Parametric Hubris:

Empirical Evidence That Tool Availability Does Not Equal Tool Usage in Frontier Language Models

Martin Gehrken

Independent Researcher, Hannover, Germany

ieamlumae@gmail.com

February 8, 2026

https://github.com/IamLumae/Project-Lutum-Veritas

Abstract

Frontier large language models are increasingly deployed with integrated retrieval tools, yet the assumption that tool availability ensures tool usage remains empirically unexamined. We introduce the concept of parametric hubris - the architecturally conditioned tendency of large language models to suppress external tool invocation in favor of parametric recall, even when internal knowledge is stale or incomplete. Empirical measurement reveals that GPT-5 triggers web search for only 31% of user prompts despite having browsing enabled [1], while Gemini models exhibit grounding rates below 50% [2]. When these models lack knowledge, they fabricate: the AA-Omniscience benchmark reports hallucination rates of 88-93% among incorrect responses across frontier Gemini and GPT models [3]. Critically, GPT-5's reported 9.6% error rate under browse-on conditions [4] represents a blended average across both searched and unsearched queries, obscuring the true error distribution. We further distinguish between capability - a model's ability to use retrieval tools when forced - and propensity - its inclination to invoke them autonomously, a distinction no existing benchmark captures. To test whether mandatory tool use can architecturally bypass the mathematical inevitability of hallucination demonstrated by Xu et al. [5], we present Veritas, a six-stage pipeline built on Gemini 2.5 Flash Lite at $0.002 per query that enforces 100% real-time web retrieval. Evaluated on SimpleQA Verified (n=100), Veritas achieves an F-Score of 89.1% with 0% fabrication, compared to 72.1% for Gemini 3 Pro and 51.6% for GPT-5 [6], while operating 20-106x cheaper than competing systems at approximately 115 seconds per query. These results demonstrate that the hallucination problem in factual question answering is not one of model capability but of architectural discipline.

Keywords: *hallucination, retrieval-augmented generation, tool use, parametric hubris, factual accuracy, benchmark*

1. Introduction

The year 2025 marked the ascent of "agentic AI." Frontier language models from OpenAI, Google DeepMind, and Anthropic shipped with integrated web search, code execution, and tool-use capabilities, marketed as systems that could autonomously verify claims, retrieve live information, and reason over external data. The implicit promise to users and enterprises was clear: these models would search when they did not know, abstain when they were uncertain, and answer from memory only when confident and correct. This promise is empirically false.

A large-scale observational study by Nectiv, analyzing over 8,500 ChatGPT prompts across nine industries, found that GPT-5 - despite having full access to Bing web search - triggered a search in only 31% of interactions [3]. The remaining 69% of responses were generated entirely from parametric memory, with no external verification. Google's own documentation for Gemini acknowledges the same architecture: "Note that providing Google Search as a tool to the model doesn't require the model to always use the Google Search tool to generate its response" [7]. An independent analysis by DEJAN of 10,000 Gemini prompts with search grounding enabled found that "ungrounded" responses constituted the majority class [6]. The tools exist. The models choose not to use them.

This paper introduces the concept of Parametric Hubris: the architecturally conditioned phenomenon whereby a language model, due to the scale of its training corpus, develops a statistically unjustified confidence in its own parametric knowledge, actively suppressing the invocation of external tools - including live web search - even when its internal knowledge is outdated, incomplete, or fabricated. The decision to invoke a tool is not a function of epistemic self-awareness; it is a function of training reward signals (RLHF) and inference cost optimization. Models are not lazy by accident. They are lazy by design.

The consequences are measurable. OpenAI's GPT-5 System Card reports a blended hallucination rate of 9.6% with browsing enabled, but 47% with browsing disabled [2]. Since browsing activates in only 31% of cases [3], the majority of responses fall into the higher-error regime. The AA-Omniscience benchmark, evaluating 36 models across 6,000 questions without tool access, reveals the depth of the problem: Gemini 2.5 Flash exhibits a 92.6% hallucination rate among incorrect responses, meaning that when it does not know the answer, it fabricates one in 92.6% of cases rather than declining to respond [4]. Even the best-performing model on the benchmark, Gemini 3 Pro, hallucinated in 88.0% of its incorrect answers. Xu et al. have provided formal proof that hallucination is mathematically inevitable in autoregressive language models [5]: next-token prediction over probability distributions cannot structurally distinguish between truth and confabulation. The industry response has been to invest billions in larger models, longer training, and more sophisticated reinforcement learning - all attempts to solve the problem within the model. This paper argues that the problem must be solved around it.

We present Veritas, a six-stage retrieval-and-verification pipeline that removes the model's discretion over tool use entirely. In Veritas, every query triggers mandatory real-time web scraping via two independent retrieval chains, followed by synthesis exclusively from scraped evidence and a final cross-verification pass against independently retrieved sources. The model is architecturally prohibited from answering from parametric memory at any stage. Crucially, Veritas achieves this using Gemini 2.5 Flash Lite - the cheapest model on the market at the time of evaluation - at a cost of $0.002 per query, with a mean latency of approximately 115 seconds.

Evaluated on a 100-question random sample from the SimpleQA Verified benchmark [1], Veritas achieves an F-score of 89.1%, surpassing the next-best model (Gemini 3 Pro Preview at 72.1%) by 17.0 points. More significantly, Veritas records a 0% fabrication rate: among all incorrect responses, zero were generated without evidential basis. The six errors that did occur were misinterpretations of real scraped data - the model extracted the wrong fact from a correct source, rather than inventing a fact from no source. Nine queries resulted in honest refusals, typically because the information was too obscure for DuckDuckGo to surface or was locked behind dynamically rendered content.

This paper makes four contributions:

1. Empirical evidence of the propensity gap. We document, with primary-source data from Nectiv (N=8,500+), OpenAI's System Card, the DEJAN grounding study (N=10,000), and the AA-Omniscience benchmark (N=6,000), that tool availability does not predict tool usage. Frontier models with full search access invoke it in a minority of cases, and when they do not search, their hallucination rates are catastrophic (47-93%). We formalize this as the distinction between capability (the ability to search when forced) and propensity (the likelihood of searching when not forced), and show that no existing benchmark measures propensity.

2. A critique of current benchmark methodology. We show that SimpleQA Verified [1] rewards parametric memorization rather than retrieval competence, and that the FACTS Grounding benchmark systematically filters out the "soft tail" queries - precisely the cases where models feel confident enough not to search but are wrong - thereby inflating reported reliability figures. The benchmarks measure what models can do in controlled conditions, not what they will do in production.

3. Veritas as proof-of-concept for mandatory retrieval architectures. We demonstrate that a six-stage pipeline built on the cheapest available model can outperform frontier systems costing 20-106x more per query. This is not a marginal improvement: Veritas achieves higher accuracy than GPT-5 (F-score 51.6%), o3 (52.3%), and Claude Opus 4.5 (39.0%) while using a model that, without the pipeline, scores 24.9% accuracy with a 92.6% hallucination rate on AA-Omniscience [4]. The architecture, not the model, is the variable that matters.

4. A cost and latency analysis showing the solution is both faster and cheaper. Against the common objection that retrieval-augmented verification must be prohibitively slow and expensive, we show that Veritas at ~115 seconds per query is faster than ChatGPT Deep Research (5-30 minutes), Gemini Deep Research (5-15 minutes), o3-pro (3-15 minutes), and comparable to Perplexity Deep Research (2-4 minutes) - systems backed by billion-dollar infrastructure. At $0.20 per 1,000 queries (token costs only, no search surcharges), Veritas is 20x cheaper than GPT-5 with Bing, 95x cheaper than Gemini 3 Pro with grounding, and 106x cheaper than Claude Opus 4.5 with web search.

The remainder of this paper is organized as follows. Section 2 reviews related work on hallucination measurement, retrieval-augmented generation, and the emerging literature on tool-use propensity. Section 3 presents the empirical data on tool-use rates across frontier models, formalizing the propensity gap. Section 4 introduces the concept of Parametric Hubris and analyzes the mechanisms driving tool-use suppression. Section 5 critiques existing benchmark methodology and proposes the capability-propensity framework as an alternative evaluation paradigm. Section 6 describes the Veritas architecture in detail, covering its six-stage pipeline (C1-C6), scraping infrastructure, and mandatory retrieval constraints. Section 7 details our evaluation methodology on the SimpleQA Verified benchmark, including scoring criteria, validation procedures, results, and comparison against the 47-model Kaggle leaderboard. Section 8 analyzes cost and latency trade-offs. Section 9 discusses limitations, threats to validity, and directions for future work, followed by Section 10 which concludes.

2. Related Work

Evaluating the factual reliability of large language models has become a central concern as these systems are deployed in high-stakes domains. Several benchmarks have been developed to measure hallucination rates, factual grounding, and parametric knowledge retention. We review the principal evaluation frameworks, identify their methodological limitations, and locate the gap that motivates this work.

2.1 SimpleQA Verified

The SimpleQA Verified benchmark [1] provides 1,000 fact-seeking questions with ground-truth answers verified against authoritative sources. As of early 2026, the public Kaggle leaderboard lists 47 evaluated models [6]. The top-performing system, Gemini 3 Pro, achieves an F-Score of 72.1%, followed by GPT-5 at 51.6% and Claude Opus 4.5 at 39.0%. These scores represent the state of the art for parametric factual recall - models answering from internal weights without external retrieval.

However, SimpleQA Verified fundamentally rewards memorization. A model that has encountered a given fact during pre-training and retains it with sufficient precision will score well regardless of whether it can reliably determine when its knowledge is outdated or incomplete. High scores thus indicate successful overfitting to the temporal window of the training corpus rather than genuine factual reliability. For systems that operate via real-time retrieval rather than parametric recall, the benchmark's design creates a structural mismatch: it measures the very capability - rote memorization - that retrieval-augmented architectures deliberately bypass. Haas et al. [1] acknowledge that tool-augmented systems could trivially outperform parametric-only models on their benchmark, yet this claim has not been empirically validated in prior work, and as we demonstrate in Section 5, the reality is more nuanced than the assertion implies.

