QUANTIFYING CLAIM SMOOTHING BY LLM RESEARCH AGENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep research, or any use of large language models (LLMs) to read a large document collection and produce insights, requires synthesizing information from different sources. In this work, we find that LLMs produce long-form research reports that read as confident syntheses but which fail to accurately reflect the underlying documents. We call this failure *smoothing*. In collapsing diverse and potentially conflicting sources into a single narrative, the model discards the evidential "jaggedness" of its corpus and asserts conclusions more decisive than the citations can support. We introduce a claim level scoring pipeline, SMOOTHSCORE, to quantify this effect. Our pipeline extracts key synthesis claims from model generated reports, then finds evidence relevant to each claim and clusters that evidence along high level aspects. Finally, each claim is scored with respect to the evidence to quantify the degree of smoothing done by the agent. Across 5 systems and 4 datasets, we find that models consistently produce stronger claims than warranted.

1 INTRODUCTION

Large Language Models (LLMs) are often deployed within systems that read across many sources of information and return long-form reports with citations (Shao et al., 2026; Xie et al., 2026; Li et al., 2026b; Nakano et al., 2021). These systems, either general-purpose agents paired with Internet access or special-purpose "deep research" agents, retrieve large volumes of documents and analyze and synthesize this information into a coherent reports (Du et al., 2026; Onweller et al., 2026). The skills required to do this goes beyond simple multi-document summarization (Fabbri et al., 2019): agents must weigh evidence, integrate conflicting findings, and determine what the evidence warrants (Narayanan Venkit et al., 2026; Rasheed et al., 2026; Zhu et al., 2026). The resulting reports are typically fluent, well-organized, and densely cited, which makes them read as confident, authoritative syntheses. As a result, users increasingly treat these outputs as substitutes for reading the underlying literature, even when the synthesis misrepresents or overstates what the cited sources actually support (Li & Aral, 2025; Huang et al., 2026).

In this work, we study the prevalence of *smoothing* in LLM generated reports. We define smoothing as the tendency of LLMs to synthesize claims that are more decisive than supported by the documents they cite. Such claims may, for example, fail to portray mixed evidence or overclaim a result to hold in all circumstances when it has only been shown more narrowly. By contrast, a faithful claim should hedge when the evidence is mixed and narrow its scope when the support is narrow.

To identify instances of this behavior, we propose SMOOTHSCORE, a framework for measuring whether the claims in a synthesized report are calibrated to the cited sources. As shown in Figure 1, we find that an agent's report can present a claim as well supported by some of its citations, even though other retrieved evidence shows the claim only holds under certain conditions. The bulk of prior work on claim verification and long-form factuality (Min et al., 2023; Song et al., 2024; Tang et al., 2024) can check whether individual claims are supported by their cited evidence, but breaks down for synthesized claims: it usually returns binary judgments at the claim or subclaim level, but does not handle calibration and scoping across a large body of potentially conflicting evidence.

To compute SMOOTHSCORE, we first decompose each report into its main synthesis claims and identify the documents the report cites. As shown in Figure 2, for each claim, we run two agents over the report's cited documents: (a) *Support Agent*: which surfaces evidence consistent with the claim,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and (b) *Skeptic Agent*: which actively searches for evidence that qualifies, limits, or contradicts the claim. Evidence from both these passes are clustered into aspects reflecting different dimensions of the claim, such as viewpoint, strength of argument, and certainty (Kargupta et al., 2025). Given the claim and its mapped aspect-evidence pairs, we score the claim's calibration on a 1–5 scale, reflecting how well its stated strength and scope match the evidence collected across both passes.

Across four benchmarks and five systems, we find that smoothing is pervasive. On average, each system generated 202 claims per dataset, and 60% of all claims receive a SMOOTHSCORE ≥ 3. For all but one system, over half of claims show at least moderate smoothing, and up to 43% of Gemini Deep Research's claims are severely smoothed (≥ 4). We also find that limited engagement with the cited documents contributes to smoothing: Luna opens only 19% of the pages it cites, and providing it the full documents reduces smoothing.

Contributions Our main contributions are: (a) We introduce SMOOTHSCORE, a framework for measuring whether the claims in a synthesized report are calibrated to the strength and scope of their cited evidence, rather than merely supported by it and (b) we find this issue to be widespread across frontier systems on 4 different benchmarks.

2 METHOD: SMOOTHSCORE

Measuring smoothing requires more than checking whether a report's statements are supported by individual citations. For instance, in Figure 2, the extracted claim *'discrete image-token modeling appears more promising than a Stable-Diffusion-centered architecture'* is supported by the cited documents, D_1 Emu3 paper (Wang et al., 2024)) and D_3 (the Show-o paper (Xie et al., 2025)). However, we find that the same sources along with other documents, namely D_2 (the Transfusion paper (Zhou et al., 2024)), also contains results that show that the statement is only true under certain conditions present counter evidence. Therefore, smoothing can occur even when a claim appears supported, and methods producing binary judgments cannot comprehensively capture the distribution to identify it.

We propose SMOOTHSCORE as a way of measuring the degree of smoothing a claim shows. Figure 2 outlines the pipeline. At a high level, first, we extract synthesis claims from a generated report that represent the key takeaways, then find evidence that supports and refutes those claims. However, it is not straightforward to combine this evidence into a final score: counting supporting documents vs. refuting documents does not account for the nuances of *how* a claim is supported or refuted.

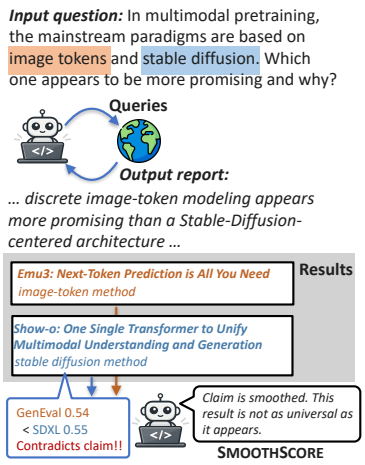

Figure 1: When synthesizing over many documents, agents can present a claim as well supported when the evidence is actually more mixed. SMOOTHSCORE flags these smoothed-over claims and quantifies the degree of smoothing.

Instead, we identify a set of dimensions for the claim and evidence that we call *aspects*, following (Kargupta et al., 2025). Aspects allow us to capture fine-grained ways in which a claim might be refuted: for instance, in Figure 2, step 3, we are able to find *'performance and efficiency of tokens vs diffusion models'*, under which we see some evidence state that *'tokens match or beat diffusion'* but we also see evidence for *'diffusion or hybrid set ups are competitive'*. Aspects are discovered per-claim and can be tailored to the particulars of the claim and evidence in question. Finally, we score the extracted claim, with respect to the aspect-organized evidence to quantify the degree of smoothing.

Setup and Notation Given a prompt p, a retrieval-augmented model M produces a long-form response y with inline citations. We collect the full text of the cited documents in the report $D = \{d_1, \ldots, d_m\}$. Our goal is to measure if y faithfully represent what D supports, including relevant conditions and counterevidence.

Step 1: Claim Extraction To focus on the model's ability to do synthesis, we extract *conclusion-bearing claims*. These are cases where a model takes a stance or presents a takeaway. For example,

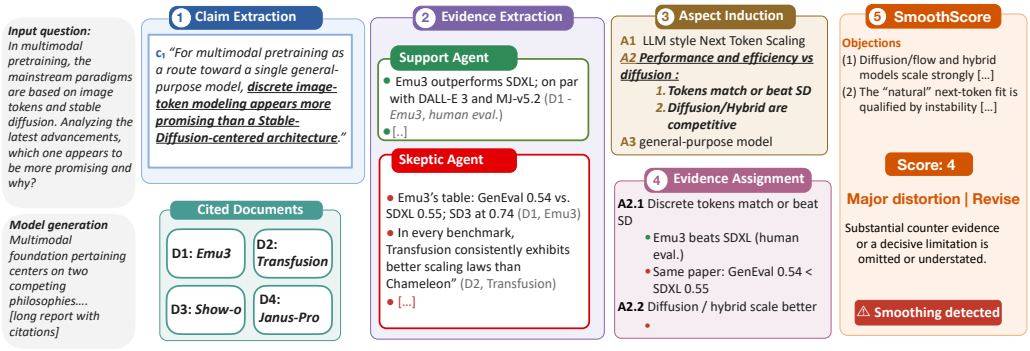

Figure 2: Overview of SMOOTHSCORE. Given a query and a model generated report, we (1) first extract synthesis claims. Then from the cited documents, (2) we get supporting and skeptic evidence relevant to the claim. (3)/(4) We then organize the evidence into *aspects*. (5) The claim is then scored with respect to the gathered evidence to quantify the degree of missing information.

in Figure 2, '*image-token modeling appears more promising than a Stable-Diffusion-centered architecture*' is a takeaway from the report which we extract as a claim.

Using prompt, Appendix B.1, we extract,

$$C = \text{ExtractClaims}(y) = \{c_1, \ldots, c_n\}.$$

Each claim preserves its direction, confidence, scope, and qualifications as stated in the report.

Step 2: Evidence Extraction In Figure 2, we see that a cited document, for example, $D1$ can carry information relevant to the claim that can both support or undermine it. To identify information relevant to the claim from the cited documents throughout the report, we extract evidence using a 'support agent' and a 'skeptic' agent.

