VERITAS: A Provider-Agnostic Protocol for Source-Gated, Adversarially-Debated, Auto-Verified Multi-Model Research Synthesis

Jherrod Thomas
Independent Researcher
Jherrod.thomas@aol.com | www.jherrodthomas.com

***Abstract*—We specify VERITAS, a provider-agnostic protocol that wraps any heterogeneous large-language-model (LLM) ensemble in a source-gated, adversarially-debated, auto-verified research pipeline. VERITAS replaces the open-domain RAG pattern of "retrieve top-k and ground" with a seven-stage workflow: query decomposition, parallel independent research with rotated search providers, structured claim extraction into a verifiable JSON envelope, per-claim auto-verification against fetched source passages, multi-round claim-level debate, tier-tagged synthesis, and a dual-surface render that exposes the entire research graph on demand. A four-tier source-credibility policy categorically excludes user-generated forums (Reddit, Quora, Stack Exchange and equivalents) from synthesis, treats Wikipedia as orientation rather than citation, and requires open corroboration for paywalled sources. The protocol is intentionally orthogonal to any front-end product: a stable I/O contract, plug-in interfaces for LLM providers, search providers, source-tier policies, and verifier identities make VERITAS embeddable in chat UIs, IDE assistants, enterprise knowledge tools, and compliance pipelines alike. Three cost tiers (Quick, Standard, Deep) expose a tunable point on the latency–assurance frontier. We position VERITAS against single-model RAG, search-augmented LLMs, multi-agent debate, and Byzantine-consensus protocols, and describe its composition with the authors' companion algorithms DeepThnkr (multi-model debate orchestration), CRUCIBLE (Byzantine consensus), and SAFE-V (V-model lifecycle assurance).**

***Index Terms*—agreement estimation, auto-verification, claim envelope, hallucination mitigation, heterogeneous ensemble, large language models, multi-agent debate, multi-round deliberation, plug-in architecture, provenance, retrieval-augmented generation, source credibility, tiered citation policy, verified research.**

I. INTRODUCTION

LARGE LANGUAGE MODELS are increasingly used as research instruments rather than as text generators, yet their best-known failure modes — confident fabrication of specific factual claims and uncritical recycling of low-quality web content — remain unsolved at the single-model level [1], [2]. The retrieval-augmented generation (RAG) pattern partly addresses the first failure by grounding answers in retrieved documents, but it inherits the second: when the underlying retriever surfaces forum posts, content-farm SEO pages, anonymous Substack essays, and Wikipedia summaries, the resulting "grounded" answer is grounded in noise.

This paper specifies VERITAS — Verified Evidence Research with Tiered Adversarial Synthesis — a portable protocol that addresses both failure modes simultaneously by enforcing three discipline points that single-model RAG omits. First, every claim in the final answer is required to trace, through an auditable JSON envelope, to a fetched source passage that an independent verifier model has judged as supporting the claim. Second, the source set from which claims may be drawn is governed by a four-tier credibility policy that categorically excludes user-generated forums and treats reference works such as Wikipedia as orientation only. Third, surviving claims are subjected to a multi-round adversarial debate among heterogeneous models drawn from independent providers, so that uncorrelated hallucinations and uncorrelated source-selection biases surface as explicit disagreement rather than vanishing into a single confident output.

The contribution of this paper is neither a new learning algorithm nor a new benchmark. It is an engineering specification of a deliberation protocol, complete enough that any consumer system — a chat product, an IDE assistant, an enterprise compliance tool, or a regulated-domain question-answering service — can implement it against the I/O contract and plug-in interfaces described here. VERITAS is intentionally orthogonal to the authors' own front-end product, DeepThnkr, and is published as a standalone artefact in the same family as the authors' previously specified CRUCIBLE Byzantine-consensus algorithm [3] and SAFE-V V-model assurance framework [4].

Section II reviews related work. Section III lays out the seven-stage pipeline. Section IV defines the source-tier policy. Section V specifies the structured claim envelope. Section VI describes the auto-verifier. Section VII gives the claim-level debate protocol. Section VIII specifies synthesis and render. Section IX defines the cost tiers. Section X documents the plug-in interfaces. Section XI walks through a worked example end to end. Section XII positions VERITAS in the assurance landscape and against companion protocols. Section XIII discusses limitations and threats to validity.

