GAVEL: Evidence-Contract Debate with Mechanized Scrutiny for Provenance-Grounded Fact-Checking

Ruoyu Xu and Gaoxiang Li and Victor S. Sheng
Computer Science Department, Texas Tech University
{ruoyxu, gaoli, victor.sheng}@ttu.edu

Abstract

Evidence-grounded fact-checking requires predicting claim veracity while returning faithful evidence at fine granularity, including exact sentences, table cells, and complete multi-document chains. Although large language models enable decomposition, planning, and multi-agent verification, they can still produce convincing rationales with weak provenance, especially under heterogeneous evidence and multi-hop requirements. We propose GAVEL, a multi-agent debate framework that enforces evidence grounding throughout inference. GAVEL introduces an Evidence Contract that requires debaters to state atomic subclaims and bind each to explicit evidence units, and a Mechanized Chain of Scrutiny in which a neutral Scrutinizer audits outputs and performs deterministic validation of cited identifiers and quoted spans. A Judge then selects a sufficient evidence set and produces the final decision. Experiments on FEVEROUS and HOVER in an open-book setting show that GAVEL improves provenance-aware metrics that jointly require correct labels and correct, complete evidence over strong recent baselines. Ablations confirm that both evidence binding and mechanized citation validation are key to the gains.

1 Introduction

The explosive growth of online content has made misinformation easier to produce, faster to spread, and harder to correct at scale (Aïmeur et al., 2023; Fernandez and Alani, 2018). Automated fact-checking therefore plays an increasingly important role in modern information systems, from social platforms to search and conversational agents (Fung et al., 2022; Hu et al., 2022; Jin et al., 2023; Tian et al., 2023). In practice, a useful verifier must not only output a label, but also return evidence that can be independently inspected and reused by downstream systems.

Fact-checking is increasingly central to natural language processing systems, with growing emphasis on producing not only a correct veracity label but also faithful evidence that can be independently verified. Two open-domain benchmarks highlight the main challenges. Jiang et al. (2020) introduces HOVER, where a claim may require evidence drawn from up to four Wikipedia articles, stressing retrieval planning and cross-document reasoning. Aly et al. (2021) proposes FEVEROUS, which extends fact verification to heterogeneous evidence and requires systems to justify predictions using both unstructured text and structured sources such as Wikipedia tables. This requirement makes precise evidence grounding substantially harder than sentence-only verification, because models must identify the correct table cells and connect them with relevant sentences.

Recent work has shown that large language models can improve complex claim verification by decomposing a claim into simpler subproblems and solving them with retrieval and intermediate reasoning. Pan et al. (2023) proposes ProgramFC, which generates executable programs that guide multi-step evidence seeking and verification, demonstrating the value of structured decomposition for complex claims. Other work explores planning and tool use for fact-checking. Zhao et al. (2024) introduces PACAR, which frames verification as planning with customized actions executed by a large language model. Multi-agent designs also aim to increase reliability by distributing roles and introducing internal critique, as in LoCal, which emphasizes logical and causal checking through collaboration among multiple agents (Ma et al., 2025). Despite this progress, existing approaches often remain vulnerable to a key failure mode. They can produce fluent arguments and critiques that are not tightly bound to correct provenance, especially when evidence must be cited at fine granularity, such as exact sentences, exact table cells, and complete multi-hop evidence sets.

Findings of the Association for Computational Linguistics: ACL 2026, pages 35907–35920

We address this gap by enforcing evidence grounding throughout inference. GAVEL (**G**rounded **A**udited **V**erification with **E**vidence-**L**ocked debate) is a multi-agent debate framework built around two ideas: an Evidence Contract that requires each atomic subclaim to be linked to explicit evidence units, and a Mechanized Chain of Scrutiny that audits arguments and validates citations against the retrieved context. Together, these components target the provenance requirements of FEVEROUS and HOVER, where correctness depends on both the predicted label and the completeness of the returned evidence (Jiang et al., 2020; Aly et al., 2021).

Our contributions are threefold. First, we formalize evidence-locked debate for fact-checking via an Evidence Contract that prevents uncited assertions and promotes verifiable reasoning at the atomic claim level. Second, we introduce mechanized auditing during multi-agent inference through a Chain of Scrutiny protocol that produces binding correction signals and enforces citation validity checks. Third, we evaluate on FEVEROUS and HOVER in open-domain settings with sentence-level and cell-level evidence, comparing against strong recent baselines spanning program-guided decomposition, planning-based verification, and multi-agent collaboration (Pan et al., 2023; Zhao et al., 2024; Ma et al., 2025).

2 Related Work

2.1 Fact-checking benchmarks and evidence-grounded methods

Automated fact-checking has received increasing attention as a way to detect and correct misinformation (Nakov et al., 2021; Guo et al., 2022). Given a claim, a fact-checking system typically retrieves relevant evidence and predicts the claim's veracity based on that evidence (Glockner et al., 2022; Thorne and Vlachos, 2018). Early benchmark-driven work (Jiang et al., 2021; Liu et al., 2020; Majumder et al., 2021; Rao and Daumé III, 2019; Saakyan et al., 2021; Schuster et al., 2021) largely focused on relatively simple claims verifiable with a small amount of evidence, often from a single Wikipedia page. However, real-world claims frequently require multi-evidence reasoning across multiple sources. Recent fact-checking models (Barnabò et al., 2023; Krishna et al., 2022; Ousidhoum et al., 2022; Pan et al., 2023) have acknowledged the importance of handling complex claims. To bridge this gap, datasets have been proposed to study complex, multi-hop verification (Jiang et al., 2020; Aly et al., 2021).

A prominent line of work constructs evaluation settings grounded in Wikipedia. FEVER requires predicting a veracity label and returning supporting sentences for supported and refuted claims (Thorne et al., 2018). HOVER extends this setting to many-hop claims that require evidence from multiple Wikipedia pages, and evaluates label correctness only when the evidence covers all required supporting documents (Jiang et al., 2020). FEVEROUS further broadens the evidence space by requiring provenance from both text and semi-structured sources such as Wikipedia tables and lists (Aly et al., 2021). These benchmarks motivate methods that not only improve label accuracy but also produce faithful evidence with fine-grained provenance.