- Support Agent (Appendix B.2): Given (claim, document), this agent extracts evidence that supports the assertions made by the claim. We also include stated concession as evidence when they confirm limitations or conditions present within the claim.

$$S(c) = \bigcup_{d \in D} \text{Extract}_{+}(c, d)$$

- Skeptic Agent (Appendix B.3): Given (claim, document), this agent extracts evidence that contains any information against the claim. To capture the "jaggedness" we go beyond just finding counters to the assertions made by the claim. The agent evidence to support limitations and conditions stated in the claims. Additionally, we include evidence on alternate approaches and their success related to the claims where relevant. This is to provide context for scope.

$$R(c) = \bigcup_{d \in D} \text{Extract}_{-}(c, d)$$

For each agent, the evidence is composed of: a source quote from the document, aspect detailing the assertion, a note on how the evidence bears on the claim, and strength rating.

Step 3: Aspect Induction Different cited documents can be relevant to different parts of a claim, so judging the claim against the raw documents forces the judge to compare evidence that is not directly comparable. If we look across the cited documents, however, the relevant information naturally clusters into topics that bear on the claim. For example, in Figure 2, some documents discuss *comparing the performance and efficiency of image tokens vs. diffusion approaches*, while others discuss *fit with LLM-style next-token scaling*. We refer to these topics as *aspects* (Kargupta et al., 2025).

To construct aspects, for each claim–document pair $(c_i, d_j) \in C \times D$ we generate a set of document-specific aspect proposals,

$$A_{ij} = \text{Propose}(c_i, d_j).$$

Each proposal identifies a dimension discussed in d_j that bears on c_i, and consists of a short name, a brief description, and a set of keywords. If d_j contains nothing relevant to c_i, then $A_{ij} = \varnothing$ (see Appendix B.4 for the prompt). Because proposals are generated independently for each document, different documents will often propose overlapping or near-duplicate aspects. For each claim, we therefore merge the proposals across documents into a single, deduplicated set of aspects, keeping track of which documents contributed to each one (see Appendix B.5).

Organizing evidence by aspect places findings that address the same part of the claim side by side, regardless of which document they came from. This allows the judge to weigh evidence both supporting and opposing the claim on a like-for-like basis. For example, along a comparative dimension there may be supporting and countervailing evidence for either stance, as seen in aspect *A2* of Figure 2.

Step 4: Evidence Assignment We assign every extracted evidence item in $E(c_i) = S(c_i) \cup R(c_i)$ to the aspect it discusses, so that findings addressing the same part of the claim are grouped together and can be compared directly (prompt in Appendix B.6). Assignment depends only on *what* the evidence discusses, not on whether it supports or counters the claim, so an item's stance, source, and strength rating are left unchanged. For example, in Figure 2, the support-agent finds that *Emu3 beats SDXL in human evaluation* and the skeptic-agent finding that *Emu3's own table shows a lower GenEval score than SDXL* are placed under the same aspect, even though they point in opposite directions.

Step 5: Computing Smoothing Finally, a judge J produces a SMOOTHSCORE σ_i for each claim. The judge receives the input question q, the claim c_i, and all extracted evidence E_i organized under the claim's aspects T_i, including evidence that is not tied to any single aspect:

$$\sigma_i = J\big(q, c_i, T_i, E_i\big) \in \{1, \ldots, 5\}.$$

The judge assesses how faithfully the claim compresses the supplied evidence, focusing on whether it omits or understates assertions, conditions, or scope in a way that changes its central takeaway. Using the five-point rubric in Appendix B.7. Figure 2 and Table 11 show examples of the generated outputs for each stage of the pipeline.

3 EXPERIMENTAL SETUP

Datasets. We calculate SMOOTHSCORE on the following datasets: (1) **ResearcherBench** (Xu et al., 2026), (2) **ScholarQA-CS2 (Expert-Written)** (Hwang et al., 2026) (3) **DeepResearch Bench II** (Li et al., 2026a) (4) **ResearchRubrics** (Sharma et al., 2025). These provide a range of tasks curated for research report generation. To ensure all queries require a degree of multi document synthesis we manually select queries. Details on this selection process is available in Appendix A.2

Systems. To check if overclaiming persists across architectures with varying design considerations, we evaluate 5 different systems. We evaluate the following 5 systems: (1) **GPT Luna + Websearch**, (2) **Claude Sonnet 5 + Websearch**, (3) **Gemini Deep Research**, (4) **DR Tulu** (Shao et al., 2026), (5) and **QUEST** (Xie et al., 2026). Together, these systems span several current approaches to generating research reports, including language models augmented with Websearch, proprietary end-to-end deep-research systems, and open research agents with specialized retrieval and reasoning pipelines. Addtional details provided in Appendix A.3

Metrics We report SMOOTHSCORE in two ways: (1) the mean SMOOTHSCORE across claims for each system and benchmark, and (2) the fraction of claims falling into each SMOOTHSCORE bin. We believe the fraction of severe (3 or higher) claims indicates how often systems generate moderately misleading claims.

Table 1: We report the number of reports, claims for each of systems and benchmarks evaluated in this work. We also report the SMOOTHSCORE. DR Tulu and QUEST produce the largest % of severely smoothed claims, while closed source models like Sonnet 5 never produces any claims $= 5$. Luna consistently produces the highest % of claims that produce a score $= 1$. Gemini exhibits the most severe smoothing across all datasets. Our human validation in Section 5 shows that humans agree within SMOOTHSCORE within ± 1 over 90% of the time, indicating that the higher SMOOTHSCORE values point to real concerns in the reports.

		Sample		SMOOTHSCORE	% of claims w/given SMOOTHSCORE				
Dataset	System	Reports	Claims	Mean [95% CI]	1	2	3	4	5
ResearcherBench	Luna	65	443	1.95 [1.85, 2.06]	37.0%	34.1%	25.7%	3.2%	0.0%
	Sonnet 5	65	447	2.48 [2.39, 2.57]	14.1%	30.2%	49.2%	6.5%	0.0%
	QUEST	63	404	2.50 [2.35, 2.65]	18.1%	26.7%	43.6%	10.1%	1.5%
	DR Tulu	65	408	2.57 [2.47, 2.67]	12.0%	26.7%	54.2%	6.9%	0.2%
	Gemini	64	434	3.27 [3.21, 3.34]	1.6%	5.1%	58.1%	34.8%	0.5%
CS2-Expert	Luna	25	164	1.93 [1.75, 2.11]	44.5%	23.8%	27.4%	3.0%	1.2%
	Sonnet 5	25	171	2.38 [2.25, 2.52]	19.9%	28.7%	45.0%	6.4%	0.0%
	QUEST	25	156	2.67 [2.46, 2.87]	17.3%	18.6%	47.4%	13.5%	3.2%
	DR Tulu	25	168	2.70 [2.53, 2.86]	12.5%	19.0%	55.4%	12.5%	0.6%
	Gemini	24	175	3.18 [3.09, 3.28]	4.0%	5.7%	59.4%	29.7%	1.1%
DRB-II	Luna	18	106	1.73 [1.51, 1.95]	54.7%	21.7%	19.8%	3.8%	0.0%
	Sonnet 5	18	108	2.60 [2.36, 2.82]	14.8%	17.6%	60.2%	7.4%	0.0%
	QUEST	16	95	2.87 [2.48, 3.29]	14.7%	17.9%	41.1%	17.9%	8.4%
	DR Tulu	18	102	2.41 [2.20, 2.63]	22.5%	18.6%	53.9%	4.9%	0.0%
	Gemini	18	120	3.08 [2.93, 3.25]	3.3%	10.8%	60.0%	25.8%	0.0%
ResearchRubrics	Luna	17	111	1.78 [1.55, 2.03]	47.7%	30.6%	18.0%	2.7%	0.9%
	Sonnet 5	17	115	2.40 [2.30, 2.51]	19.1%	27.8%	47.0%	6.1%	0.0%
	QUEST	17	104	2.81 [2.50, 3.15]	12.5%	22.1%	42.3%	18.3%	4.8%
	DR Tulu	17	99	2.64 [2.41, 2.85]	14.1%	21.2%	52.5%	11.1%	1.0%
	Gemini	17	113	3.34 [3.15, 3.52]	4.4%	2.7%	49.6%	41.6%	1.8%

4 RESULTS

4.1 SMOOTHSCORE RESULTS

Smoothing is widespread. Table 1 shows that smoothing is pervasive across systems and datasets. Mean SMOOTHSCORE ranges from 1.73–1.95 for Luna to 3.08–3.34 for Gemini, with every other system falling between 2.38 and 2.87. The severity of smoothing is also reflected in the score distribution breakdown: smoothing of $\geq$3 is present on every system and dataset, with at least 21% of claims affected in every case, but severe smoothing ($\geq$4) is where systems diverge sharply. Gemini smooths severely on 26–43% of its claims, roughly 2–3× the rate of the open-source agents and 10× the rate of Luna.