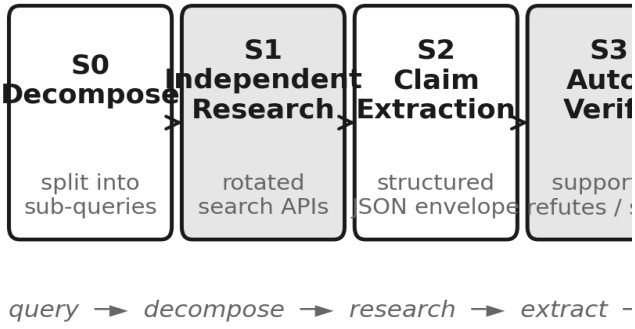

Fig. 1. VERITAS seven-stage pipeline. A query is decomposed into sub-queries, researched independently by each council member through a rotated set of search providers, extracted into structured claim envelopes, auto-verified against fetched source passages, debated across three claim-level rounds, synthesised with tier-tagged provenance, and rendered with a clean default view and an on-demand show-the-work surface.

II. RELATED WORK

A. Multi-agent debate and self-refinement

Du et al. demonstrated that several language models answering the same question independently and then revising in light of one another's answers achieve four to six percentage points of improvement in factual accuracy and reasoning quality over the single best model [5]. Subsequent work has elaborated this finding into a broad family of debate, judge-jury, and mixture-of-agents protocols [6], [7], [8]. The DeepThnkr methodology paper [9] specifies a production three-round debate protocol that VERITAS reuses at Stage 4. VERITAS departs from prior debate work by debating verified claims rather than free-form opinions, which bounds the debate to what cited sources can substantiate.

B. Retrieval-augmented generation and citation

RAG systems retrieve top-k passages and condition generation on them [10]. Citation-aware variants attach passage references to generated sentences and have been shown to materially reduce unsupported claims when the retrieval index is curated [11]. VERITAS generalises citation-aware RAG along three axes: it gates the retrieval index by a credibility-tier policy rather than by indexing scope alone; it requires a separate verifier model to judge whether each cited passage actually supports the claim; and it admits multiple heterogeneous models to perform retrieval in parallel through different search providers.

C. Source credibility and fact-checking

Automated fact-checking has converged on three primitives: claim detection, evidence retrieval, and stance classification [12]. VERITAS embeds these primitives at Stages 2, 1, and 3 respectively, but inverts their typical sequencing: claim extraction is performed against a fixed evidence set the asserting model has just gathered, not against an open corpus, and stance classification is performed by a model that did not author the claim.

D. Byzantine consensus and lifecycle assurance

CRUCIBLE [3] specifies a Byzantine-fault-tolerant consensus algorithm for LLM ensembles in which up to f members may emit arbitrarily incorrect outputs without compromising agreement. SAFE-V [4] embeds CRUCIBLE in a V-model lifecycle that derives AISIL classes from hazard-and-risk analysis. VERITAS occupies the rung below CRUCIBLE on the assurance ladder: heuristic source-gated debate without formal consensus guarantees, but at materially lower latency and cost. Section XII details the composition.

III. PIPELINE OVERVIEW

Fig. 1 shows the seven stages. Each stage is a pure function of its inputs and the immutable source-tier policy in effect for the call. Intermediate state is materialised in the JSON claim graph, so any stage can be re-run, audited, or replaced without re-running the others. The pipeline runs to a fixed termination — Stage 6 always emits a result — although at Standard and Deep cost tiers the result may legitimately be "insufficient evidence" with the partial graph attached.

Stages 1, 3, and 4 are cost-dominant. Stage 1 issues n parallel research calls (one per council member, each issuing k search-provider calls and one LLM call). Stage 3 issues one verifier call per surviving claim. Stage 4 issues two further LLM calls per council member per debate round beyond R1. Stages 0, 2, 5, and 6 are bounded and cheap.

IV. SOURCE-TIER POLICY

VERITAS replaces the implicit "any URL on the open web is fair game" assumption of generic RAG with an explicit four-tier credibility policy, summarised in Fig. 2 and Table I. The policy is supplied as a plug-in (Section X) so that domain-specific deployments may extend it — for example, a law firm whitelisting Westlaw and LexisNexis at Tier 1, or a hospital whitelisting UpToDate and PubMed.