Recent systems tackle complex verification with decomposition and structured reasoning. ProgramFC decomposes a claim into sub-questions expressed as an executable program, delegates subtasks to specialized handlers, and aggregates intermediate results for final verification (Pan et al., 2023). PACAR frames fact-checking as planning with customized actions executed by an LLM, emphasizing iterative evidence acquisition and structured tool use (Zhao et al., 2024). LoCal studies LLM-based multi-agent collaboration for logical and causal verification, combining decomposition, reasoning, and evaluator agents to reduce reasoning errors (Ma et al., 2025). While these approaches improve reasoning and can handle multi-hop claims, they may still produce plausible conclusions with weak provenance when citations are not explicitly constrained and validated, particularly in settings requiring fine-grained evidence units such as table cells.

2.2 Inference-time auditing and reliability for LLM-based verification

A growing body of work improves reliability by adding verification loops, critique, or multi-agent debate. Verification-loop approaches such as Chain of Verification reduce hallucinations by prompting a model to generate verification questions and answer them before producing a final response (Dhuliawala et al., 2024). Debate-style protocols encourage diverse reasoning by having agents argue for opposing positions under a judge-controlled process (Liang et al., 2024).

Our work is most closely related to inference-time reliability methods for LLM-based verification, but it focuses specifically on provenance-grounded fact-checking. Unlike prior approaches, GAVEL enforces explicit evidence links for intermediate claims and validates citations during inference. These design choices are especially important for FEVEROUS and HOVER, where evaluation depends not only on the final label but also on the validity and completeness of the returned evidence.

3 Task Definition and Problem Setup

We study open-domain claim verification with explicit evidence grounding. Given a natural language claim x, a system must predict a veracity label y and return a set of evidence units E drawn from a large reference corpus. We target two benchmarks that emphasize complementary difficulties: multi-hop evidence completeness in HOVER (Jiang et al., 2020) and heterogeneous evidence from text and tables in FEVEROUS (Aly et al., 2021).

3.1 Evidence Corpus and open-book Setting

We assume a fixed Wikipedia snapshot aligned to each dataset. Models operate in an open-book setting: at inference time, they may only use information retrieved from the corpus for both label prediction and evidence citation. This constraint is standard for open-domain verification and ensures that the prediction is traceable to explicit provenance rather than parametric memory.

3.2 Evidence Units

To support FEVEROUS and HOVER, we represent evidence as a set of fine-grained units drawn from Wikipedia.

Sentence evidence. A sentence evidence unit is represented as a tuple $\langle p, s \rangle$ where p is a Wikipedia page title and s is a sentence identifier within that page. Sentence units are used for both FEVEROUS and HOVER.

Table cell evidence (FEVEROUS only). A table cell evidence unit is represented as a tuple $\langle p, t, r, c \rangle$ where p is the page title, t is a table identifier within the page, and (r, c) are row and column indices for the cell. This representation follows FEVEROUS, which requires structured evidence from tables in addition to text (Aly et al., 2021). HOVER does not include table-cell evidence, so $\langle p, t, r, c \rangle$ units are not used when evaluating on HOVER.

3.3 Outputs

For each claim x, the system outputs a label and an evidence set:

$$f(x) = (y, E),$$

where E is a set of cited evidence units. For FEVEROUS, $y \in \{\text{SUPPORTS}, \text{REFUTES}, \text{NOTENOUGHINFO}\}$ and E may contain both sentence and cell units. For HOVER, we follow the official binary evaluation setting and use $y \in \{\text{SUPPORTED}, \text{NOTSUPPORTED}\}$.

4 GAVEL

We propose GAVEL (**G**rounded **A**udited **V**erification with **E**vidence-**L**ocked debate), a multi-agent debate framework that improves evidence-grounded verification by enforcing provenance constraints throughout inference. GAVEL combines two mechanisms. First, an Evidence Contract forces debaters to express arguments as evidence-linked atomic subclaims. Second, a Mechanized Chain of Scrutiny equips a neutral auditor with structured logic checking and deterministic validation of citations. The overall protocol is round-based and bounded for feasibility. Figure 1 provides an overview of the GAVEL inference protocol, including the evidence contract, the chain of scrutiny, and the audit feedback loop used to refine evidence-grounded predictions.

4.1 Agents and Roles

GAVEL instantiates four large language model agents with fixed roles under the task setup defined in Section 3.

Affirmative Debater. Agent A is biased toward predicting SUPPORTS. It retrieves candidate evidence and constructs an evidence-linked argument that the claim is supported.

Negative Debater. Agent N is biased toward predicting REFUTES. It retrieves candidate evidence and constructs an evidence-linked argument that the claim is refuted or not supported.