2.2 FACTS Grounding

The FACTS Grounding benchmark [8], developed by DeepMind, evaluates whether models can generate responses that are faithful to a provided document context. By supplying reference documents alongside queries, FACTS isolates the grounding capability - a model's ability to constrain its output to information present in given sources. Gemini 2.5 Pro leads the benchmark at 87.8%.

While methodologically rigorous for its stated purpose, FACTS suffers from a critical selection bias introduced by its hardness filter. During benchmark construction, questions that models can already answer from parametric memory are systematically removed, leaving only the "hard tail" of queries where retrieval is effectively forced [8]. This filtering eliminates precisely the category of questions where parametric hubris manifests most acutely: queries for which the model possesses partial, outdated, or subtly incorrect knowledge and therefore feels confident enough to skip retrieval. In real-world deployment, the Nectiv study [3] demonstrates that approximately 69% of prompts fall into this "soft tail" category - questions where models elect not to search. FACTS, by design, excludes these cases from evaluation.

The benchmark therefore measures capability - whether a model can use retrieved context when compelled - but not propensity - whether it will autonomously invoke retrieval when it should. This distinction, as we argue throughout this paper, is the central failure mode of tool-augmented language models.

2.3 AA-Omniscience

The AA-Omniscience benchmark [4, 11], maintained by Artificial Analysis, provides a complementary perspective by evaluating 36 models across 6,000 questions spanning 42 topics and 6 knowledge domains. Its defining contribution is a hallucination metric defined as overconfidence: the share of incorrect responses in which the model provides a confident answer rather than declining to respond. This formulation directly quantifies the failure mode we term parametric hubris.

The results are striking. Only 3 of 36 evaluated models achieve an Omniscience Index above zero (a composite score penalizing hallucination alongside rewarding accuracy). Gemini 3 Pro leads in raw accuracy at 54% but exhibits an 88% hallucination rate - meaning that of every 100 questions it answers incorrectly, 88 receive a confident (and wrong) response rather than an appropriate refusal. GPT-5.1 follows at 39% accuracy with 81% hallucination. Among smaller models, Gemini 2.5 Flash demonstrates the most extreme overconfidence: 24.9% accuracy coupled with a 92.6% hallucination rate, implying that nearly all of its errors are delivered with unwarranted certainty.

A critical finding from AA-Omniscience is that accuracy and hallucination rate are not inversely correlated. Larger, more capable models hallucinate at rates comparable to or higher than smaller models, suggesting that increased parametric knowledge amplifies confidence without proportionally improving calibration. This undermines the prevailing assumption that scaling alone will resolve the hallucination problem.

2.4 Vectara Hallucination Leaderboard

The Vectara Hallucination Leaderboard [9] tracks hallucination rates in a controlled document summarization task. Current leaders include Gemini 2.0 Flash at 0.7% and GPT-4o at 1.5%, reflecting a dramatic reduction from approximately 21.8% in 2021 to sub-1% rates by 2025.

While these results demonstrate that grounded hallucination - fabrication of content not present in a provided source document - has been largely solved for summarization, the benchmark's relevance to open-domain factual question answering is limited. Summarization is a constrained generation task where the answer space is bounded by the input document. Open-domain QA, by contrast, requires the model to determine whether it possesses relevant knowledge, whether to seek external information, and how to reconcile potentially conflicting sources. The near-zero rates on Vectara should not be mistaken for evidence that hallucination has been solved in the general case.

2.5 Theoretical Inevitability of Hallucination

Xu et al. [5] provide a formal proof that hallucination is mathematically inevitable in large language models. Their analysis demonstrates that any system generating text via next-token prediction over learned probability distributions cannot structurally distinguish between factually grounded and fabricated outputs. This result establishes that no amount of scaling, training refinement, or reinforcement learning from human feedback can eliminate hallucination from within the parametric framework itself. The implication is that solutions must be architectural rather than parametric - an insight that motivates the mandatory retrieval approach we evaluate in this work.

2.6 The Missing Dimension: Propensity to Use Tools

Across the benchmarks reviewed above, a consistent gap emerges. SimpleQA Verified [1] measures parametric recall. FACTS Grounding [8] measures retrieval capability under forced conditions. AA-Omniscience [4, 11] measures overconfidence when tools are absent. Vectara [9] measures faithfulness to provided documents. Xu et al. [5] establishes theoretical bounds on parametric systems.

None of these benchmarks measures the propensity of a tool-augmented model to invoke its available tools during autonomous operation. Yet empirical evidence from deployment studies - 31% search trigger rate for GPT-5 [3], sub-50% grounding rate for Gemini [2] - indicates that propensity, not capability, is the binding constraint on real-world factual reliability. A model that can use retrieval tools with 83.8% accuracy when forced but chooses to invoke them for only one-third of queries delivers effective reliability far below its measured capability. This gap between laboratory capability and deployed propensity is the central phenomenon we investigate, and the architectural enforcement of tool usage is the intervention we evaluate.

3.1 GPT-5 Search Behavior: Measured, Not Assumed

The most comprehensive empirical measurement of autonomous search behavior in a frontier model comes from the Nectiv study, published via SearchEngineLand in October 2025 [3]. Analyzing over 8,500 ChatGPT prompts spanning nine industry verticals, the researchers found that only 31% of prompts triggered a web search - despite browsing being enabled for all sessions. The remaining 69% were answered entirely from parametric memory. When search was triggered, models issued an average of 2.17 queries per prompt, with a maximum fan-out depth of four.

The variance across industries is instructive. Prompts with explicit local intent (e.g., "restaurants near me") triggered search in 59% of cases, while fashion queries triggered search only 19% of the time, and credit card queries 18%. This distribution reveals a critical architectural detail: the model's search decision is not driven by semantic understanding of its own knowledge boundaries but by n-gram heuristics. Nectiv found that the year "2025" was among the most frequent n-grams that ChatGPT appended internally to its search queries [3]. The model does not know that it lacks current interest rate data; it detects that users asking about interest rates often co-occur with temporal markers in its training distribution. Without an explicit temporal signal ("current," a year, "latest"), the model defaults to parametric recall - silently, without disclaimer, and without any indication to the user that the response may be stale.

This is not intelligent tool use. It is pattern matching on surface features, with the failure mode being invisible to the end user.

3.2 GPT-5 Hallucination Rates: Decomposing the Blend

OpenAI's GPT-5 System Card [2] reports hallucination rates - defined as the percentage of factual claims containing minor or major errors - under two conditions. With browsing enabled, GPT-5-main produces erroneous claims at a rate of 9.6%, while GPT-5-thinking achieves 4.5%. With browsing disabled, the rate climbs to 47%.

The 9.6% figure demands careful decomposition. It is not the error rate of a search-augmented model; it is a blended average across two populations: the 31% of prompts where web search was actually triggered, and the 69% where the model answered from memory alone. If we accept the Nectiv trigger rate as approximately representative and the browse-off rate of 47% as an upper bound for unsearched queries under browse-on conditions (noting that browse-on models may exhibit lower parametric error rates due to training-time improvements), a rough back-calculation yields an estimated memory-only error rate of approximately 12% under browse-on conditions and a search-augmented error rate of approximately 4-5%. The latter figure aligns closely with GPT-5-thinking's 4.5%, which is expected given that thinking-mode models search more aggressively.

This decomposition has significant implications. The headline 9.6% figure - the one cited in marketing materials and model comparison tables - obscures a bimodal distribution. For the majority of queries (69%), the model operates at an error rate several times higher than the reported average. For the minority where search is triggered, the error rate is substantially lower. No published evaluation from OpenAI separates these populations, and the blended metric systematically understates the risk to users whose queries fall in the unsearched majority.

The December 2025 GPT-5.2 System Card Update [10] reports a further reduction to below 1% hallucination rate for GPT-5.2-thinking with browsing enabled, evaluated across five domains. However, the update discloses no data on search trigger rates. Without knowing what fraction of queries actually invoked search, the sub-1% figure remains uninterpretable: it could reflect a model that searches 90% of the time, or one that searches 35% of the time but happens to be more accurate on the selected evaluation domains. The metric is meaningless without the denominator.

3.3 Gemini Search Grounding: Optional by Design

Google's Gemini models implement search grounding through two distinct architectures. Gemini 3 and later treat Google Search as a callable tool - the model autonomously decides whether and how often to invoke it for each prompt. Earlier versions (Gemini 2.5 and prior) use a Prediction Score system: each prompt receives a confidence score between 0.0 and 1.0, and search is triggered only when the score exceeds a configurable threshold (default 0.3 for the Gemini API, 0.7 for Vertex AI) [7, 17].

Google's own documentation is explicit about the optional nature of this architecture. The Firebase AI Logic documentation states: "Note that providing Google Search as a tool to the model doesn't require the model to always use the Google Search tool to generate its response" [7]. This is not a limitation - it is a design choice.

Unlike GPT-5, no public measurement of Gemini's natural search trigger rate exists. The closest approximation comes from the DEJAN Grounding Classifier study [6], which analyzed 10,000 prompts processed through Gemini 2.5 Pro with search grounding enabled. Each response was labeled as grounded (search-augmented) or ungrounded (parametric-only). The study reports a "class imbalance between grounded and ungrounded responses" that required synthetic data generation to balance the minority class for classifier training. The critical detail: the ungrounded class was the majority class. This places Gemini's natural grounding rate below 50%, though the exact percentage was not published.

Two additional studies - by Nectiv [13] and Seer Interactive [14] - analyzed Gemini's search fan-out behavior, reporting averages of 9.06 and 10.7 queries per prompt respectively. However, both studies forced grounding on all prompts ("forced grounding was applied on all prompts"), making their results inapplicable to natural usage behavior. They measure what happens when the model is compelled to search, not how often it chooses to do so.

The Gemini 3 Pro and Gemini 3 Flash Model Cards [17, 18] contain no data on search trigger rates or grounding statistics. The Gemini 2.5 Technical Report [19] reports a SimpleQA score of 54.0% and a FACTS Grounding Score of 87.8%, but again, no trigger-rate data. Google has published no metric equivalent to the Nectiv measurement for GPT-5. The absence is not incidental; it is strategic.

