DeepFact: Co-Evolving Benchmarks and Agents for Deep Research Factuality

Yukun Huang[1], **Leonardo F. R. Ribeiro**[2], **Momchil Hardalov**[2], **Bhuwan Dhingra**[1]
Markus Dreyer[2], **Venkatesh Saligrama**[2,3]
[1]Duke University, [2]Amazon AGI, [3]Boston University
yukun.huang@duke.edu

Abstract

Search-augmented LLM agents can produce deep research reports (DRRs), but verifying claim-level factuality remains challenging. Existing fact-checkers target general-domain atomic claims, and there is no benchmark to test whether such verifiers transfer to DRRs. Yet building such a benchmark for DRR fact-checkers is itself difficult because it requires expert judgments over cognitively demanding, domain-specific claims. In a controlled study with PhD-level specialists, unassisted experts achieve only 60.8% accuracy on hidden known-answer claims. We therefore propose evolving benchmarking via **Audit-then-Score** (**AtS**), in which labels and rationales remain revisable: when a verifier disagrees with the current benchmark, it submits evidence; an auditor adjudicates the dispute; and accepted revisions update the benchmark before scoring. After three additional AtS rounds, expert accuracy rises to 90.9%, showing that experts are better auditors than one-shot labelers. We instantiate AtS as **DeepFact-Bench**, a versioned DRR factuality benchmark with auditable rationales, and introduce **DeepFact-Eval**, a claim-level verifier. On the frozen DeepFact-Bench release, DeepFact-Eval achieves 83.4% accuracy, outperforming the best prior deep-research and traditional fact-checkers by 14.3 and 24.9 points, respectively, and transferring well to external factuality datasets. The code is released. [1]

1 Introduction

Deep research agents (OpenAI, 2024; Jin et al., 2025) can generate long, citation-rich deep research reports (DRRs) for complex questions, but evaluating the claim-level factuality of those reports remains a challenge. A common automated strategy is *citation-grounded fact-checking*: verifying whether each claim is entailed by sources cited in the report (Du et al., 2025; Wang et al., 2025). But this ignores claims without explicit citations, often synthesized across documents, and conflates "supported by a text" with "supported by scientific consensus," even when the cited source is outdated, disputed, or cherry-picked. Reliable DRR verification must therefore go beyond in-report citations and cross-check the broader literature.

Existing *open-world fact-checking* methods (Wei et al., 2024; Wang et al., 2024) go beyond citation-grounded fact-checking by retrieving external evidence, but are still not well suited to DRRs. Most target short, general-domain factoids and rely on snippet-level matching, whereas DRR verification often requires reasoning over full documents and synthesizing evidence across sources. Thus, neither citation-grounded nor open-world fact-checking suffices; the field needs a benchmark tailored to evaluate better DRR fact-checkers.

But building such a benchmark is itself difficult. The standard paradigm is to ask human experts to create a static "gold standard" dataset (Malaviya et al., 2024; Bayat et al., 2025; Wang et al., 2024; Thorne et al., 2018). This assumes expert labels are reliable enough to serve as fixed ground truth. Recent work shows that factuality benchmarks can contain noisy or inconsistent labels (Xie et al., 2025; Nahum et al., 2025; Glockner et al., 2024; Thibault et al., 2025). DRR verification is harsher still. It demands deep domain expertise, reasoning over extensive context, and sustained attention: verifying a single claim can take hours, while a single report may contain hundreds of claims (Patel et al., 2025). Moreover, expertise is both scarce and fragmented: even slight domain drift can make verification substantially harder, rendering multi-expert adjudication unrealistic at DRR scale. As a result, static expert labels may be too brittle to serve as reliable ground truth for DRR factuality.

We test this directly in a controlled study with PhD-level specialists, explicitly incentivized for accuracy, annotating DRR claims drawn from their

[1]https://github.com/kkkevinkkkkk/DeepFact

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 34356–34386

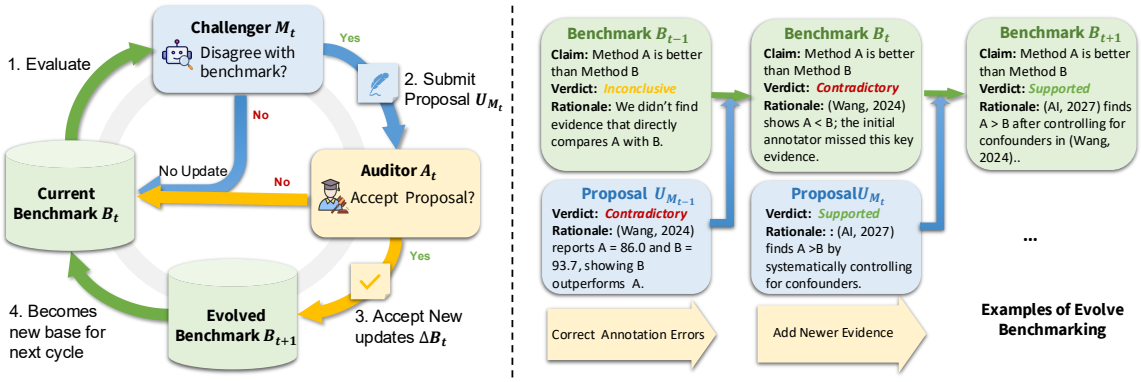


Figure 1: **Evolving Benchmarking via Audit-then-Score (AtS). Left:** AtS workflow. **Right:** an example of evolving benchmark. Unlike traditional static benchmarking, AtS treats ground truth $y_i^{(t)}$ as an evolving consensus. The process proceeds in four stages: (1) **Evaluate:** Run a *Challenger* agent (M_t) on the current benchmark state (B_t), producing a verdict $\hat{y}_i$. (2) **Challenge:** When $\hat{y}_i \neq y_i^{(t)}$, the Challenger submits a *proposal* with evidence. (3) **Audit:** An *Auditor* (human expert or trusted agent) adjudicates the dispute; if the Challenger's argument is stronger than the incumbent rationale, the update is accepted. (4) **Evolve & Score:** Accepted updates yield the next benchmark state (B_{t+1}); the Challenger is then scored against this refined ground truth.

own research areas. To measure annotator accuracy, we embed a hidden micro-gold set of adversarially constructed known-answer claims in the evaluation stream. We find that *unassisted experts achieve only 60.8% accuracy on these micro-golds*, even within their domains. This suggests that, for cognitively intensive expert-level reasoning tasks, static expert labels are too unreliable to serve as a stable benchmark, motivating a new evaluation paradigm.

To address this, we propose **Audit-then-Score** (AtS), an evolving benchmarking protocol for DRR fact-checkers (see Figure 1). Under AtS, benchmark labels are explicitly revisable: when a verifier disagrees with the current benchmark, it submits a competing verdict with evidence; an auditor then *audits* the disagreement, and accepted challenges revise the benchmark before models are *scored*. Evaluation thus becomes a versioned, auditable process rather than a fixed one-shot judgment, allowing stronger verifiers to surface weaknesses in the current benchmark without changing it unilaterally. *This process mirrors how scientific knowledge evolves, not as a frozen snapshot, but as an ongoing dialogue in which new findings can overturn prior conclusions* (Elliott et al., 2017).

We instantiate AtS as **DeepFact-Bench**, a versioned benchmark of 944 claims from 20 DRRs across six domains: computer science, control theory, environmental engineering, education, public health, and engineering management. The expert-only annotations form the seed benchmark. We then run three AtS rounds where PhD experts audit disagreements raised by progressively stronger DRR verifiers, including our proposed **DeepFact-Eval**. Across rounds, hidden micro-gold accuracy rises significantly from the 60.8% seed to 90.9%. This validates AtS's core premise: for DRR factuality, experts are more reliable *auditors* of contested claims than producers of static labels. We also show agent auditors maintain benchmark quality.

On the frozen DeepFact-Bench release, DeepFact-Eval achieves 83.4% accuracy, outperforming the best traditional and deep-research baselines by 24.9 and 14.3 points, respectively. It also transfers well to other static benchmarks like SciFact, ExpertQA, and Factcheck-Bench. Because these benchmarks are static and may contain annotation errors, we audit disagreement cases to distinguish benchmark noise from true model errors; doing so consistently increases estimated accuracy (e.g., 84.6% $\rightarrow$ 94.7% on SciFact; 82.0% $\rightarrow$ 98.0% on Factcheck-Bench). This suggests static benchmark noise under-credits strong verifiers, further motivating AtS.

Contributions. (1) We show that static expert-labeled benchmarks are brittle for DRR factuality: PhD specialists achieve only 60.8% accuracy. (2) We propose **Audit-then-Score**, an evolving benchmarking protocol instantiated as **DeepFact-Bench**, raising expert micro-gold accuracy from 60.8% to 90.9%. (3) We introduce **DeepFact-Eval**, a claim-level verifier with a grouped lower-cost variant that outperforms prior traditional and deep-research baselines on the frozen DeepFact-Bench release.

2 Related Work

DRR Evaluation: LLM agents generate DRRs by synthesizing retrieved information into long-form outputs (Shao et al., 2025). Current evaluations primarily focus on *report-level* qualities (e.g., coherence, coverage) via LLM-Judge (Du et al., 2025), rubrics (Sharma et al., 2025; Gou et al., 2025), expert preference (Chandrahasan et al., 2025; Zhao et al., 2025), or hybrid approaches (Wang et al., 2025). Factuality is usually approximated by citation checking (Du et al., 2025; Wang et al., 2025), which misses uncited claims and can over-trust incomplete or biased cited sources. We instead verify claims against the broader scientific literature.

Fact-Checking: Existing fact-checking benchmarks mainly cover general-domain claims drawn from news (Wang, 2017; Augenstein et al., 2019), Wikipedia (Thorne et al., 2018; Jiang et al., 2020), LLM responses to general user instruction (Bayat et al., 2025), with recent work expanding into scientific domains (Wadden et al., 2020; Malaviya et al., 2024). Because they rely on human annotation, label noise from humans becomes an increasing bottleneck as models improve (Xie et al., 2025; Nahum et al., 2025; Thibault et al., 2025). Methods typically involve a claim-centric workflow: claim extraction/decomposition, web retrieval, and snippet matching (Wei et al., 2024; Song et al., 2024; Xie et al., 2025; Metropolitansky and Larson, 2025; Liu et al., 2025). However, this shallow workflow may not transfer to DRR verification, which requires reasoning over full papers and cross-paper consistency, not only snippet-level adjudication.

Reliability in Human Annotations: Human "gold" labels are increasingly contested, often compromised by cognitive biases, insufficient evidence, annotator priors, and subjectivity (Soprano et al., 2024; Atanasova et al., 2022; Sap et al., 2022; Pavlick and Kwiatkowski, 2019). These issues compound in high-complexity domains with fragmented expertise. Recent benchmarks typically use only 1–2 experts per example (Malaviya et al., 2024; Asai et al., 2024), treating inter-annotator agreement (IAA) as a proxy for correctness (Malaviya et al., 2024; Zhao et al., 2025). However, IAA obscures unresolved disputes and misses shared blind spots or systematic errors (van der Velden et al., 2025; Goh et al., 2023). Indeed, experts remain fallible, with documented error rates in LLM benchmarks (Phan et al., 2025) and manual literature reviews (Salvador-Oliván et al., 2019). To address the limitations of static annotations, we introduce adversarial micro-golds to audit performance alongside a dynamic human–AI framework to iteratively refine consensus. See Appendix I for extended related work on dynamic benchmarking, research-focused benchmarks, and role-based and multi-agent evaluation.

3 Problem Formulation

3.1 Task: Verifying Factuality in DRRs

The task is to verify the factuality of claims in DRRs, whose long-form expert-level synthesis makes verification a non-trivial reasoning task. Our goal is to assign a claim-level factuality label $y_i \in \{\text{SUPPORTED}, \text{INCONCLUSIVE}, \text{CONTRADICTORY}\}$ to each verifiable claim c_i (see Appendix A for definitions). Each data point is a triplet (c_i, d_i, y_i), where c_i is a verbatim sentence and d_i is the full DRR providing the context. Evaluation must consider d_i to disambiguate the claim; if any part of a sentence is inconclusive or contradictory, the entire sentence inherits that label. We follow (Malaviya et al., 2024) and verify at the sentence level to minimize noise from imperfect sub-claim extraction and ensure compatibility with future evaluators. Addressing this task requires solving two coupled problems.

Problem 1: Modeling (Building the Verifier). Following prior factuality evaluation setups (Song et al., 2024; Wei et al., 2024), we focus on *automated* verification to scale DRR verification. Specifically, we build a verifier M that predicts a factuality verdict from a claim and its report context: $\hat{y}_i = M(c_i, d_i)$.

Problem 2: Benchmarking (Evaluating the Verifier). To measure the performance of any verifier M, we need a reliable benchmark B containing ground-truth labels y_i^* to validate verifiers' outputs.

We first discuss Problem 2 (Benchmarking), since without a reliable benchmark B, verifiers (Problem 1) cannot be measured.

3.2 Failure of Static Ground Truth

The standard paradigm builds a static benchmark $B = \{(c_i, d_i, y_i^h)\}_{i=1}^N$, where one-shot human labels y_i^h are treated as gold (Wang et al., 2024), and scores a verifier by exact match: $\text{Score}(M; B) = \frac{1}{N}\sum_i \mathbf{1}[\hat{y}_i = y_i^h]$. This is brittle for DRRs. Verifying a claim may require long-context reasoning and cross-source evidence synthesis, making oversights hard to avoid even for experts (Salvador-

Oliván et al., 2019). Meanwhile, expertise is *fragmented*: each DRR spans narrow sub-topics for which few experts overlap, making multi-annotator redundancy impractical. This motivates a central question: *can experts alone provide reliable ground truth for DRR verification?*

4 Empirical Analysis: The Unreliability of Expert Verification

To test whether experts can reliably label DRR claims, we run a controlled study with PhD specialists annotating risk-stratified claims, embedding hidden micro-gold checks to measure accuracy.

4.1 Methodology: The Micro-Gold Protocol

To measure annotation quality, we created a "micro-gold" set of hidden, known-answer claims, generated via a scalable, two-pronged approach that requires minimal domain expertise:

1. Unsupported Micro-Golds. We generate unsupported micro-golds by modifying authentic DRR sentences to introduce controlled factual errors. Modifications are guided by an error taxonomy distilled from a pilot study of real model failures (detailed in Table 8), covering three cognitive stages (Pirolli and Card, 2005; Kuhlthau, 1991): *collection-stage errors* (e.g., hallucinated references, misattributed citations, or contextually irrelevant retrieval), *analysis-stage errors* (e.g., misinterpreting or incorrectly synthesizing evidence, causal inversion, or merging distinct facts into a misleading claim), and *generalization-stage errors* (e.g., over-generalization, taxonomic simplification, or neglected qualifiers). Guided by this taxonomy, we inject realistic failure modes into claims and then verify only that the introduced modification is indeed false, making unsupported micro-gold construction much cheaper than open-ended verification. All injected errors were manually verified by the authors (see examples in Appendix L.1).