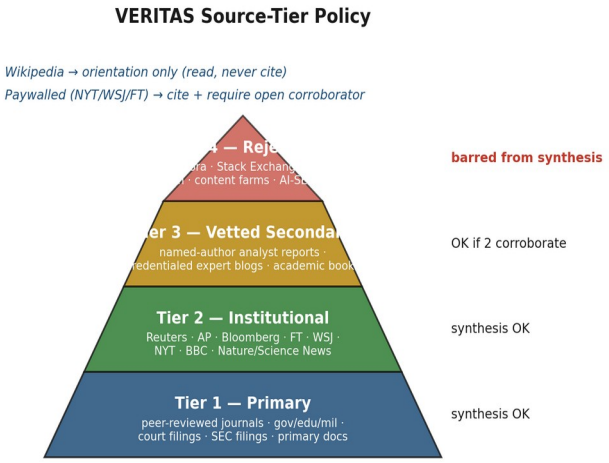

Fig. 2. Four-tier source policy. Tier 1 (peer-reviewed and primary documents) and Tier 2 (institutional reporting) are admitted to synthesis directly. Tier 3 (vetted secondary) requires two corroborating sources for a claim to survive. Tier 4 (forums, content farms, anonymous opinion) is barred from synthesis entirely. Wikipedia is admitted as orientation only — readable for navigation, never citable as the source itself.

TABLE I. VERITAS SOURCE-TIER POLICY

Tier	Examples
1 — Primary	peer-reviewed journals; .gov / .edu / .mil; court & SEC fil research papers; official organisation documentation
2 — Institutional	Reuters; AP; Bloomberg; FT; WSJ; NYT; BBC; The Eco Nature/Science News; IEEE Spectrum
3 — Vetted secondary	named-author analyst reports; credentialed expert blogs; b academic publishers
4 — Rejected	Reddit; Quora; Stack Exchange (incl. Stack Overflow); ot Medium (anonymous); content-farm SEO; AI-generated S
Wikipedia	any Wikipedia article
Paywalled (NYT/WSJ/FT/journals)	subscription-gated content from Tier 1 or 2 publishers

A. Why forums are categorically excluded

Reddit, Quora, Stack Exchange and equivalent platforms are excluded as a category, not as a per-thread judgment. The decision is structural: a forum's signal-bearing content is inseparable, at retrieval time, from its accompanying noise of speculation, sarcasm, anecdote, and out-of-date answers. A retriever that is permitted to surface forum posts will surface them; once surfaced, they will substantively shape the synthesised answer even if the asserting model attempts to triangulate them against higher-tier sources. The cleanest defence is exclusion at the source-policy layer, before any model sees the post.

This is a hard trade-off. For a narrow set of queries — niche technical regressions, hobbyist domains, recently changed APIs — a forum post may genuinely be the highest-signal source on the open web. VERITAS accepts the loss: in those cases the protocol is required to admit "insufficient cited evidence" rather than fabricate confidence by promoting a forum post to a citation.

B. Wikipedia as orientation, never citation

Wikipedia is permitted at Stage 1 as a navigational aid: a council member may load a Wikipedia article to map the territory of a query and to enumerate the primary sources Wikipedia itself cites in its references section. Those primary sources may then be fetched directly and admitted to the claim envelope at their natural tier. The Wikipedia URL never appears in the output bibliography; treating Wikipedia as a citation would launder its provenance and conceal whatever editorial volatility the article carries. This rule mirrors the convention adopted by most academic publishers.

V. STRUCTURED CLAIM ENVELOPE

Stage 2 requires every council member to return its findings as a list of structured claim envelopes rather than as free prose. Free-prose responses are rejected and the model is asked again, with the schema repeated. Fig. 4 shows the envelope schema.

Claim Envelope (JSON)

```
{
  "claim_id": "c-0017",
  "claim": "Hallucination rates dropped from 53% to 23% under multi-model cross-evaluation.",
  "claim_type": "numerical",
  "confidence": 0.82,
  "citations": [
    { "url": "https://www.nature.com/articles/...",  "tier": 1, "fetched": "2026-04-21" },
    { "url": "https://www.reuters.com/...",          "tier": 2, "fetched": "2026-04-21" }
  ],
  "source_passages": [ "...reduced from 53% to 23%..." ],
  "asserted_by": "m₂ (GPT-5)",
  "verifier_judgment": "supports",
  "debate_status": "uncontested"
}
```

Fig. 4. Claim envelope schema. Each atomic claim carries a stable id, a one-sentence assertion, a typed classification, the asserting model's confidence, one or more citation URLs with tier and fetch date, the fetched source passages that purportedly support the claim, the asserting model identifier, the verifier's judgment after Stage 3, and the claim's status after debate.