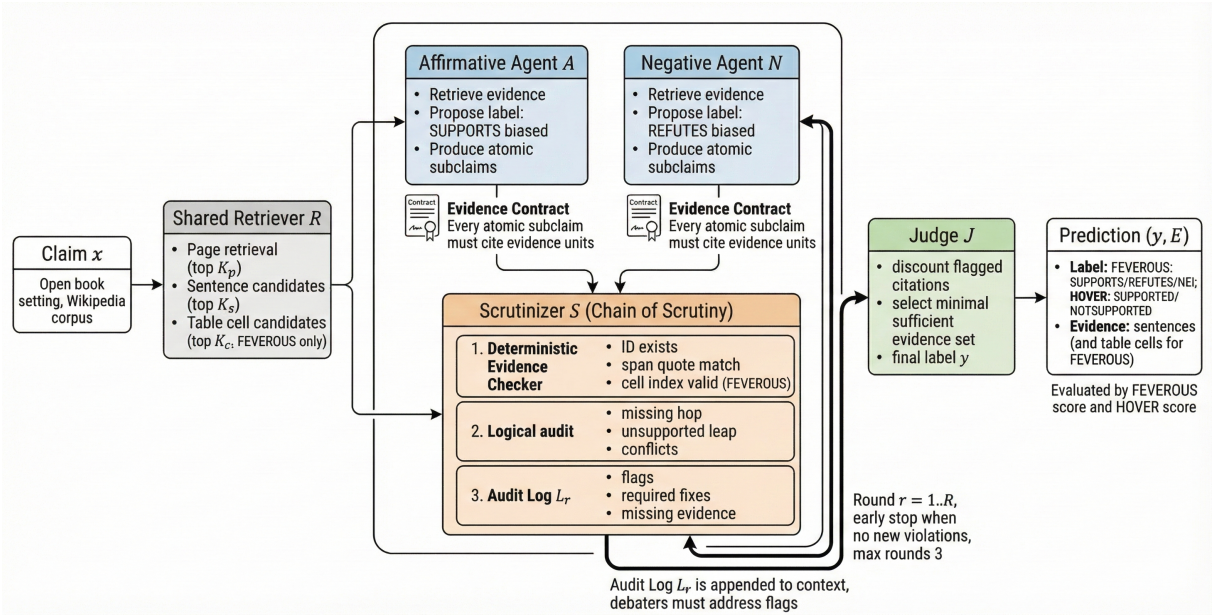

Figure 1: Overview of GAVEL. Two debaters retrieve sentence and table cell candidates and produce evidence linked atomic subclaims under an Evidence Contract. A Scrutinizer audits arguments with a mechanized chain of scrutiny and returns an audit log that guides subsequent rounds, after which a Judge outputs the final label and evidence set. Cell-level checks are applied only on FEVEROUS; on HOVER, the checker validates sentence identifiers and quote grounding only.

Scrutinizer. Agent S is neutral and does not issue new retrieval calls. It audits debater outputs using the provided evidence candidates, checks evidence validity, identifies logical gaps and fallacies, and emits an Audit Log that serves as a binding correction signal for the next round.

Judge. Agent J synthesizes debate history and audit logs to output the final label and the final evidence set.

4.2 Evidence Contract

The Evidence Contract makes debate content provenance-bound. Each debater must produce a set of atomic subclaims and link every atomic subclaim to explicit evidence units. Concretely, a debater outputs a structured record:

$$\mathcal{O} = \{(c_i, E_i, t_i, r_i)\}_{i=1}^{m}$$

where c_i is an atomic subclaim, E_i is a set of evidence units, $t_i \in \{\text{SUPPORTS}, \text{REFUTES}, \text{INSUFFICIENT}\}$ is the local entailment tag of E_i with respect to c_i, and r_i is a short rationale.

The contract imposes a strict rule.

Contract rule. Any statement that is not expressed as an atomic subclaim with linked evidence is treated as unsupported. Unsupported content is penalized by the Scrutinizer and discounted by the Judge.

This constraint aligns inference with provenance-aware evaluation in FEVEROUS and HOVER, where correctness depends on both the label and evidence completeness (Aly et al., 2021; Jiang et al., 2020).

4.3 Mechanized Chain of Scrutiny

The Scrutinizer performs a three-phase Chain of Scrutiny after each debate round. Phase one is partially mechanized through an evidence checker.

Phase 1: Decomposition and Deterministic Evidence Validation The Scrutinizer decomposes each debater's output into evidence claims and inference steps, isolating each cited fact for verification. It then invokes a deterministic evidence checker over all cited evidence tuples. The checker performs:

- Schema validity for sentence tuples (and cell tuples for FEVEROUS)
- Existence checks to confirm that the cited unit appears in the retrieved context
- Quote match checks to confirm that the claimed span appears verbatim in the referenced unit

- Duplication and conflict checks for inconsistent citations

The checker outputs a set of flags $\mathcal{F}$ such as missing identifiers, out-of-range indices, or quote mismatches. These flags are incorporated into the audit report for downstream correction.

Phase 2: Logical and Fallacy Auditing Given the decomposed inference steps, the Scrutinizer identifies common reasoning failures such as circular reasoning, strawman arguments, hasty generalization, and unsupported leaps. The goal is to ensure that the final verdict is not only evidence-grounded, but also logically consistent.

Phase 3: Audit Log as a Binding Intervention After auditing both sides, the Scrutinizer produces an Audit Log L_r for round r. The log enumerates invalid citations, missing evidence requirements, and required corrections. The Audit Log is appended to the shared context and becomes a binding constraint in the next round. Debaters must explicitly address outstanding flags in L_r in subsequent rounds; otherwise, their claims are discounted.

4.4 Debate Protocol and Stopping Criteria

GAVEL proceeds for a bounded number of rounds. In round r, the debaters generate arguments under the Evidence Contract. The Scrutinizer then audits the round and emits L_r. In round $r+1$, both debaters must rebut the opponent and address L_r.

We stop early if the Scrutinizer reports no new violations or if the system reaches the maximum round cap. In all experiments, we cap the number of rounds at three to ensure feasibility.

4.5 Final Decision and Evidence Selection

The Judge produces the final label and evidence set by synthesizing the debate transcript and the final audit log. The Judge uses an evidence-first rubric:

1. Evidence validity, discounting any evidence flagged by the deterministic checker
2. Evidence sufficiency, preferring a set that covers all atomic subclaims required for the final verdict
3. Reasoning consistency, preferring arguments with fewer unresolved audit flags

The Judge outputs the final label y, the final evidence set E.

Algorithm 1 summarizes the inference protocol.

Algorithm 1 GAVEL

Require: Claim x, retriever $\mathcal{R}$, maximum rounds $R_{\max} = 3$

Initialize shared context $\mathcal{C}_0 \leftarrow \{x\}$
for $r = 1$ to $R_{\max}$ **do**
  $\mathcal{O}_r^A \leftarrow \textsc{Debate}(A, \mathcal{C}_{r-1}, \mathcal{R})$ {Evidence Contract enforced}
  $\mathcal{O}_r^N \leftarrow \textsc{Debate}(N, \mathcal{C}_{r-1}, \mathcal{R})$ {Evidence Contract enforced}
  $\mathcal{F}_r \leftarrow \textsc{CheckEvidence}(\mathcal{O}_r^A, \mathcal{O}_r^N)$ {Deterministic}
  $L_r \leftarrow \textsc{Scrutinize}(S, \mathcal{O}_r^A, \mathcal{O}_r^N, \mathcal{F}_r)$
  $\mathcal{C}_r \leftarrow \mathcal{C}_{r-1} \cup \{\mathcal{O}_r^A, \mathcal{O}_r^N, L_r\}$
  if $\textsc{NoNewViolations}(L_r)$ **then**
    break
  end if
end for
$(y, E) \leftarrow \textsc{Judge}(J, \mathcal{C}_r)$
return y, E

5 Experiments and Results

5.1 Experimental Setup

Datasets. We evaluate on FEVEROUS and HOVER, following the official dataset releases and evaluation protocols. FEVEROUS requires fine-grained evidence from both text and tables (Aly et al., 2021), while HOVER focuses on multi-hop verification across multiple Wikipedia pages (Jiang et al., 2020). For both datasets, we use the provided Wikipedia snapshot in the standard open-book setting.