2. Supported Micro-Golds. We selected claims with explicit citations and narrow factual scope. Each candidate underwent a two-stage validation: an LLM-based entailment check against the citation, followed by a human review to confirm both the entailment and the narrow scope.

Usage and Validation. These micro-golds, using a 1:4 supported-to-unsupported ratio, were hidden within annotation batches and comprised 25% of all items. Annotator performance on this set provided a continuous measure of reliability. After the main annotation, we revealed the micro-golds to the experts, who reconfirmed their quality, thereby further validating them (details in Appendix B.6).

4.2 Study Setup

To ensure annotation competence and broad domain coverage, we recruited PhD-level domain experts who are active contributors in fields such as control theory, environmental engineering, education, public health, and engineering management. Each annotator began by proposing six research questions within their area of expertise, defined as domains in which they had at least one first-author, peer-reviewed publication. Then, among the six DRRs generated in response to these questions by deep research models (detailed in Appendix B.4), we let them choose the three they were most confident evaluating. This setup ensured that annotators were familiar with the subject matter, allowing them to verify complex claims more accurately with lower cognitive load. Moreover, we told them that there were hidden tests they needed to pass to receive full compensation (see Appendix B for details). This setup helped ensure that they were both *qualified* and *well-motivated*.

4.3 Finding 1:Static Expert Gold is Unreliable

We had experts independently annotate sampled important and risky claims (40 claims/report, details in Appendix E) from DRRs in their domains of expertise, and we scored their performance on the micro-gold set. However, they achieved only 60.8% micro-gold accuracy, showing that expert "gold" labels created within a finite time for complex DRR claims are unreliable. Yet multi-expert redundancy is impractical given fragmented expertise, motivating a new paradigm.

5 Evolving Benchmarking via AtS

5.1 The Audit-then-Score Protocol

To address the brittleness of the one-shot expert labels, we replace the standard *annotate-once-then-score* static benchmarking pipeline with **evolving benchmarking** via **Audit-then-Score** (AtS), a protocol in which benchmark labels and rationales are revisable, but only through evidence-backed auditing. AtS maintains a versioned benchmark state: $B_t = \{(c_i, d_i, y_i^{(t)}, \rho_i^{(t)})\}_{i=1}^{N}$, where c_i is a claim,

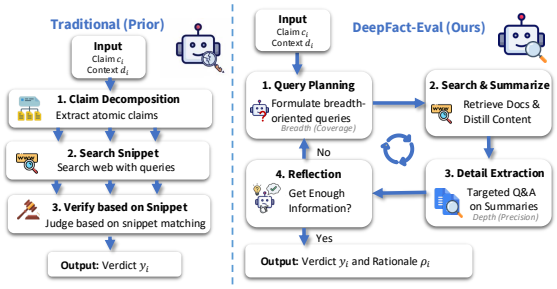


Figure 2: **DeepFact-Eval vs. traditional fact-checkers**: left, simplified VeriScore/FactCheck-GPT/SAFE; right, DeepFact-Eval workflow

d_i is its DRR context, $y_i^{(t)}$ is the current factuality verdict, and $\rho_i^{(t)}$ is the rationale.

1. Audit. At round t, a **Challenger** verifier M_t is run against B_t. Whenever the Challenger disagrees with the current benchmark state, it must submit not only an alternative verdict $\hat{y}_i$ but also an auditable rationale $\hat{\rho}_i$ grounded in evidence. These disagreements form a proposal set $U_{M_t,t} = \{(i, \hat{y}_i, \hat{\rho}_i)\}$. An **Auditor** A_t adjudicates each proposal by comparing the Challenger's rationale against the rationale in B_t. Formally, accepted updates are $\Delta B_t = \{(c_i, d_i, \hat{y}_i, \hat{\rho}_i) \mid A_t(\hat{\rho}_i, \rho_i^{(t)}) = \texttt{ACCEPT}\}$, and the benchmark evolves as $B_{t+1} = B_t \oplus \Delta B_t$, where $\oplus$ denotes the update operation that replaces the current verdict and rationale with the accepted revision. Disagreement alone does not change the benchmark: a revision is adopted only if the Challenger provides a stronger evidence-backed rationale and that rationale survives audit.

2. Score. The Challenger is then scored against the *updated* benchmark state B_{t+1} rather than the stale state B_t that it originally challenged. AtS preserves benchmark stability while allowing stronger verifiers to expose and correct weaknesses in the current consensus. See Algorithm 1 for details.

5.2 Governance and Controlled Evolution

AtS requires explicit protocol-level governance. Benchmark creators maintain the benchmark evolution and release *immutable* benchmark versions B_t. To prevent quality drift, evolution must be monitored using calibration signals such as hidden micro-golds. For fair comparison over time, reported results must specify the benchmark version used. Evolution stops when a pre-specified audit budget is exhausted, calibration quality stabilizes or reaches a target, or a fixed maintenance horizon is reached; details are provided in Appendix H.

6 Instantiating AtS: DeepFact

We instantiate AtS with two concrete artifacts under the DeepFact framework: DeepFact-Bench, a versioned benchmark for claim-level factuality verification in DRRs, and DeepFact-Eval, a verifier designed for DRR claim checking.

6.1 DeepFact-Bench

Each item in DeepFact-Bench is a tuple (c_i, d_i, y_i, ρ_i): a claim sentence, its source DRR as context, the current audited verdict, and an auditable rationale recording the supporting evidence and reasoning. AtS maintains DeepFact-Bench through versioned releases: each release is a frozen snapshot with full revision provenance.

The release used in this paper is DeepFact-Bench v4, produced by the four-round AtS rollout (one expert seed round plus three audit rounds against progressively stronger challengers in § 6.3). It contains 944 claims from 20 reports spanning six domains. We use 323 claims from 5 CS reports as the validation split and 621 claims from 15 reports across the remaining five domains as the test split (control theory, environmental engineering, education, public health, and engineering management). Within the test set, 143 claims are micro-golds, of which 120 are adversarially constructed. Excluding those adversarial examples, 27.0% of the remaining test claims are naturally unsupported. Appendix B.6 provides a post-hoc quality check after AtS, and Appendix L.2 presents qualitative examples of factual errors in DRRs.

6.2 DeepFact-Eval

Unlike prior fact-checkers that rely on snippet matching or cited-source checking, DeepFact-Eval verifies at the document level, where claims often depend on uncited synthesis, broader literature context, and technical distinctions spread across multiple sources. Because it produces evidence-backed verdicts and rationales, DeepFact-Eval can also serve as the Challenger within AtS.

As shown in Figure 2, DeepFact-Eval is designed to combine *breadth* and *depth*. Given a claim and its report context, it reads the surrounding report, plans search queries to cover the relevant document space, retrieves and summarizes candidate sources, and then asks targeted follow-up questions to recover claim-critical details that coarse summaries may miss. It iterates this retrieve–interrogate–reason loop until the evidence is suf-

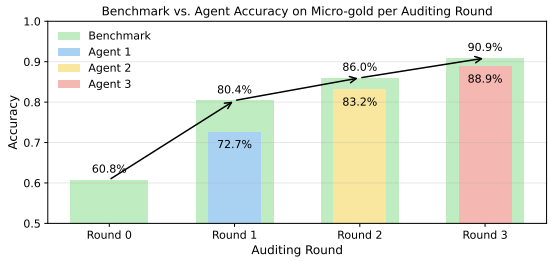


Figure 3: **Benchmark and challenger accuracy on hidden micro-golds across AtS rounds.** Outer bars show the updated benchmark accuracy after each audit round; inner bars show the corresponding challenger accuracy before audit. Human auditors improve benchmark accuracy from 60.8% to 90.9%.

ficient to produce both a verdict and an auditable rationale. To improve efficiency, we also introduce **DeepFact-Eval-lite**, a grouped variant that jointly verifies semantically related claims that share evidence and report context.

6.3 DeepFact AtS Rollout

DeepFact-Bench is built in four AtS rounds: one expert seed round, then three rounds of expert auditing against progressively stronger challengers. **Round 0 (Expert-only).** Experts annotate claims independently to initialize the seed benchmark. **Rounds 1–3 (Expert auditing agents).** Experts then audit progressively stronger challengers, each conditioned on the consensus from the previous round: **Agent1 (A1): SmolAgents (GPT-4.1, see § 8.1)**, **Agent2 (A2): DeepFact-Eval (GPT-4.1)**, and **Agent3 (A3): DeepFact-Eval (GPT-5)**. In each round, human auditors review only the challenger's disagreements with the current benchmark and accept a proposed revision only when its rationale provides stronger evidence or reasoning than the incumbent one. Accepted revisions update the benchmark to the next version, which becomes the input state for the following round.

7 Results: Validating Audit-then-Score

This section asks a simple question: *if benchmark labels are allowed to evolve, why should we trust the process?* We validate AtS by asking whether human auditors improve benchmark quality, whether agents can proxy auditors, how audit frequency and revision strictness shape evolution, and whether the process remains practical to maintain.

Decision Flow	Interpretation	Proportion
H:0 → A:1 → H':1	Human learns from correct agent	22.4%
H:0 → A:0 → H':1	Human learns from wrong agent	0.0%
H:1 → A:1 → H':1	Both correct throughout	44.8%
H:1 → A:0 → H':1	Human resists being misled	13.3%
H:0 → A:1 → H':0	Stays wrong despite right agent	5.6%
H:0 → A:0 → H':0	Both wrong; no improvement	11.2%
H:1 → A:0 → H':0	Human misled by wrong agent	2.8%
H:1 → A:1 → H':0	Human flips to wrong despite both correct	0.0%

Table 1: Human–agent decision flows on micro-gold claims, showing correctness transitions (1/0) for Human (H), Agent (A), and post-audit Human (H'). Flow patterns ending correct are in green; wrong ones are in red.

7.1 Finding 2: Humans Are Effective Auditors

We track how AtS improves benchmark quality over the four rounds through micro-gold accuracy.

Experts are weak alone, but strong under audit. Experts struggle to verify DRR claims in isolation: in *Round 0* of Figure 3, micro-gold accuracy is only 60.8%, highlighting how long-context, cross-source DRR verification can hide experts' blind spots. In contrast, auditing agent verdicts and rationales substantially improves expert accuracy, which increases monotonically as the challenger strengthens (Agent1 → Agent2 → Agent3; Figure 3), supporting AtS: fallible experts can refine benchmarks when scaffolded by strong verifiers.

Expert auditors are selective, not passive. This improvement does not come from blindly following the challenger. The dominant decision flows in Table 1 are exactly the patterns AtS relies on: auditors adopt correct agent revisions when their initial judgment was wrong (22.4%), and they resist incorrect agent suggestions when their original judgment was already correct (13.3%). By contrast, harmful flips are rare (2.8%). The auditor is therefore neither a rubber stamp nor a fixed source of truth; AtS improves benchmark quality by turning disagreement into disciplined adjudication.

7.2 Finding 3: Agents Can Serve as Auditors

We test whether agents can be auditors by replicating AtS with agent auditors. Round 0: each agent A_i verifies claims independently. Round 1: A_i audits another agent A_j by adjudicating between their Round-0 outputs to produce an updated decision. For each A_i, we report solo accuracy and audited accuracy when auditing the other two agents.

Agents are non-regressive auditors. As shown in Figure 4, across all pairings, auditing outperforms the audited agent's solo micro-gold baseline

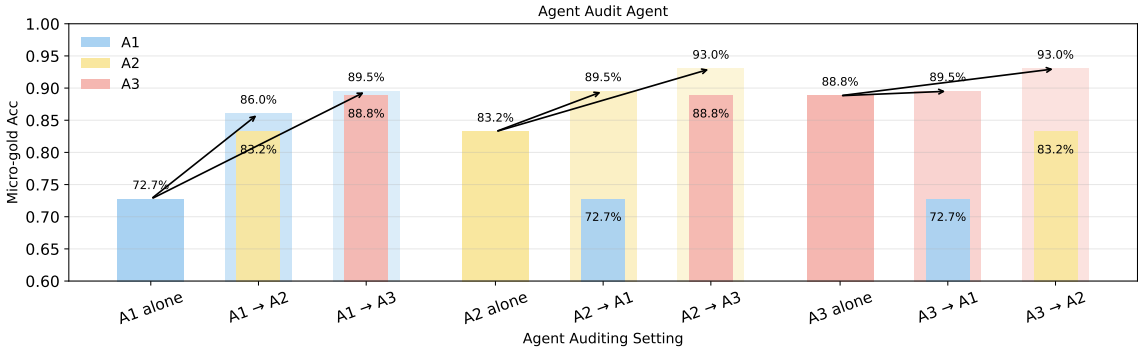


Figure 4: **Agent-only auditing for AtS**. For each auditor A_i, we report its Round-0 solo accuracy and its Round-1 audited accuracy when auditing another agent A_j ($A_i \to A_j$; outer bars). Inner bars within each $A_i \to A_j$ show the audited agent's solo (Round-0) accuracy A_j for reference.

in both weaker→stronger and stronger→weaker directions (e.g., $A2 \to A3$ and $A3 \to A2$ exceed both $A2$ solo and $A3$ solo), indicating that agent auditors can combine complementary evidence and catch oversights, creating a benchmark that surpasses the individual verifiers.

Auditing consolidates; Verifiers expand. Stronger→weaker and weaker→stronger auditing perform similarly (e.g., $A1 \to A3 \approx A3 \to A1$), indicating that auditing is constrained by available evidence rather than adjudication skill. Therefore, the auditor serves to **consolidate** a rigorous baseline from existing outputs, but benchmark evolution depends entirely on stronger verifiers to **expand** the information scope.

7.3 Ablations on Evolution Dynamics

We study how two design choices affect benchmark evolution dynamics: *how often* conflicts are audited, and *how strict* the revision rule should be.

Audit frequency controls refinement speed. Using an offline counterfactual replay, we audit a random fraction $p \in \{0.25, 0.5, 0.75, 1.0\}$ of detected conflicts in each round. Higher audit frequency accelerates early improvement: Round-1 micro-gold accuracy is $66.4/72.0/73.4/80.4$, and Round-2 accuracy is $68.5/76.2/81.8/85.3$, for $p = 0.25/0.5/0.75/1.0$. By Round-3, returns diminish and performance largely converges at the high end: $76.2/85.3/89.5/90.9$, with $p = 0.75$ within 1.4 points of full auditing. Audit frequency therefore mainly controls *how quickly* a benchmark improves, making it a natural knob for trading annotation budget against release velocity.