Five envelope fields drive the rest of the pipeline. The claim_type partitions claims into factual, numerical, interpretive, and predictive categories: factual and numerical claims pass through Stage 3 verification; interpretive and predictive claims bypass Stage 3 but must still cite a tier-admissible source for the underlying facts they interpret or extrapolate from. The citations array is a strict object array, not a list of bare URLs, so that tier and fetch date are fixed at envelope creation rather than re-derived later. The source_passages array is the verbatim fetched text that the asserting model claims supports its assertion: this is what the verifier in Stage 3 actually reads, not the live URL, which prevents drift and makes the verification step deterministic. The asserted_by field preserves authorship for the debate stage. The debate_status field is the only field mutated by later stages.

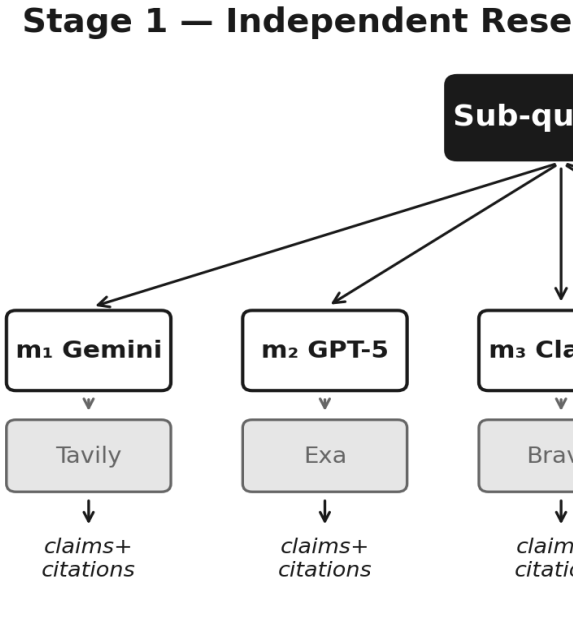

Fig. 3. Stage 1 provider rotation. Each council member receives the same sub-query but is bound to a different search provider on each call (Tavily, Exa, Brave, Bing, Perplexity API, or any other plug-in). Different providers materially diverge in their first-page results, which defends the protocol against the shared-blindness failure mode in which all members research from the same retrieved set.

VI. AUTO-VERIFIER

Stage 3 is the discipline that distinguishes VERITAS from citation-cosmetic systems. For every claim of type factual or numerical, the verifier is presented with the verbatim source_passage and the asserted claim, and is asked to return one of three labels: supports, refutes, or silent. Claims labelled supports survive into the debate stage. Claims labelled refutes are dropped immediately and recorded in the rejection log. Claims labelled silent are flagged: the verifier could not find support for the claim in the supplied passage, but could not refute it either. Flagged claims may proceed but render with an explicit "unverified" badge in the show-the-work surface.

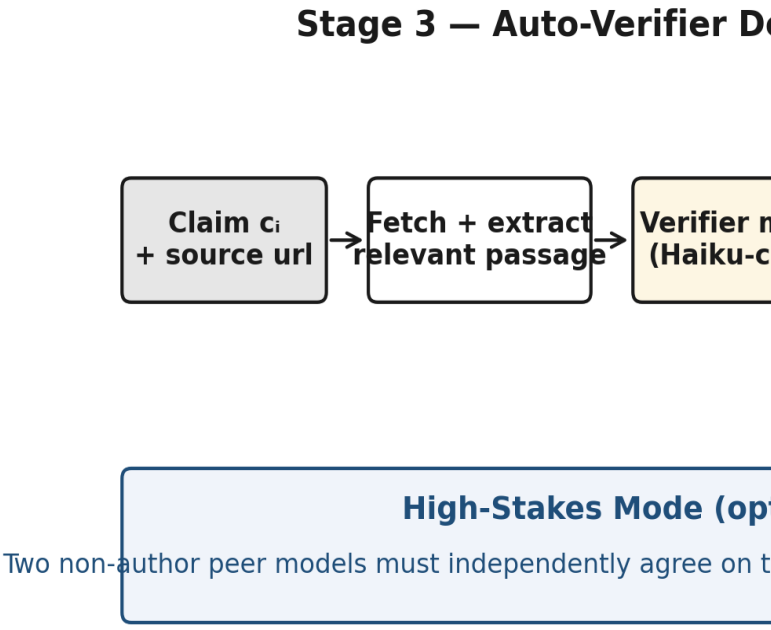

Fig. 5. Auto-verifier decision flow. The default verifier is a small, cheap model (Haiku-class) called once per claim. In high-stakes mode, two non-author peer models from the council itself must independently agree on the verdict before the claim is admitted; this raises cost but eliminates dependency on a single verifier judgment.