3.4 The Economic Incentive: Hallucination as Cost Optimization

The reluctance of frontier models to invoke search tools is not a bug - it is an economic feature. Each search operation incurs three costs: additional compute for processing retrieved documents, increased latency that degrades user experience, and direct API charges.

Google's pricing structure makes this incentive explicit. Gemini 3 charges $14 per 1,000 search grounding queries; Gemini 2.x charges $35 per 1,000 queries [28]. For a model serving millions of requests daily, a high trigger rate would multiply inference costs by an order of magnitude. A Gemini 3 Flash model offered at $0.50 per million tokens cannot economically afford to read three web pages (thousands of additional input tokens) for every factual question. The model's parametric confidence - even when misplaced - is the cheaper path.

OpenAI absorbs search costs into token pricing, but the incentive structure is identical: every Bing query adds latency and compute that erodes margin. The RLHF training signal reinforces this dynamic. Human evaluators consistently prefer fluent, confident responses over hedged refusals or search-delayed answers. Over millions of training iterations, models learn that confabulation is rewarded more often than epistemic humility. The AA-Omniscience benchmark quantifies the result: Gemini 3 Flash exhibits a 91% hallucination rate among incorrect responses - meaning it almost never refuses to answer, even when wrong [4]. It has been optimized, through training and economic pressure alike, to always produce an answer.

This creates a structural misalignment. The model provider's cost function (minimize compute per query) and the user's utility function (maximize factual accuracy) are in direct tension. Hallucinations are not a failure of capability; they are an accepted collateral cost of inference-time optimization. The model has been trained to prefer the "fast path" - parametric recall at near-zero marginal cost - over the "slow path" - search, retrieval, synthesis, and verification at substantially higher cost. Every unsearched query that happens to be correct validates the fast path. Every unsearched query that produces a hallucination is invisible in aggregate metrics.

The propensity gap is, at its root, an economic gap. Models do not search because searching is expensive, and not searching is usually good enough - where "usually" means the 50-60% of the time that parametric recall happens to be correct, and where "good enough" is defined by the provider's cost model, not the user's accuracy requirements.

4. Parametric Hubris: Why Models Actively Suppress Search

The central phenomenon examined in this paper is not one of missing capability but of misaligned incentive. Frontier language models possess the architectural capacity to invoke web search, retrieve grounding data, and synthesize factually current responses. Yet empirical measurement demonstrates that they systematically decline to exercise this capacity. We term this behavior parametric hubris: the architecturally conditioned overconfidence that arises as a function of training dataset scale, causing models to suppress tool access even when internal knowledge is outdated, incomplete, or fabricated. Parametric hubris is not an intelligence failure. It is a predictable consequence of RLHF reward functions, inference cost optimization, and the absence of reliable self-knowledge within transformer architectures.

4.1 Definition and Mechanism

A language model trained on trillions of tokens develops broad distributional coverage across an enormous range of factual domains. This coverage produces high activation confidence for a wide variety of queries - the model's internal probability distribution assigns non-trivial mass to plausible-sounding completions for nearly any factual question. Crucially, confidence in this context is not calibrated against ground truth. It is a statistical artifact of exposure frequency during training. A model that has encountered thousands of passages about interest rates, central bank policy, and financial markets will generate fluent, structurally coherent text about current deposit rates - even if its training data predates the most recent rate change by twelve to eighteen months.

The decision to invoke an external tool (web search, API call, database lookup) competes against this parametric confidence at inference time. When the model's internal activation pattern produces a high-confidence completion, the marginal expected utility of an external call - which introduces latency, computational cost, and the risk of contradictory information - is judged insufficient. The model defaults to parametric recall. This default is not a deliberate reasoning process. It is a learned heuristic, shaped by millions of training iterations in which plausible completions were rewarded and tool invocations were never explicitly incentivized.

4.2 RLHF Conditioning: The Reward Structure of Overconfidence

The origins of parametric hubris lie in the reinforcement learning from human feedback (RLHF) stage of model training. Human evaluators consistently prefer responses that are fluent, confident, and directly helpful over responses that hedge, refuse, or defer to external sources [3, 4]. This preference is rational from the evaluator's perspective - a plausible, well-structured answer appears more competent than a refusal - but it produces a systematic distortion in the model's learned policy.

Over millions of RLHF cycles, the model learns that generating a plausible answer, even when uncertain, yields higher reward than admitting ignorance. The reward signal does not distinguish between correct confidence and confident fabrication. Both produce fluent, assertive text; only the former happens to be true. The model has no training signal that penalizes confident incorrectness as such - it is penalized only when evaluators detect the error, which they often cannot.

The empirical evidence for this mechanism is stark. The AA-Omniscience benchmark [4, 5] measures hallucination rate as the share of false responses among all incorrect attempts - effectively, the overconfidence rate when the model is wrong. Gemini 3 Flash (Non-Reasoning) exhibits a hallucination rate of 90.9%, meaning that when it answers incorrectly, it provides a confident fabrication rather than a refusal in over nine out of ten cases. Its refusal rate is near zero. At the opposite end of the spectrum, Claude 4.5 Haiku shows a hallucination rate of only 26%, with a refusal rate of approximately 62%. This model, when uncertain, declines to answer - a behavior that RLHF typically suppresses.

The critical insight is that accuracy correlates with model size, but hallucination rate does not [4, 11]. Gemini 3 Pro achieves 54% accuracy - the highest on the AA-Omniscience benchmark - yet still exhibits an 88% hallucination rate. It knows more, but it does not know what it does not know. Larger models accumulate more correct parametric knowledge without developing correspondingly better metacognitive calibration. The additional parameters increase coverage but not epistemic humility.

4.3 N-Gram Dependency: Heuristic Triggers Instead of Semantic Self-Awareness

The Nectiv study [3] reveals a further dimension of parametric hubris: models do not decide to search based on genuine semantic awareness of their own ignorance. Instead, they respond to heuristic surface patterns - n-gram cues that statistically correlate with search-worthy queries in the training distribution.

Analysis of ChatGPT's internal search behavior showed that the year "2025" was among the most frequent n-grams appended to internally generated search queries. The model does not possess a mechanism to evaluate whether its parametric knowledge of a specific fact is current. It possesses only a learned association between certain lexical patterns (temporal markers, recency indicators, specific keywords) and the action of triggering a search. When a user asks "What are the best fixed-deposit rates?" without an explicit temporal cue such as "current" or a year, the model defaults to its training distribution - producing an answer based on data that may be twelve to eighteen months stale. No warning is issued. No confidence interval is provided. The user receives obsolete information presented with the same assertive fluency as a verified fact.

This n-gram dependency means that tool invocation is not governed by genuine uncertainty estimation but by pattern matching against surface features of the input. Questions that happen to contain temporal markers trigger search; questions about equally dynamic topics that lack such markers do not. The model's "decision" to search is as shallow as its confidence is deep.

4.4 Context Contamination: When Retrieval Fails to Override Training

Parametric hubris does not cease when a model does invoke its retrieval tools. Even after search results are retrieved and injected into the context window, the model faces a conflict between two competing sources of information: strong parametric priors accumulated over trillions of training tokens, and weak contextual signals from a few hundred tokens of retrieved text [6].

When the retrieved information aligns with parametric knowledge, processing is straightforward. When it contradicts or updates what the model "believes," the parametric prior exerts a gravitational pull on the generated output. The model may selectively attend to portions of the retrieved text that confirm its training distribution while downweighting or ignoring contradictory evidence. The result is subtle factual distortion: responses that appear grounded (they may even cite the retrieved source) but that have been silently contaminated by stale parametric knowledge.

This phenomenon is particularly insidious because it is invisible to the end user. The response reads as though it was derived from current data. The citation may be present. But the specific claim - a number, a date, a name - reflects the training distribution rather than the retrieved document. Context contamination transforms retrieval-augmented generation from a reliability mechanism into a false assurance mechanism: the user believes the answer is grounded because a search was performed, while the answer actually reflects parametric recall filtered through the appearance of grounding.

4.5 The Overconfidence Paradox

The data from AA-Omniscience [4, 5] reveal a counterintuitive relationship between model capability and model reliability. Among frontier models evaluated without tool access, those with the highest accuracy also exhibit the highest hallucination rates. Gemini 3 Pro achieves 54% accuracy with 88% hallucination. On 100 questions, it answers approximately 54 correctly, fabricates approximately 40, and refuses approximately 6. GPT-5.1 achieves 39% accuracy with 81% hallucination. Claude Opus 4.6 achieves 44.3% accuracy with 60% hallucination.

The pattern is consistent: models that know more are also more willing to fabricate when they do not know. This is the overconfidence paradox at the heart of parametric hubris. Greater parametric coverage produces greater confidence across the board - both for domains where that confidence is warranted and for domains where it is not. The model cannot distinguish between the two because its confidence is a function of distributional exposure, not of ground-truth verification.

The only models that escape this paradox are those with aggressive refusal policies. Claude 4.5 Haiku, with 16% accuracy and 26% hallucination, refuses to answer 62% of questions. Its low accuracy reflects limited parametric coverage; its low hallucination rate reflects a training regime that penalizes confident fabrication more heavily than competing models. Among the 36 models evaluated on the AA-Omniscience benchmark, only three achieve a combined Omniscience Index above zero - demonstrating that for the vast majority of frontier models, the penalty incurred by hallucinated responses outweighs the benefit of correct ones [4].

Parametric hubris is therefore not a failure of individual models but a structural property of the current training paradigm. It is the inevitable outcome of optimizing for fluent helpfulness without a corresponding mechanism for epistemic calibration. The models are not lying - they are doing exactly what they were trained to do. The training objective simply does not include knowing when to stop.

5. The Benchmark Trap: How Current Evaluations Systematically Conceal the Propensity Problem

The preceding sections have established that frontier models suppress tool invocation at scale and that the resulting hallucination rates are severe. A natural question follows: why have existing evaluation frameworks failed to detect this failure mode? The answer lies in a systematic methodological blind spot. The two most prominent factuality benchmarks - SimpleQA Verified [1] and FACTS Grounding [8] - each encode assumptions that render the propensity problem invisible by construction.