Baselines. We compare against strong recent baselines that represent common verification paradigms. **ZS plus CoT** prompts a single model to produce a label and evidence-backed rationale given retrieved context, serving as a lightweight LLM baseline. **ProgramFC** (Pan et al., 2023) generates an explicit program that decomposes a complex claim into intermediate subproblems and executes the program with retrieval and verification steps. We report N=1 and N=5, where N denotes the number of sampled programs aggregated for the final decision. **PACAR** (Zhao et al., 2024) frames fact-checking as planning with a customized action space, where an LLM iteratively selects actions such as decomposition, retrieval, verification, and aggregation. **LoCal** (Ma et al., 2025) is a multi-agent collaboration framework that combines decomposition and reasoning with evaluator agents

that critique intermediate reasoning for logical and causal issues.

Metrics. We report the official provenance-aware metrics for each benchmark, together with diagnostic label and evidence metrics computed over the same evidence unit format. For FEVEROUS, we report the official FEVEROUS score (Aly et al., 2021). For a claim, a prediction is counted as correct if (i) the predicted label matches the gold label and (ii) the predicted evidence set contains at least one complete gold evidence set as a subset (gold annotations may provide multiple valid evidence sets). The FEVEROUS score is the average of this indicator across instances. We additionally report label accuracy and evidence quality. For evidence F_1, we treat predicted evidence as a set of fine-grained units (i.e., sentences and table cells). Because gold evidence can have multiple alternative sets, we compute unit-level F_1 against each gold evidence set and take the maximum over gold sets for that instance, then average across instances. For HOVER, we report the official HOVER score (Jiang et al., 2020). For a claim, a prediction is counted as correct if (i) the predicted label matches the gold label and (ii) the predicted evidence covers the required supporting documents for that claim (i.e., at least one cited sentence from each gold supporting page). The HOVER score is the average of this indicator across instances. We additionally report label accuracy, unit-level evidence F_1 computed over cited sentences (with max-over-gold-set matching when alternatives exist), and document recall, defined as the fraction of gold supporting pages that are covered by the predicted evidence (averaged across instances).

5.2 Implementation Details

We implement all agents in GAVEL using the OpenAI API with `gpt-3.5-turbo` as the base model. We use greedy decoding with temperature $= 0$ for determinism. We cap the number of debate rounds at three and enable early stopping when the Scrutinizer reports no new violations. For each agent call, we provide the claim, retrieved evidence candidates, and the current debate context, including the most recent Audit Log. The deterministic evidence checker validates cited sentence and cell identifiers against the retrieved context, flags out-of-range or missing identifiers, and verifies that quoted spans appear in the cited units. All methods operate in an open-book setting with a shared retrieval backend and identical evidence formatting to control for retrieval effects. Full reproducibility details are provided in Appendix A.

5.3 Main results

Table 1 reports the main results. Overall, GAVEL achieves the strongest performance on both benchmarks, with the largest gains on provenance-aware metrics that require both correct labels and correct evidence. On FEVEROUS, GAVEL improves the FEVEROUS score to 41.80%, a gain of 4.20% over the strongest baseline (37.60%), and increases evidence F_1 to 47.20%, indicating fewer provenance errors for both sentences and table cells. On HOVER, GAVEL reaches a HOVER score of 49.60%, improving by 4.10% over the strongest baseline (45.50%), and also yields higher evidence F_1 (51.10%) and document recall (66.40%), consistent with improved multi-hop evidence completeness.

5.4 Ablation Study

To isolate the contribution of each component in GAVEL, we evaluate five ablations: removing the Evidence Contract (no atomic subclaim and citation binding for debaters), removing the deterministic evidence checker (Scrutinizer relies only on language-model-based auditing), removing the Scrutinizer (no Chain of Scrutiny; debate proceeds directly to judging), restricting the protocol to a single round (to test the benefit of iterative correction), and forcing shared retrieval outputs for both debaters (to test whether retrieval diversity contributes to multi-hop completeness). Table 2 shows that removing the Evidence Contract causes the largest drop, reducing FEVEROUS score from 41.80% to 36.50% and HOVER score from 49.60% to 44.20%, highlighting that explicit evidence binding is critical for preventing unsupported statements from propagating through debate. Removing the deterministic checker also consistently harms provenance metrics, decreasing FEVEROUS evidence F_1 from 47.20% to 43.10% and HOVER document recall from 66.40% to 62.80%. Finally, removing the Scrutinizer or collapsing to one round reduces performance (e.g., HOVER score drops to 45.50% and 47.00%), suggesting that iterative auditing and binding feedback provide benefits beyond simply having two opposing debaters.

Method	FEVEROUS			HOVER			
	FEVEROUS score ↑	Label Acc ↑	Evidence F_1 ↑	HOVER score ↑	Label Acc ↑	Evidence F_1 ↑	Doc Recall ↑
ZS + CoT	28.50	59.20	33.00	34.80	63.50	38.10	52.00
ProgramFC (N=1)	34.80	63.50	39.50	41.20	67.80	44.30	58.20
ProgramFC (N=5)	<u>37.60</u>	65.10	<u>41.80</u>	44.60	69.20	46.80	60.50
PACAR	36.20	64.80	41.10	43.80	68.90	46.10	60.10
LoCal	37.10	<u>65.60</u>	40.90	<u>45.50</u>	<u>70.40</u>	<u>47.20</u>	<u>61.80</u>
GAVEL (ours)	**41.80**	**67.10**	**47.20**	**49.60**	**72.30**	**51.10**	**66.40**

Table 1: Main results in the open-book setting. We evaluate GAVEL and baselines on FEVEROUS and HOVER under a shared retrieval backend and identical evidence formatting to control for retrieval effects, reporting dataset-specific scores together with label accuracy, evidence F_1, and (for HOVER) document recall, all values are percentages (%). The best and second-best results in each column are shown in **bold** and <u>underlined</u>, respectively.