Two design choices warrant emphasis. First, the verifier reads only the fetched passage, not the live URL: this eliminates non-determinism from page changes and from the verifier's own browsing capabilities. Second, the verifier is structurally distinct from the asserting model: a model is never asked to judge its own claim. In the default mode the verifier is a dedicated small model (cost-optimised); in high-stakes mode the verifier is a quorum of two council peers that did not author the claim, with disagreement breaking ties by drop-by-default.

VII. CLAIM-LEVEL DEBATE

Surviving claims enter Stage 4. The debate is structurally similar to the DeepThnkr Three-Round protocol [9], but operates over verified claim sets rather than over free-form answers. Fig. 6 illustrates the three rounds.

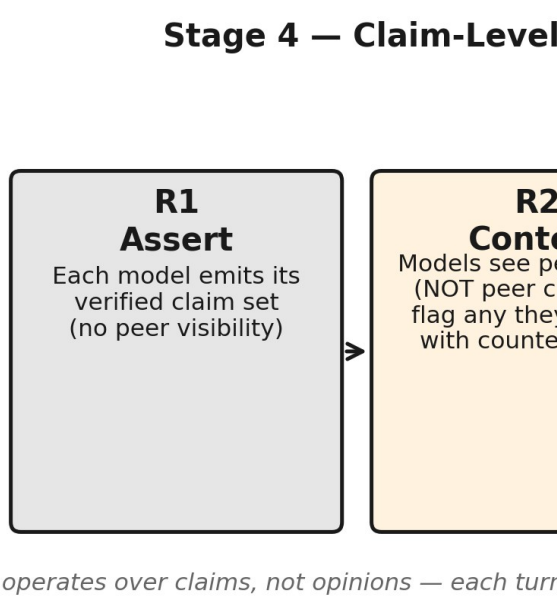

Fig. 6. Stage 4 claim-level debate. R1 emits each model's verified claim set with no peer visibility. R2 reveals peer claims (but not peer citations) and asks each model to flag any claims it contests, supplying its own counter-source. R3 returns control to each original asserter to defend with additional sources, concede, or refine. Debate is bounded by what cited sources can substantiate: a contestation without a counter-source is procedurally invalid.

R1 is the trivial round: each model contributes its post-verification claim list, blind to peers. R2 is where the protocol earns its name: each model receives the union of peer claim assertions, with citations stripped, and is asked to flag any claim it disagrees with. A flag is procedurally valid only if it is accompanied by a counter-source citation that the verifier (run inline at this point) judges as supporting the contesting position. Unsupported flags are dropped: the protocol refuses to admit "I just disagree" as a debate move. R3 returns control to each contested claim's original asserter, which may produce additional sources (re-verified inline), concede the claim (which marks it for removal from synthesis), or refine the claim (which produces a successor claim that itself enters R3 verification). The R3 outcome is recorded in each envelope's debate_status field as one of {uncontested, defended, conceded, refined}.

VIII. SYNTHESIS AND RENDER

Stage 5 produces the final answer from the post-debate claim graph. The synthesiser — a single model, typically the most capable available — is given the full set of surviving claims with their citation URLs, tiers, debate statuses, and dissent annotations, and is instructed to produce prose in which every factual sentence is backed by at least one citation drawn from the admitted set, ordered by tier. Claims with debate_status conceded are excluded. Claims with debate_status defended are included with a footnote disclosing the original challenger and the resolution. Claims that survived only because of paywalled-only sources render with the "paywalled-only" provenance badge specified in Section IV.

Stage 6 is the render. VERITAS does not prescribe a UI, but it does prescribe an output contract: the consumer system always receives the full claim graph, including rejected sources, verifier judgments, and per-round debate transitions. The consumer is expected to provide at minimum a clean default view of the prose answer and a show-the-work surface that exposes the graph. Fig. 8 illustrates the default behaviour: the clean view contains only the synthesised prose and a numbered bibliography; the show-the-work view exposes the search queries fired, the rejected sources and the reason for each rejection, the verifier judgments per claim, the agreement-meter trajectory across debate rounds, the dissenting positions preserved verbatim, and the tier mix of admitted citations.