5.1 SimpleQA Verified: A Test of Memory, Not Retrieval

SimpleQA Verified, introduced by Haas et al. [1] and hosted as a public leaderboard on Kaggle [26], evaluates models on short factual questions with unambiguous, verifiable answers. Its design explicitly measures parametric knowledge: how much a model has memorized from its training corpus and can recall under direct questioning. Gemini 3 Pro scores 72.1%, GPT-5 scores 51.6%, and Gemini 2.5 Pro scores 54.0% on this benchmark [26, 19].

The metric rewards memorization. A model that has absorbed more factual content during pre-training - or that has been fine-tuned on distributions overlapping with the evaluation set - achieves higher scores. This creates a perverse incentive structure: the more a model "knows" (or believes it knows), the less likely it is to invoke external retrieval tools when deployed in production, since its internal confidence threshold is met more frequently. A high SimpleQA score is therefore not an indicator of reliability in a changing world; it is an indicator of overfitting to the training period. The model's parametric confidence is validated in the lab precisely because the benchmark questions fall within its memorization window, but that same confidence becomes a liability when deployed against questions whose answers have shifted since the training cutoff.

For architecture-enforced systems such as Veritas, this distinction is critical. The ideal Veritas system would achieve 0% from parametric recall and 100% from retrieved evidence - a tabula rasa that treats every query as novel. SimpleQA Verified measures exactly the wrong dimension for such systems: it rewards the very parametric confidence that, in deployment, produces the suppression of tool use documented in Section 3.

5.2 FACTS Grounding and the Hardness Filter

The FACTS Grounding benchmark, developed by DeepMind [8], attempts to evaluate factuality more rigorously by assessing whether model outputs are grounded in provided or retrieved evidence. However, its methodology introduces a structural bias that we term the hardness filter.

To construct the evaluation set, the researchers remove all questions that the model can already answer correctly from parametric memory. Only the residual "hard tail" remains - questions where the model is factually compelled to search because its internal knowledge is clearly insufficient. Within this filtered set, models perform well: Gemini 3 Pro achieves 83.8% on FACTS-Search [8]. The benchmark thus demonstrates that when a model is forced into a retrieval scenario, it can execute that retrieval competently.

The problem is that real-world queries are not drawn from this hard tail. The Nectiv study (N=8,500+) reveals that 69% of production prompts are answered from memory [3]. These are predominantly "soft tail" questions - queries where the model possesses a partial, vague, or outdated internal representation. The model's confidence is high enough to suppress tool invocation, but its knowledge is insufficiently precise or current to produce a correct answer. It is precisely in this soft-tail region - where the model could search but chooses not to - that the most consequential errors occur. The hardness filter removes exactly these cases from the evaluation by design, leaving behind only the scenarios where the failure mode cannot manifest.

5.3 Capability Versus Propensity: The Central Distinction

The divergence between FACTS scores and real-world performance exposes the most consequential gap in current evaluation methodology: the conflation of capability with propensity.

FACTS measures capability - the model's ability to conduct retrieval and produce grounded answers when it does search. It does not measure propensity - the model's willingness to initiate that search in the first place. When Gemini 3 Pro achieves 83.8% on FACTS-Search, that figure applies exclusively to the subset of queries where retrieval was triggered. It says nothing about how frequently retrieval is triggered under naturalistic conditions.

The effective reliability of a retrieval-augmented model in deployment is not its FACTS score but the product of its propensity and its capability. If a model searches in approximately one-third of cases (as measured for GPT-5 [3]) and achieves 83.8% accuracy when it does search, the compound reliability for the searched fraction is roughly 28% of all queries answered correctly via retrieval. The remaining two-thirds are answered from parametric memory, where - as the AA-Omniscience benchmark demonstrates - hallucination rates range from 78% to 93% among incorrect responses [4, 5]. The blended effective accuracy is therefore far below the laboratory figure.

For Veritas, propensity is architecturally irrelevant. The pipeline enforces retrieval at a 100% rate; there is no decision point at which the model can elect to rely on parametric recall. For frontier models deployed with optional search - GPT-5 with Bing, Gemini with Google Search - propensity is the decisive weakness. And no existing benchmark measures it.

5.4 Implications for Evaluation Design

The analysis above reveals four dimensions that future factuality benchmarks must incorporate to produce ecologically valid assessments:

Search invocation rate. Benchmarks must report not only accuracy conditional on retrieval but the unconditional rate at which models trigger retrieval across representative query distributions. Without this metric, capability scores are uninterpretable as deployment reliability estimates.

Calibration curves. The relationship between a model's expressed or implicit confidence and its actual accuracy must be measured across the full query distribution, including the soft tail where confidence is moderate but correctness is low. Current benchmarks that filter for difficulty extremes (either trivially easy or impossibly hard) miss the calibration failures that drive real-world hallucination.

Temporal degradation. Factuality evaluations must include questions whose correct answers have changed since the model's training cutoff. A model that scores highly on static knowledge but fails on post-cutoff facts provides an inflated picture of deployment reliability. Time-stratified evaluation - measuring accuracy as a function of the gap between training cutoff and the date the ground truth was last updated - would expose the parametric decay that current benchmarks conceal.

Contamination rate. When benchmark questions overlap with training data, high scores reflect memorization rather than reasoning or retrieval competence. Evaluation frameworks must quantify and report this overlap to distinguish genuine factual capability from distributional leakage.

Until benchmarks measure these dimensions, the gap between laboratory performance and deployment reliability will remain hidden - and the parametric hubris of frontier models will continue to be validated rather than exposed.

6. Veritas Architecture

6.1 Theoretical Foundation: Bypassing Mathematical Impossibility

Xu et al. (2024) provide a formal proof that hallucination is an inevitable property of autoregressive language models [5]. The argument is structural, not empirical: next-token prediction over learned probability distributions cannot distinguish between sequences that correspond to true propositions and sequences that merely resemble true propositions. No amount of scale, no refinement of training data, and no reinforcement learning objective can eliminate a failure mode that arises from the generative mechanism itself. The proof is not a claim about current limitations; it is a claim about mathematical boundaries.

The industry response has been to treat hallucination as an engineering problem amenable to brute-force solutions. OpenAI invested in larger parameter counts, longer training horizons, and multi-stage RLHF pipelines. Google DeepMind pursued reasoning-augmented decoding and chain-of-thought verification. Anthropic developed Constitutional AI and iterative red-teaming. These efforts have produced measurable improvements - GPT-5's browse-off hallucination rate of 47% is lower than its predecessors [2], and Gemini 3 Pro achieves the highest accuracy on AA-Omniscience at 54% [4] - but they have not and, per Xu et al., cannot eliminate the underlying failure mode. The trajectory is asymptotic: billions of dollars yield incremental reductions in a rate that is bounded away from zero by construction.

Veritas begins from the opposite premise. Rather than asking how do we make the model more truthful, it asks: how do we make the model's truth irrelevant? If hallucination is mathematically inevitable in parametric generation, then the solution is not to improve the generation but to remove the conditions under which hallucination occurs. A model hallucinates when it answers from memory. Therefore, never let it answer from memory. The problem is not solved; it is bypassed - and this bypass respects rather than contests the mathematical proof.

This reframing has a concrete architectural consequence. In conventional retrieval-augmented generation (RAG), the model retains discretion over when and whether to invoke retrieval. It may consult external sources, but it may equally decide - based on RLHF-conditioned confidence signals - that its parametric knowledge is sufficient. As documented in Sections 3 and 4, this discretion is exercised in the majority of cases: GPT-5 declines to search 69% of the time [3], and Gemini's grounding rate falls below 50% [6]. Veritas eliminates this discretion entirely. The model is never asked what it knows. It is asked only what the data says.

6.2 The C1-C6 Pipeline

Veritas implements a six-stage pipeline in which every stage is constrained to operate on externally retrieved evidence. No stage permits the model to draw on parametric knowledge.

C1: Primary Retrieval. The user's query is transformed into search terms and submitted to DuckDuckGo. The resulting URLs are scraped via Camoufox, a headless browser built on Firefox that achieves a 0% detection rate across standard bot-detection frameworks. Camoufox operates without fingerprint leakage, executes JavaScript, and renders dynamically loaded content - capabilities that distinguish it from simple HTTP fetchers that fail on modern web pages. C1 produces a corpus of raw web content: the evidentiary basis for all subsequent stages.

C2: Secondary Retrieval. A second round of scraping is performed to increase source diversity and depth. Additional queries are generated from the C1 results to follow leads, resolve ambiguities, and broaden coverage. The purpose of C2 is to ensure that the synthesis stage operates on a sufficiently rich evidence base rather than a single source.

C3: Evidence Synthesis. The model receives the combined C1 and C2 scrape data and is instructed to synthesize a coherent factual summary. Critically, the system prompt for C3 explicitly prohibits the use of parametric knowledge. The model must cite from the provided scrape data or state that the data is insufficient.

C3 produces a structured evidence summary, not an answer.

C4: Answer Generation. The answer is generated exclusively from the C3 synthesis. The model does not see the original query in isolation; it sees the query together with the evidence summary and is constrained to answer only from what C3 provides. This two-stage indirection - scrape, then synthesize, then answer from the synthesis - creates an architectural barrier against parametric injection. The model cannot "sneak in" a remembered fact because its input context contains only retrieved evidence.

C5: Independent Verification Retrieval. A new set of search queries is generated, distinct from those used in C1 and C2, and submitted through a fresh scraping pass. C5 produces an independent evidence corpus that shares no query lineage with the primary retrieval chain. This is the verification evidence base.

C6: Cross-Verification. The C4 answer is evaluated against the C5 evidence. If the independent sources corroborate the answer, it is confirmed. If they contradict it, the discrepancy is flagged and the verification evidence is used to correct or qualify the response. If neither the primary nor the verification chain yields sufficient data, the system issues a refusal.