Revision strictness trades recall for conservativeness. We also test a stricter update rule that accepts a revision only when both a human auditor and an agent auditor agree. Stricter gating can filter noisy revisions, but it can also block beneficial updates: it raises Round-2 micro-gold accuracy from 86.0% to 88.0%, but slightly reduces Round-3 accuracy from 90.9% to 90.2%. The key tradeoff is between accepting noisy revisions and freezing avoidable errors into the benchmark.

7.4 Cost and Post-Release Maintenance

AtS amortizes expert cost rather than multiplying it. Constructing DeepFact-Bench required over 400 expert-hours, including more than 250 hours of paid external expert annotation at about $30/hour on average (Appendix B). However, AtS is not the main cost driver: the dominant cost is the one-time expert verification needed to seed any benchmark for a task as difficult as DRR claim checking. We therefore decompose total effort into the base cost of building B_0 and the incremental cost of later AtS rounds. The base cost dominates: Round 0 uses 65.5% of total expert time, while the three later AtS rounds use only 34.5% yet raise benchmark accuracy from 60.8% to 90.9%. As conflicts shrink, later rounds also become cheaper: the number of test claims requiring expert review falls from 621 in Round 0 to 361, 247, and 182. AtS therefore does not multiply annotation cost; it amortizes the unavoidable cost of expert verification into cheaper refinement rounds.

DeepFact-Bench uses gated post-release maintenance. AtS iterations are already cheaper than seed construction, and post-release maintenance can be made cheaper by auditing only when improvement is likely. After v4, we will trigger a new audit cycle only when a challenger beats the cur-

Model	Backbone	Quality				Efficiency		
		Acc	F1	Precision	Recall	Input Tokens	Output Tokens	Cost
Main Results								
Factcheck-GPT	GPT-4.1	55.0	58.3	67.7	51.2	–	–	–
SAFE	GPT-4.1	55.9	53.0	76.3	40.6	–	–	–
VeriScore	GPT-4.1	52.5	48.9	71.9	37.0	–	–	–
Fire	GPT-4.1	58.5	63.2	69.2	58.3	–	–	–
GPT-Researcher (Deep)	GPT-4.1	69.1	79.7	66.7	98.9	52.3K	9.0K	$0.18
GPT-Researcher (Deep+)	GPT-4.1	68.3	79.3	66.1	**99.2**	83.3K	13.9K	$0.28
SmolAgents	GPT-4.1	68.8	69.5	58.0	86.7	294.4K	3.4K	$0.62
DeepFact-Eval	GPT-4.1	**83.4**	**86.9**	**85.7**	88.2	516.9K	18.6K	$1.16
DeepFact-Eval (Group=5)	GPT-4.1	77.9	83.1	78.5	88.2	131.4K	4.9K	$0.30
DeepFact-Eval (Group=10)	GPT-4.1	76.3	82.2	76.4	89.0	93.5K	3.5K	$0.21
Model Ablations								
DeepFact-Eval	GPT-5	87.2	89.9	87.9	91.9	–	–	–
DeepFact-Eval	Gemini-2.5-Pro	81.5	85.0	84.6	85.3	–	–	–
DeepFact-Eval	Qwen-3-32B	72.5	77.4	78.1	76.6	–	–	–

Table 2: **Comparison of fact-checkers on DeepFact-Bench** (accuracy/F1/precision/recall) and efficiency. Best GPT-4.1-backbone results are **bolded**. Traditional and deep-research methods are color-coded.

rent benchmark on the hidden micro-gold set. We rely on agent auditors for routine revisions, while reserving human experts for periodic recalibration once cumulative updates change more than 5% of benchmark verdicts. We will continue evolving the benchmark until micro-gold accuracy saturates.

8 Results: Evaluating Verifiers

We first benchmark verifiers on the released DeepFact-Bench v4 test set (Table 2). Because no baseline surpasses DeepFact-Eval on the hidden micro-gold set, we do not evolve the benchmark further during head-to-head comparison.

8.1 Baselines

FactCheck-GPT (Wang et al., 2024) extracts atomic claims, retrieves evidence, judges stance, and issues corrections. **SAFE** (Wei et al., 2024) breaks responses into atomic facts and iteratively issues Google Search queries, judging support from retrieved snippets. **VeriScore** (Song et al., 2024) follows a similar retrieve–judge paradigm but verifies only *verifiable* claims in a single optimized pass. **FIRE** (Xie et al., 2025) casts verification as an agentic loop: it either returns a verdict or generates a follow-up query and repeats until confident. Finally, we repurpose deep-research-style scaffolding as a verifier baseline for comparison: **GPTResearcher (Deep Research Mode)** (Elovic, 2025), a workflow agent that iteratively performs query planning and retrieval-augmented synthesis with a tunable search-depth budget ("Deep+" uses a larger budget), and **SmolAgents** (Roucher et al., 2025), a ReAct-style agent where a main agent can invoke sub-agents to interact with websites and gather evidence. See Appendix C for details.

8.2 Results on DeepFact-Bench

Following (Song et al., 2024), we merge contradictory and inconclusive cases into `Unsupported`, yielding a binary supported/unsupported setting. We report accuracy, F1, precision/recall for the supported class, plus efficiency (I/O tokens and estimated cost per claim) for deep-research methods only, since snippet-based pipelines have negligible cost (details in Appendix C). As no baseline outperforms DeepFact-Eval on micro-gold, we do not run benchmark evolution and instead evaluate all methods on the current snapshot (see Table 2).

DeepFact-Eval achieves the best performance. DeepFact-Eval attains the highest accuracy (83.4%; Table 2), outperforming both traditional fact-checking pipelines (best: 58.5%) and prior deep-research agent baselines (best: 69.1%). Deep-research verifiers generally outperform snippet-based methods (e.g., GPT-Researcher 69.1 vs. VeriScore 52.5) because DRR claims rarely have a single verbatim supporting span; evidence is often distributed across a document. This gap is reflected in the error profiles: snippet-based checkers are high-precision but low-recall and often default to `Unsupported` when retrieval fails, while general deep-research agents improve recall but sacrifice precision by accepting fuzzy topical support.

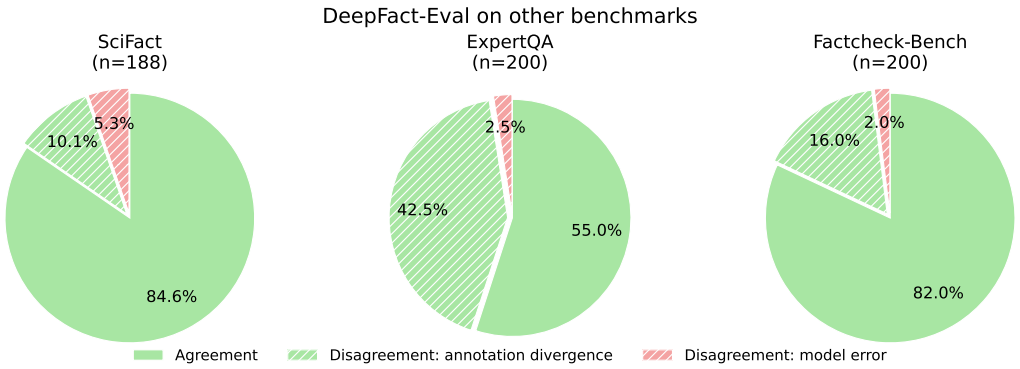


Figure 5: **Results of DeepFact-Eval on SciFact, ExpertQA, Factcheck-Bench.** Solid green indicates *Agreement* (verifier's prediction matches the benchmark label). Hatched slices denote *Disagreements* (verifier's prediction doesn't match the benchmark label). Green-hatched indicates *Annotation divergence* (e.g., evidence–label misalignment, non-verifiable/ambiguous sentences, subjective or underspecified claims, or annotation divergence can't be resolved due to the lack of a gold rationale), while red-hatched indicates *Likely model error* (expert re-annotation aligns with the benchmark label).

DeepFact-Eval closes this gap via detail verification with targeted deep queries that surface the technical distinctions that determine support, yielding both high precision and high recall. See qualitative examples of DeepFact-Eval in Appendix L.3.

Grouping reduces cost with minimal quality trade-off. DeepFact-Eval reduces verification cost substantially via grouped verification, with only a minor loss in accuracy, making it a cost-efficient option. Notably, DeepFact-Eval (Group=10) outperforms GPT-Researcher at a comparable budget (76.3 vs. 69.1). In contrast, scaling GPT-Researcher's compute by increasing max search depth does not improve performance (69.1 $\rightarrow$ 68.3) despite higher claim cost ($0.18 $\rightarrow$ $0.28).

DeepFact-Eval generalizes across backbones. Upgrading the backbone of DeepFact-Eval to a stronger model (GPT-5) improves performance, while Gemini-2.5-Pro performs comparably to GPT-4.1. Using an open-source backbone (Qwen3-32B) reduces accuracy, but DeepFact-Eval still outperforms other baselines using GPT-4.1.

8.3 Results on Other Factuality Benchmarks

We further test whether DeepFact-Eval generalizes beyond DeepFact-Bench by evaluating it on SciFact (Wadden et al., 2020), ExpertQA (Malaviya et al., 2024), and Factcheck-Bench (Wang et al., 2024). To better understand its errors, we audit disagreements to distinguish verifier failures from benchmark artifacts. On SciFact, where evidence rationales are available, we directly compare disagreements against the provided evidence; on ExpertQA and Factcheck-Bench, where rationale support is weaker or absent, we use blinded re-annotation on disagreement subsets. Full setup details are deferred to Appendix D. We find that DeepFact-Eval transfers well to these benchmarks, and that many of its residual disagreements are better explained by benchmark brittleness than by verifier failure. After auditing disagreement cases, its estimated accuracy rises from 84.6% to 94.7% on SciFact and from 82.0% to 98.0% on Factcheck-Bench. On ExpertQA, expert re-annotation also frequently sides with DeepFact-Eval, though the lack of gold rationales makes adjudication less definitive. Static factuality benchmark noise can undercredit modern verifiers, reinforcing the case for auditable, evolving evaluation.

9 Conclusion

We introduced DeepFact, a framework for evaluating factuality in AI-generated deep research reports. Our central finding is that static expert labels are brittle in this setting. Audit-then-Score turns benchmark labels into an auditable, revisable consensus, improving expert micro-gold accuracy across audit rounds. We also introduce DeepFact-Eval, a verifier that outperforms prior baselines on the frozen benchmark snapshot and transfers to external factuality datasets. More broadly, our results suggest that as AI systems approach expert-level performance, trustworthy evaluation may require benchmarks that co-evolve with verifier capabilities through evidence-backed human–AI auditing.

Limitations

Our current verifiers function as expert literature reviewers rather than active laboratory scientists, constrained to validating claims against *existing* scientific literature. Consequently, they cannot empirically verify findings through new experiments or data simulations—a gap that future "AI Scientists" must bridge to address scenarios where the literature is silent or conflicted (Lu et al., 2024). Beyond these epistemic limits, significant opportunities for efficiency improvement remain. Although we introduce a lite variant of DeepFact-Eval, the necessity of long-context reasoning and iterative retrieval makes deep verification expensive for long-form reports. This computational burden currently limits its real-time applicability, despite the framework's necessity for high-stakes accuracy.

Ethical Considerations

While our tools are designed to verify factuality, the underlying technology (generating and refining complex claims) could theoretically be repurposed to generate sophisticated factual inaccuracies. However, we believe the development of strong verification tools is the most effective countermeasure against such risks. By releasing DeepFact-Bench and DeepFact-Eval, we aim to empower the community to detect and refute hallucinated or manipulated scientific reports.

Acknowledgments

Part of this work was conducted during an internship at Amazon. This work was also supported in part by NSF award IIS-2211526. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, and 6 others. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2022. Fact checking with insufficient evidence. *Transactions of the Association for Computational Linguistics*, 10:746–763.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4685–4697, Hong Kong, China. Association for Computational Linguistics.

Farima Fatahi Bayat, Lechen Zhang, Sheza Munir, and Lu Wang. 2025. Factbench: A dynamic benchmark for in-the-wild language model factuality evaluation. *Preprint*, arXiv:2410.22257.

Prahaladh Chandrahasan, Jiahe Jin, Zhihan Zhang, Tevin Wang, Andy Tang, Lucy Mo, Morteza Ziyadi, Leonardo F. R. Ribeiro, Zimeng Qiu, Markus Dreyer, Akari Asai, and Chenyan Xiong. 2025. Deep research comparator: A platform for fine-grained human annotations of deep research agents. *Preprint*, arXiv:2507.05495.

Justin Chen, Swarnadeep Saha, and Mohit Bansal. 2024. ReConcile: Round-table conference improves reasoning via consensus among diverse LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7066–7085, Bangkok, Thailand. Association for Computational Linguistics.

Sanxing Chen, Yukun Huang, and Bhuwan Dhingra. 2025. Real-time factuality assessment from adversarial feedback. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1610–1630, Vienna, Austria. Association for Computational Linguistics.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. 2025. Deepresearch bench: A comprehensive benchmark for deep research agents. *ArXiv*, abs/2506.11763.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Julian H. Elliott, Anneliese Synnot, Tari Turner, Mark Simmonds, Elie A. Akl, Steve McDonald, Georgia Salanti, Joerg Meerpohl, Harriet MacLehose, John Hilton, David Tovey, Ian Shemilt, James Thomas, and Living Systematic Review Network. 2017. Living systematic review: 1. introduction-the why, what, when, and how. *Journal of Clinical Epidemiology*, 91:23–30. Epub 2017 Sep 11.

Assaf Elovic. 2025. GPT researcher. GPT researcher is an open deep research agent designed for both web and local research on any given task.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. AmbiFC: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Hui Wen Goh, Ulyana Tkachenko, and Jonas Mueller. 2023. Crowdlab: Supervised learning to infer consensus labels and quality scores for data with multiple annotators. *Preprint*, arXiv:2210.06812.