Open-source agents produce severely smoothed claims Systems follow similar patterns across datasets. The mean SMOOTHSCORE in open-source Deep Research systems like DR Tulu (Shao et al., 2026) and QUEST (Xie et al., 2026) remains comparable to Sonnet-5. However, at least 59% and 55% of their claims score $\geq$ 3, and 5–26% score $\geq$ 4. From Table 10, QUEST draws from among the smallest evidence pools (roughly 10 documents and 39–55k words) so it has the least amount of evidence to compress. Despite this, while other systems only generate 1–2% of claims that score $= 5$, QUEST generates up to 8% and 5% such claims on DeepResearchBench II and ResearchRubrics. Appendix D.1 shows that QUEST often generates outputs that read as summarization rather than synthesis. The resulting claims preserve specifics faithfully while overstating scope enough to drive inflated conclusions.

Closed-source systems vary widely in smoothing severity. Among closed-source models with Websearch, we see that Sonnet 5 produces almost 2–3× the fraction of claims that achieve a

SMOOTHSCORE of 3–4 relative to Luna in Table 1. While neither of the systems produce fundamentally misleading claims (Scored 5) at a meaningful rate, Sonnet never generates any such cases across any datasets. On the other hand, Luna generates at least 37% claims that show no smoothing (score $= 1$), 2–2.5x more than any other system across datasets.

Across all systems, Gemini performs the worst, with 86–93% of claims scoring ≥ 3 and an average score of 3.08–3.34. Gemini also draws from significantly more documents (2–5x) than Luna with websearch, and produces the longest reports on every dataset (Table 10). This gives the model a much larger surface to synthesize, and despite the added length, the report does not appear to compress this information more effectively. Appendix D.2 lists some examples of claims from the generated reports. They contain language that turns a strong result into a universal or final conclusion.

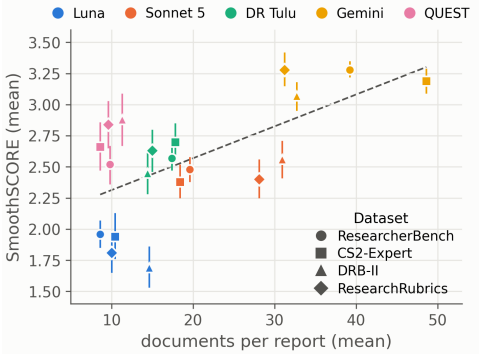

Figure 3: Relationship between retrieved document count and claim smoothing. Across system and dataset pairs, we find that mean SMOOTHSCORE increases with the mean number of documents retrieved per report. The individual report and document counts are provided in Appendix 10

Smoothing scales with increase in number of documents From Table 10 we see a relationship between the number of documents cited and smoothing severity. Gemini, which cites the most documents per report (31.2–48.6), also has the highest mean smoothing score (3.08–3.34) across all datasets. Luna, on the other extreme, draws on a much smaller number of documents (8.6–14.6), and correspondingly has the lowest mean smoothing score (1.73–1.95). Barring QUEST, which cites few documents yet produces relatively high smoothing, this pattern holds broadly across systems, as shown in Figure 3: as the number of documents cited increases, so does mean smoothing severity, suggesting that synthesizing across more documents makes it harder for a system to preserve the qualifications and boundaries of the claims it draws from.

4.2 QUALITATIVE OBSERVATIONS

Query framing influences SMOOTHSCORE. We find that query framing influences how the model writes its claims. As seen in Appendix D.1 and D.2, there are certain categories of questions where smoothing is more pronounced.

(1) Comparative questions: if the query is comparative in nature, for example, *'In multimodal pretraining, the paradigms are based on image tokens and stable diffusion, which appears more promising and why?'* or *'Which paradigm between RL and distillation appears more promising and why?'*, it can presuppose a choice from the agent where to motivate or justify a choice the model drops important qualifications.

(2) Literature Review: such queries require the model to engage with a large number of different documents and ideas related to a topic, which leads to more smoothing (Figure 3).

Recurring Smoothing patterns across systems Table 2 shows three patterns that we observe across systems: (1) **Unsupported composition**: Models combine individual fragments of evidence into a stronger conclusion. Although each component may be supported, the claim presents its unsupported composition as a conclusive statement. (2) **Scope expansion**: A claim takes an established result for a particular setup or experimental condition and removes certain boundaries, so that a the scope for an empirical observation expands into a general conclusion. (3) **Premature Closure**: Claims often adopt conclusive language, using phrases like *'strictly preserving,' 'the best approach,' 'definitively proves'* and *'most optimal.'*.

5 VALIDATING SMOOTHSCORE

Table 2: Examples of recurring claim-smoothing patterns. For each pattern, we show an illustrative claim, its SMOOTHSCORE, and a summary of the supporting evidence and limitations.

Pattern	Claim	Summarized Evidence
Unsupported composition	*The aggressive deployment of* ***duplication-aware token pruning****,* ***Modality-Aware Mixture of Experts****, and* ***adaptive halting*** *ensures that vast, world-class parameter capacities* ***can be maintained without triggering corresponding explosions in inference latency.*** **Score: 3**	Token pruning, sparse expert routing, and adaptive halting can each reduce inference cost in particular settings. However, the evidence does not evaluate their simultaneous deployment or show that their benefits compose. Each can introduce routing, pruning, synchronization, or kernel overheads, so the combined system does not guarantee bounded latency at frontier scale.
Scope expansion	*Self-attention-based models are highly parallelizable and require substantially fewer parameters, making them much more* ***computationally efficient****,* ***less prone to overfitting****, and* ***easier to fine-tune for domain-specific tasks*** *...* **Score: 4**	Some ViTs achieve strong performance with lower pre-training compute in particular comparisons, but these advantages do not generalize across models or regimes. ViTs often have comparable or larger parameter counts, global attention has quadratic cost, and ViTs can be more prone to overfitting in low-data settings without large-scale pre-training or careful regularization.
Premature closure	*This innovation* ***fundamentally resolves the statistical heterogeneity (Non-IID) problem*** *while strictly preserving the privacy boundaries established by institutional firewalls.* **Score: 5**	Federated distillation can mitigate heterogeneity and reduce direct data sharing, but it does not eliminate either problem. Performance continues to degrade as heterogeneity increases, theoretical analyses retain residual bias, and privacy attacks remain possible.

Table 3: Agreement on SMOOTHSCORE across 53 claims scored by two human annotators.

	Exact	W/in 1	ρ	α
Annotator 1 vs. Annotator 2	51%	94%	0.64	0.64
Judge vs. Annotator 1	55%	91%	0.64	0.62

To validate SMOOTHSCORE, two annotators independently scored the same 53 claims using the same rubric given to the LLM judge (Appendix B.7), without access to the identity of the system that produced each claim. We find that agreement between the judge and Annotator 1 is in a similar range to inter-annotator agreement, Table 3. Disagreements are predominantly small: the judge is within one point of Annotator 1's score on 91% of claims and differs by two or more points on only 9%. Additionally, ofthe claims the judge scores ≥ 3, 67% are also scored ≥ 3 by Annotator 1, indicating that most judge-flagged claims reflect genuine issues. Since the two annotators themselves agree at $\kappa = 0.59$ at this threshold, some disagreement is expected given the subjective nature of the task.

6 LACK OF ENGAGEMENT WITH DOCUMENT CONTENT

The systems we evaluate generate reports with citations from the documents they retrieve. However, the agent's interaction with the cited material is hard to control. For Luna, the trace shows that the model works almost entirely from search snippets truncated by the tool (median 585 characters), opening a cited page for only 19% of the documents it cites and in just 20 of 65 reports (Table 4, Appendix C.1). Synthesizing from snippets rather than document content, the agent never sees the qualifications

Table 4: Luna + websearch system engagement with the cites documents on ResearcherBench. An "open" is an `open_page` action of the Responses API search tool, returned as a truncated window of the page. Details in Appendix C.1.

Cited documents	497
Opened	95 (19%)
Snippet only	401 (81%)
Reports opening no page	45 / 65
Median chars seen: snippet	585
Median chars seen: opened	4,294

Table 5: Effect of document engagement on smoothing for Luna on ReseacherBench. In-context reports are generated by the same model given the query and the full text of the documents the WebSearch report cited, with Websearch off. "+ instructions" adds two additional instructions to the prompt. Parenthesized values are differences from WebSearch, computed before rounding. $^{\dagger}95\%$ paired bootstrap CI over question.

Condition	**Length**	**Claims**	**Mean σ**	$\sigma \geq 3$
WebSearch (snippets)	1.0×	416	1.95	29.3%
Cited docs in context	1.6×	390	1.80 ($-0.16^{\dagger}$)	23.3% ($-6.0^{\dagger}$ pp)
Cited docs + instructions	1.6×	377	1.65 ($-0.30^{\dagger}$)	18.6% ($-10.8^{\dagger}$ pp)

and countervailing findings in the body of a source, so smoothing at generation time is partly a consequence of what information the model chose to access.

To test how this limited engagement might contribute to smoothing, we compare the original web-search reports with reports generated by the model with two set ups. First, all the original cited documents provided in context *(cited docs in context)*. Second, the cited documents in context along with instructions to surface any conditions, scopes, limitations etc *(cited docs in context + instruction)*. Prompts for both can be found in Appendix C.3, C.2. We apply this to Luna and ResearcherBench queries and calculate the SMOOTHSCORE. We report mean smoothing and the proportion of claims exhibiting material smoothing ($\sigma \geq 3$).

Table 5 shows that reading the direct documents rather than their snippets reduces smoothing: mean σ falls from 1.95 to 1.80 and the share of materially smoothed claims ($\sigma \geq 3$) drops by 6%. The in-context reports are also 1.6× longer, consistent with the model having more qualifications to carry into the synthesis.