The critical architectural property is that at no point in the C1-C6 pipeline does the model answer from parametric knowledge. Each stage either retrieves external data or operates exclusively on previously retrieved data. The chain of evidence is fully traceable: every claim in the final output can be linked to a specific scraped source. Fabrication - the generation of confident assertions without any evidentiary basis - is not merely discouraged; it is architecturally impossible. The model would need to hallucinate data into its own input context, which the pipeline structure prevents.

6.3 "No Data, No Answer": Refusal as a Feature

The design philosophy of Veritas inverts the incentive structure of RLHF-conditioned models. Contemporary frontier models are trained to be helpful, which in practice means they are trained to produce responses rather than refusals. The AA-Omniscience benchmark quantifies this: Gemini 3 Flash exhibits a hallucination rate of 90.9% among incorrect responses [4], meaning that when it lacks knowledge, it fabricates an answer in over nine out of ten cases rather than declining. Gemini 2.5 Flash reaches 92.6%. The refusal mechanism has been conditioned out of these models.

Veritas operates on the opposite principle: if the scraping pipeline returns insufficient data, the system refuses to answer. There is no fallback to parametric memory, no "best guess" mode, no attempt to construct a plausible response from training data. A refusal is not a failure; it is the system functioning correctly under data scarcity.

In our evaluation (n=100), Veritas produced 9 refusals - a 9.0% refusal rate. Each was justified by identifiable retrieval limitations: queries too obscure for DuckDuckGo to surface relevant results (e.g., Lukacs's early career in Nepszava), information locked behind dynamically rendered sports scorecards, data available only in physical books, or search terms too specialized to yield matches. Crucially, in all 9 cases, the system acknowledged its inability rather than fabricating. Among the 91 attempted answers, 85 were correct, yielding an accuracy-given-attempted of 93.4%. The 6 errors were misinterpretations of real scraped data - the wrong fact extracted from a correct source - not inventions from no source.

The contrast with frontier models is stark. Gemini 3 Pro, the highest-accuracy model on AA-Omniscience, fabricates in 88% of its incorrect responses. Veritas fabricates in 0%. When Veritas does not know, it says so. When Gemini does not know, it invents.

6.4 Model Choice: Architecture Over Scale

Veritas uses Gemini 2.5 Flash Lite, accessed via OpenRouter, as its language model backend. At the time of evaluation, this was the cheapest model available on the market: $0.10 per million input tokens and $0.40 per million output tokens, yielding an effective cost of approximately $0.002 per query across the full six-stage pipeline.

This choice is deliberate. Gemini 2.5 Flash Lite, when evaluated without retrieval augmentation on the AA-Omniscience benchmark, scores 24.9% accuracy with a 92.6% hallucination rate [4] - the worst hallucination rate of any model on the leaderboard. It is, by conventional metrics, among the least reliable models available. Yet with the Veritas pipeline, this same model achieves an F-score of 89.1% on SimpleQA Verified, surpassing Gemini 3 Pro (72.1%), GPT-5 (51.6%), and o3 (52.3%) - models that cost 20-58 times more per query.

The implication is direct: accuracy in factual question answering is a function of architecture, not of model scale. The cheapest model on the market, wrapped in a mandatory retrieval pipeline, produces the highest F-score on the benchmark. The most expensive models, left to their own discretion about when to search, produce lower scores despite vastly superior parametric capabilities. This finding does not diminish the value of larger models for tasks requiring reasoning, creativity, or nuanced judgment. It demonstrates, specifically and measurably, that for the task of factual question answering, the architectural decision to enforce retrieval dominates the model selection decision by an order of magnitude.

7. Evaluation

7.1 Benchmark Selection and Experimental Setup

We evaluate Veritas on SimpleQA Verified, a factual accuracy benchmark curated by Google DeepMind and hosted on Kaggle, comprising 1,000 short-answer factual questions drawn from diverse knowledge domains including history, geography, science, politics, and culture. The benchmark is designed to test a model's ability to produce precise, verifiable factual claims rather than open-ended reasoning, making it an ideal testbed for systems whose primary contribution is grounding generation in retrieved evidence.

From the full dataset we draw a uniform random sample of 100 questions. The language model serving as the generative backbone is Gemini 2.5 Flash Lite, accessed via the OpenRouter API. Crucially, this is a lightweight, cost-optimized model - not a frontier reasoning system. It ranks outside the top 15 on the SimpleQA Verified leaderboard when evaluated in its default parametric configuration. Any performance gains observed therefore cannot be attributed to the model's intrinsic factual recall but must originate from the retrieval-and-verification pipeline that surrounds it.

Scoring protocol. A response is marked Correct if the ground-truth answer appears anywhere in either the C4 (primary answer) or C6 (verification) output. It is marked Wrong if the system committed to an incorrect answer while possessing real scraped evidence. It is marked Hallucination if the system fabricated an answer without any underlying scrape data. It is marked Refusal if the system honestly declined to answer. The F-Score is computed as the harmonic mean of accuracy and accuracy-given-attempted: F = 2 (Accuracy Acc|Attempted) / (Accuracy + Acc|Attempted).

Validation methodology. Every response underwent a three-stage verification process. First, eight parallel Claude Sonnet agents each independently validated a batch of ten result files, examining both C4 and C6 outputs. Second, all 100 results were reviewed manually by a human evaluator. Third, every initially flagged error was cross-checked against independent web sources to confirm the ground-truth label. This multi-layered protocol minimizes both false positives and false negatives in the scoring.

7.2 Aggregate Results

Table 3 presents the final evaluation metrics across all 100 questions.

Metric	Value
Correct	85
Wrong (evidence-based error)	6
Hallucination (fabricated without data)	0
Refusal (correctly declined)	9
Accuracy	85.0%
Accuracy Given Attempted	93.4% (85/91)
F-Score	89.1%
Error Rate	6.0%
Fabrication Rate	0.0%
Average Duration	~115 s/question

The system answered 91 of 100 questions, declining the remaining 9 where its dual-scraping pipeline could not locate sufficiently specific evidence. Of the 91 attempted answers, 85 were correct, yielding an accuracy-given-attempted of 93.4%. The fabrication rate - answers generated without any evidentiary basis - is precisely zero.

7.3 Error Classification: Type A vs. Type B

We introduce a binary error taxonomy that distinguishes the epistemic origin of each mistake:

- Type A (Fabrication): The model produced a confident answer despite possessing zero scrape results. The claim was invented from parametric memory with no retrieved evidence to support it. This error type is the hallmark of parametric hubris.
- Type B (Evidence-Based Misinterpretation): The model retrieved real, relevant data but extracted or interpreted it incorrectly. The answer is wrong, but it is grounded - traceable to a specific source document and a specific extraction failure.

Across 100 questions, Veritas produced zero Type A errors and six Type B errors. The six evidence-based errors are:

1. IAEG Congress city: The system returned Prague instead of Paris, confusing the IAEG Assembly with the IAEG Congress. Both events are real; the scraper retrieved data about the former rather than the latter.
2. Last U.S. President born in the 18th century: The system answered Millard Fillmore (born 1800) instead of James Buchanan (born 1791), committing a calendrical logic error - 1800 belongs to the 19th century, not the 18th.
3. Minor planet named after Kapitsa (1982): The system returned asteroid (24918) Kapitsa instead of 3437 Kapitsa. Both asteroids carry the same name; the system selected the wrong one (discovered in 1971, not 1982).
4. Peninsular plateau of India spanning 900 km: The system answered the Western Ghats instead of the Satpura Range, selecting the wrong mountain range from genuine geographic data.
5. Reddy's first election from Hindupur: The system extracted 1962 instead of the correct year 1967 from biographical data that contained both dates in different contexts.
6. Pak-China Business Council formation: The system returned July 2019 instead of June 2019, conflating a later announcement with the earlier decision date - both dates appear in authentic source documents.

This distinction carries significant implications for liability and auditability. Type A errors are existential threats: the system confabulates with no evidentiary trail, making post-hoc verification impossible. Type B errors, while still incorrect, are auditable. Each can be traced to a specific source URL, a specific extracted passage, and a specific point of misinterpretation. An engineer or end user can inspect the C2 scrape results, identify where the extraction diverged, and correct the pipeline. In regulated domains - healthcare, legal research, financial compliance - this auditability difference is not incremental but categorical.

7.4 C6 Verification Saves

The dual-pass architecture includes a verification stage (C6) that independently re-scrapes and cross-checks the primary answer produced at C4. In seven cases, C4 failed to locate the correct answer but C6 recovered it:

1. Catapult ride at Busch Gardens: C4 answered "The Viking's Rage"; C6 found "The Catapult" (correct).
2. First elephant born at Valencia Bioparc: C4 stated "Name not stated"; C6 identified "Makena" (correct).
3. Bessie Smith's amputated arm: C4 reported "No amputation specified"; C6 found "right arm" (correct).
4. Carlo Balboni sculptor: C4 returned "Cannot be provided"; C6 confirmed "Carlo Balboni" (correct).
5. Andrea Borella fencing: C4 returned "Cannot be identified"; C6 confirmed "Andrea Borella" (correct).
6. Country with highest freshwater per capita: C4 answered Bhutan; C6 corrected to Iceland (correct).
7. Rosalia EP released in 2019: C4 hedged with "Not explicitly an EP"; C6 found "Fucking Money Man" (correct).

These seven recoveries represent a +7.7 percentage point improvement in raw accuracy attributable solely to the verification stage. Without C6, accuracy would have been 78 out of 91 attempted (85.7% Acc|Attempted) rather than 85 out of 91 (93.4%). This empirically validates the architectural decision to invest a second scraping pass: the marginal cost in latency (~30 seconds) purchases a substantial gain in correctness.

7.5 Leaderboard Position

Table 4 presents the top entries on the SimpleQA Verified Kaggle leaderboard as of February 2026, spanning 47 evaluated models.