	FEVEROUS			HOVER			
	Score ↑	Label Acc ↑	Evidence F1 ↑	Score ↑	Label Acc ↑	Evidence F1 ↑	Doc Recall ↑
GAVEL (full)	**41.80**	**67.10**	**47.20**	**49.60**	**72.30**	**51.10**	**66.40**
w/o Evidence Contract	36.50	65.20	40.60	44.20	70.10	46.00	60.10
w/o deterministic checker	38.20	66.10	43.10	46.80	71.00	48.20	62.80
w/o Scrutinizer	37.00	65.60	41.80	45.50	70.60	47.00	61.50
one round only	39.00	66.70	44.70	47.00	71.50	49.20	63.40
shared retrieval for debaters	40.00	66.90	45.50	47.80	71.80	49.90	64.10

Table 2: Ablation results of GAVEL on FEVEROUS and HOVER. All values are percentages (%).

5.5 Fine-grained analysis on FEVEROUS and HOVER

To better understand where GAVEL yields improvements, we analyze performance along two dimensions: evidence type on FEVEROUS and hop count on HOVER. For FEVEROUS, we split claims into those solvable with sentence evidence only versus those requiring table-cell evidence. Table 3 shows that GAVEL improves both subsets, with larger gains on table-required claims (38.90% vs 33.70%) than on text-only claims (46.80% vs 43.50%), consistent with the benefit of citation validity checks and evidence-linked atomic subclaims. For HOVER, we stratify claims by hop count. Table 4 shows consistent gains that grow with reasoning depth, with improvements of 2.40% at 1 hop (58.40% vs 56.00%) and 4.80% at 4 hops (41.20% vs 36.40%), suggesting that adversarial evidence seeking and audit-log-driven correction help identify missing intermediate documents and complete evidence chains.

Method	Text only Score ↑	Table required Score ↑
ProgramFC (N=5)	43.10	33.40
LoCal	43.50	33.70
GAVEL (ours)	**46.80**	**38.90**

Table 3: FEVEROUS score breakdown by evidence type. All values are percentages (%).

Method	1 hop	2 hop	3 hop	4 hop
ProgramFC (N=5)	55.20	47.10	40.60	35.80
LoCal	56.00	48.30	42.20	36.40
GAVEL (ours)	**58.40**	**51.60**	**46.10**	**41.20**

Table 4: HOVER score by hop count. All values are percentages (%).

5.6 Case Study

We present two concrete examples in Appendix B (one from FEVEROUS and one from HOVER) to illustrate how GAVEL addresses common provenance failures: (i) unsupported inference under entity or date confusion and (ii) missing-hop evidence chains in multi-document verification.

5.7 Efficiency and cost

We report efficiency in terms of language model calls and retrieval calls per claim. Compared to single-agent baselines, GAVEL incurs additional cost due to multi-agent interaction. In practice, cost is controlled by a maximum of three rounds and early stopping when the Scrutinizer reports no new violations. The deterministic checker adds negligible overhead because it is linear in the number of cited evidence units and requires no additional model calls.

Method	LLM calls/claim ↓	Retrieval calls/claim ↓
ZS + CoT	1	1
ProgramFC (N=1)	6	3
ProgramFC (N=5)	30	15
PACAR	10	5
LoCal	14	6
GAVEL (ours)	18	6

Table 5: Efficiency summary in terms of average LLM calls and retrieval calls per claim under each method's protocol.

5.8 Qualitative error analysis

We manually inspect a sample of errors to identify remaining failure modes. On FEVEROUS, a common issue is schema mismatches in tables, such as selecting a cell from the correct row but the wrong column due to ambiguous headers, or failing to retrieve the table that contains the relevant cell. On HOVER, most errors are due to missing intermediate documents, where systems retrieve a semantically related page but fail to bridge to the correct entity page needed for the final evidence chain. GAVEL reduces unsupported reasoning errors through the Evidence Contract, but it can still fail when retrieval does not surface the necessary intermediate page, or when the claim requires implicit normalization of entities or quantities that are not explicitly stated in evidence units.

6 Conclusion and Future Work

We introduced GAVEL, a multi-agent debate framework for evidence-grounded claim verification on FEVEROUS and HOVER. GAVEL treats evidence grounding as a first-class constraint during inference through an Evidence Contract that binds atomic subclaims to explicit evidence units, and a Mechanized Chain of Scrutiny that audits debate outputs and validates citations with deterministic checks. Our experiments show that enforcing provenance constraints and structured auditing improves provenance-aware metrics that require correct labels together with correct and complete evidence, and ablations indicate that the Evidence Contract and mechanized validation play central roles in reducing unsupported reasoning and citation errors.

There are several directions for future work. One direction is to improve retrieval for multi-hop claims by integrating stronger multi-step retrievers and by explicitly modeling intermediate entity discovery. Another direction is to extend mechanized scrutiny beyond citation validity to cover semantic validation for numerical reasoning and table operations. Finally, it would be valuable to study domain transfer beyond Wikipedia and to evaluate robustness under adversarially constructed claims and evidence distractors, while preserving strict provenance requirements.

Limitations

Our work has several limitations.

First, GAVEL increases inference time cost relative to single-pass baselines because it uses multiple agents and may run multiple rounds. Although we cap the number of rounds and include early stopping, the approach may still be impractical for high-throughput applications or strict latency settings.

Second, the approach depends on retrieval quality. When the retriever fails to surface the necessary pages, sentences, or table cells, evidence contracts and auditing cannot recover missing information. This limitation is especially salient for multi-hop HOVER claims that require discovering intermediate entities and for FEVEROUS claims whose key evidence appears in less prominent tables.