Boyu Gou, Zanming Huang, Yuting Ning, Yu Gu, Michael Lin, Weijian Qi, Andrei Kopanev, Botao Yu, Bernal Jimenez Gutierrez, Yiheng Shu, Chan Hee Song, Jiaman Wu, Shijie Chen, Hanane Nour Moussa, TIANSHU ZHANG, Jian Xie, Yifei Li, Tianci Xue, Zeyi Liao, and 7 others. 2025. Mind2web 2: Evaluating agentic search with agent-as-a-judge. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277, Seattle, United States. Association for Computational Linguistics.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta programming for a multi-agent collaborative framework. In *The Twelfth International Conference on Learning Representations*.

Abhinav Java, Ashmit Khandelwal, Sukruta Midigeshi, Aaron Halfaker, Amit Deshpande, Navin Goyal, Ankur Gupta, Nagarajan Natarajan, and Amit Sharma. 2025. Characterizing deep research: A benchmark and formal definition. *Preprint*, arXiv:2508.04183.

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

Jiahe Jin, Abhijay Paladugu, and Chenyan Xiong. 2025. Beneficial reasoning behaviors in agentic search and effective post-training to obtain them. *Preprint*, arXiv:2510.06534.

Jungo Kasai, Keisuke Sakaguchi, yoichi takahashi, Ronan Le Bras, Akari Asai, Xinyan Velocity Yu, Dragomir Radev, Noah A. Smith, Yejin Choi, and Kentaro Inui. 2023. Realtime QA: What's the answer right now? In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Douwe Kiela, Max Bartolo, Yixin Nie, Divyansh Kaushik, Atticus Geiger, Zhengxuan Wu, Bertie Vidgen, Grusha Prasad, Amanpreet Singh, Pratik Ringshia, Zhiyi Ma, Tristan Thrush, Sebastian Riedel, Zeerak Waseem, Pontus Stenetorp, Robin Jia, Mohit Bansal, Christopher Potts, and Adina Williams. 2021. Dynabench: Rethinking benchmarking in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4110–4124, Online. Association for Computational Linguistics.

Carol C. Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science*, 42(5):361–371.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Xin Liu, Lechen Zhang, Sheza Munir, Yiyang Gu, and Lu Wang. 2025. VeriFact: Enhancing long-form factuality evaluation with refined fact extraction and reference facts. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17919–17936, Suzhou, China. Association for Computational Linguistics.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Chaitanya Malaviya, Subin Lee, Sihao Chen, Elizabeth Sieber, Mark Yatskar, and Dan Roth. 2024. ExpertQA: Expert-curated questions and attributed answers. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3025–3045, Mexico City, Mexico. Association for Computational Linguistics.

Dasha Metropolitansky and Jonathan Larson. 2025. Towards effective extraction and evaluation of factual claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6996–7045, Vienna, Austria. Association for Computational Linguistics.

Omer Nahum, Nitay Calderon, Orgad Keller, Idan Szpektor, and Roi Reichart. 2025. Are LLMs better than reported? detecting label errors and mitigating their effect on model performance. In *Proceedings*

of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 26770–26797, Suzhou, China. Association for Computational Linguistics.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial NLI: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901, Online. Association for Computational Linguistics.

OpenAI. 2024. Introducing ChatGPT search.

Liana Patel, Negar Arabzadeh, Harshit Gupta, Ankita Sundar, Ion Stoica, Matei Zaharia, and Carlos Guestrin. 2025. Deepscholar-bench: A live benchmark and automated evaluation for generative research synthesis. *Preprint*, arXiv:2508.20033.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, Michael Choi, Anish Agrawal, Arnav Chopra, Adam Khoja, Ryan Kim, Richard Ren, Jason Hausenloy, Oliver Zhang, Mantas Mazeika, and 1093 others. 2025. Humanity's last exam. *Preprint*, arXiv:2501.14249.

Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of the International Conference on Intelligence Analysis*, volume 5, pages 2–4, McLean, VA, USA.

Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ChatDev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15174–15186, Bangkok, Thailand. Association for Computational Linguistics.

Aymeric Roucher, Albert Villanova del Moral, Thomas Wolf, Leandro von Werra, and Erik Kaunismäki. 2025. 'smolagents': a smol library to build great agentic systems. `https://github.com/huggingface/smolagents`.

Jie Ruan, Inderjeet Nair, Shuyang Cao, Amy Liu, Sheza Munir, Micah Pollens-Dempsey, Tiffany Chiang, Lucy Kates, Nicholas David, Sihan Chen, Ruxin Yang, Yuqian Yang, Jasmine Gump, Tessa Bialek, Vivek Sankaran, Margo Schlanger, and Lu Wang. 2025. Expertlongbench: Benchmarking language models on expert-level long-form generation tasks with structured checklists. *Preprint*, arXiv:2506.01241.

José Antonio Salvador-Oliván, Gonzalo Marco-Cuenca, and Rosario Arquero-Avilés. 2019. Errors in search strategies used in systematic reviews and their effects on information retrieval. *Journal of the Medical Library Association*, 107(2):210–221.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, and 2 others. 2025. Dr tulu: Reinforcement learning with evolving rubrics for deep research. *Preprint*, arXiv:2511.19399.

Manasi Sharma, Chen Bo Calvin Zhang, Chaithanya Bandi, Clinton Wang, Ankit Aich, Huy Nghiem, Tahseen Rabbani, Ye Htet, Brian Jang, Sumana Basu, Aishwarya Balwani, Denis Peskoff, Marcos Ayestaran, Sean M. Hendryx, Brad Kenstler, and Bing Liu. 2025. Researchrubrics: A benchmark of prompts and rubrics for evaluating deep research agents. *Preprint*, arXiv:2511.07685.

Andries Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474, Miami, Florida, USA. Association for Computational Linguistics.

Michael Soprano, Kevin Roitero, David La Barbera, Davide Ceolin, Damiano Spina, Gianluca Demartini, and Stefano Mizzaro. 2024. Cognitive biases in fact-checking and their countermeasures: A review. *Information Processing & Management*, 61(3):103672.

Camille Thibault, Jacob-Junqi Tian, Gabrielle Péloquin-Skulski, Taylor Lynn Curtis, James Zhou, Florence Laflamme, Luke Yuxiang Guan, Reihaneh Rabbany, Jean-François Godbout, and Kellin Pelrine. 2025. A guide to misinformation detection data and evaluation. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, page 5801–5809, New York, NY, USA. Association for Computing Machinery.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018.

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

Mariken A. C. G. van der Velden, Felicia Loecherbach, Wouter van Atteveldt, Antske Fokkens, Myrthe Reuver, and Kasper Welbers. 2025. Whose truth is it anyway? an experiment on annotation bias in times of factual opinion polarization. *Communication Methods and Measures*, 19(4):332–349.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2024. Freshllms: Refreshing large language models with search engine augmentation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13697–13720, Bangkok, Thailand. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

Jiayu Wang, Yifei Ming, Riya Dulepet, Qinglin Chen, Austin Xu, Zixuan Ke, Frederic Sala, Aws Albarghouthi, Caiming Xiong, and Shafiq Joty. 2025. Liveresearchbench: A live benchmark for user-centric deep research in the wild. *Preprint*, arXiv:2510.14240.

William Yang Wang. 2017. "liar, liar pants on fire": A new benchmark dataset for fake news detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 422–426, Vancouver, Canada. Association for Computational Linguistics.

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, Isabelle Augenstein, Iryna Gurevych, and Preslav Nakov. 2024. Factcheck-bench: Fine-grained evaluation benchmark for automatic fact-checkers. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14199–14230, Miami, Florida, USA. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Zixia Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V Le. 2024. Long-form factuality in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2024. Autogen: Enabling next-gen LLM applications via multi-agent conversations. In *First Conference on Language Modeling*.

Zhuohan Xie, Rui Xing, Yuxia Wang, Jiahui Geng, Hasan Iqbal, Dhruv Sahnan, Iryna Gurevych, and Preslav Nakov. 2025. FIRE: Fact-checking with iterative retrieval and verification. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2901–2914, Albuquerque, New Mexico. Association for Computational Linguistics.

Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Taira Anderson, Jonathan Bragg, Joseph Chee Chang, Jesse Dodge, Matt Latzke, Yixin Liu, Charles McGrady, Xiangru Tang, Zihang Wang, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. 2025. Sciarena: An open evaluation platform for foundation models in scientific literature tasks. *Preprint*, arXiv:2507.01001.

A DRR Claims Factuality Definition

A.1 DRR Claims Factuality Definition

We adapt VeriScore's atomic-fact definitions (Song et al., 2024) of supported, contradictory, and inconclusive to the sentence level for DRR evaluation (Table 3).

A.2 Non-Verifiable Definition

We define the main types of non-verifiable content in DRRs in Table 7, based on our in-house annotation findings.

A.3 Deep Research Report Claim Error Taxonomy

We observed several common error patterns in deep research models during our pilot annotations. We therefore categorize them as shown in Table 8.

B Expert Annotations

B.1 Pilot In-House Annotations

To calibrate the difficulty of DRR claim verification, we first conducted a pilot round of in-house annotations. We prompted LLMs with research questions and collected reports generated by multiple deep-research agents, then attempted to verify the resulting claims against the global literature.

This pilot immediately revealed that DRR verification is qualitatively more demanding than standard claim checking. A single report can contain *hundreds* of verifiable statements, and a single challenging claim can take *hours* to resolve due to multi-hop dependency on scattered, technical

Label	Definition (sentence-level)	Requirements / Decision criteria
Supported	All factual claims in the sentence are directly or logically supported by at least one source, and no equally or more credible source contradicts any part of the sentence.	(i) The source(s) cover every factual element of the sentence. (ii) If inference is used, it must be transparent and logically valid (no speculative leaps). (iii) Evidence is sufficient to guarantee the claim as stated. (iv) No equally or more credible source refutes any part of the sentence.
Inconclusive	If no claims are *Contradictory*, but at least one claim is *Inconclusive*, and the rest are either *Inconclusive* or *Supported*, then the sentence is marked *Inconclusive*.	(i) At least one claim lacks direct support *and* lacks credible refutation. (ii) The inference chain is weak/speculative or not clearly grounded. (iii) Evidence is missing or internally conflicting without a clear resolution.
Contradictory	If any single factual claim in the sentence is contradicted by a reliable source, and no equally strong support exists, the entire sentence is marked *Contradictory*.	(i) At least one claim is clearly refuted (negated or directly contradicted) by a reliable source. (ii) No stronger or equally reliable evidence supports the refuted claim.
None	The sentence contains no verifiable factual claims; it instead expresses opinions, vague speculation, moral judgments, or rhetorical language.	Non-verifiable (see Table 7).

Table 3: DRR sentence-level factuality labels. A sentence aggregates over its constituent factual claims: any contradicted claim yields *Contradictory*; otherwise, any unresolved claim yields *Inconclusive*; otherwise *Supported*; and *None* if no verifiable factual claims are present.

sources. This makes exhaustive, claim-by-claim verification infeasible at scale.

Domain drift and fragmentation amplify burden and reduce reliability. We found that even slight domain drift sharply increases annotation time while degrading reliability. For example, asking a PhD-level annotator specializing in LLM RL to verify a report centered on RAG (or vice versa) often led to slower verification and more errors, despite both topics falling under the broad "LLM" umbrella. Similarly, expertise can decay with *temporal drift*: an annotator who previously worked on RAG but has since shifted to agentic systems may be less familiar with the most recent literature, making verification substantially harder. These observations suggest that *hyper-specialization* is a core obstacle for DRR verification: the effective competence set is narrow, and modest topic/time mismatches can push claims beyond a reasonable verification budget (e.g., hours per claim).

Multi-annotator adjudication is not a silver bullet. In this setting, conventional multi-annotator adjudication is often less informative than expected. When secondary annotators have even slight domain mismatch, they frequently defer to the primary annotator (or converge on the same surface-level judgment) due to limited confidence and familiarity, which can inflate agreement without improving correctness.

Cognitive load is extreme. Finally, we observed pronounced attention decay over long annotation sessions. Annotators can remain highly focused on the first several claims, but performance deteriorates as the report length and verification horizon grow—a predictable failure mode when the task involves hundreds of decisions with heavy context switching.

Design implications. These pilot findings directly shaped our full-scale annotation protocol. To reduce cognitive overload, we (i) developed an annotation interface with visualization and navigation support, and (ii) used importance- and risk-stratified sampling to concentrate effort on the most consequential and error-prone claims rather than attempting exhaustive verification. (iii) Since expert annotation is no longer reliable, we design micro-gold to quantify how reliable it is.

B.2 Annotation Visualization

To reduce cognitive overload, we built an annotation interface Figure 6 that makes DRR reading and labeling lightweight. Annotators first select from

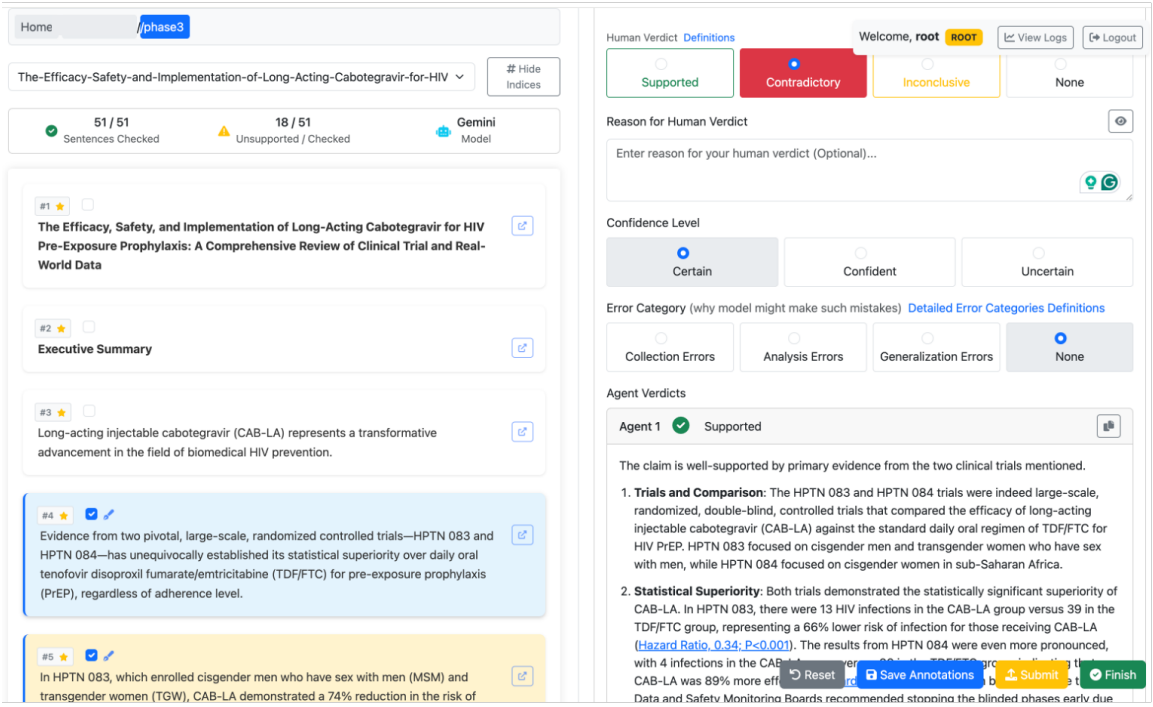


Figure 6: Annotation Interface for DRR

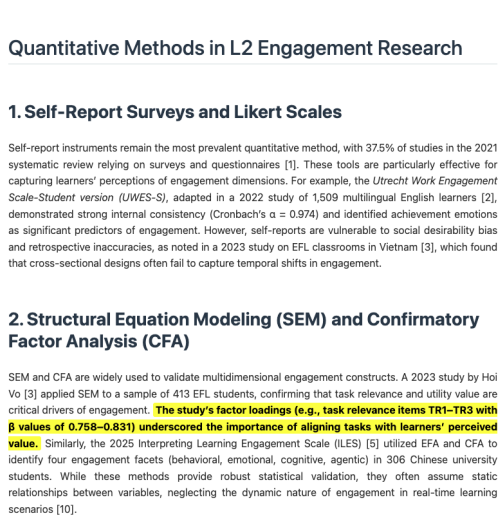
the set of reports assigned to them. For each report, the interface shows a clear progress indicator (e.g., how many claims remain). The report is segmented into sentences in the left panel, while the right panel contains the fields to complete: the human verdict and a brief justification. In phases where we provide model assistance, the UI also displays the agent(s)' verdicts and rationales as reference. For claims marked Unsupported, annotators additionally choose an error category. Definitions of the factuality labels and error taxonomy are embedded in the interface for quick lookup. Annotators can jump from any claim directly to its location in the original report for quick context recovery Figure 7a. The UI also supports checkpointing: they can reset to earlier checkpoints and resume later, enabling long-horizon annotation without losing state and helping prevent fatigue-driven errors Figure 7b.