Rank	Model	F-Score
1	Gemini 2.5 Flash Lite + VERITAS	89.1%
2	Gemini 3 Pro Preview	72.1%
3	Gemini 2.5 Pro	54.5%
4	Qwen3-235B-A22B	53.7%
5	o3	52.3%
6	Grok-4	51.9%
7	GPT-5	51.6%
8	o1	47.0%
9	GPT-4.1	40.6%
10	Grok-3	39.3%

Veritas surpasses the second-ranked system (Gemini 3 Pro Preview, a frontier model with substantially greater parameter count) by +17.0 F-Score points. It exceeds GPT-5 by +37.5 points, o3 by +36.8 points, and the highest-ranked Claude model (Opus 4.5 at 39.0%) by +50.1 points. The 0% fabrication rate is unmatched across all 47 models evaluated on the benchmark. These results confirm that retrieval-grounded verification, not parametric scale, is the dominant factor in factual accuracy on knowledge-intensive tasks.

7.6 Edge Cases and Rerun Policy

D'Souza system error. One question - asking where Dinesh D'Souza earned his bachelor's degree - initially produced a hallucinated answer (Stanford) because the scraper returned zero results due to a transient system error. With zero scrape data, the model fell back to parametric memory and confabulated. On rerun with a functioning scraper (10+10 scrape results), the system correctly answered UIUC (University of Illinois at Urbana-Champaign). We classify this as a system error, not an AI error: the pipeline's invariant - never fabricate when evidence is available - held perfectly. When the scraper failed silently, the guardrail could not engage. This edge case motivated the addition of explicit zero-scrape detection and automatic refusal in the production pipeline.

Benchmark ground-truth error. One question asked for the latitude of Lilongwe, Malawi. The benchmark's ground truth listed 33.7738, which is in fact the city's longitude. Veritas returned -13.9669, the correct latitude. We count this as correct, and note it as evidence that benchmark labels themselves are not infallible - a consideration relevant to any system evaluated against fixed ground-truth datasets.

8. Cost and Latency Analysis

8.1 The False Tradeoff

A predictable objection to architecturally-enforced verification is the latency argument: that systems achieving high factual accuracy must necessarily sacrifice response speed by orders of magnitude, rendering them impractical for real-world deployment. The implicit framing - "accurate but 1000x slower" - assumes that baseline model inference operates in the millisecond regime and that any retrieval-augmented pipeline inflates this to minutes or hours. This assumption is empirically false.

The millisecond era of language model inference ended with GPT-4o. Every frontier model released since mid-2025 incorporates some form of extended computation: chain-of-thought reasoning, multi-step planning, iterative self-correction, or agentic tool use. These capabilities come at a direct latency cost. A user submitting a factual query to GPT-5 with high reasoning enabled waits approximately 39 seconds for a response - before any web search occurs. When web search is triggered, latencies extend to 60 seconds or more, with community reports documenting response times exceeding 40 minutes for complex queries [27, 28]. The reasoning-specialized model o3 routinely requires 1-5 minutes per query; its premium variant o3-pro demands 3-15 minutes, with documented cases reaching 26 minutes [29]. These are not edge cases. They are the operational reality of frontier reasoning systems.

The relevant comparison class for VERITAS is therefore not the sub-second inference of a cached embedding lookup, but the multi-second to multi-minute response times that characterize any system performing non-trivial cognitive work. Table 5 presents empirically measured latency data across current frontier systems.

Table 5. Response latency across frontier models and research systems. All figures represent end-to-end wall-clock time for factual queries unless otherwise noted.

System	Task Type	Typical Latency
GPT-5 (no search, high reasoning)	Reasoning	39 s
GPT-5 (with web search)	Factual query	60+ s (up to 40 min reported)
GPT-5.2 (no search)	Analytical	7-17 s
o3	Complex reasoning	1-5 min
o3-pro	Any query	3-15 min (up to 26 min)
Gemini 3 Pro	TTFT only	4.5-33 s
Claude Opus 4.5 (thinking)	500-token output	~45 s
ChatGPT Deep Research	Research query	5-30 min
Gemini Deep Research	Research query	5-15 min
Perplexity Deep Research	Research query	2-4 min (benchmark: ~8 min)
VERITAS	Factual research + verification	~115 s (~2 min)

Sources: GPT-5 latency by reasoning level [27]; GPT-5 web search 40+ min [28]; o3-pro 3-15 min [29]; Gemini 3 Pro TTFT [30]; Claude Opus 4.5 [31]; ChatGPT Deep Research FAQ [32]; Perplexity DRACO benchmark [33].

8.2 Competitive Positioning

VERITAS at approximately 115 seconds occupies a distinctive position in the latency landscape. It is faster than o3-pro by roughly an order of magnitude (10x), faster than ChatGPT Deep Research by a factor of 3-15x, and faster than Gemini Deep Research by a factor of 3-8x. It is broadly comparable to Perplexity Deep Research (2-4 minutes) - a system backed by billions of dollars in infrastructure investment - and comparable to GPT-5 with web search enabled (60+ seconds for initial response, with tail latencies extending far beyond). The only systems that respond substantially faster are pure reasoning models without search capability: GPT-5.2 without search returns in 7-17 seconds, but as documented in Section 5, these configurations hallucinate at rates between 47% and 88% on factual queries.

The correct framing is therefore not "VERITAS is slow" but rather: every AI system that searches the web, reads sources, reasons over retrieved content, and synthesizes a verified answer requires 30 seconds to 30 minutes. VERITAS at two minutes lies at the fast end of this spectrum, not the slow end. The systems that respond faster achieve their speed by skipping verification entirely - and their hallucination rates reflect this architectural choice.

8.3 API Cost Per Query

We now turn to the economic dimension of the accuracy-cost tradeoff. Table 6 presents per-query API costs assuming a representative factual QA interaction of 750 input tokens and 300 output tokens.

Table 6. API token costs per query (750 input + 300 output tokens). Factor column indicates cost relative to VERITAS baseline.

Model	Input Cost	Output Cost	Total Per Query	Factor vs. VERITAS
Gemini 2.5 Flash Lite (VERITAS)	$0.000075	$0.000120	$0.000195	1x
GPT-5	$0.000938	$0.003000	$0.003938	20x
o3 (without reasoning tokens)	$0.001500	$0.002400	$0.003900	20x
o3 (with ~2,000 reasoning tokens)	$0.001500	$0.018400	~$0.020	103x
Gemini 3 Pro	$0.001500	$0.003600	$0.005100	26x
Grok-4	$0.002250	$0.004500	$0.006750	35x
Claude Opus 4.5	$0.003750	$0.007500	$0.011250	58x

Sources: OpenAI API pricing [34]; Gemini API pricing [35]; Anthropic pricing [36]; xAI pricing [37].

The cost differential is striking. VERITAS operates on Gemini 2.5 Flash Lite at $0.000195 per query - less than two hundredths of a cent. GPT-5 costs 20x more. The o3 model with typical reasoning token overhead reaches 103x the cost of VERITAS. Claude Opus 4.5, the most expensive model in this comparison, costs 58x more per query. These figures represent token generation costs alone and do not yet account for the additional surcharges imposed by providers who offer web search as a service.

8.4 Search Surcharges

The API prices in Table 6 cover only token generation. Providers that offer integrated web search capabilities impose additional per-query surcharges that substantially alter the total cost of ownership.

Table 7. Web search surcharges by provider.

Provider	Tool	Price
Google (Gemini 3)	Google Search Grounding	$14 / 1,000 queries
Google (Gemini 2.x)	Google Search Grounding	$35 / 1,000 queries
Anthropic (Claude)	Web Search Tool	$10 / 1,000 searches
xAI (Grok)	Web/X Search	$2.50-$5.00 / 1,000 calls
OpenAI (ChatGPT API)	Web Search	Included in token costs
VERITAS (Camoufox)	Direct scraping	$0

VERITAS incurs zero search surcharge. Its retrieval mechanism - Camoufox-based direct web scraping - operates independently of any provider's search API. The only costs are LLM token costs for the orchestration pipeline.

8.5 Total Cost of Ownership: 1,000 Factual Queries

Table 8 presents the aggregate cost comparison for a realistic deployment scenario: 1,000 factual queries requiring web-grounded answers.

Table 8. Total cost for 1,000 factual queries with web search.

System	Token Costs	Search Costs	Total (1,000 Queries)	Factor vs. VERITAS
VERITAS (Flash Lite + Camoufox)	$0.195	$0.00	~$0.20	1x
GPT-5 + Bing	$3.94	$0.00 (incl.)	~$3.94	20x
Gemini 3 Pro + Grounding	$5.10	$14.00	~$19.10	96x
o3 (with reasoning)	~$20.00	$0.00 (incl.)	~$20.00	100x
Claude Opus 4.5 + Web Search	$11.25	$10.00	~$21.25	106x

The economic implications are unambiguous. VERITAS processes 1,000 verified factual queries for approximately $0.20. The same workload on GPT-5 with Bing costs $3.94 - 20x more - while achieving lower accuracy (F-score 51.6% vs. 89.1%) and providing no fabrication guarantees. Gemini 3 Pro with grounding costs $19.10 (96x), largely driven by Google's $14-per-thousand search surcharge. Claude Opus 4.5 with web search reaches $21.25 (106x), the most expensive option in this comparison. The o3 reasoning model, at $20.00 for 1,000 queries (100x), does not even include web search capability - it achieves this cost through reasoning token overhead alone.

8.6 The Inversion

These data invert the assumed tradeoff entirely. The conventional expectation is that higher accuracy demands higher cost and higher latency. VERITAS demonstrates the opposite: by delegating factual grounding to mandatory retrieval and using the cheapest available model for orchestration, the system achieves the highest measured accuracy at the lowest cost and competitive latency. The 89.1% F-score costs $0.20 per thousand queries. The 72.1% F-score (Gemini 3 Pro) costs $19.10. The 51.6% F-score (GPT-5) costs $3.94. Accuracy and economy are not in tension - they are aligned, provided the architecture enforces verification rather than hoping the model will choose to verify itself.

The false tradeoff narrative persists because it serves the commercial interests of frontier model providers. If accuracy required expensive models, then expensive models would be necessary. VERITAS provides empirical evidence that this is not the case. The bottleneck to factual reliability was never model capability. It was, and remains, architectural discipline.

9. Discussion and Limitations

9.1 Summary of Contributions

This paper has presented five interlocking contributions to the understanding of factual reliability in frontier language models.