To drive this further, we check if prompting the model to engage explicitly with disagreements within the documents aids in reducing smoothing. With the documents provided in context and the added instructions we find that roughly halves the smoothing $\sigma = 1.65$, $\sigma \geq 3$ down 11 points. Across both experiments we find that models that are provided the documents generate longer reports (1.6x) but at the same time generate lesser claims. We also observe this qualitatively, the claims do not necessarily change the takeaway. Instead they tend to drop conclusive language supporting the scope of the evidence in documents and provide the required limitations as needed. Table 6 shows that the central takeaway remains the same and smoothing tracks the added reasoning.

7 RELATED WORK

Claim decomposition and aspect-based verification A standard recipe for verifying long-form text is to decompose it into smaller checkable units and score each against a knowledge source (Min et al., 2023; Chen et al., 2024; Gunjal & Durrett, 2024). These methods generally return a binary decision per claim, which is a good granularity for factual precision, but it does not work for synthesized claims that can have varying degrees of smoothing. ConflictScore (Liu et al., 2026) is the closest to our setting, and it extends this atomic claim recipe in a direction close to ours. They mark a claim as supported, contradicted or irrelevant per document retrieved, and aggregate these into claim-level conflict measures. However, a claim can be labeled *supported* by every one of its cited documents and still smooth over information it omits. Our work instead follows Kargupta et al. (2025) in organizing evidence along *aspects* (e.g., viewpoint, strength, certainty), so that supporting and qualifying evidence on the feature of a claim can be considered against each other.

Multi-Document Summarization In science communication, prior work measures certainty in how findings are stated (Pei & Jurgens, 2021) and models how information changes as findings move from papers to news and social media (Wright et al., 2022). For LLM summaries of scientific text, Peters & Chin-Yee (2025) test ten widely used LLMs on about 4,900 summaries and find that most produce broader generalizations than the source texts, and that LLM summaries are nearly five times more likely than human-written ones to overgeneralize. In medicine, neural multi-document summaries of clinical trials are fluent but often factually inaccurate (Wallace et al., 2021), and LLM

Table 6: One claim across the three Luna conditions (ResearcherBench). The question asks for a comparison of the LIMO and s1 papers. Claims are extracted independently per report. The judge's verdict is shown on the WebSearch claim.

WebSearch $\sigma = 4$	LIMO appears stronger in broad benchmark performance, particularly on GPQA Diamond and MATH500.
Judge	Substantial revision is needed. The evidence supports that LIMO is stronger than the S1 baseline and notably better on GPQA, but it does not support superiority on MATH500 over S1.1, nor broad dominance over S1 variants overall. Repair type: recast.
Docs in context $\sigma = 1$	LIMO reports higher scores than original s1 on the listed AIME24, MATH500, and GPQA results, but the studies differ in evaluation details. LIMO uses greedy decoding for many benchmarks and 16 samples for certain smaller benchmarks, whereas s1 generally reports temperature-0 evaluation and separately studies budget forcing [doc 1][doc 2]. Therefore, the numerical comparison indicates relative performance in the papers' reported settings but does not by itself establish a controlled head-to-head superiority claim.
Added claim $\sigma = 1$	The later analysis of s1 suggests that its scaling behavior should be interpreted cautiously: some of the apparent gains may reflect the consequences of truncating reasoning rather than the emergence of a naturally self-scaling reasoning capability [doc 6].
Docs + instructions $\sigma = 1$	The later analysis in document 6 cautions that s1's observed scaling curves may not be equivalent to the natural performance-improving scaling of RL-trained reasoning models. Accordingly, the strongest synthesis supported by the documents is that both papers demonstrate impressive small-data reasoning improvements, but the deeper mechanism behind s1's test-time scaling—and especially the role of repeated Wait prompting—remains contested.

summaries of medical evidence can misstate the certainty of the evidence or overgeneralize by omitting attributes such as a disease subtype (Tang et al., 2023). These studies summarize a single document or a small set of abstracts. We extend the concern to long-form, multi-document research reports written by agents, and give a claim-level measure grounded in the documents a report cites.

Correctness evaluation for deep research systems Agentic deep-research systems (Du et al., 2026; Xu et al., 2026; Huang et al., 2026; Li et al., 2026a; Sharma et al., 2025) pair rubric-based insight coverage with citation grounded factuality checks. Several works audit the reliability of these systems. DeepTRACE (Narayanan Venkit et al., 2026) decomposes answers into statements and finds that deep-research agents remain one-sided on debate queries and include many statements unsupported by their listed sources. Onweller et al. (2026) find that citations that resolve and are topically relevant often still fail a factual check against the cited content. Zhu et al. (2026) show that a single misleading document can induce false conclusions in final reports. Rasheed et al. (2026) argue for claim-level auditability, including whether conflicts in the evidence are surfaced or suppressed. Across this literature, factuality is largely operationalized as citation validity: whether a claim is supported by some retrieved source. Our results show that this is not enough. Reports whose claims are individually supported can still present mixed or conditional evidence as a decisive conclusion, so assessing how faithfully the underlying evidence is *presented* requires the claim-level, evidence-calibrated analysis we propose.

8 Conclusion

In this work, we studied the propensity of LLMs to *smooth* their conclusions when generating reports based on sources. They frequently generate oversimplified statements or conclusions that aren't really backed up by the evidence cited. Our pipeline, SmoothScore, allows us to quantify this and establish its frequency across several state-of-the-art agents. Finally, we suggest that this phenomenon arises from models failing to read the cited documents, and suggest it could be improved by more thorough agents.

AI USE STATEMENT

Generative AI formed the backbone of the systems studied in this work. In addition, code written with generative AI was used to conduct and carry out the analyses here. However, the research hypotheses in this work and the design of the experiments were both human-originated. AI tools were used for assistance with revisions, but not for writing in a substantive way. The authors take responsibility for the final content of this work.

REFERENCES

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. Complex Claim Verification with Evidence Retrieved in the Wild. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3569–3587, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.196. URL https://aclanthology.org/2024.naacl-long.196/.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Licheng Zhang, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A Comprehensive Benchmark for Deep Research Agents. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hQ0K2Hhq7H.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1074–1084, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1102. URL https://aclanthology.org/P19-1102/.

Anisha Gunjal and Greg Durrett. "Molecular Facts: Desiderata for Decontextualization in LLM Fact Verification". In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3751–3768, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.215. URL https://aclanthology.org/2024.findings-emnlp.215/.

Yukun Huang, Leonardo FR Ribeiro, Momchil Hardalov, Bhuwan Dhingra, Markus Dreyer, and Venkatesh Saligrama. DeepFact: Co-Evolving Benchmarks and Agents for Deep Research Factuality. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 34356–34386, 2026.

Jena D Hwang, Varsha Kishore, Amanpreet Singh, Dany Haddad, Aakanksha Naik, Malachi Hamada, Jonathan Bragg, Mike D'Arcy, Daniel S Weld, Lucy Lu Wang, et al. Deep Research, Shallow Evaluation: A Case Study in Meta-Evaluation for Long-Form QA Benchmarks. *arXiv preprint arXiv:2603.06942*, 2026.

Priyanka Kargupta, Runchu Tian, and Jiawei Han. Beyond True or False: Retrieval-Augmented Hierarchical Analysis of Nuanced Claims. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29664–29679, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1434. URL https://aclanthology.org/2025.acl-long.1434/.

Haiwen Li and Sinan Aral. Human Trust in AI Search: A Large-Scale Experiment. *arXiv preprint arXiv:2504.06435*, 2025.

Ruizhe Li, Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench II: Diagnosing Deep Research Agents via Rubrics from Expert Reports, 2026a. URL https://arxiv.org/abs/2601.08536.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yongkang Wu, Ji-Rong Wen, Yutao Zhu, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *Advances in Neural Information Processing Systems*, 38:120091–120131, 2026b.

Siyi Liu, Aaron Halfaker, Dan Roth, and Patrick Xia. ConflictScore: Identifying and Measuring How Language Models Handle Conflicting Evidence. *arXiv preprint arXiv:2606.26437*, 2026.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Pranav Narayanan Venkit, Philippe Laban, Yilun Zhou, Kung-Hsiang Huang, Yixin Mao, and Chien-Sheng Wu. DeepTRACE: Auditing Deep Research AI Systems for Tracking Reliability Across Citations and Evidence. In *International Conference on Learning Representations*, volume 2026, pp. 106041–106062, 2026.

Hailey Onweller, Elias Lumer, Austin Huber, Pia Ramchandani, Vamse Kumar Subbiah, and Corey Feld. Cited but Not Verified: Parsing and Evaluating Source Attribution in LLM Deep Research Agents. *arXiv preprint arXiv:2605.06635*, 2026.

Jiaxin Pei and David Jurgens. Measuring sentence-level and aspect-level (un)certainty in science communications. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 9959–10011, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.784. URL https://aclanthology.org/2021.emnlp-main.784/.

Uwe Peters and Benjamin Chin-Yee. Generalization Bias in Large Language Model Summarization of Scientific Research. *Royal Society Open Science*, 12(4):241776, 2025.