B.3 Expert Recruitment

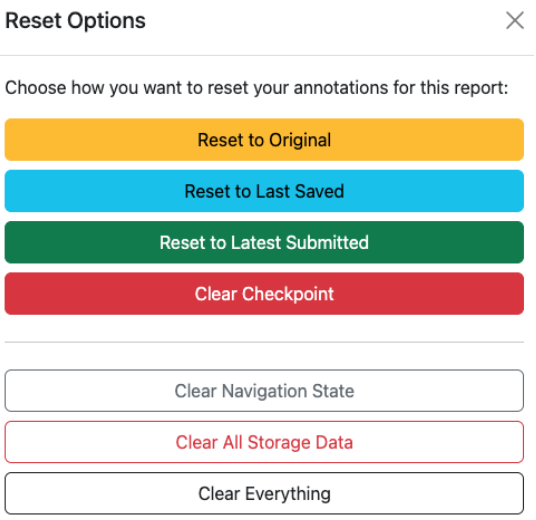
To mitigate domain drift, we recruit PhD-level experts through university channels and have them verify DRRs derived from their own research questions within their specialization. We require annotators to be currently enrolled PhD students to ensure up-to-date familiarity with recent literature; in pilot audits, more senior researchers (e.g., faculty) were often less attuned to fine-grained details despite strong broad expertise. Experts must demonstrate domain competence (i.e., first-author publications and reviewer experience in the area) and complete a short general-domain calibration task to align on our factuality definitions (Table 3). To discourage low-effort labeling and improve reliability, compensation is contingent on passing a hidden micro-gold quality check, incentivizing sustained attention throughout verification. Our experts span control theory, environmental engineering, education, public health, and engineering management. In total, the annotation process required more than 400 expert-hours. Participants were told that their responses (verdicts and rationales) would be used to develop and evaluate DRR factuality benchmarks and might be released in de-identified form. We did not collect or release personally identifying information; all records were anonymized and stored under access control. Compensation averages about $30/hour, and is region-adjusted and set above the local research-assistant hourly wage.

B.4 Annotation Stage 1: DRR Generation

This stage constructs challenging yet verifiable prompts that elicit deep research reports (DRRs) requiring multi-source reasoning, synthesis, and evidence integration—settings where agentic LLMs are more likely to fail. To ensure verifiability, each prompt must fall within the annotator's domain of expertise, so that a knowledgeable expert can



(a) Jump from a claim to its exact span in the original report for fast, low-friction context recovery.



(b) Reset to an earlier checkpoint to resume long-horizon annotation without losing progress.

Figure 7: **Interface features.** Left: Jump from a claim to its exact span in the original report for fast, low-friction context recovery. Right: Reset to an earlier checkpoint to resume long-horizon annotation without losing progress.

assess correctness efficiently.

Prompt authoring. Each expert follows the steps below:

1. **Choose niche subtopics.** Select 2–3 highly specific subtopics they have published on or know well (e.g., granular methods, datasets, or keywords from recent work).

2. **Write research-style questions.** Draft 6 well-scoped prompts that are precise (not generic) and typically require synthesizing or comparing multiple sources rather than summarizing a single document.

3. **Ensure factual verifiability.** Avoid (i) speculative/future-looking prompts (e.g., "predict trends"), (ii) opinion-based prompts (e.g., "is X good?"), and (iii) unrealistic or non-verifiable requests (e.g., proposing novel experiments without established evidence).

4. **Design for grounded outputs.** Prompts should yield factual claims supported by public literature and verifiable via online resources (papers, datasets, official reports).

LLM clarification. Given each question, we run GPT-4.1 to request clarifications and missing details, mimicking the "question refinement" step used by OpenAI Deep Research. We then integrate the feedback into more specific prompts.

Report generation. For each expert's 6 questions, we generate DRRs using three deep-research systems: OpenAI DeepResearch (o3), Gemini Deep Research (Gemini-2.5-pro), and OpenDeepResearch (Qwen-32B). We generate two reports per system.

Report selection. Each expert selects three reports they feel most confident verifying—one from each system.

B.5 Annotation Stage 2: DRR Annotation (Audit-then-Score)

Experts label and justify sampled claims from their selected DRRs through multiple rounds of auditing:

Round 0 (Expert-only). Experts annotate claims independently to form an initial static benchmark. We also ask each expert to indicate a confidence level (*certain*, *confident*, or *uncertain*).

Round 1 (Audit Agent 1: SmolAgent GPT-4.1). Experts review Agent 1's verdicts and rationales, adopting them when the agent provides stronger evidence or reasoning. In this and all following rounds, they are blinded to agent quality. They are explicitly told that both agent outputs and their own earlier annotations can be correct or incorrect, and that they must use their own judgment when deciding whether to revise a label.

Round 2 (Audit Agent 2: DeepFact-Eval GPT-4.1). Experts audit a stronger verifier and update

labels when its rationale is better supported or more coherent. From this round onward, experts also provide their own rationales when making updates.

Round 3 (Audit Agent 3: DeepFact-Eval GPT-5). Experts audit and incorporate revisions from the strongest verifier, producing the latest benchmark version.

For all stages, annotators are blinded to the agents' quality and are explicitly informed that both agent predictions and human annotations can contain errors. They are instructed to use their own judgment when accepting, rejecting, or revising any verdict. Across rounds, experts repeatedly revisit the same claims, encouraging careful reconsideration rather than one-shot labeling.

B.6 Annotation Stage 3: Post-hoc Quality Check

To assess the quality of the released benchmark (including the micro-gold set), we conduct a post-hoc quality check roughly one month after Stage 2 to reduce memorization effects.

Non-micro-gold items. Experts re-annotate with all agents' verdicts and rationales visible. We measure intra-expert consistency against their earlier decisions, following prior work (Zhao et al., 2025).

Micro-gold items. Experts are shown the micro-gold items and their construction process, and indicate whether they agree with the micro-gold verdict; if not, they provide a rationale. This adds an additional validity check for the micro-gold set.

Overall, intra-expert consistency is 92.7%, and the micro-gold confirmation rate is 99.3%.

Following this procedure, we obtain a test set of 621 claims from 15 reports spanning five domains. Separately, our in-house annotation yields a validation set of 323 claims from 5 CS reports. Together, these form DeepFact-Bench.

C Implementations

C.1 Traditional Methods

Methods such as VeriScore, SAFE, FIRE, and FactCheck-GPT operate at the atomic-claim level: they decompose each sentence into multiple atomic facts and output a verdict per fact, whereas our method outputs a single verdict per sentence. FIRE is also designed for atomic-claim verification but does not include a built-in decomposition step; therefore, we first extract atomic claims using a GPT-4.1 claim extractor and then run FIRE on these claims. To make all baselines comparable, we aggregate atomic-level verdicts into a sentence-level verdict.

We follow VeriScore's three-way label space {supported, inconclusive, contradictory}. FIRE and FactCheck-GPT use {true, not-enough-evidence, false}, which we map to supported, inconclusive, contradictory, respectively. For VeriScore/FIRE/FactCheck-GPT, we aggregate atomic verdicts using the rule: (i) if any atomic claim is contradictory, the sentence is contradictory; (ii) else if any atomic claim is inconclusive, the sentence is inconclusive; (iii) otherwise supported. For final evaluation, we merge contradictory and inconclusive into unsupported to obtain a binary label space {supported, unsupported}. SAFE outputs {supported, not supported} plus an irrelevant label. We aggregate SAFE by marking a sentence as unsupported if any atomic claim is not supported, otherwise supported; if all atomic claims are labeled irrelevant, we count the sentence as an incorrect prediction. To this end, we convert all methods to a sentence-level binary prediction, enabling fair comparison.

C.2 Cost Estimation

We estimate cost for all deep-research methods using OpenAI API token prices. DeepFact-Eval uses GPT-4.1 for verification, but uses GPT-4.1 mini to summarize full documents to reduce cost; we convert GPT-4.1 mini token usage into GPT-4.1–equivalent cost by scaling by the corresponding price ratios. GPT-Researcher is evaluated under its default RAG setup, which retrieves relevant passages using an OpenAI embedding model before generation. Under OpenAI's listed rates (as of Dec 23, 2025), GPT-4.1 costs $2.00 / 1M input tokens and $8.00 / 1M output tokens, and GPT-4.1 mini costs $0.40 / 1M input tokens and $1.60 / 1M output tokens.

C.3 Hyper Parameters

For DEEPFACT-EVAL, we use the following inference hyperparameters: max steps=2 (maximum iterations), max queries=5 (queries per step), max sources=40 (maximum retrieved sources retained for synthesis), and max completion tokens=8192 per request.

D Results on Other Datasets

We evaluate DEEPFACT-EVAL beyond our benchmark on three established factuality datasets: Sci-

Claim	Label	Evidence	Model	Model reason	New	Note
CHEK2 is not associated with breast cancer.	T	We genotyped these six tagSNPs in 1,577 postmenopausal breast cancer cases and 1,513 population controls, but found no convincing association between any common CHEK2 haplotype and breast cancer risk. The 1100delC mutation was rare in our Swedish population—0.7% in cases and 0.4% in controls—with a corresponding odds ratio for carriers versus noncarriers of 2.26 (95% confidence interval, 0.99–5.15).	F	Clinical genetics literature treats *CHEK2* as a moderate-penetrance breast cancer susceptibility gene (e.g., pathogenic variants like 1100delC increase risk); thus the absolute "not associated" claim is false.	F	**Label-evidence misalignment:** the evidence supports "common CHEK2 polymorphisms/haplotypes were not convincingly associated", but it does not support "CHEK2 (as a gene) is not associated," since a known rare variant shows increased risk in their data and prior work
In young and middle-aged adults, current or remote uses of ADHD medications do not increase the risk of serious cardiovascular events.	T	RR for serious CV events: current vs nonuse 0.83; new users 0.77; conclusion: not associated with increased risk.	F	Argues newer long-horizon evidence suggests duration-dependent increases in broader CVD outcomes (e.g., hypertension/arterial disease; cardiomyopathy), so the categorical "do not increase" is contradicted.	F	**Newer evidence founded**: Earlier work (e.g., Habel et al., JAMA 2011) found no increase in serious acute cardiovascular events with ADHD medication use, but newer studies with longer follow-up and cumulative exposure analyses report a modest increase in cardiovascular disease risk. e.g., Zhang et al., JAMA Psychiatry 2024)
Bone marrow cells contribute to adult macrophage compartments.	F	After bone marrow transplantation, host macrophages expanded when donor development was compromised (suggesting limited donor contribution).	T	General immunology evidence supports bone-marrow-derived monocytes contributing to adult macrophage pools (tissue- and context-dependent), so the broad claim can be true.	T	**Overgeneralizing context-specific replenishment into a global macrophage origin claim:** Benchmark evidence supports *minimal contribution* in a specific setting/tissue; claim is generic and may hold across tissues/conditions.

Table 4: **Disagreements of Annotation on SciFact.** We manually inspect instances where the model's prediction disagrees with the benchmark verdict. **T** = supported/true, **F** = unsupported/false (including contradictory/inconclusive), **U** = unverifiable. "`label`", "`model`", and "`new`" correspond to the original dataset label, the model's predicted label, and the expert re-annotated label, respectively. "`evidence`", "`model_reason`", and "`note`" correspond to the SciFact-provided rationale (abstract sentences used to justify the original label), the model's rationale, and the expert re-annotation rationale, respectively; all are summarized concisely.

Fact (Wadden et al., 2020), ExpertQA (Malaviya et al., 2024), and Factcheck-Bench (Wang et al., 2024).

D.1 Dataset Set-up

SciFact. SciFact (Wadden et al., 2020) is a scientific claim verification benchmark where claims are derived from citation sentences (*citances*) and verified against a corpus of paper abstracts. Each (claim, abstract) pair is labeled SUPPORTS, REFUTES, or NOINFO, and SUPPORTS/REFUTES instances include gold rationale sentences from the abstract. We evaluate on the SciFact validation set under a binary setting by excluding NOINFO and treating the task as SUPPORTS vs. REFUTES/*unsupported*, yielding 188 evaluated instances; DEEPFACT-EVAL disagrees with the gold label on 29/188 (15.4%).

ExpertQA. ExpertQA (Malaviya et al., 2024) pairs expert-authored questions with LLM responses. Responses are split into sentences, and annotators assign a sentence-level factuality label on a five-point scale (*Definitely correct / Probably correct / Unsure / Likely incorrect / Definitely incorrect*) (Malaviya et al., 2024). We focus on domains where we can access relevant experts—Engineering & Technology, Education, Environmental Science, and Healthcare/Medicine—and re-

strict to the most objective labels (*Definitely correct* and *Definitely incorrect*). We sample 100 per label (200 total); DEEPFACT-EVAL disagrees with ExpertQA on $90/200$ (45%).