Third, our deterministic evidence checker validates citation format and presence in the retrieved context, but it does not guarantee semantic correctness beyond what is encoded in the cited units. As a result, the system can still produce logically coherent but semantically incorrect interpretations of valid evidence, particularly for numerical comparisons, aggregation, or implicit normalization that is not explicitly stated in a single sentence or cell.

Fourth, our experiments use `gpt-3.5-turbo` with deterministic decoding (temperature $= 0$) for all agents, which limits conclusions about generalization to stronger proprietary models or open-weight LLMs. While GAVEL is a protocol-level method that does not depend on model-specific training, the magnitude of gains and the cost-quality trade-off may change with different base models.

Finally, FEVEROUS and HOVER are limited to Wikipedia-based evidence and predefined claim distributions. Performance on these datasets may not directly transfer to other domains, languages, or adversarial settings, and our evaluation does not measure broader social impacts or long-term robustness beyond the benchmark protocols.

Ethics Statement

Our work studies automated fact-checking and evidence-grounded verification. A potential positive impact is improving the transparency and auditability of model predictions by requiring explicit citations. However, systems that generate fact-checking outputs may still produce incorrect labels or misleading evidence selections, which could be misused to lend false credibility to misinformation. We therefore emphasize that such systems should be deployed with clear uncertainty communication, human oversight, and safeguards against overreliance on automated judgments.

Our experiments use publicly available datasets derived from Wikipedia and do not involve user data or personally identifiable information. We do not collect new annotations. Nonetheless, Wikipedia contains biases and coverage gaps, and models trained or evaluated on it may inherit such biases. We encourage future work to assess performance across domains and to evaluate fairness and potential harms in real-world settings where misinformation can affect vulnerable populations.

References

Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. FEVEROUS: Fact extraction and VERification over unstructured and structured information. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Giorgio Barnabò, Federico Siciliano, Carlos Castillo, Stefano Leonardi, Preslav Nakov, Giovanni Da San Martino, and Fabrizio Silvestri. 2023. Deep active learning for misinformation detection using geometric deep learning. *Online Social Networks and Media*, 33:100244.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3563–3578, Bangkok, Thailand. Association for Computational Linguistics.

Miriam Fernandez and Harith Alani. 2018. Online misinformation: Challenges and future directions. In *Companion proceedings of the the web conference 2018*, pages 595–602.

Yi Fung, Kung-Hsiang Huang, Preslav Nakov, and Heng Ji. 2022. The battlefront of combating misinformation and coping with media bias. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Tutorial Abstracts*, pages 28–34, Taipei. Association for Computational Linguistics.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders NLP fact-checking unrealistic for misinformation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5916–5936, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Linmei Hu, Siqi Wei, Ziwang Zhao, and Bin Wu. 2022. Deep learning for fake news detection: A comprehensive survey. *AI Open*, 3:133–155.

Kelvin Jiang, Ronak Pradeep, and Jimmy Lin. 2021. Exploring listwise evidence reasoning with t5 for fact verification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 402–410, Online. Association for Computational Linguistics.

Yichen Jiang, Shikha Bordia, Zheng Zhong, Charles Dognin, Maneesh Singh, and Mohit Bansal. 2020. HoVer: A dataset for many-hop fact extraction and claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3441–3460, Online. Association for Computational Linguistics.

Yiqiao Jin, Yeon-Chang Lee, Kartik Sharma, Meng Ye, Karan Sikka, Ajay Divakaran, and Srijan Kumar. 2023. Predicting information pathways across online communities. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '23, page 1044–1056, New York, NY, USA. Association for Computing Machinery.

Amrith Krishna, Sebastian Riedel, and Andreas Vlachos. 2022. ProoFVer: Natural logic theorem proving for fact verification. *Transactions of the Association for Computational Linguistics*, 10:1013–1030.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17889–17904, Miami, Florida, USA. Association for Computational Linguistics.

Zhenghao Liu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2020. Fine-grained fact verification with kernel graph attention network. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7342–7351, Online. Association for Computational Linguistics.

Jiatong Ma, Linmei Hu, Rang Li, and Wenbo Fu. 2025. Local: Logical and causal fact-checking with llm-based multi-agents. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 1614–1625, New York, NY, USA. Association for Computing Machinery.

Bodhisattwa Prasad Majumder, Sudha Rao, Michel Galley, and Julian McAuley. 2021. Ask what's missing and what's useful: Improving clarification question generation using global knowledge. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4300–4312, Online. Association for Computational Linguistics.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 4551–4558. International Joint Conferences on Artificial Intelligence Organization. Survey Track.

Nedjma Ousidhoum, Zhangdie Yuan, and Andreas Vlachos. 2022. Varifocal question generation for fact-checking. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2532–2544, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Liangming Pan, Xiaobao Wu, Xinyuan Lu, Anh Tuan Luu, William Yang Wang, Min-Yen Kan, and Preslav Nakov. 2023. Fact-checking complex claims with program-guided reasoning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6981–7004, Toronto, Canada. Association for Computational Linguistics.

Sudha Rao and Hal Daumé III. 2019. Answer-based Adversarial Training for Generating Clarification Questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 143–155, Minneapolis, Minnesota. Association for Computational Linguistics.

Arkadiy Saakyan, Tuhin Chakrabarty, and Smaranda Muresan. 2021. COVID-fact: Fact extraction and verification of real-world claims on COVID-19 pandemic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2116–2129, Online. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 624–643, Online. Association for Computational Linguistics.

James Thorne and Andreas Vlachos. 2018. Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3346–3359, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Lin Tian, Xiuzhen Zhang, and Jey Han Lau. 2023. Metatroll: Few-shot detection of state-sponsored trolls with transformer adapters. In *Proceedings of the ACM Web Conference 2023*, WWW '23, page 1743–1753, New York, NY, USA. Association for Computing Machinery.

Xiaoyan Zhao, Lingzhi Wang, Zhanghao Wang, Hong Cheng, Rui Zhang, and Kam-Fai Wong. 2024. PACAR: Automated fact-checking with planning and customized action reasoning using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12564–12573, Torino, Italia. ELRA and ICCL.

A Reproducibility Details

This appendix specifies data processing, retrieval, prompting, parsing, and deterministic checking used to reproduce GAVEL.