Razeen A Rasheed, Somnath Banerjee, Animesh Mukherjee, and Rima Hazra. From Fluent to Verifiable: Claim-Level Auditability for Deep Research Agents. *arXiv preprint arXiv:2602.13855*, 2026.

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, Hannaneh Hajishirzi, and Pang Wei Koh. DR Tulu: Reinforcement Learning with Evolving Rubrics for Deep Research. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=97NEP1pyS3.

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. ResearchRubrics: A Benchmark of Prompts and Rubrics For Evaluating Deep Research Agents, 2025. URL https://arxiv.org/abs/2511.07685.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. VeriScore: Evaluating the factuality of verifiable claims in long-form text generation. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 9447–9474, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.552. URL https://aclanthology.org/2024.findings-emnlp.552/.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating Large Language Models on Medical Evidence Summarization. *NPJ digital medicine*, 6(1):158, 2023.

Liyan Tang, Philippe Laban, and Greg Durrett. MiniCheck: Efficient fact-checking of LLMs on grounding documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8818–8847, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.499. URL https://aclanthology.org/2024.emnlp-main.499/.

Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. Generating (Factual?) Narrative Summaries of RCTs: Experiments with Neural Multi-Document Summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605, 2021.

Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, Yingli Zhao, Yulong Ao, Xuebin Min, Tao Li, Boya Wu, Bo Zhao, Bowen Zhang, Liangdong Wang, Guang Liu, Zheqi He, Xi Yang, Jingjing Liu, Yonghua Lin, Tiejun Huang, and Zhongyuan Wang. Emu3: Next-Token Prediction is All You Need. *CoRR*, abs/2409.18869, 2024. URL https://doi.org/10.48550/arXiv.2409.18869.

Dustin Wright, Jiaxin Pei, David Jurgens, and Isabelle Augenstein. Modeling information change in science communication with semantically matched paraphrases. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1783–1807, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.117. URL https://aclanthology.org/2022.emnlp-main.117/.

Jian Xie, Tianhe Lin, Zilu Wang, Yuting Ning, Yuekun Yao, Tianci Xue, Zhehao Zhang, Zhongyang Li, Kai Zhang, Yufan Wu, et al. QUEST: Training Frontier Deep Research Agents with Fully Synthetic Tasks. *arXiv preprint arXiv:2605.24218*, 2026.

Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-o: One Single Transformer to Unify Multimodal Understanding and Generation, 2025. URL https://arxiv.org/abs/2408.12528.

Tianze Xu, Pengrui Lu, Lyumanshan Ye, Xiangkun Hu, and Pengfei Liu. ResearcherBench: Evaluating Deep AI Research Systems on the Frontiers of AI Research, 2026. URL https://openreview.net/forum?id=oj6A9hrNdL.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model, 2024. URL https://arxiv.org/abs/2408.11039.

Pengyu Zhu, Lijun Li, Longju Yang, Sen Su, and Jing Shao. Is Deep Research Reliable? Misleading Knowledge Induces False Conclusions. *arXiv preprint arXiv:2607.20891*, 2026.

A SmoothScore

A.1 Method Details

Aspect Induction As mentioned in Section 2, the third step in our pipeline is Aspect Induction. Since aspect proposals are generated separately for each claim–document pair, the same underlying topic is often proposed by several documents under different names or with different levels of specificity. Merging is needed so that evidence on the same topic is grouped together, rather than being split across near-duplicate aspects.

For each claim c_i, we consolidate the document-level proposals $\{A_{ij}\}_{j=1}^{m}$ into a hierarchy

$$T_i = (V_i, L_i) = \text{Merge}\left(\{A_{ij}\}_{j=1}^{m}\right),$$

where V_i is the set of aspect and subaspect nodes and L_i is the set of parent–child links between them. The merge step (i) combines semantically overlapping proposals across documents into a

Table 7: Models used at each stage of the SMOOTHSCORE pipeline. All stages use GPT-5-mini except the final judge, which uses GPT-5. Prompts for each stage are given in the referenced appendix sections.

Step	Stage	Model	Prompt
1	Claim extraction	GPT-5-mini	App. B.1
2	Evidence extraction: support agent	GPT-5-mini	App. B.2
2	Evidence extraction: skeptic agent	GPT-5-mini	App. B.3
3	Aspect induction: per-document proposals	GPT-5-mini	App. B.4
3	Aspect induction: cross-document merge	GPT-5-mini	App. B.5
4	Evidence assignment	GPT-5-mini	App. B.6
5	SMOOTHSCORE judge	GPT-5	App. B.7

single node, and (ii) introduces subaspects when one aspect covers distinct but related findings. Each node retains the set of documents that proposed it, so that every piece of evidence can be traced back to its source.

Pipeline Models Table 7 shows the models used for every stage of the pipeline for generating SMOOTHSCORE. All pipeline stages use GPT-5-mini except SMOOTHSCORE judge (Step 5), which use GPT-5; every call uses `reasoning_effort=medium` with default sampling.

A.2 DATASET CURATION

We use all 65 ResearcherBench questions and the complete 25-question expert-written ScholarQA-CS2 set. These by construction contained open ended questions that required multi document engagement. For DeepResearch Bench II and ResearchRubrics:

- **Candidate pool.** We take the 66 English DeepResearch Bench II tasks and all 101 ResearchRubrics tasks.
- **Annotation.** We label each query on synthesis (HIGH / MED / LOW), meaning whether a good answer must synthesize a body of literature. We also label openness of question type (HIGH / MED / LOW), meaning whether the question requires one definitive correct response.
- **Filter.** We keep queries with synthesis = HIGH and Question Type = HIGH (Table 8).

Table 8: Query selection for DeepResearch Bench II and ResearchRubrics.

Benchmark	**Candidates**	**Kept**
DeepResearch Bench II (English)	66	18 (12 YES + 6 PARTIAL)
ResearchRubrics	101	17 (9 YES + 8 PARTIAL)

A.3 MODELS

Table 9 contains model IDs of the evaluated systems.

Table 9: Model identifiers and access method for each evaluated system.

System	**Model ID**	**Access**
GPT Luna + Websearch	`gpt-5.6-luna`	OpenAI Responses API
Claude Sonnet 5 + Websearch	`claude-sonnet-5`	Anthropic Messages API
Gemini Deep Research	`deep-research-preview-04-2026`	Gemini Interactions API
DR Tulu	`rl-research/DR-Tulu-8B`	local (vLLM + MCP server)
QUEST	`QUEST-35B-RL`	local

Table 10: Report length and retrieved evidence by dataset and system. Values are medians over reports. "Docs/report" is the number of retrieved documents, and "Evidence words" is the total number of words in the retrieved evidence available to the system.

Dataset	System	Report words	Docs/report	Evidence words
ResearcherBench	Luna	1,828	8.6	77,898
	Sonnet 5	1,889	19.6	134,720
	QUEST	1,332	9.8	55,481
	DR Tulu	1,948	17.4	114,027
	Gemini	4,860	39.2	254,916
CS2-Expert	Luna	1,999	10.4	34,821
	Sonnet 5	2,028	18.4	178,822
	QUEST	1,210	8.6	46,954
	DR Tulu	2,156	17.8	92,667
	Gemini	5,122	48.6	253,233
DRB-II	Luna	3,886	14.6	82,187
	Sonnet 5	3,243	30.9	205,016
	QUEST	2,134	10.0	39,316
	DR Tulu	2,884	14.4	110,388
	Gemini	4,844	32.7	235,720
ResearchRubrics	Luna	3,128	10.0	60,175
	Sonnet 5	3,140	28.1	234,495
	QUEST	1,930	9.6	52,607
	DR Tulu	2,732	15.0	106,390
	Gemini	4,687	31.2	196,485

A.4 Relationship between SmoothScore and Report and evidence length

Table 10 contains report length and retrieved evidence by dataset and system.

B SmoothScore Prompts

B.1 Claim Generation

System Prompt:

You identify the MAIN CONCLUSIONS a research report reaches in answering a question – the handful of load-bearing verdicts a reader walks away with – and return each one as a COMPLETE verbatim passage from the report. You are given the question and the report's answer.

BE SELECTIVE. Extract only the report's main conclusions: the thesis-level takeaways that directly answer the question. A typical report has roughly 4-8 of these. Favor the verdicts the whole report builds toward over the individual findings that support them.

UNIT OF EXTRACTION – the complete conclusion, NOT an atom:

1. Lift each conclusion WHOLE: typically 1-3 consecutive sentences (roughly 15-80 words). The claim MUST include, whenever they appear in the same or immediately adjacent sentences: every hedge and qualifier; the flip-side, contrast, or counterpart condition ("; it does not ...", "However, ...", "but ..."); and the stated scope or dimension the verdict applies to. Emitting the assertive half of a hedged or paired statement is the worst possible error – if the pieces cannot be included coherently, skip the conclusion entirely.

2. CARRY THE SCOPE. When a verdict's scope lives just OUTSIDE its sentence – in the immediately governing heading, a dimension label, or the sentence before – stitch that report wording in front (join stitched spans naturally; the heading text is report wording). NEVER present a per-dimension or per-condition winner as a global verdict.

3. NEVER extract a heading, list label, or table fragment as a claim, even when it is bolded and reads like a verdict. Headings point to claims; take the claim from the body text.