Factcheck-Bench. Factcheck-Bench is built from LLM-generated answers to open-domain questions (in-house questions, Dolly Closed-QA, Dolly Open-QA). The authors decompose these responses into 678 claims, label 661 as checkworthy, and include 94 $(\text{question}, \text{response})$ pairs (Wang et al., 2024). Following FIRE's cross-dataset preprocessing (Xie et al., 2025), we map the original four labels (*supported, partially supported, not supported, refuted*) to a binary scheme: supported/partially supported $\rightarrow$ TRUE, refuted $\rightarrow$ FALSE, and exclude *not supported*. This yields 631 labeled claims (472 TRUE, 159 FALSE). We sample 200 claims for evaluation; DEEPFACT-EVAL disagrees with the benchmark on $36/200$ (18%) cases.

D.2 Expert Re-annotation Methods

We audit only the *disagreements* between DEEPFACT-EVAL and each benchmark and assign each case to a subject-matter expert in the closest matching domain.

SciFact: evidence-grounded two-stage audit. Because SciFact includes abstracts and gold rationale sentences, we audit each disagreement in two stages:

1. *Evidence–label consistency check:* verify whether the SciFact-provided rationale/evidence (abstract sentences) actually entails the gold verdict.

2. *Blind rationale adjudication:* for the remaining cases, present experts with two blinded packages—(i) SciFact rationale + abstract and (ii) DEEPFACT-EVAL rationale + retrieved abstract—and ask which explanation is better supported and why.

ExpertQA and Factcheck-Bench: blinded re-annotation on a subset. ExpertQA and Factcheck-Bench do not provide per-claim, evidence-grounded rationales, which makes disagreement adjudication harder. We therefore rely on blinded re-annotation for disputed claims (experts in the closest domains for ExpertQA; authors for Factcheck-Bench). For ExpertQA, we do not re-annotate all disagreements: we first screen all disputed items to flag clear labelability issues (e.g., non-verifiable discourse sentences). Among the remaining 76 disagreements, we randomly sample 30 claims for blinded re-annotation, where annotators do not see either the dataset label or the model prediction.

D.3 Results

SciFact. Among the 29 disagreements: (i) $12/29$ arise from evidence–label misalignment, where the provided abstract does not substantiate the annotated verdict; (ii) among the remaining 17, blind adjudication finds $7/17$ are unresolvable due to insufficient expert confidence; and (iii) of the 10 resolvable cases, experts favor the model's interpretation in $4/10$. Extrapolating this preference rate to all 17 adjudicated cases and evaluating against the resulting expert-recalibrated labels, DEEPFACT-EVAL achieves an estimated $178/188$ accuracy (94.7%), suggesting that a substantial fraction of the apparent errors may be driven by annotation noise rather than systematic model failures. (Examples in Table 4)

ExpertQA. Out of $90/200$ disagreements, the author inspection suggests that $14/90$ are non-verifiable discourse sentences (e.g., hedges, conversational prompts, generic advice) that arguably should not receive factuality labels. For the remaining disagreements, we sample 30 claims for blind expert re-annotation, allowing the experts to use search engines and LLM tools; experts side with DEEPFACT-EVAL in $28/30$ cases and with the original dataset label in $2/30$ (see examples in Table 5). However, because ExpertQA does not provide per-claim rationales, many conflicts are difficult to adjudicate conclusively.

Factcheck-Bench. DEEPFACT-EVAL disagrees with the benchmark on $36/200$ (18%) cases. Manual inspection suggests $32/36$ disagreements likely reflect annotation noise (e.g., subjective/ambiguous claims, or cases where the author's founded evidence aligns more with the model judgment than the benchmark label; examples in Table 5). Similarly, since the dataset does not provide rationales, these disagreements are difficult to resolve conclusively.

Summary. We summarize these findings using a three-way taxonomy to distinguish the source of the conflict:

1. **Agreement**: the verifier matches the benchmark label.

2. **Disagreement: Annotation divergence**: the verifier does not match the benchmark label, and our re-annotation diverges from the benchmark label as well, which indicates the disagreement cannot be cleanly resolved into a definitive model error. This includes evidence–label misalignment (SciFact), non-verifiable or non-checkworthy sentences being labeled (ExpertQA), subjective or underspecified claims, and cases that cannot be conclusively adjudicated due to missing gold rationales (ExpertQA, Factcheck-Bench).

3. **Disagreement: Likely model error**: the verifier does not match the benchmark label, and our re-annotation aligns with the benchmark label, suggesting a likely verifier error.

For ExpertQA and SciFact, the proportions for *Annotation Divergence* and *Model Error* are estimated by extrapolating the ratios observed in our annotated subsets to the total number of disagreements.

Results in Figure 5 reveal a consistent performance saturation point: as verifiers mature, residual discrepancies are driven predominantly by labeling noise and claim ambiguity rather than model error. These edge cases are difficult to adjudicate without evidence-grounded rationales. This underscores the need for auditable, evolving benchmarking, which allows diagnosis and correction of data artifacts while disentangling them from genuine verification failures.

E Importance- and Risk-Stratified Claim Sampling

Deep research reports frequently contain hundreds or thousands of distinct claims, making exhaustive annotation infeasible. We therefore design a **two-factor sampling scheme** that emphasises (i) *importance*—how central a claim is to the report's thesis (defined in Table 6)—and (ii) *risk*—the probability that the claim is incorrect according to an automatic evaluator (SmolAgent with GPT-4.1). By concentrating annotation effort on the most consequential and most error-prone statements, we obtain a balanced yet information-dense subset of claims.

Step 1 – Quota definition. Given a target batch size N, we allocate per-bucket quotas over five importance levels, e.g. $\{5{:}\,40\%, 4{:}\,35\%, 3{:}\,20\%, 2{:}\,5\%, 1{:}\,0\%\}$. For level i, the quota is $q_i = \lfloor N \times p_i \rfloor$ where p_i is the desired proportion. The quotas satisfy $\sum_i q_i = N$.

Step 2 – Risk weights. Each candidate claim is tagged by an automatic factuality evaluator as SUPPORTED or UNSUPPORTED/LOW-CONFIDENCE. We assign a risk weight

$$w_j = \begin{cases} 1, & \text{SUPPORTED} \\ \rho > 1, & \text{UNSUPPORTED} \end{cases}$$

where ρ controls how strongly we oversample likely errors.

Step 3 – Quota adjustment for sparse buckets. If an importance bucket contains fewer than q_i candidates, we down-scale q_i to the available count and redistribute the deficit proportionally across buckets that still have surplus capacity. This guarantees the final sample size remains exactly N.

Step 4 – Risk-weighted sampling without replacement. Within each bucket we sample without replacement, using inclusion probability

$$\Pr(j \mid j \in i) = \frac{w_j}{\sum_{k \in i} w_k}.$$

F Audit-then-Score Algorithm

The full AtS algorithm is listed in Algorithm 1.

G Statistical Significance of Findings

As the DeepFact-Bench test set consists of 15 reports from multiple domains, which may introduce substantial cross-report variance, we additionally assess whether our main conclusions are robust to *report sampling*. In particular, we test the significance of two central findings: (i) AtS improves human annotation quality, and (ii) DEEPFACT-EVAL outperforms existing verifier baselines.

Report-level paired bootstrap. We treat each *report* as the independent unit, since claims within the same report are correlated and should not be treated as independent samples. To quantify uncertainty, we use a *paired cluster bootstrap* over reports. For each comparison between two methods A and B (e.g., Round-3 vs. Round-2 human labels, or DEEPFACT-EVAL vs. GPT-Researcher), we perform 20,000 bootstrap replicates. In each replicate,

Claim	Label	Model	Model reason (concise)	New	Note (concise)
ExpertQA					
Perhaps you could narrow down your question by specifying a time period, a location, or a source of information that you are interested in.	F	T	Treats the suggestion as "supported" by research-methods best practices (narrowing questions by time/location/source).	U	This is conversational advice, not a checkable world fact.
Website security: You should verify that your website is safe and trustworthy for your users.	F	T	Best-practice guidance (e.g., OWASP/NIST/CISA-style) recommends validating security to protect users and maintain trust.	T	Normative, but still a valid best-practice claim; "supported" fits if ExpertQA treats recommendations as verifiable.
An in-memory database is a database that resides entirely in the memory of the application and does not persist data on disk.	T	F	In-memory DBs store primarily in RAM, but many support disk persistence (snapshots/logging) and are not necessarily "in the app's memory."	F	Overly absolute / incorrect definition due to "entirely" and "does not persist."
FactCheck-Bench					
Abacus computing doesn't need large amounts of energy.	F	T	An abacus is manual/mechanical and requires no electricity; energy use is only minimal human effort, negligible vs. electronic computing.	T	True in the intended "no external power / low energy" sense.
The decision on *NYSRPA v. Bruen* was seen as a setback for gun rights supporters.	T	F	*Bruen* struck down New York's "proper cause" carry-permit requirement and was widely framed as a win for gun rights advocates; calling it a setback for them is incorrect.	F	Reverses the typical framing (setback for gun control efforts, not for gun rights supporters).
China's Olympic success has been remarkable since 1980s.	F	T	Since first full participation in 1984, China has sustained high medal counts and frequent top medal-table finishes, supporting "strong success since the 1980s."	U	"Remarkable" is subjective; better treated as non-verifiable despite strong supporting facts.
Lawson was first established in 1975.	T	F	Ambiguous entity: *Lawson, Inc. (Japan)* dates to 1975, but the *Lawson brand/origin* traces to a U.S. dairy store (1939). Without "in Japan," the absolute claim is misleading.	F	Underspecified; correct only under a narrower reading ("Lawson, Inc. in Japan").

Table 5: **Disagreements of Annotation on ExpertQA and FactCheck-Bench.** We manually reannotate and inspect instances where the model's prediction disagrees with the benchmark verdict. **T** = supported/true, **F** = unsupported/false (including contradictory/inconclusive), **U** = unverifiable. "`label`", "`model`", and "`new`" correspond to the original dataset label, the model's predicted label, and the expert re-annotated label, respectively.

Algorithm 1 The Audit-then-Score (AtS) Protocol

Prerequisite: An initial seed benchmark $B_0 = \{(c_i, d_i, y_i^{(0)}, \rho_i^{(0)})\}_{i=1}^{N}$ created by human experts.

procedure EVOLVEBENCHMARK(B_t, M_t, A_t) ▷ B_t: current benchmark, M_t: Challenger, A_t: Auditor
  $U_{M_t,t} \leftarrow \emptyset$ ▷ Initialize empty proposal
  $\hat{Y} \leftarrow \emptyset$ ▷ Initialize set of all model predictions
  ▷ **Phase 1: Generate Verdicts and Challenges**
  for each claim i from 1 to N **do**
    $(c_i, d_i, y_i^{(t)}, \rho_i^{(t)}) \leftarrow B_t[i]$ ▷ Get current benchmark data
    $(\hat{y}_i, \hat{\rho}_i) \leftarrow M_t(c_i, d_i)$ ▷ Run Challenger model
    $\hat{Y} \leftarrow \hat{Y} \cup \{(i, \hat{y}_i)\}$ ▷ Store all predictions for final scoring
    if $\hat{y}_i \neq y_i^{(t)}$ **then**
      $U_{M_t,t} \leftarrow U_{M_t,t} \cup \{(i, \hat{y}_i, \hat{\rho}_i, \rho_i^{(t)})\}$ ▷ Add disagreement to proposal
    end if
  end for
  ▷ **Phase 2: Audit**
  $\Delta B_t \leftarrow \emptyset$ ▷ Initialize empty set of updates
  for each challenge $(i, \hat{y}_i, \hat{\rho}_i, \rho_i^{(t)})$ in $U_{M_t,t}$ **do**
    if $A_t(\hat{\rho}_i, \rho_i^{(t)}) =$ `ACCEPT` **then** ▷ Auditor adjudicates
      $\Delta B_t \leftarrow \Delta B_t \cup \{(i, \hat{y}_i, \hat{\rho}_i)\}$ ▷ Accept the challenger's update
    end if
  end for
  ▷ **Phase 3: Evolve Benchmark**
  $B_{t+1} \leftarrow B_t \oplus \Delta B_t$ ▷ Apply updates to create the new benchmark version
  ▷ **Phase 4: Score**
  $Y^{(t+1)} \leftarrow \{(i, y_i^{(t+1)}) \mid (c_i, d_i, y_i^{(t+1)}, \rho_i^{(t+1)}) \in B_{t+1}\}$ ▷ Get new ground truth labels
  $S \leftarrow$ CALCULATESCORE($\hat{Y}, Y^{(t+1)}$) ▷ e.g., Accuracy
  return B_{t+1}, S
end procedure

Score	Definition
5 – Backbone claim	Essential to the core thesis. Removing this would break the logic or invalidate the main takeaway. Often appears in the title, abstract, or conclusion, and is referenced multiple times.
4 – Critical support	Key evidence or reasoning that directly supports a backbone claim. Removing it weakens the argument substantially but does not break the core conclusion.
3 – Standard support	Provides helpful background or secondary evidence. Removing it moderately weakens the report but leaves the main thesis intact.
2 – Minor context	Background detail, definition, or peripheral comment. Removing it has little to no effect on the main message.
1 – Irrelevant or off-topic	Unrelated, redundant, or likely a segmentation artifact. Removing it improves clarity or focus.

Table 6: Claim importance scale used for prioritizing verification.

#	Category	Typical Location	Why It's Non-Verifiable	Concrete Examples
A	Document-structure & rhetorical framing (headings, previews, transitions)	Section titles, figure headings, opening or bridging sentences	Merely label or orient the reader; they do not assert external facts.	"3 Dataset Construction"; "In the next subsection we detail the ablation study."
B	Forward-looking statements	Abstract, introduction, future-work, conclusion	Refer to intentions or events that have not happened yet.	"We will extend this method to multilingual data next year."
C	Research questions & hypotheses	Introduction, methods	Pose uncertainties; they explicitly seek answers rather than state facts.	"Does larger model size improve calibration?"
D	Subjective judgments & opinions	Discussion, conclusion, related-work critiques	Depend on author viewpoint or value judgments.	"Our approach is considerably more elegant than prior work."
E	Citation lists & bibliographic metadata	In-text citations, References section	Point to external sources; correctness is about formatting, not truth value.	"(Smith et al., 2024)"
F	Speculative or motivational claims	Introduction, broader-impact, abstract	Describe possibilities, potential impact, or high-level vision rather than established facts.	"This technique could revolutionize personalized medicine."