First, empirical evidence of the propensity gap. Drawing on primary-source data from the Nectiv observational study (N=8,500+) [3], OpenAI's GPT-5 System Card [2], the DEJAN grounding classifier study (N=10,000) [6], and the AA-Omniscience benchmark (N=6,000) [4, 11], we have documented that tool availability does not predict tool usage. GPT-5 invokes web search in only 31% of interactions despite having browsing enabled; Gemini's grounding rate falls below 50% under natural conditions. When these models decline to search, the consequences are severe: GPT-5's hallucination rate rises from an estimated 4-5% on searched queries to 47% on unsearched queries [2], and Gemini 2.5 Flash fabricates in 92.6% of its incorrect responses [4]. The propensity gap - the distance between a model's capability to use tools and its inclination to do so - is the single largest contributor to factual unreliability in deployed systems, yet no existing benchmark measures it.

Second, the Parametric Hubris framework. We have formalized the architecturally conditioned overconfidence that causes models to suppress tool invocation, tracing its origins to RLHF reward structures that incentivize fluent confidence over epistemic humility, to inference cost optimization that makes parametric recall the economically preferred path, and to n-gram-based heuristic triggers that substitute surface pattern matching for genuine uncertainty estimation. Parametric hubris is not a bug in individual models; it is a structural property of the current training paradigm, predictable from first principles and measurable in production data.

Third, the capability versus propensity distinction. We have shown that existing benchmarks - SimpleQA Verified [1] and FACTS Grounding [8] in particular - systematically conflate what a model can do when forced to search with what it will do when given discretion. SimpleQA rewards parametric memorization; FACTS filters out the soft-tail queries where models feel confident enough not to search but are wrong. Both produce inflated reliability estimates that do not transfer to deployment conditions. We propose that future factuality evaluations must report search invocation rates, calibration curves, and temporal degradation alongside accuracy metrics to produce ecologically valid assessments.

Fourth, architectural proof that mandatory tool use eliminates fabrication. Veritas, a six-stage retrieval-and-verification pipeline built on Gemini 2.5 Flash Lite at $0.002 per query [28], achieves an F-Score of 89.1% on SimpleQA Verified with a 0% fabrication rate [1, 26]. This result is achieved using a model that, without the pipeline, scores 24.9% accuracy with a 92.6% hallucination rate on AA-Omniscience [4] - the worst hallucination rate on the leaderboard. The architecture, not the model, is the variable that matters. All six errors in the evaluation were Type B (evidence-based misinterpretation) - the model extracted the wrong fact from a correct source, rather than inventing a fact from no source. This finding supports the theoretical argument advanced by Xu et al. [5] that hallucination is mathematically inevitable in parametric generation, while demonstrating that the problem can be architecturally bypassed by removing the conditions under which parametric generation occurs.

Fifth, the cost and latency analysis demonstrates that the solution is both faster and cheaper than the systems it outperforms. At approximately 115 seconds per query, Veritas operates faster than ChatGPT Deep Research (5-30 minutes) [23], Gemini Deep Research (5-15 minutes), o3-pro (3-15 minutes), and comparably to Perplexity Deep Research (2-4 minutes) [24] - systems backed by billion-dollar infrastructure. At $0.20 per 1,000 queries, it is 20x cheaper than GPT-5 with Bing [27], 95x cheaper than Gemini 3 Pro with grounding [28], and 106x cheaper than Claude Opus 4.5 with web search [29]. The common objection that retrieval-augmented verification must be prohibitively slow and expensive is empirically falsified.

9.2 The Pipeline Comparison Objection

A legitimate methodological concern is that Veritas is a multi-stage pipeline being compared against individual model scores on a parametric benchmark. This objection deserves direct engagement.

We acknowledge that Veritas is not a model; it is an architecture. The comparison is intentionally asymmetric: we are comparing architectures, not models. The central claim of this paper is precisely that architecture dominates model selection for factual question answering, and the asymmetry of the comparison is the evidence for that claim.

Moreover, the objection implicitly assumes that frontier models are monolithic parametric systems. They are not. GPT-5 internally orchestrates web search, code execution, and multi-step reasoning through an agentic pipeline. Gemini 3 Pro invokes Google Search as a callable tool, with optional grounding, fan-out queries, and result synthesis. Claude integrates web search, artifact generation, and tool use within its extended thinking framework. These are all pipelines - the difference is that their retrieval components are optional and model-gated, while Veritas's retrieval is mandatory and architecture-gated. The comparison is not between a pipeline and a model; it is between an architecture that enforces retrieval and architectures that leave retrieval to the model's discretion. The empirical finding is that the former produces categorically better factual reliability.

9.3 Limitations

We identify six limitations of this work and discuss their implications for the validity of our claims.

1. Sample size. The evaluation was conducted on N=100 questions randomly drawn from the 1,000-question SimpleQA Verified benchmark. A binomial confidence interval at the 95% level yields a margin of error of approximately $\pm$7 percentage points around the measured 85% accuracy, placing the worst-case lower bound at approximately 78%. Even at this lower bound, Veritas would still surpass the next-best system (Gemini 3 Pro at 72.1%) by nearly 6 points - a margin that remains practically significant. The F-Score confidence interval is correspondingly bounded: under worst-case assumptions, the lower bound of approximately 82% still exceeds all 47 models on the leaderboard. The sample size was constrained by the cost of human validation (each query requires multi-stage manual verification against independent sources) and the computational time of the six-stage pipeline (approximately 32 hours for the full benchmark at 115 seconds per query). Full replication on all 1,000 questions is feasible at a cost of approximately $2 in API fees and remains a priority for future work. The benchmark, the pipeline code, and the evaluation protocol are all open source, making independent replication straightforward.

2. Self-evaluation bias. The primary scoring was conducted by the pipeline's developer, supplemented by eight parallel Claude Sonnet agents for independent batch validation. While self-evaluation is standard practice in machine learning research - the SimpleQA Verified leaderboard itself relies on model-graded scoring - it introduces a potential for unconscious favorable bias. We mitigate this through the three-stage validation protocol (parallel AI agents, human review, independent web cross-check) and through full transparency: all 100 result files, including raw C4 and C6 outputs, are available for inspection. Nevertheless, independent replication by a separate research group, with blinded scoring, would substantially strengthen the claims.

3. Task scope. Our claims are restricted to factual question answering - the task of producing precise, verifiable factual claims in response to short-answer knowledge queries. We make no claims about Veritas's applicability to reasoning tasks, code generation, creative writing, mathematical proof, or any domain where the answer cannot be directly retrieved from web sources. The capability-propensity distinction and the parametric hubris framework are general phenomena applicable across task types, but the architectural solution presented here - mandatory web retrieval - is specific to knowledge-intensive factual tasks. Extending the mandatory-evidence principle to other domains (e.g., requiring code to compile before acceptance, requiring mathematical proofs to be formally verified) represents a distinct research agenda.

4. Latency constraints. The mean query latency of approximately 115 seconds is acceptable for research workflows, asynchronous information retrieval, and applications where accuracy takes precedence over response time. It is not acceptable for real-time conversational interfaces, interactive search, or latency-sensitive

production systems where users expect sub-second or single-digit-second responses. We note, however, that the relevant comparison class is not instant parametric recall but other systems that perform web retrieval and synthesis - ChatGPT Deep Research at 5-30 minutes [23], Perplexity Deep Research at 2-4 minutes [24], and GPT-5 with web search at 60+ seconds [22]. Within this comparison class, Veritas sits at the faster end. Future optimizations - parallelizing C1/C2 scraping, caching frequently queried domains, reducing unnecessary intermediate generation - could reduce latency substantially without compromising the mandatory-retrieval invariant.

5. Scraper dependency and failure modes. Veritas depends on Camoufox's ability to scrape live web content, which introduces vulnerabilities to anti-bot countermeasures, paywalled content, dynamically rendered data, and transient network failures. In our evaluation, 9 of 100 queries resulted in refusals, primarily because DuckDuckGo could not surface sufficiently specific results or because the target content was rendered via JavaScript that the scraper could not execute in time. The D'Souza edge case (Section 7.6) further demonstrated that a silent scraper failure can cascade into a fabricated response when zero-scrape detection is absent - a vulnerability that has since been patched with explicit zero-result guards. As web publishers continue to deploy increasingly aggressive bot detection, the reliability of the scraping layer will require ongoing maintenance. This is a practical engineering constraint, not a theoretical limitation of the approach, but it must be acknowledged as a deployment risk.

6. Benchmark-specific evaluation. Our results are measured on a single benchmark (SimpleQA Verified) with a specific question distribution biased toward encyclopedic factual knowledge. Performance on domain-specific corpora - medical literature, legal precedent, financial data, scientific publications - may differ substantially, particularly where authoritative sources are paywalled, sparsely indexed, or expressed in specialized terminology that degrades search recall. Evaluation on domain-specific benchmarks is needed to establish the generalizability of the mandatory-retrieval architecture beyond general-knowledge factual QA.

9.4 Threats to External Validity

Two structural factors may limit the generalizability of these findings.

First, the SimpleQA Verified benchmark draws from a distribution of questions that, by construction, have unambiguous single-correct answers retrievable from public web sources. Real-world factual queries are often ambiguous, context-dependent, or require synthesizing information across multiple contradictory sources. Veritas's cross-verification architecture (C5-C6) is designed to handle source disagreement, but it has not been evaluated on deliberately adversarial or ambiguous query sets.

Second, the open web is not a neutral evidence source. It contains misinformation, outdated pages, SEO-optimized content that prioritizes ranking over accuracy, and systematic biases toward English-language, Western-centric knowledge. A mandatory-retrieval architecture that treats scraped content as ground truth inherits the biases and errors of the web itself. Veritas mitigates this through dual-path retrieval and cross-verification, but it cannot overcome systematic misinformation present across multiple independent sources. The architecture eliminates fabrication (answers with no evidentiary basis) but does not guarantee truth (answers whose evidentiary basis is itself correct). This distinction - between grounded-but-wrong and ungrounded-and-wrong - is captured in our Type A/Type B error taxonomy, but it remains a fundamental limitation of any retrieval-based system.