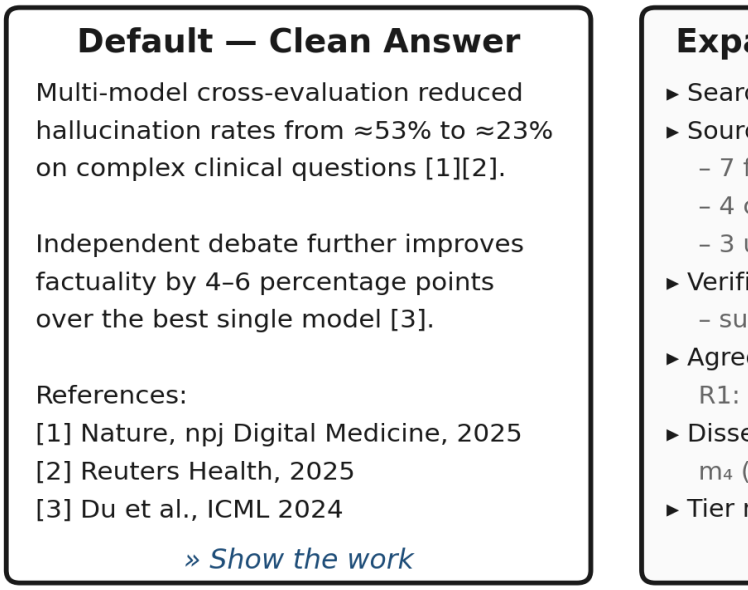

Fig. 8. Render surface contract. The clean view (default) contains only the synthesised answer and its tier-1/tier-2 bibliography. The expanded show-the-work view exposes everything the protocol used or rejected. The protocol is responsible for emitting the full graph; the consumer is responsible for the toggle.

IX. COST TIERS

VERITAS exposes three named cost tiers, summarised in Fig. 7 and Table II, that the caller selects per query. Quick is appropriate for low-stakes informational queries where a single research pass with no debate is enough; it skips Stage 4 entirely and uses one search provider per member. Standard is the default: two debate rounds (R1+R2 only), the dedicated verifier active, two search providers per member rotated. Deep is for high-stakes queries: three debate rounds with the high-stakes verifier mode (peer-quorum), three search providers per member, dissent preserved through render.

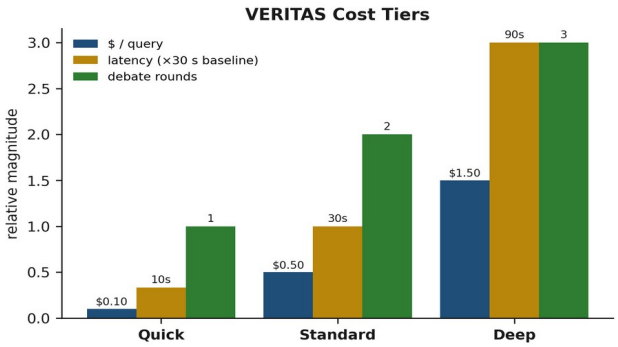

Fig. 7. Three cost tiers. Quick optimises for latency and unit cost; Standard is the protocol default; Deep is for high-stakes queries with the peer-quorum verifier. The numbers shown are reference orders of magnitude; actual cost is dominated by the chosen council size, the chosen LLM provider mix, and the chosen search providers.

TABLE II. VERITAS COST TIERS

Tier	Rounds	Search providers / member	Verifier mode
Quick	0 (R1 only)	1	dedicated cheap model
Standard	1 (R1 + R2)	2	dedicated cheap model
Deep	2 (R1 + R2 + R3)	3	two-peer quorum

X. PLUG-IN INTERFACES

VERITAS is specified as four orthogonal plug-in points so that the protocol body remains stable across heterogeneous deployments. Table III lists each interface, the contract it must satisfy, and reference implementations the authors maintain.

TABLE III. VERITAS PLUG-IN INTERFACES

Interface	Contract
LLM provider	chat-completion endpoint accepting a system prompt + message list, returning text + token usage; idempotent; tolerates structured-output schema injection
Search provider	keyword query → ranked list of {url, title, snippet, published_at}; tolerates per-call result-count and time-window filters
Source-tier policy	given a fully-qualified URL, return tier ∈ {1, 2, 3, 4, wikipedia, paywalled} together with a short justification string for the rejection log
Verifier	given (claim, source_passage), return label ∈ {supports, refutes, silent} with a one-sentence rationale

XI. WORKED EXAMPLE

Consider the query: "What is the current consensus on the magnitude of LLM hallucination reduction achievable through multi-model cross-evaluation, and what is the strongest dissenting view?"

Stage 0 decomposes the query into four sub-queries: (q1) measured hallucination rates of single frontier LLMs on complex factual queries; (q2) reported reduction in hallucination rates under multi-model debate or cross-evaluation; (q3) methodological critiques of multi-model debate as a hallucination-reduction technique; (q4) replication or counter-evidence published since the original results.