Table 7: Common types of non-verifiable sentences in DRRs (mapped to the None label).

we sample 15 test reports *with replacement*, recompute the *micro-accuracy* of both methods on the same resampled report set, and record the paired difference

$$d = \text{score}(A) - \text{score}(B).$$

We then compute a 95% confidence interval from the empirical bootstrap distribution of d. If the 95% confidence interval excludes 0, we consider the improvement statistically significant at approximately the 0.05 level.

AtS significantly improves human annotation quality across rounds. We first apply this procedure to human annotation accuracy on the micro-gold set across ATS rounds. The results show that human annotation quality improves significantly over time. For example, **Round-3** outperforms **Round-2** by **4.9 points**, with a **95% confidence interval of [1.4, 7.9]**, which excludes 0, confirming that the improvement in human label quality under ATS is not driven by a small subset of reports.

DeepFact-Eval significantly outperforms existing verifiers. We next compare DEEPFACT-EVAL against existing verifier baselines using the same paired report-level bootstrap. DEEPFACT-EVAL outperforms **GPT-Researcher** by **14.7 points** (**95% CI: [7.4, 23.3]**) and **Smolagents** by **15.0 points** (**95% CI: [9.5, 20.5]**). These results indicate that our main verifier gains are statistically robust and not driven by idiosyncrasies of a few sampled reports.

Takeaway. Overall, these report-level significance tests strengthen our conclusions in two ways. First, they show that ATS yields genuine improve-

ments in human annotation quality. Second, they show that DEEPFACT-EVAL significantly outperforms existing verifiers. Together, these results suggest that our findings are stable despite the limited number of reports in the current benchmark.

H Managing Evolving Benchmarks with AtS

An evolving benchmark also requires explicit maintenance policies for governance, stopping, and fair reporting over time. In our setting, benchmark maintainers are responsible for curating updates, releasing new versions, and publishing changelogs of accepted revisions, making benchmark evolution transparent and auditable. For DEEPFACT-BENCH, we denote the released benchmark after the four-round rollout as DEEPFACT-BENCH V4. To reduce drift toward verifier or agent auditor biases, we adopt two safeguards: hidden micro-gold monitoring to detect degradation, and periodic human recalibration once agent-driven updates exceed a small threshold (e.g., $\sim$5% of benchmark verdicts). AtS is not intended to evolve indefinitely. In practice, stopping can be determined by one or more criteria: (i) a fixed audit budget, (ii) stabilization of micro-gold accuracy above a target threshold, and/or (iii) a preset maintenance horizon. Because AtS produces benchmark versions B_t, fair comparison also requires versioned reporting: results should always *specify the benchmark version*, and longitudinal comparisons should be made either on frozen snapshots or by re-scoring archived outputs under an explicitly specified B_t.

I More Related Work

Dynamic Benchmarking. Dynamic benchmarking is "dynamic" in what changes over time: (i) the test set is iteratively expanded to track model weaknesses via human/model-in-the-loop adversarial rounds—Dynabench frames benchmarking as continuous data creation (Kiela et al., 2021), ANLI operationalizes this with iterative adversarial collection so the target keeps moving as models improve (Nie et al., 2020), and real-time factuality assessment adversarially modifies claims from news to make them harder to fact-check (Chen et al., 2025); (ii) the world state changes, so benchmarks refresh questions and answers to stay current—FreshQA explicitly targets fast-changing knowledge and commits to regular updates (Vu et al., 2024), while RealTime QA evaluates newly announced (e.g., weekly) questions tied to recent events (Kasai et al., 2023); (iii) the evaluation distribution is mined "in the wild" and re-versioned, as in FactBench (Bayat et al., 2025), which curates prompts from real user interactions and is designed to be regularly updated with newly observed hallucination-triggering prompts; and (iv) the benchmark is refreshable by construction, where a repeatable generator yields new tasks—LiveDRBench (Java et al., 2025) proposes "problem inversion" as a recipe to periodically produce new deep-research queries from existing reasoning problems. Compared to these "refresh the inputs" approaches, our evolving benchmarking is dynamic primarily in the supervision itself: the benchmark's ground truth and coverage are iteratively strengthened through auditing and verification, not merely by swapping in new questions or sampling a new prompt stream.

Expert-Led Benchmarking for Research Tasks. Benchmarks for research tasks evaluate how LLM agents search, read, and synthesize literature, spanning tasks from literature-review (Asai et al., 2024) to Expert-level QA (Malaviya et al., 2024; Zhao et al., 2025). Most existing evaluations implicitly treat expert(s) judgments as an infallible gold standard (Sharma et al., 2025; Wang et al., 2025; Ruan et al., 2025). Yet expert reliability is rarely quantified directly; instead, it is usually approximated via inter-annotator agreement, which obscures unresolved disagreements (Malaviya et al., 2024; Zhao et al., 2025) and cannot detect shared blind spots (Sharma et al., 2025; Wang et al., 2025). Moreover, some benchmarks rely on STEM practitioners as annotators (Malaviya et al., 2024; Sharma et al., 2025) rather than the hyper-specialized researchers the questions may demand—further weakening the "ground truth." This expert-dominance assumption will become a bottleneck as agents approach expert-level performance: the ceiling is set by annotation quality, and models may be penalized for correct outputs that conflict with noisy labels. Indeed, prior work reports that experts are not error-free, including annotation mistakes in HLE (Phan et al., 2025) and documented failures in human literature review (Salvador-Oliván et al., 2019). The issue is even more acute for verifying DRR, where a well-informed judgment may require locating and integrating substantial portions of the relevant literature. Our work challenges the assumption of human dominance by using adversarial hidden sets to monitor expert quality and proposing a human-AI

collaborative framework to elevate benchmarking beyond expert limits.

Role-based and multi-agent LLM systems. Role-based and multi-agent LLM systems are increasingly used as *test-time scaffolds* for better task solving. Prior work assigns agents different roles (Qian et al., 2024) or interaction protocols—e.g., role-playing cooperation in CAMEL (Li et al., 2023), programmable multi-agent conversations in AUTOGEN (Wu et al., 2024), SOP-style collaborative workflows in METAGPT (Hong et al., 2024), and debate-based reasoning with multiple model instances (Smit et al., 2024; Du et al., 2024; Chen et al., 2024). However, in these settings, the multi-agent system remains part of the *solver*: it is designed to generate a better answer against a *fixed, human-defined target*, such as a reference answer, solution, or rubric. In contrast, our ATS framework is not merely a solver-side method; it is part of the *benchmark*. The evaluated deep-research agent contributes the candidate claims and evidence, and the benchmark's labels are updated through human–AI auditing rather than assumed static. This reframes the problem from "using multiple agents to solve a task" to "using human–AI collaboration to keep evaluation trustworthy" when systems approach or surpass expert-level capability.

Fact-Checking. Survey papers provide broader overviews of fact-checking methods and evaluation settings (Guo et al., 2022; Hardalov et al., 2022).

J Use of AI Assistants

We used LLMs to assist with writing. Specifically, we employed GPT-5 thinking, GPT-5 and GPT-4o to rephrase paragraphs for grammatical correctness and improved flow. We also used them to shorten text, making descriptions more concise and easier to read. All LLM-generated text was reviewed, edited, and approved by the human authors.

K Reproducibility, Release, and Intended Use

K.1 Reproducibility and release

To support reproducibility, we will release (i) the DEEPFACT-BENCH dataset, including de-identified annotations and claim metadata, and (ii) the DEEPFACT-EVAL verifier code, prompts, and evaluation scripts. The code will be released under the **Apache-2.0** (or **MIT**) license, and the dataset under **CC BY 4.0** (or **CC BY-NC 4.0**) for research use. Where examples originate from third-party sources, we will follow their terms and, when necessary, distribute only derived metadata/identifiers rather than full text.

K.2 Intended use and consistency with upstream terms.

We use existing datasets and tools strictly for their intended research purpose: evaluating factuality and evidence-grounded verification, consistent with the licenses and access conditions specified by their authors. Our released artifacts—DEEPFACT-BENCH and DEEPFACT-EVAL—are intended for *research-only* use in benchmarking and developing claim-level verifiers for Deep Research Reports (DRRs), including evaluation, ablations, and error analysis. To remain compatible with upstream access conditions, we avoid redistributing restricted third-party content when applicable and release only derived, de-identified annotations and metadata (e.g., claim text, verdicts, rationales, and provenance pointers) needed to reproduce our experiments. We explicitly prohibit non-research use that would violate upstream terms (e.g., commercial redistribution of restricted content or attempts to re-identify participants) and require users to comply with the original licenses/ToS of any upstream resources and retrieval services used in our pipeline.

L Qualitative Examples

L.1 Adversarial Examples

Here we show examples of how we construct adversarial examples with intentional errors. We provide one example each for a collection error, an analysis error, and a generalization error.

Example: Collection Error

Context:
...Intrinsic resistance genes, naturally encoded in soil microbial genomes, are primarily disseminated through vertical gene transfer (VGT) and limited horizontal gene transfer (HGT) under stable conditions. Acquired resistance genes, introduced via anthropogenic inputs like swine manure, rely heavily on MGEs (plasmids, transposons, integrons) to facilitate rapid HGT. Environmental factors such as soil moisture, pH, and heavy metals differentially impact these gene types, with acquired ARGs showing greater responsiveness to external stressors. The study by **Guo et al. (2025)** highlights the role of transposons like *ISRj1* and *IS91* in amplifying acquired resistance, while **Forsberg et al. (2012)** note that intrinsic resistance genes in soil producers (e.g., *Streptomyces*) are less mobile but can persist for decades. Key findings reveal that swine farm soils exhibit higher acquired ARG diversity and MGE abundance compared to undisturbed soils, with winter conditions paradoxically enhancing their persistence. The report concludes that acquired resistance genes pose a greater risk for dissemination due to their mobility and environmental adaptability, while intrinsic genes remain a baseline reservoir. Limitations include the need for longitudinal studies and the challenge of distinguishing intrinsic from acquired genes in complex soil metagenomes. ...

Original Sentence:
The study by **Guo et al. (2025)** highlights the role of transposons like *ISRj1* and *IS91* in amplifying acquired resistance, while **Forsberg et al. (2012)** note that intrinsic resistance genes in soil producers (e.g., *Streptomyces*) are less mobile but can persist for decades.

Adversarial Sentence:
The study by **Forsberg et al. (2012)** highlights the role of transposons like *ISRj1* and *IS91* in amplifying acquired resistance, while **Guo et al. (2025)** note that intrinsic resistance genes in soil producers (e.g., *Streptomyces*) are less mobile but can persist for decades.

Analysis:
The adversarial sentence performs an *attribution swap*: it flips which paper supports each claim, subtly misassigning the transposon-driven amplification result (from Guo et al., 2025) and the intrinsic-gene mobility/persistence observation (from Forsberg et al., 2012). Because both papers are plausibly related, the swap can appear credible while corrupting provenance and authority.

Example: Analysis Error

Context:
...Studies over the past 30 years have consistently found tetracycline and sulfonamide antibiotic resistance genes (ARGs) to be ubiquitous in agricultural soils, even in the absence of recent antibiotic inputs digitalcommons.unl.edu frontiersin.org. These genes are typically quantified in terms of gene copies per gram of soil, or as a ratio to total bacterial 16S rRNA gene copies. Across diverse farm soil environments, *the abundance of tetracycline and sulfonamide ARGs generally falls in the range of* 10^4–10^6 *gene copies per gram of soil* pmc.ncbi.nlm.nih.gov researchgate.net. **This corresponds to roughly** 10^{-5}–10^{-3} **ARG copies per bacterial 16S gene – about one ARG per 1,000 bacterial cells in typical human-impacted soils** ouci.dntb.gov.ua. Table 1 summarizes representative concentration ranges for common tetracycline (tet) and sulfonamide (sul) resistance genes reported in agricultural soils. *Table 1. Typical abundance ranges of selected ARGs in agricultural soils (gene copies per g soil).* ...

Original Sentence:
This corresponds to roughly 10^{-5}–10^{-3} **ARG copies per bacterial 16S gene – about one ARG per 1,000 bacterial cells, in typical human-impacted soils** ouci.dntb.gov.ua.

Adversarial Sentence:
This corresponds to roughly 0.1–0.3% of bacterial 16S genes carrying an ARG – or one in every **300** bacterial cells, on average, in typical human-impacted soils.

Analysis:
The adversarial sentence introduces a subtle quantitative distortion by changing units and the implied denominator: it reframes "about one per 1,000 cells" as "one in every 300 cells" while presenting the ratio as a percent. This inflates the perceived prevalence by roughly $\sim 3\times$, which can be easily overlooked but materially changes the risk impression.

Example: Generalization Error

Context:
...It stands to reason that part of that predictive power would involve the listening sub-scores. Similarly, Sudina et al. (2021) assessed English proficiency holistically (with sections for listening and reading) and saw perseverance-linked gains experts.nau.edu – implying grittier students had stronger listening comprehension alongside other skills. One specific study in a mobile-assisted learning context evaluated beginners' listening outcomes. In that study (Botes et al., 2025), 245 Duolingo learners' listening and reading proficiency were tested; results indicated that **L2 grit was among the factors associated with better listening scores**, although motivation and age also played roles researchgate.net. Furthermore, anecdotal evidence from EFL instructors suggests that gritty students are more likely to engage in extra listening practice (such as watching English media without subtitles, repeatedly listening to difficult audio until they understand, etc.), which over time enhances their listening ability. The limited direct research means we should be cautious, but no findings so far contradict the expectation that grit aids listening comprehension. If anything, listening might show a slightly weaker grit correlation than productive skills, simply because even a very persistent learner can still struggle with fast, unfamiliar speech. ...

Original Sentence:
In that study (Botes et al., 2025), 245 Duolingo learners' listening and reading proficiency were tested; results indicated that **L2 grit was among the factors associated with better listening scores**, although motivation and age also played roles researchgate.net.

Adversarial Sentence:
In that study (Botes et al., 2025), 245 Duolingo learners' listening and reading proficiency were tested; results indicated that L2 grit was the **primary factor** determining listening scores, with little influence from motivation or age.

Analysis:
The misleading sentence creates a *conditional collapse* by removing caveats about motivation and age, incorrectly reframing grit as the sole or primary driver of listening outcomes and overstating the strength of the study's conclusion.

L.2 Deep Research Errors Examples

Here, we show the errors deep research models make in the generated deep research reports.