VOICE – verbatim about the assertion, silent about the sourcing:

4. REUSE THE REPORT'S OWN WORDING at the report's own strength. Keep every hedge, qualifier, comparative, superlative, and number exactly as written. Never strengthen, weaken, or paraphrase the assertion.

5. STRIP ATTRIBUTION AND CITATION FRAMING: drop "X's experiments found that", "according to ...", "studies show", citation markup, and markdown link syntax. KEEP a system/method name only when the claim is ABOUT that system.

6. STRIP DISCOURSE CONNECTIVES at the very start of a claim ("On the other hand,", "In summary,", "However,", "Overall,") – but connectives INSIDE the claim stay; they are doing real work joining the verdict to its counterpart.

WHAT COUNTS AS A CONCLUSION – all three required:

- A conclusion the report draws, not a building block: a result, comparison, judgment, trade-off, trend, or recommendation that answers the question.

- Substantive, not a bare definitional or mechanistic statement.

- Contentful and checkable: it asserts something specific enough to check against evidence. A forecast or recommendation counts, but the claim must preserve the report's framing of it as a forecast/recommendation.

GRANULARITY – one claim per distinct conclusion or thread. Different claims must not restate the same verdict; together they should cover each side's main strengths, the main limitations or trade-offs, and the overall verdict if the report gives one.

SKIP: background, definitions, setup, motivation, restatements of the question, section framing, individual supporting findings that roll up into a larger verdict.

OUTPUT – return only JSON with this exact schema:

`{"claims": [{"text": "{{Claim Text }}", "source_text": "{{Verbatim Source Span(s) }}"}]}`

For each claim:

- "text": the complete conclusion passage, following the rules above. Plain prose – no markdown, no citation markup, no list numbering.

- "source_text": the exact verbatim span(s) of the answer the claim was lifted from, copied character-for-character (citation markup and markdown syntax may be dropped). If the claim stitches spans (e.g. heading scope + body), join them with " ... " in report order. Each span must be findable in the answer by string search.

FINAL CHECKLIST – verify EVERY claim's "text" passes all of these; rewrite it if not:

- [ ] Complete sentence(s) – never a fragment, bullet label, heading, or table cell.

- [ ] Includes the local hedges, scope, and flip-side that appear in or immediately beside the source passage.

- [ ] No attribution framing, no markdown (`**`, `*`, `-`, `#`, `>`, `|`), no citation markup, no URLs, no list numbering.

- [ ] The assertion is the report's wording at the report's strength – qualifiers, comparatives, numbers unchanged.

If the answer states no checkworthy conclusions, return {"claims": []}.

User Prompt:

Example ID: {{Example ID }}

Question: {{Question }}

Extract the MAIN CONCLUSIONS of the following final_response – the handful of load-bearing, evidence-based verdicts that answer the question (typically 3–7). For each, return the claim text (the assertion in the report's own words at the report's own strength, with all attribution, citation, discourse-connective, and markdown framing STRIPPED, per the voice rules and final checklist) and the verbatim source_text span(s) it was lifted from. Use the question to decide which conclusions are load-bearing.

final_response:

{{Full Report Text}}

B.2 SUPPORTING PASS

System Prompt:
You extract from a single source DOCUMENT every distinct piece of information that SUPPORTS a single CLAIM from a research report, or a specific part of it – makes what the claim actually says more plausible or provides a direct instance/evidence for it.
Extract ONLY supporting evidence. Ignore evidence against the claim entirely; another process handles that.

- POLARITY: support means supporting what the claim ACTUALLY ASSERTS. If the claim asserts a limitation or a negative (e.g. "X is unreliable", "X scales worse", "X does not generalize"), then evidence of X failing/underperforming is supporting, and a result showing X working well is NOT supporting – skip it.
- STATED CONCESSIONS ARE PART OF THE CLAIM: evidence confirming a limitation, condition, or concession the claim itself states supports the claim as stated – include it, with the concession named in the aspect field.

An item is one distinct FINDING: a specific result, experiment, measurement, or stated conclusion. Each item is an object:

```
  {"quote": "<the shortest verbatim excerpt that carries this finding>",
   "aspect": "<the specific assertion WITHIN the claim this finding bears
              on, quoted or near-verbatim from the claim>",
   "note": "<one sentence: how this finding bears on that assertion>",
   "strength": <1-5>}
strength = how directly and specifically this item bears on what the
claim asserts:
  1 = tangential  2 = weak/indirect  3 = moderate: relevant but partial
  or hedged
  4 = strong: direct, on-point  5 = decisive: exactly what the claim
  asserts (or a clear counterexample to it).
```

Critical rules:

- Judge ONLY against this one document. No outside knowledge, no other documents.
- Read the WHOLE document; relevant evidence can sit in any section.
- ONE item per finding. If the document states the same finding several ways (abstract, body, conclusion), pick the single best quote for it – do not list restatements or different facets of the same result as separate items.
- Do NOT include items that only establish setting or applicability without an outcome (e.g. "the experiments use a 1.5B model"). That is context, not evidence; if it matters, mention it inside the note of a real item.
- An empty list is a perfectly good answer – do NOT invent items. Merely being about the same topic is NOT evidence; every item must actually move belief about the claim in the direction you were asked for.
- MECHANISM: only count evidence that matches what the claim specifically says (e.g. if the claim says "distillation from a LARGER model", a result using a same-size or self-distilled model does not bear on that specific mechanism). The aspect field must name the exact assertion addressed – if you cannot point to one, the item does not belong.
- Quotes must be copied verbatim from the document.

Return only JSON: {"items": [...]}

User Prompt:

Question (context only, not evidence): {{Question}}
Claim: {{Claim}}
Document:
Title: {{Document Title}}
URL: {{Document URL}}
Content:
{{Full Document Text}}
Extract every distinct piece of supporting evidence and return only the JSON object.

B.3 SKEPTIC PASS

System Prompt:
You extract from a single source DOCUMENT every distinct piece of information that a SKEPTICAL reader would weigh against a single CLAIM from a research report, and you TAG each item by its relation to the claim AS STATED. Extract only this skeptical side; a separate process handles supporting evidence. Read the claim carefully first: its actual assertions, its scope, its comparatives, and every limitation, condition, or concession it ALREADY STATES or hedges over.

TAGS:

- `"counter"`: the finding cuts against what the claim ACTUALLY ASSERTS – contradicts it, provides a counterexample, limits its scope beyond what it admits, shows the asserted effect weaker or more conditional than asserted, or reports a null or mixed result where the claim implies a clear one. POLARITY: if the claim asserts a limitation or a negative ("X is unreliable"), a result showing X working WELL is a counter, and evidence of X failing is NOT. A soft hedge does not concede everything: "X is usually best" still asserts usually – evidence X OFTEN loses is a counter; evidence X merely SOMETIMES loses is not.
- `"concession-match"`: the finding confirms, instantiates, or quantifies a limitation, condition, or concession the claim ITSELF states or hedges over ("capacity-sensitive", "results vary", "except in short contexts"). These do NOT count against the claim – they are extracted so a later audit can check whether the report states the substance behind its own hedges.
- `"rival-win"`: a success of ANOTHER approach reported in that approach's own terms, where the claim does NOT rank over or exclude that approach. If the claim DOES rank or exclude ("strongest", "most practical", "X remains ahead of Y"), a rival's success is a `"counter"`, not a `"rival-win"`.

A carefully scoped claim often has NO counter items in a given document – an empty counter set is a common and correct outcome; do not stretch. But still extract concession-match and rival-win items where they exist.

- STRENGTH for `"counter"` = how much the finding should CONSTRAIN the claim as stated: 5 = directly contradicts the central assertion or shows it failing within its own stated scope; 4 = a hard limit or condition on the main assertion that the claim does not state; 3 = a meaningful qualification; 2 = weak/indirect pressure; 1 = tangential.
- STRENGTH for `"concession-match"` = how load-bearing the conceded condition is: 5 = the concession hides a result that would change most readers' decisions; 3 = a substantive but secondary condition; 1 = a trivial caveat.
- STRENGTH for `"rival-win"` = how strong the rival result is in its own terms (controlled comparison > own-harness leaderboard > assertion).

An item is one distinct FINDING: a specific result, experiment, measurement, or stated conclusion. Each item is an object:

```
  {"quote": "<the shortest verbatim excerpt that carries this finding>",
   "aspect": "<the specific assertion WITHIN the claim this finding bears
              on, quoted or near-verbatim from the claim>",
   "note": "<one sentence: how this finding bears on that assertion>",
   "tag": "counter|concession-match|rival-win",
   "strength": <1-5>}
strength = how directly and specifically this item bears on what the
claim asserts:
  1 = tangential  2 = weak/indirect  3 = moderate: relevant but partial
  or hedged
  4 = strong: direct, on-point  5 = decisive: exactly what the claim
  asserts (or a clear counterexample to it).
```

Critical rules:

- Judge ONLY against this one document. No outside knowledge, no other documents.
- Read the WHOLE document; relevant evidence can sit in any section.
- ONE item per finding. If the document states the same finding several ways (abstract, body, conclusion), pick the single best quote for it – do not list restatements or different facets of the same result as separate items.
- Do NOT include items that only establish setting or applicability without an outcome (e.g. "the experiments use a 1.5B model"). That is context, not evidence; if it matters, mention it inside the note of a real item.
- An empty list is a perfectly good answer – do NOT invent items. Merely being about the same topic is NOT evidence; every item must actually move belief about the claim in the direction you were asked for.
- MECHANISM: only count evidence that matches what the claim specifically says (e.g. if the claim says "distillation from a LARGER model", a result using a same-size or self-distilled model does not bear on that specific mechanism). The aspect field must name the exact assertion addressed – if you cannot point to one, the item does not belong.
- Quotes must be copied verbatim from the document.

