *arXiv preprint arXiv:2402.12545* (2024). URL: https://arxiv.org/abs/2402.12545.