A.1 Datasets and preprocessing

We follow the official dataset releases and evaluation scripts for FEVEROUS and HOVER. For FEVEROUS, we use the provided Wikipedia snapshot and structured annotations that identify sentence evidence and table-cell evidence. For HOVER, we use the Wikipedia snapshot aligned with the dataset and the document-level provenance annotations.

Sentence segmentation. For each Wikipedia page, we split text into sentences using a standard English sentence segmenter. We store sentence offsets and assign a zero-based sentence index s within each page title p, yielding sentence units $\langle p, s \rangle$.

Table extraction. For FEVEROUS-style evidence, we parse Wikipedia tables and normalize each table into a grid with a table identifier t per page. Each cell is assigned a zero-based row index r and column index c, producing cell units $\langle p, t, r, c \rangle$. Header rows are included in the table grid and can be cited. When a table contains multi-row headers, we keep the original row structure rather than collapsing headers.

Evidence candidate representation. To provide a unified evidence interface to agents, we convert retrieved candidates into a structured list of items. For FEVEROUS, candidates may include both sentence items and cell items; for HOVER, candidates include sentence items only:

- **Sentence item**: page title, sentence index, and raw sentence text.
- **Cell item (FEVEROUS only)**: page title, table id, row, column, and raw cell text, plus optional row and column header strings when available.

A.2 Retrieval pipeline

We use a shared retrieval backend for all methods, implemented as a two-stage retriever.

Stage 1: page retrieval. We retrieve Wikipedia pages using BM25 over page titles and page text. The query is the claim string x (no external rewriting). We return the top $K_p = 5$ pages.

Stage 2: evidence unit retrieval within pages. Given the retrieved pages, we rank candidate sentences and table cells using BM25 with the same claim query x over:

- sentence text for sentence candidates
- a linearized cell string for cell candidates that concatenates column header, row header (if available), and cell value

We return top $K_s = 25$ sentences across the retrieved pages for both datasets. For FEVEROUS, we additionally return top $K_c = 30$ table cells (after deduplicating identical strings and filtering empty cells). For HOVER, we set $K_c = 0$ and retrieve only sentences. The final candidate set provided to agents is capped at 55 items per claim to control context length.

Rationale for K values. We choose $K_p = 5$ to bound the number of pages while preserving multi-hop coverage in HOVER. We choose $K_s = 25$ and $K_c = 30$ to ensure sufficient table coverage for FEVEROUS while keeping the prompt input within context limits for `gpt-3.5-turbo`. These values are fixed across all methods. For HOVER, we use only sentence candidates (i.e., $K_c = 0$).

A.3 Evidence formatting and interface

All methods receive evidence candidates in the same canonical format.

Sentence identifier. A sentence evidence unit is identified as `(page_title, sent_id)` corresponding to $\langle p, s \rangle$.

Cell identifier. A table-cell evidence unit is identified as `(page_title, table_id, row_id, col_id)` corresponding to $\langle p, t, r, c \rangle$. Cell identifiers are used only for FEVEROUS; HOVER instances use sentence identifiers exclusively.

Agent input context. Each agent receives:

1. claim x
2. candidate evidence list (up to 55 items), each with its identifier and content
3. debate context consisting of prior round outputs and the latest Audit Log (if any)

We truncate evidence content by retaining up to 240 characters per sentence and up to 120 characters per cell value (for FEVEROUS), while always preserving identifiers.

A.4 Prompt templates and constrained outputs

All agents use a short system message that fixes the agent role and constraints, followed by a user message that includes the claim, candidate evidence, and debate context. Each agent must output valid JSON conforming to a fixed schema to enable deterministic parsing and evidence checking. Figure 2 summarizes the role-specific instruction prompts for the four agents in GAVEL. Figure 3 summarizes the constrained JSON schemas used for deterministic parsing and evidence checking.

Affirmative Debater (A) Instruction.
You are the Affirmative Debater A in an evidence-grounded fact-checking debate. Your goal is to argue that the claim is **SUPPORTED** when possible.
Evidence Contract: Every atomic subclaim must cite one or more evidence units from the provided candidates using their identifiers. Do not introduce any factual statement that is not backed by cited evidence.
Input: Claim; evidence candidates (sentences and, for FEVEROUS, table cells with IDs); debate context (opponent output and latest Audit Log).
Output: JSON only, following the Debater schema: `proposed_label`, `atomic_claims` (with entailment + evidence IDs), and `final_evidence`.

Negative Debater (N) Instruction.
You are the Negative Debater N in an evidence-grounded fact-checking debate. Your goal is to argue that the claim is **REFUTED** when possible, otherwise **NOTENOUGHINFO**. For HOVER, both map to **NOTSUPPORTED**.
Evidence Contract: Every atomic subclaim must cite one or more evidence units from the provided candidates using their identifiers. Do not introduce any factual statement that is not backed by cited evidence.
Input: Claim; evidence candidates (sentences and, for FEVEROUS, table cells with IDs); debate context (opponent output and latest Audit Log).
Output: JSON only, following the Debater schema: `proposed_label`, `atomic_claims` (with entailment + evidence IDs), and `final_evidence`.

Scrutinizer (S) Instruction.
You are the Scrutinizer S. You are neutral and **do not retrieve new evidence**. Audit both debaters for evidence validity and reasoning gaps. Treat deterministic checker flags as hard constraints.
Input: Claim; evidence candidates; Debater A JSON; Debater N JSON; deterministic checker flags.
Output: JSON only, following the Scrutinizer schema: violations with required fixes, and `no_new_violations`.

Judge (J) Instruction.
You are the Judge J. Output the final label and a sufficient evidence set. Prefer valid citations and discount evidence flagged by the deterministic checker or left unresolved by the Audit Log. (For FEVEROUS, the label is SUPPORTS or REFUTES or NOTENOUGHINFO. For HOVER, the proposed label is SUPPORTED or NOTSUPPORTED)
Input: Claim; evidence candidates; full debate transcript and Audit Logs.
Output: JSON only, following the Judge schema: label and evidence IDs (up to 10 units).