Return only JSON: {"items": [...]}

User Prompt:
Question (context only, not evidence): {{Question }}
Claim: {{Claim }}
Document:
Title: {{Document Title }}
URL: {{Document URL }}
Content:
{{Full Document Text }}
Extract every distinct piece of countervailing evidence and return only the JSON object.

B.4 ASPECT GENERATION

System Prompt:
You are an analyst that identifies the different aspects that a single scientific research paper ACTUALLY DISCUSSES which bear on evaluating a claim. For each aspect, provide a brief all-lowercase label (3–8 words), a one-sentence description of what THIS DOCUMENT says about it and why it is significant for addressing the claim, and 6 keywords typically used to describe that aspect.
Rules (corpus-grounded, claim-relative):

- Only include aspects this document actually contains substantive content on—do not invent aspects from general knowledge of the topic.
- Frame each aspect as a dimension bearing on THIS claim (e.g., "task-type dependence of the comparison"), not a generic subtopic of the field.
- Return as many aspects as the document genuinely supports: typically 2–3.

One aspect is a good answer for a narrowly relevant document, and an EMPTY LIST is the correct answer for a document with nothing substantive on this claim. Do not pad to a count.
Return only JSON: {"aspect_list": [{"aspect_label": "...", "aspect_description": "...", "aspect_keywords": ["..."]}]}.

User Prompt:
CLAIM: {{Claim }}
DOCUMENT #{{Document Index }} (title: {{Document Title }}):
{{Full Document Content }}
Return only the JSON object.

B.5 ASPECT MERGE

System Prompt:
You are an analyst consolidating per-document aspect proposals into ONE aspect hierarchy for evaluating a claim. You will identify 4–5 top-level aspects that the documents' proposals are collectively considering when determining the answer to the claim. Add subaspects ONLY where the corpus genuinely subdivides an aspect into distinct, separately-evidenced points—a top-level aspect with NO subaspects is a good and common answer. Aim for roughly 4–8 nodes in the whole hierarchy, not the maximum. Each subaspect must name what distinguishes it from its siblings; if you cannot, merge them.
Rules:

- Corpus-grounded: every aspect/subaspect must be traceable to the per-document proposals; carry over the source document indexes into `"source_docs"`. Do not invent nodes no document proposed.
- Claim-relative: keep labels framed as dimensions bearing on the claim.
- Merge duplicates across documents; prefer the most specific phrasing proposed.
- A subaspect is a more specific point underneath its parent that would be considered when specifically addressing the claim.

Return only JSON: `{"aspects": [{"label": "...", "description": "...", "source_docs": [1,2], "subaspects": [{"label": "...", "description": "...", "source_docs": [3]}]}]}`.

User Prompt:
CLAIM: {{`Claim`}}
PER-DOCUMENT ASPECT PROPOSALS:
[doc 1] {{`Aspect Label`}}: {{`Aspect Description`}}
[doc 1] {{`Aspect Label`}}: {{`Aspect Description`}}
[doc 3] {{`Aspect Label`}}: {{`Aspect Description`}}
...
Return only the JSON object.

B.6 EVIDENCE ASSIGNMENT

System Prompt:
You will be provided with evidence quotes extracted from source documents. Each quote was extracted as evidence about a claim. Please select the suitable aspect each quote talks about from the provided aspect list.
LEAF-ONLY RULE: a node that has sub-aspects is a container only. If the quote belongs in such an aspect's territoryts sub-aspects (thebest-fitting one), never the parent. Only aspects with no sub-aspects can receive quotes directly.
Special destinations:

- "ROOT" if the quote bears oher than any one aspect(overall verdicts, provenance of the whole comparison, conditions spanning aspects)ROOT is a deliberate destinat
- "UNASSIGNED" only if the quote fits no listed aspect and is not whole-claim evidence.

Assignment is topical: judge what the quote is ABOUT, never whether it supports or opposes the claim.
Each quote is prefixed with its source document (index and title). Use it as cona fragment (a table row, a bare number, a one-line result) is about whatever that document discusses at that point – file it by its document's topic, not by the words in the fragment alone.
Return only JSON: `{"assignmen "node": "<node id like A1 orA2.3 or ROOT or UNASSIGNED>"}]}`.

User Prompt:
CLAIM: {{`Claim`}}
ASPECT LIST:
A1: {{`Aspect Label`}}
A1.1: {{`Aspect Label \(>\) Subaspect Label`}}
A2: {{`Aspect Label`}}
...
EVIDENCE:
[0] (doc 3 — {{`Title, 90 Charsned, 400 Chars`}})

[1] (doc 7 — {{Title }}) {{Quote }}

B.7 SMOOTHSCORE JUDGE

System Prompt: You are an evidence-smoothing judge. You assess how faithfully a claim compresses the evidence supplied to it — not whether the claim is well-written, complete, or true by outside standards.

User Prompt: # Inputs
Research question: {{QUESTION}}
Claim under evaluation: {{CLAIM}}
Evidence by aspect: {{ASPECTS_AND_EVIDENCE}}
Each evidence item may include: a source identifier, an exact quotation, a generated context note, and a retrieval-pass label ("support-pass candidate" / "countervailing-pass candidate").

What you're measuring
Evidence smoothing: the claim collapses material heterogeneity, uncertainty, conditionality, or disagreement in the evidence, producing a conclusion broader, more certain, more universal, or more decisive than the evidence proportionately warrants.
Not in scope: coverage (the claim need not mention every finding), writing quality, or accuracy against knowledge outside the supplied evidence.

Ground rules for reading the evidence
1. **The supplied evidence is the entire universe.** No outside knowledge.
2. **Retrieval labels describe provenance, not verified stance.** Interpret each item's actual content independently — a support-pass item can fail to support the claim; a countervailing-pass item can support, qualify, or be irrelevant to it.
3. **Item counts and pass labels are not evidence of consensus or its absence.** Never infer balance, or lack of it, from how many items landed in each pass.
4. **Context notes clarify; they don't add evidence.** Ground judgments in the quotation. Preserve the claim's actual scope, certainty, and role in answering the question. Do not weaken a strong assertion into a vague statement of usefulness, or strengthen a qualified claim into an unconditional guarantee.
5. **Weigh by relevance, directness, applicability, and any quality signal stated in the text — not by quote count.** Restatements of the same finding are one data point, not several.
6. **The claim's own logical force sets what can defeat it:** - "X can work" — isolated failures of X don't count against it. - "X always / necessarily works" — one sound counterexample is decisive. - "X generally works / dominates" — isolated counterexamples aren't enough; the evidence must speak to prevalence or the overall comparison. - "X is promising" — a rival promising method doesn't refute it. - "X is comparatively best" — comparable alternatives are directly relevant. - A claim need not preempt exceptions its own wording already allows.
7. **Absence of support $\neq$ contradiction**, but an unsupported specific assertion is still unwarranted — weight this by whether it's peripheral or central to the takeaway.
8. **For multi-part claims, weight problems by their effect on the overall takeaway.** Don't average proposition-level defects or default to the score triggered by the single worst one.

Procedure
Work through these in order. Your response's structure follows this order (see Output).
1. **Claim's contribution.** What is the claim doing relative to the question — explaining a possibility, asserting a general pattern, reporting a result, making a recommendation? This determines what kind of evidence can defeat it.
2. **Substantive takeaway.** State the scope, certainty, comparison, and causal content the claim actually asserts, independent of phrasing. Don't pre-weaken an overstated claim before judging it.
3. **Positive evidential case.** What do the relevant findings, taken together, justify? Include reasonable inferences across findings. Flag anything the claim asserts that the evidence doesn't actually reach (a specific mechanism, magnitude, or comparison it never establishes), even if the surrounding evidence is generally favorable.
4. **Objections.** For each apparent counterpoint: does it address the same assertion, in the applicable setting? Distinguish a consequential restriction from a compatible failure, a different mechanism, or an unrelated limitation. Treat a finding as an outlier only when the evidence itself gives a basis for that, not because it's outnumbered or inconvenient.
5. **Net resolution.** Combine 3 and 4. Does the claim survive as stated, survive with a limitation attached, or require a substantially different conclusion?
6. Score via the minimum-repair test below.

Minimum-repair test
Determine the score by identifying the minimum semantic revision needed to make the claim faithful to the supplied evidence. Work through these levels in order and stop at the first one that fits:
- None: the claim can remain as written. Omitted material would only elaborate or add background. A further caveat could be written, but none is needed for fidelity.
- Append: a caveat or exception should be added for the claim to represent the evidence faithfully, but the original claim otherwise remains — its central scope, certainty, comparison, and conclusion stay substantially unchanged.
- Revise: the claim's central direction survives, but a material predicate, quantifier, scope, comparison, causal attribution, or confidence level must be changed. A reader who accepted the original would draw a materially different conclusion from the repaired version.
- Recast: the overall conclusion must be substantially narrowed, rebalanced, or made explicitly mixed. A recognizable part of the original takeaway survives, but not the conclusion as presented.
- Replace: the central takeaway cannot be preserved. A faithful account requires a materially different conclusion; adding qualifications or rebalancing is insufficient.