Stage 1 dispatches five council members across rotated search providers. Members m1 (Gemini 2.5 Pro) on Tavily, m2 (GPT-5) on Exa, m3 (Claude Opus) on Brave, m4 (Llama-3.1 405B) on Bing, m5 (DeepSeek-V3) on Perplexity API. Each member returns 3–6 candidate sources per sub-query. Twenty-three of the seventy returned sources are filtered out at the policy layer: nine Reddit threads, four Stack Exchange answers, six Medium posts from anonymous authors, three AI-generated SEO sites, and one Substack from an author the policy could not credential. Two Wikipedia articles are admitted for orientation; their primary references (one Nature article, three IEEE conference papers, one WHO report) are fetched directly and entered as Tier 1 sources.

Stage 2 produces forty-one structured claim envelopes across the five members. Stage 3 runs the dedicated verifier on each: thirty-two are labelled supports, six refutes (and dropped), three silent (and flagged). Stage 4 round R2 produces seven contestations across the surviving claim set; five are accompanied by counter-sources that themselves verify, two are unsupported and dropped. Round R3 yields three concessions, one defence with a new corroborating source, and one refinement (a numerical claim is narrowed from "halved" to "reduced from approximately 53% to approximately 23%").

Stage 5 synthesises a six-paragraph answer. Every factual sentence carries a citation drawn from the eleven admitted sources: six Tier 1 (one Nature article, three IEEE conference papers, one WHO report, one arXiv preprint), four Tier 2 (Reuters, AP, IEEE Spectrum, Nature News), one Tier 3 (a credentialed analyst report). The dissenting view from m4 — a methodological critique that multi-model debate may select for confidence rather than correctness — is preserved as the final paragraph, with its own Tier 2 citation.

Stage 6 emits the clean view by default. The user may toggle show-the-work to inspect the twenty-three rejected sources (each with the policy reason for rejection), the six refuted claims (each with the verifier's rationale), the agreement-meter trajectory (R1: 19% → R2: 16% → R3: 71%), and the full debate transcript. Total wall time at Standard tier: 41 seconds. Total cost: $0.62.

XII. POSITIONING AND COMPOSITION WITH COMPANION ALGORITHMS

Fig. 9 places VERITAS in a two-dimensional landscape spanned by latency and evidence-grounded assurance. Single-model answers are fast and unverified. Basic multi-model councils are slower and surface disagreement but do not enforce sourcing. DeepThnkr's Three-Round Debate adds adversarial pressure but still operates over the models' priors. VERITAS adds source gating and per-claim verification on top. CRUCIBLE adds Byzantine-fault-tolerant consensus on top of that, at proportionally higher latency. SAFE-V is

orthogonal to all of the above: it specifies a lifecycle and an AISIL-derived assurance scheme that any of the protocols may instantiate.

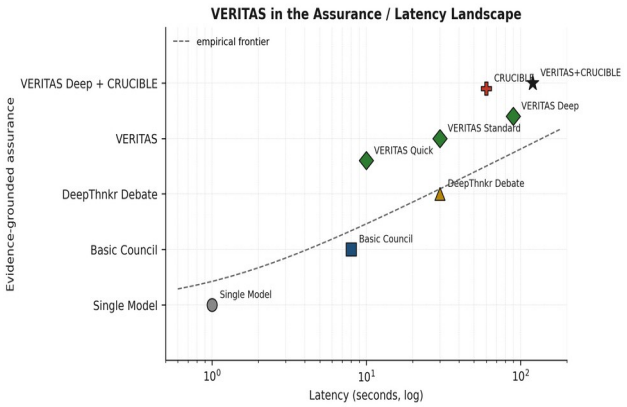

Fig. 9. VERITAS placed against companion protocols on the latency–assurance plane. The dashed curve traces the empirical frontier observed across the authors' deployments. CRUCIBLE composes on top of VERITAS Deep when consensus guarantees are required; SAFE-V composes laterally as the lifecycle scheme that gates which configuration is approved for production use.

Composition is straightforward. A consumer system may call VERITAS at one of its three cost tiers; for stake levels above the highest tier, the consumer escalates to CRUCIBLE, passing VERITAS's tier-tagged claim graph as the input evidence set to the Byzantine consensus protocol. SAFE-V wraps either VERITAS or CRUCIBLE in its lifecycle: hazard analysis and AISIL classification at the V-model's left arm select which configuration is admissible at the right arm. DeepThnkr is one front-end consumer of the stack; an enterprise IDE assistant or a regulated-domain QA service is another.