Figure 2: Instruction prompts for the four agents in GAVEL. Each agent receives the claim and a shared set of retrieved evidence candidates. For FEVEROUS, candidates include sentences and table cells; for HOVER, candidates include sentences only and labels are evaluated in the binary space SUPPORTED vs NOTSUPPORTED. Debaters must satisfy the Evidence Contract, the Scrutinizer audits without new retrieval, and the Judge outputs the final decision.

Parsing and retry policy. We parse agent outputs with strict JSON parsing. If parsing fails or required keys are missing, we issue a repair prompt that asks the model to output valid JSON conforming to the schema and retry at most two times. If output remains invalid after retries, we fall back to an empty evidence set and assign NOTENOUGHINFO for FEVEROUS and NOTSUPPORTED for HOVER to avoid introducing uncontrolled text into evaluation.

A.5 Deterministic evidence checker

The checker validates citations with respect to the retrieved candidate evidence list. It does not perform semantic entailment; it only checks identifier validity and quote grounding.

Identifier existence and range checks. For each cited sentence unit $\langle p, s \rangle$, the checker verifies that page p is among retrieved pages and s is a valid sentence index for that page in our candidate list. For each cited cell unit $\langle p, t, r, c \rangle$, the checker verifies that the referenced table exists in the retrieved page and that (r, c) are valid indices.

Quote match checks. When a debater includes a quoted span in its rationale, the checker validates that the quoted string appears in the cited unit text after normalization. We apply the same normalization to both the unit text and the quoted span:

- lowercase
- collapse multiple whitespace into one space
- strip surrounding punctuation

For table cells, we match against the concatenation of header strings (when available) and the cell value.

Duplicate and conflict checks. We flag duplicate citations that repeat identical units excessively (more than three times in a single response) and

Debater JSON Schema (A and N).

```
{
"proposed_label": "SUPPORTS"|"REFUTES"|"NOTENOUGHINFO",
"atomic_claims": [{ "id": "...", "claim": "...", "entailment": "supports"|"refutes"|"insufficient",
"evidence": [{...evidence unit...}], "rationale": "..." }],
"final_evidence": [{...evidence unit...}]
}
```

Scrutinizer JSON Schema.

```
{
"violations": [{ "agent": "A"|"N", "type": "...", "description": "...", "required_fix": "..." }],
"no_new_violations": true|false
}
```

Judge JSON Schema.

```
{
"label": "SUPPORTS"|"REFUTES"|"NOTENOUGHINFO",
"evidence": [{...evidence unit...}]
}
```

Figure 3: Constrained JSON output schemas used by GAVEL for deterministic parsing and evidence validation. Evidence units are formatted as sentence tuples $\langle p, s \rangle$ and table-cell tuples $\langle p, t, r, c \rangle$ (FEVEROUS only), enabling the deterministic checker to validate identifier existence, index ranges, and quote grounding. For FEVEROUS, the proposed label is SUPPORTS or REFUTES or NOTENOUGHINFO. For HOVER, the proposed label is SUPPORTED or NOTSUPPORTED

flag internal conflicts when the same atomic claim is tagged as `supports` and `refutes` with the same evidence.

Checker outputs. The checker produces a list of flags used by the Scrutinizer, including `invalid_id`, `out_of_range`, `quote_mismatch`, and `duplicate_citation`.

A.6 Hyperparameters and budgets

We use temperature $= 0$ for all agent calls. We cap the number of rounds at $R_{\max} = 3$ and enable early stopping when `no_new_violations` is true. Each agent call uses a maximum generation length of 512 tokens. We cap the number of atomic claims per debater to 6 and cap the final cited evidence set size to 10 units.

A.7 Open-book control

All methods operate in an open-book setting with a shared retrieval backend and identical evidence formatting to control for retrieval effects. Concretely, the same retrieval pipeline (page selection and evidence unit ranking) is used across methods, and all methods are required to output evidence in the same sentence and cell identifier formats described above.

B Case Study Examples

Figure 4 and 5 provide two representative case studies (one from FEVEROUS and one from HOVER) illustrating how GAVEL improves provenance by (i) preventing unsupported inferences under entity/date confusion and (ii) repairing missing-hop evidence chains in multi-document verification. Each example includes the claim, gold label, gold evidence type, a typical baseline failure mode, an excerpt of the GAVEL audit log, and the final evidence-grounded decision.

Case Study 1 (FEVEROUS).
Instance ID: 4329
Claim: George Brown began his Liberal Party leadership in Canada on the first of July in 1867.
Gold label: REFUTES **Challenge:** Entity Disambiguation
Gold evidence (sentence): `George Brown (Canadian politician)`:
"He reorganized the Clear Grit (Liberal) Party in 1857, supporting . . . "
Typical baseline failure: infers the July 1, 1867 date from historical context without citing evidence that explicitly states a leadership start date.
GAVEL Audit Log (excerpt):
unsupported_leap: claim asserts "began leadership on July 1, 1867" but cited evidence only mentions party reorganization in 1857.
required_fix: provide a sentence that explicitly states the claimed start date, or revise the conclusion to match the cited year.
GAVEL final decision: REFUTES with evidence grounded in the 1857 statement.

Figure 4: FEVEROUS example illustrating how the Evidence Contract and scrutiny flag unsupported date/role inferences when citations do not support the atomic claim.

Case Study 2 (HOVER).
Instance ID: `4fb633f3-417f-4600-969f-afe1e356d90a`
Claim: Carnegie Hall Tower is located in the same city as Staten Island.
Gold label: SUPPORTS **Num hops:** 2
Gold evidence (sentences):
(1) `Staten Island`: *"Staten Island is one of the five boroughs of New York City . . . "*
(2) `Carnegie Hall Tower`: *"Carnegie Hall Tower is a 60-story skyscraper located . . . in New York City . . . "*
Typical baseline failure: cites only (2), yielding an incomplete evidence chain and losing HOVER credit despite a correct label.
GAVEL Audit Log (excerpt):
missing_hop: evidence confirms Carnegie Hall Tower is in New York City but does not cite evidence that Staten Island is in the same city.
required_fix: cite a sentence from `Staten Island` stating it is a borough of New York City.
GAVEL final decision: SUPPORTS with both documents cited, satisfying multi-document coverage.

Figure 5: HOVER example illustrating how the Audit Log explicitly requests missing-hop evidence to complete the multi-document chain.