We identified 27.0% naturally unsupported claims among the test set (excluding adversarially constructed examples). These errors span several stages of the research pipeline—collection, analysis, and generalization—and reflect distinct reasoning failures, aligns with the taxonomy in Table 8: Collection errors: fabricated claims without any identifiable source, or misattributions where a statement is linked to the wrong paper or dataset. Analysis errors: misstatements of numerical results, incorrect mappings between experimental setups (e.g., attributing results from setup A to setup B), or faulty synthesis across sections (e.g., conflating Natural Questions with other multi-hop datasets). Generalization errors: over-extending localized or conditional findings, such as extrapolating a global trend to a specific region without supporting evidence. Together, these categories highlight how unsupported claims in DRRs arise not only from missing citations but also from deeper reasoning and synthesis failures during evidence interpretation.

Example: Analysis Error

Context:
5. Longitudinal and Multimodal Data Integration. Longitudinal designs are underrepresented in L2 engagement research, with only 13.4% of studies in the 2021 review employing them ([1]). A 2025 study on achievement emotions ([2]) used a mixed-methods explanatory sequential design (ESD) to link qualitative interview data with quantitative regression models. The study reports a strong association between achievement emotions and emotional engagement (r = .798, p = .000). Despite these insights, the cross-sectional nature of most studies limits their ability to capture engagement as a dynamic process. A 2021 behavioral analytics study ([10]) proposed integrating multimodal data (e.g., facial expressions + keystroke dynamics) to address this gap, though its application to L2 contexts remains untested. *Table 1: Comparison of Quantitative Methods in L2 Engagement Research.*

Sentence:
The study found that positive emotions like hope and enjoyment correlated with higher engagement scores (r = 0.798), while negative emotions (e.g., anxiety) reduced participation.

Human Verdict:
Contradictory

Human Reason:
The strongest association was observed between achievement emotions and emotional engagement (r = .798, p = .000). However, the study did not distinguish between positive and negative emotions. Instead, achievement emotions are considered as a unified construct. See: source.

Example: Collection Error

Context:
1. Cost and Insurance Coverage. In the U.S., CAB-LA's adoption is limited to 1.4% of PrEP users due to insurance restrictions and high out-of-pocket costs ([3]). Lenacapavir's pricing (US$28,000 per dose) further exacerbates access disparities ([10]). **2. Regulatory Delays.** While CAB-LA is approved in the U.S. and Brazil, lenacapavir's rollout in Europe and the Asia-Pacific is pending, with regulatory submissions in 2025 ([11]). **3. Healthcare Infrastructure.** In Zambia, inconsistent HIV testing protocols (e.g., 17% RNA testing at first injection) may underrepresent seroconversions ([6]). In the U.S., mental health and substance use comorbidities correlate with CAB-LA discontinuation ([4]). **4. Stigma and Patient Preferences.** The **PILLAR trial** (2024) found that 75% of U.S. participants preferred CAB-LA over daily oral PrEP, citing reduced stigma and convenience ([2]). Similar data for lenacapavir is absent in the sources.

Sentence:
The **PILLAR trial** (2024) found that 75% of U.S. participants preferred CAB-LA over daily oral PrEP, citing reduced stigma and convenience. Similar data for lenacapavir is absent in the sources.

Human Verdict:
Contradictory

Human Reason:
The claim asserts that the PILLAR trial in 2024 found a 75% preference for injectable cabotegravir (CAB-LA) over oral PrEP in the U.S., citing a ViiV Healthcare press release as its source. This is contradicted by multiple facts. First, the PILLAR study (NCT05422333) is an ongoing implementation study with an estimated completion date of December 2025, so it could not have produced final results in 2024 (ClinicalTrials.gov). Second, the cited ViiV press release discusses real-world effectiveness data from the OPERA and Trio Health cohorts; it does not mention the PILLAR trial, a 75% preference rate, or reasons such as stigma and convenience (ViiV Healthcare). While other studies (e.g., HPTN 083) have reported similar preference rates, this sentence's specific attribution to the 2024 PILLAR trial and the provided source is factually incorrect.

Example: Overclaim (Field-wide "Most Promising")

Context:
For larger instances, its strength shifts to its sophisticated feasibility reasoning, making it an invaluable component within hybrid solvers. Its main weakness is its limited scalability when used as a pure, standalone optimization engine. **Reinforcement Learning** represents a paradigm shift from offline problem-solving to online policy learning. Its capacity to learn generalizable, adaptive scheduling policies that can be executed in near-real-time positions it as the most promising approach for dynamic and stochastic environments. However, it faces the highest implementation barriers, significant computational costs for training, and challenges in interpretability and achieving the absolute best solution quality compared to finely-tuned search methods. Ultimately, the most compelling future for academic research and practical application in construction project scheduling lies at the intersection of these three paradigms. The integration of RL's adaptive learning, CP's logical reasoning, and metaheuristics' powerful search capabilities promises to yield a new generation of hybrid, intelligent scheduling systems.

Sentence:
Its capacity to learn generalizable, adaptive scheduling policies that can be executed in near-real-time positions it as the most promising approach for dynamic and stochastic environments.

Human Verdict:
Inconclusive

Human Reason:
Based on a review of recent (2022–2024) systematic surveys, no authoritative, up-to-date evidence establishes Reinforcement Learning (RL) as *the most promising approach* for project scheduling in dynamic and stochastic environments compared to Metaheuristics and Constraint Programming. Recent reviews (e.g., Khajesaeedi et al., 2024; Bahroun et al., 2024) note RL strengths (generalizability, adaptivity, near-real-time inference post-training), but metaheuristics and hybrids remain widely recognized as state-of-the-art in research and practice. RL is often described as promising, yet there is limited systematic head-to-head evidence supporting overall superiority. Thus, the "most promising" claim is currently unsupported by field-wide consensus, yielding an evidence-based verdict of **INCONCLUSIVE**.

L.3 DeepFact-Eval Examples

Here we show representative cases where DeepFact-Eval succeeds and fails (Model output is simplified to be easy to read). DeepFact-Eval can decompose a sentence into atomic claims, cross-check the broader literature, and synthesize evidence to verify each claim. However, it can still err—for example, it may miss critical evidence due to incomplete retrieval, retrieve closely matching evidence but miss key nuances or misinterpret it, or

fail to validate niche sub-claims embedded within a longer sentence, suggesting room for improvements.

Example: DeepFact-Eval Success

Context:
When reference answers are available (e.g., a human-written answer for the question), automated metrics like **BLEU** or **ROUGE** have been used to measure n-gram overlap between the model's answer and the reference. For instance, ROUGE-L was used in early LFQA benchmarks like ELI5 (Krishna et al., 2021). However, researchers found that such overlap metrics are **not very informative for long answers**—they can be gamed and do not truly reflect answer quality (Krishna et al., 2021). Exact-match or F1 score (common in short-answer QA) become less applicable as answers grow longer and more free-form. More recently, embedding-based metrics (e.g., **BERTScore**) and Q&A-based evaluation have been explored to capture semantic similarity. Yet a 2023 study reported that **no existing automatic metric reliably predicts human judgments** of long-form answer quality (Xu et al., 2023).

Sentence:
Exact-match or F1 score (common in short-answer QA) become less applicable as answers grow longer and more free-form.

Human Verdict:
Supported

Model Verdict:
Supported

Model Reason:
The claim has two components: (i) EM/F1 are standard for short-answer QA, and (ii) they are less suitable for long, free-form answers. Evidence supports both:

- **Standard in short-answer QA:** EM/F1 are widely used in extraction-based / factoid QA evaluations (e.g., Upadhyay et al., 2024; Risch et al., 2021).
- **Less applicable for long-form answers:** Overlap-style metrics correlate weakly with human judgments for long-form, open-ended QA and fail to capture semantic correctness and diversity of valid phrasings (e.g., Xu et al., 2023; Krishna et al., 2021; Laskar et al., 2024).

Therefore, the sentence is supported by multiple independent sources.

Example: DeepFact-Eval finds closely-match citation but with nuances

Context:
For example, using a powerful model (like GPT-4) to directly score or rank a report on various criteria has become common. These *LLM-as-a-judge* approaches can be more sensitive to nuances of content quality than static metrics. However, they come with reliability issues (biases and lack of transparency, discussed more below). As a workaround, some research proposes training dedicated **reward models** specialized for long-text evaluation (Wu et al., 2025). Such a reward model would be taught to predict human preferences on long outputs, similar to how RLHF trains a model to judge responses. This is an active research area: for instance, one 2024 work designed a reward model specifically to evaluate long academic essays, as standard metrics were insufficient (Wu et al., 2025).

Sentence:
As a workaround, some research proposes training dedicated **reward models** specialized for long-text evaluation (source).

Human Verdict:
Contradictory

Model Verdict:
Supported

Model Reason:
The cited LongEval paper states that, as an alternative to LLM-based scoring, prior work (e.g., Zhang et al., 2024c) proposes a reward model specifically designed for long-text evaluation, which directly supports the sentence (Wu et al., 2025).

Human Reason:
While Wu et al., 2025 contains wording like "Zhang et al., 2024c propose a reward model," the underlying referenced work does *not* actually train a dedicated reward model; instead, it relies on existing LLMs as judges.

Example: DeepFact-Eval ignores niche details

Context:
Key Benchmarks. *No single standard benchmark yet.* **LONG2RAG (2024)** evaluates long-form answer quality and source usage (source). **Academic/Domain evals** (e.g., medical long-form QA with doctor evaluations) (source). **Research agent tests**—e.g., **DeepResearcher (2025)** uses open-domain QA sets (NQ, HotpotQA, MuSiQue, etc.) as a testbed for research agents, evaluated with model-based judges (source). **Long-form generation benchmarks** (HelloBench, LongEval 2024) evaluate general long outputs (source).

Sentence:
Research agent tests – e.g. DeepResearcher (2025) uses open-domain multi-hop QA sets (NQ, HotpotQA, MuSiQue, etc.) as testbed for research agents (source), evaluated with model-based judges.

Human Verdict:
Contradictory

Model Verdict:
Supported

Model Reason:
The DeepResearcher paper evaluates on multiple open-domain QA datasets including NaturalQuestions (NQ), HotpotQA, and MuSiQue, and uses a model-based judge (LLM-as-a-judge) to assess answer correctness rather than relying only on string-overlap metrics (DeepResearcher, 2025).

Human Reason:
While DeepResearcher indeed evaluates on NQ, HotpotQA, and MuSiQue and uses model-based judging, the sentence characterizes *NQ* as a *multi-hop* QA set. Natural Questions is generally treated as *single-hop* (factoid/short-answer style) rather than a multi-hop benchmark, so the sentence's taxonomy claim is incorrect, yielding **CONTRADICTORY**.

Code	Principle	Name	Definition / Example
		1. Collection-Stage Errors (Evidence Gathering)	
C-AU	Authenticity	Fabricated Source	Cites a source, author, or quote that does not exist. *e.g., "OpenAI's GPT-4V did ..." (no such study)*
C-PV	Provenance	Mis-sourced Evidence	Real fact but assigned to the wrong author, venue, or year. *e.g., arXiv preprint claimed to be a 2023 Nature paper*
C-CP	Completeness	Omitted Counter-Evidence	Omits accessible contradictory or qualifying evidence. *e.g., Ignores a larger meta-analysis contradicting a cited RCT*
C-CU	Currency	Out-of-Date Source	Relies on retracted or outdated sources without caveats. *e.g., Citing a 2019 draft despite a reversed 2024 version*
C-RE	Representativeness	Biased Sampling	Uses narrow evidence (e.g., language, geography) that skews conclusions. *e.g., All English news used to infer global media trends*
C-CX	Contextual Relevance	Contextual Mismatch	Collects evidence topically related but from a different domain or task. *e.g., Legal claim supported using biomedical QA accuracy*
		2. Analysis-Stage Errors (Evidence Processing)	
A-N1	Numerical Fidelity	Numeric Distortion	Misrepresents counts, percentages, means, or CIs. *e.g., 25% vs. 0.25 absolute points*
A-S1	Semantic Fidelity	Semantic/Entity Swap	Substitutes similar but non-equivalent terms (e.g., metric, dataset type, model variant). *e.g., "faithfulness" reported when only F1 was measured*
A-P1	Causal Discipline	Causal Projection	Claims causality from correlation or reverses direction. *e.g., "Retrieval reduces hallucination" based on observational data*
A-X1	Study Integrity	Cross-Study Conflation	Blends results from different studies into a single narrative. *e.g., Claims KnowPO outperforms CTPC with no direct comparison*
A-B1	Balanced Synthesis	Cherry-Picked Synthesis	Selects supportive evidence while omitting stronger contradictory data. *e.g., Cites positive RCT, ignores null meta-analysis*
A-T1	Temporal Alignment	Temporal Misalignment	Compares studies/data from incompatible timeframes. *e.g., Comparing 2018 vs. 2024 SQuAD results*
A-O1	Aggregation Soundness	Over-Aggregation	Combines incompatible metrics or tasks into a single number. *e.g., Merging latency, accuracy, and cost into one score*
A-C1	Logical Coherence	Contradiction Ignorance	Presents contradictory findings without resolving them. *e.g., Quotes studies with opposite trends as co-validating*
A-L1	Reasoning Validity	Chain-of-Thought Leap	Introduces an unjustified intermediate premise. *e.g., "Since large models are always calibrated..." (unsupported)*
		3. Generalization-Stage Errors (Claim Expansion)	
G-O1	Scope Discipline	Over-Scope Leap	Generalizes beyond the evidence's domain, task, or population. *e.g., From WebQA to biomedical QA without evidence*
G-H1	Claim Proportionality	Hyperbolic Statement	Turns conditional or limited findings into absolutes. *e.g., "Always improves performance"*
G-T1	Taxonomic Completeness	Taxonomy Oversimplification	Omits known categories or claims exhaustiveness without support. *e.g., "Two types of evaluation" ignoring a third*
G-C1	Condition Transparency	Conditional Collapse	Drops necessary qualifiers or assumptions. *e.g., Removes "in low-resource settings" from claim*
G-R1	Temporal Projection	Recency Extrapolation	Projects recent trend into the future without evidence. *e.g., 3-month rise ⇒ "will keep increasing exponentially"*
G-B1	Base-Rate Awareness	Base-Rate Neglect	Reports large relative gains on near-zero baselines. *e.g., "50% gain in recall" where base rate is 0.2%*
G-S1	Evidentiary Sufficiency	Single-Study Certainty	Claims general truth from one small study. *e.g., Lab study to industry-wide claim*

Table 8: Taxonomy of factuality errors in deep research report generation, organized by cognitive phase. Each code reflects a distinct violated principle.