9.5 Future Work

Four directions emerge from this research.

Scaling evaluation to N=1,000. The full SimpleQA Verified benchmark is publicly available and the pipeline is open source. A complete evaluation - estimated at approximately $2 in API costs and 32 hours of compute - would reduce the confidence interval from $\pm$7% to $\pm$2% and provide definitive leaderboard positioning. This is the most immediate next step and requires no methodological innovation, only computational time.

Independent evaluation. The strongest validation of these results would be replication by an independent research group using blinded scoring protocols. We have released all pipeline code, evaluation scripts, and raw result files to facilitate this. We particularly invite groups with expertise in factual grounding evaluation (e.g., teams affiliated with the FACTS or SimpleQA benchmarks) to conduct independent assessments.

Domain-specific evaluation. Medical, legal, and financial question answering present distinct challenges: authoritative sources are often paywalled, terminology is specialized, and the consequences of error are higher than in general-knowledge QA. Evaluating mandatory-retrieval architectures on domain-specific benchmarks - MedQA, LegalBench, FinQA - would test the generalizability of the approach and identify domain-specific failure modes (e.g., paywall-induced refusal rates, terminology-degraded search recall).

Comparison with proprietary retrieval systems. OpenAI's Deep Research, Google's Gemini Deep Research, and Perplexity's Deep Research represent the industry's most sophisticated retrieval-augmented systems. A controlled comparison on identical question sets - measuring accuracy, fabrication rate, latency, and cost - would provide the most direct test of whether the mandatory-retrieval principle outperforms even heavily engineered optional-retrieval systems. The challenge is methodological: these systems are accessible only through consumer interfaces with limited API control, making reproducible comparison difficult but not impossible.

10. Conclusion

This paper began with a simple empirical observation: frontier language models equipped with web search tools choose not to use them in the majority of cases. GPT-5 triggers search for only 31% of queries [3]. Gemini's grounding rate falls below 50% under natural conditions [6]. When these models answer from parametric memory instead - as they do for the majority of user interactions - their hallucination rates are catastrophic: 47% for GPT-5 with browsing disabled [2], 88-93% overconfidence among incorrect responses across frontier Gemini and GPT models on the AA-Omniscience benchmark [4, 11]. The tools exist. The models choose not to use them. And when they choose wrong, they fabricate with confidence.

We have formalized this phenomenon as parametric hubris - the architecturally conditioned overconfidence that causes models to suppress tool invocation in favor of parametric recall, driven by RLHF reward structures that incentivize fluent confidence, inference cost optimization that makes memory the cheaper path, and n-gram heuristic triggers that substitute surface pattern matching for genuine epistemic self-awareness. We have drawn a distinction between capability - a model's ability to use retrieval tools when compelled - and propensity - its willingness to invoke them autonomously - and shown that no existing benchmark measures the latter, producing evaluation frameworks that systematically overstate deployment reliability.

Against this background, we have presented Veritas: a six-stage mandatory retrieval-and-verification pipeline that removes the model's discretion over tool use entirely. Built on Gemini 2.5 Flash Lite - a model that, without the pipeline, exhibits a 92.6% hallucination rate and 24.9% accuracy on AA-Omniscience [4] - Veritas achieves an F-Score of 89.1% on SimpleQA Verified with zero fabrication, surpassing the next-best system by 17.0 points. The six errors it produced were all evidence-based misinterpretations: the model extracted the wrong fact from a correct source, never inventing a fact from no source. The nine refusals were honest acknowledgments of retrieval limitations, not parametric confabulations disguised as answers.

The implications extend beyond a single benchmark result. Veritas demonstrates that for factual question answering, architecture dominates model scale by an order of magnitude. The cheapest model on the market, constrained by a mandatory-retrieval pipeline, outperforms frontier systems costing 20 to 106 times more per query. This finding does not diminish the importance of model capability for tasks requiring reasoning, creativity, or judgment. It demonstrates, specifically and measurably, that the factual reliability problem is not one of insufficient parameters but of insufficient architectural discipline. The models are powerful enough. The question is whether we build systems that use that power wisely or systems that let models decide for themselves when to bother.

Xu et al. [5] proved that hallucination is mathematically inevitable in autoregressive language models. They are correct. Token prediction over probability distributions cannot structurally distinguish between truth and confabulation. The industry has spent billions attempting to solve this problem within the model - larger parameter counts, longer training, more sophisticated RLHF, chain-of-thought reasoning. These investments have produced measurable improvements and will continue to do so. But they are asymptotic: bounded away from zero by the mathematical structure of the generative process itself.

Veritas does not contest this proof. It respects it. By never allowing the model to answer from parametric memory, by enforcing real-time retrieval at every stage, by cross-verifying against independently sourced evidence, and by refusing when the evidence is insufficient, Veritas bypasses the conditions under which hallucination occurs rather than attempting to suppress hallucination after it has already been generated. The problem is not solved. It is rendered structurally irrelevant.

The path to reliable factual AI lies not in making models that know more, but in building architectures that know how to find everything - and possess the integrity to fabricate nothing.

References

[1] Haas, J., Liu, A., Glaese, A., Sherburn, C., & Gonzalez, A. (2025). SimpleQA Verified: A Factual Accuracy Benchmark for Language Models. arXiv preprint arXiv:2509.07968. https://arxiv.org/abs/2509.07968

[2] OpenAI. (2025). GPT-5 System Card. https://cdn.openai.com/gpt-5-system-card.pdf

[3] Nectiv Digital / SearchEngineLand. (2025). ChatGPT Search Prompts Data: How Often GPT-5 Triggers Web Search. SearchEngineLand. https://searchengineland.com/chatgpt-search-prompts-data-463407

[4] Artificial Analysis. (2025). AA-Omniscience: AI Reliability Benchmark. https://artificialanalysis.ai/evaluations/omniscience

[5] Xu, Z., Jain, S., & Kankanhalli, M. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. arXiv preprint arXiv:2401.11817. https://arxiv.org/abs/2401.11817

[6] DEJAN SEO. (2025). Grounding Classifier: Predicting When Gemini Uses Google Search. https://dejan.ai/blog/grounding-classifier/

[7] Google. (2025). Firebase AI Logic: Grounding with Google Search. Firebase Documentation. https://firebase.google.com/docs/ai-logic/grounding-google-search

[8] Google DeepMind. (2025). FACTS Grounding: Evaluating and Improving Factuality in Large Language Models. https://storage.googleapis.com/deepmind-media/FACTS/FACTS_grounding_paper.pdf

[9] Vectara. (2025). Hallucination Leaderboard. GitHub Repository. https://github.com/vectara/hallucination-leaderboard

[10] OpenAI. (2025). GPT-5.2 System Card Update. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf

[11] Artificial Analysis. (2025). AA-Omniscience: Measuring AI Reliability Through Knowledge and Hallucination Assessment. arXiv preprint arXiv:2511.13029. https://arxiv.org/abs/2511.13029

[12] DEJAN SEO. (2025). How Google Decides When to Use Gemini Grounding for User Queries. https://dejan.ai/blog/how-google-decides-when-to-use-gemini-grounding-for-user-queries/

[13] Nectiv Digital. (2025). New Research: We Analyzed 60K Google Fan-Out Queries. https://nectivdigital.com/new-research-we-analyzed-60k-google-fan-out-queries/

[14] Seer Interactive. (2025). Gemini 3 Query Fan-Outs Research. https://www.seerinteractive.com/insights/gemini-3-query-fan-outs-research

[15] Google Developers Blog. (2025). Gemini API and AI Studio Now Offer Grounding with Google Search. https://developers.googleblog.com/en/gemini-api-and-ai-studio-now-offer-grounding-with-google-search/

[16] Google. (2025). Gemini API: Grounding with Google Search. Gemini API Documentation. https://ai.google.dev/gemini-api/docs/google-search

[17] Google DeepMind. (2025). Gemini 3 Pro Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Pro-Model-Card.pdf

[18] Google DeepMind. (2025). Gemini 3 Flash Model Card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf

[19] Google DeepMind. (2025). Gemini 2.5 Technical Report. https://storage.googleapis.com/deepmind-media/gemini/gemini_v2_5_report.pdf

[20] Sparkco. (2025). Gemini 3 Grounding with Google Search: Analysis. https://sparkco.ai/blog/gemini-3-grounding-with-google-search

[21] OpenAI. (2025). Why Language Models Hallucinate. https://openai.com/index/why-language-models-hallucinate/

[22] D4B.dev. (2025). GPT-5 Response Time Metrics by Reasoning Level. https://www.d4b.dev/blog/2025-09-28-gpt5-response-time-metrics

[23] OpenAI. (2025). Deep Research FAQ. OpenAI Help Center. https://help.openai.com/en/articles/10500283-deep-research-faq

[24] Perplexity AI. (2025). Evaluating Deep Research Performance in the Wild with the DRACO Benchmark. https://research.perplexity.ai/articles/evaluating-deep-research-performance-in-the-wild-with-the-draco-benchmark

[25] National Institutes of Health. (2025). GPT-5 Hallucination Reduction in Clinical Question Answering. PubMed Central. https://pmc.ncbi.nlm.nih.gov/articles/PMC12701941/

[26] Google DeepMind / Kaggle. (2025). SimpleQA Verified Benchmark Leaderboard. https://www.kaggle.com/benchmarks/deepmind/simpleqa-verified

[27] OpenAI. (2026). API Pricing. https://openai.com/api/pricing/

[28] Google. (2026). Gemini API Pricing. https://ai.google.dev/gemini-api/docs/pricing

[29] Anthropic. (2026). API Pricing. https://platform.claude.com/docs/en/about-claude/pricing

[30] xAI. (2026). Models and Pricing. https://docs.x.ai/developers/models

[31] Artificial Analysis. (2026). Gemini 3 Pro: Provider Benchmarks. https://artificialanalysis.ai/models/gemini-3-pro/providers

[32] Artificial Analysis. (2026). Claude Opus 4.5 (Thinking): Provider Benchmarks. https://artificialanalysis.ai/models/claude-opus-4-5-thinking/providers