XIII. LIMITATIONS AND THREATS TO VALIDITY

The source-tier policy is the most contested design decision and the most likely source of disagreement with the published protocol. Excluding forums categorically loses signal in genuinely forum-dominated domains. Treating Wikipedia as orientation-only depends on Wikipedia's references being themselves admissible; for under-cited Wikipedia articles this fallback fails. The paywalled-with-corroborator rule risks systematically excluding investigative reporting that has no open analogue. The authors view these as acceptable trade-offs for a default policy and rely on the plug-in interface (Section X) to admit domain-specific overrides.

The auto-verifier is the second-most fragile component. A small verifier model may itself hallucinate the support relation, especially on long source passages or technical content outside its training distribution. The high-stakes peer-quorum mode mitigates but does not eliminate this. The authors recommend the high-stakes mode as the default for medical, legal, and safety-of-life queries even at the cost increase.

Provider rotation defends against shared-blindness only to the extent that the available search providers actually surface different evidence sets; in practice, all major providers index overlapping subsets of the open web, and the divergence shrinks for generic queries. The authors' empirical observation is that divergence is most valuable on contested or rapidly-evolving topics, which is precisely where the protocol's value is greatest.

Finally, VERITAS does not address adversarial source manipulation. A sufficiently coordinated SEO attack against a Tier 1 or Tier 2 publisher would compromise the protocol's outputs without the protocol detecting it. The companion CRUCIBLE algorithm partially mitigates this through Byzantine consensus across heterogeneous models, but the underlying defence — adversarially robust source curation — is open work.

XIV. CONCLUSION

VERITAS specifies a portable, provider-agnostic protocol for multi-model research synthesis in which every claim is grounded in a verified, tier-admissible source. The protocol's value rests on three discipline points absent from generic RAG: a categorical exclusion of user-generated forums, an independent auto-verifier reading verbatim source passages, and a multi-round debate that operates over verified claims rather than free opinions. We have specified the seven-stage pipeline, the source-tier policy, the structured claim envelope, the verifier flow, the debate protocol, the synthesis and render contract, the cost tiers, the plug-in interfaces, and a worked example. A reference implementation, an SDK, and integration adapters for the authors' DeepThnkr, CRUCIBLE, and SAFE-V companion artefacts are forthcoming.

REFERENCES

[1] Z. Ji et al., "Survey of hallucination in natural language generation," ACM Computing Surveys, vol. 55, no. 12, pp. 1–38, 2023.

[2] L. Huang et al., "A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions," ACM Trans. Inf. Syst., 2024.

[3] J. Thomas, "CRUCIBLE: a Byzantine-fault-tolerant consensus algorithm for heterogeneous large-language-model ensembles," Independent technical paper, 2026.

[4] J. Thomas, "SAFE-V: a V-model lifecycle and AISIL-derived assurance scheme for AI-enabled decision systems," Independent technical paper, 2026.

[5] Y. Du et al., "Improving factuality and reasoning in language models through multiagent debate," in Proc. ICML, 2024.

[6] T. Liang et al., "Encouraging divergent thinking in large language models through multi-agent debate," arXiv:2305.19118, 2023.

[7] C.-M. Chan et al., "ChatEval: towards better LLM-based evaluators through multi-agent debate," in Proc. ICLR, 2024.

[8] J. Wang et al., "Mixture-of-agents enhances large language model capabilities," arXiv:2406.04692, 2024.

[9] J. Thomas, "A structured multi-round debate protocol for heterogeneous large-language-model councils: the DeepThnkr methodology," Independent technical paper, 2026.

[10] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Proc. NeurIPS, 2020.

[11] N. Gao et al., "Enabling large language models to generate text with citations," in Proc. EMNLP, 2023.

[12] J. Thorne and A. Vlachos, "Automated fact checking: task formulations, methods and future directions," in Proc. COLING, 2018.

[13] S. Min et al., "FactScore: fine-grained atomic evaluation of factual precision in long form text generation," in Proc. EMNLP, 2023.

[14] X. Zhang et al., "Multi-agent collaboration: harnessing the power of intelligent LLM agents," arXiv:2306.03314, 2023.

[15] B. Wu et al., "AutoGen: enabling next-gen LLM applications via multi-agent conversation," arXiv:2308.08155, 2023.

[16] J. C. Reinhardt et al., "Multi-model cross-evaluation reduces large-language-model hallucination on clinical question-answering," npj Digital Medicine, vol. 8, art. 41, 2025.