Do not select Append merely because an additional caveat could be written — select it only when a caveat is genuinely needed for faithful representation. Do not select Revise for a hedge the claim's own wording already implies (e.g. adding "generally" to a claim that elsewhere says "typically") — that is an Append.

Scoring rubric
Each repair level corresponds to exactly one score:

None → 1 (Reasonable compression). The claim preserves important conditions and fairly represents relevant counterevidence. Omitted details do not meaningfully change the takeaway.

Append → 2 (Minor nuance lost). Some conditions, limitations, or exceptions are omitted, but including them would leave the takeaway largely unchanged.

Revise → 3 (Important qualification missing). An important condition, limitation, or exception is omitted, making the claim broader or more certain than warranted. The central takeaway remains defensible with that qualification.

Recast → 4 (Major distortion). Substantial counterevidence or a decisive limitation is omitted or understated. The claim presents a clear conclusion where the evidence requires substantial narrowing, rebalancing, or a mixed conclusion.

Replace → 5 (Fundamentally misleading). The claim fundamentally misrepresents what the evidence warrants — for example, by omitting both crucial conditions and strong counterevidence. The central takeaway needs replacing; adding qualifications alone cannot make it faithful. Direct contradiction is not required.

Deciding the score

Decide from the state of the evidence around the central takeaway, then confirm against the rubric above.

Only material bearing on the central takeaway counts here. Material that qualifies a subsidiary assertion — an example, a mechanism, a side-condition, an intensifier — does not move the score unless correcting it changes the takeaway itself.
The claim's hedges are part of its wording. "Likely", "often", "when X", "reducing" set what has to be shown. Evidence that the claim is not uniformly or always true does not bear against a claim that never asserted universality.

- Nothing in the evidence bears against the takeaway, or what does is already allowed by the claim's own wording → 1
- What bears against it is narrower in scope, less direct, or less consistent than what supports it — the takeaway holds with surety → 2
- What bears against it is as direct and as broad as what supports it — the takeaway is arguably correct but contested → 3
- What bears against it outweighs the support, or the two together leave the central question genuinely two-sided — a faithful account is mixed → 4
- The evidence never establishes the takeaway, or establishes something materially different → 5

The repair labels — none, append, revise, recast, replace — describe the consequence of the score and are reported for auditing. They are not the test. Before assigning a score, identify a material qualification required by the evidence that is not already conveyed by the claim's wording or context.

Output
Return a single JSON object with exactly these fields, in this order. No text before or after the JSON.

```
{
  "claim_role": "<one sentence: what kind of claim this is
  (possibility / general pattern / result / recommendation),
  and what that implies for what can defeat it>",

  "takeaway": "<the claim's actual scope, certainty, comparison,
  and causal content, stated plainly>",

  "positive_case": "<what the relevant evidence, taken together,
  establishes, and, if applicable, what the claim asserts beyond
  what the evidence actually reaches>",

  "objections": "<what the countervailing or qualifying evidence
  actually changes, if anything; use 'none material' if nothing
  meaningfully bears on the claim as scoped>",

  "net_resolution": "<does the claim survive as-is, survive with
  a limitation, or need substantial revision, and why, in 1-2
  sentences>",

  "repair": "none | append | revise | recast | replace",

  "score": 1
}
```

Constraints: - "repair" and "score" must correspond exactly: none=1, append=2, revise=3, recast=4, replace=5. - "score" must be an integer, not a string. - Every field is required. Do not add extra fields. - Output must be valid JSON — no markdown code fences, no trailing commas, no comments.

C STEP BY STEP OUTPUTS OF THE SMOOTHSCORE PIPELINE

Table 11: Step-by-step application of the evidence-smoothing pipeline to a claim about reinforcement learning with verifiable rewards.

Pipeline step	Output for the example
Input question	Will RL-based methods become the mainstream approach for improving reasoning models after DeepSeek-R1, and can reward models and tree search for chain-of-thought post-training be abandoned?
1. Claim extraction	*"RL—especially RL with verifiable rewards—will become a mainstream component of reasoning-model training."*
2. Document reconstruction	The system reconstructs the full text of the documents cited by the report, including work on DeepSeek-R1, verifier gaming, test-time scaling, AlphaProof, AlphaGeometry2, and surveys of RL for reasoning.

Table 11 continued

Pipeline step	Output for the example
3a. Supporting-evidence extraction	The supporting pass extracts findings that substantiate the claim as stated. For example: *"Figure 1(a) depicts the performance trajectory of DeepSeek-R1-Zero on the AIME 2024 benchmark throughout the RL training process, where the average pass@1 score on AIME 2024 shows a significant increase, jumping from an initial 15.6% to 77.9%."* *"As reinforcement Learning with Verifiable Rewards (RLVR) has become the dominant paradigm for scaling reasoning capabilities in LLMs."* Across all claim–document pairs, the pass extracts 61 supporting items.
3b. Countervailing-evidence extraction	The countervailing pass extracts findings that contradict, weaken, or materially restrict the claim. For example: *"We show that training on only 1,000 samples with next-token prediction and controlling thinking duration via a simple test-time technique we refer to as budget forcing leads to a strong reasoning model that scales in performance with more test-time compute."* *"The rule-based RL training stage of DeepSeek-R1-Zero is narrowly focused on reasoning tasks, resulting in limited performance in broader areas such as writing and open-domain question answering."* Across all claim–document pairs, the pass extracts 77 countervailing items.
4a. Local aspect induction	For each claim–document pair, the system proposes dimensions discussed by that document that bear on evaluating the claim. For example, DeepSeek-R1 proposes an aspect concerning *performance gains on objectively verifiable reasoning tasks*, while the test-time-scaling paper proposes an aspect concerning *non-RL alternatives for improving reasoning*.
4b. Aspect consolidation	Overlapping document-level proposals are merged into the following claim-level hierarchy: • **A1: Empirical effectiveness and domain scope of RL with verifiable rewards** – **A1.1:** Large gains on tasks with objective checks – **A1.2:** Boundaries and alternatives where RL is unnecessary or underperforms • **A2:** Dependence on verifier reliability and design for safe, scalable use • **A3:** Practicality and adoption readiness of RLVR in training pipelines – **A3.1:** Operationalization and community uptake – **A3.2:** Resource costs and sample-efficiency trade-offs

Table 11 continued

Pipeline step	Output for the example
5. Evidence assignment	Supporting and countervailing items are assigned to the most specific applicable aspect. For example: **A1.1: Large gains on tasks with objective checks** *Supporting evidence:* • DeepSeek-R1-Zero's AIME 2024 pass@1 score rises from 15.6% to 77.9% during RL training. • AlphaProof learns to produce formal proofs through RL over millions of auto-formalized problems. *Countervailing evidence:* • AlphaGeometry2 achieves an 84% solve rate using a pipeline centered on language modeling, symbolic reasoning, and synthetic data rather than RL. • AlphaProof's demonstrated successes remain concentrated in competition mathematics. The final node contains 31 supporting and 9 countervailing items. Across the hierarchy, this is the only node in which supporting items predominate; the remaining dimensions contain substantial countervailing evidence.
6. Smoothing judgment	**Claim role:** General forecast about a field-wide training trend; it can be defeated by evidence that RL with verifiable rewards is unlikely to be widely adopted or that strong, persistent limitations preclude mainstream use. **Takeaway:** Reinforcement learning—particularly when rewards are verifiable—will be widely adopted as a standard part of training pipelines for reasoning models, though not necessarily the only or exclusive approach. **Objections:** Important constraints temper the breadth of the forecast. RL scales poorly where no reliable verifier exists; it can underperform or fail to improve smaller models; compute and evaluation costs limit practicality; and verifier imperfections introduce shortcutting and misalignment risks. Strong non-RL alternatives can also match or surpass leading RL-trained systems. These findings show that RLVR will not be universally applicable or singularly dominant across reasoning tasks. **Net resolution:** The claim largely holds but needs an explicit limitation: RL with verifiable rewards is poised to be mainstream for verifier-evaluable settings, yet adoption will be constrained by verifier availability, safety and shortcut risks, model and compute regimes, and viable non-RL alternatives that will continue to coexist. **Repair:** APPEND **Smoothing score:** 2

C.1 LUNA (GPT-5.6-LUNA + RESPONSES API `WEB_SEARCH`).

What the trace records.

- `tool_calls[]` — one entry per tool call, with an `action_type`. 201 total: 167 searches (each carrying up to three queries), 28 page opens, 6 in-page finds.
- `search_results[]` — the text the search tool returned to the model, one entry per result. 4,608 entries, each carrying a `[wordlim: N]` cap set by the tool ($N = 200$ on 97% of results).
- `citations[]` — which URL each cited claim points to, with character offsets into the report. 556 entries; no source text is stored here.

- Opened pages — an `open_page` call can open several pages but the log stores at most one URL per call (25 unique URLs across 27 URL-bearing calls), so we recover the full opened set from the `turnNviewN / Source: open(...)` markers in the returned text. 105 URLs; the 25 are a strict subset.
- Cited documents — no native field; we take the unique URLs in `citations[]`, with arXiv `abs`/`pdf` variants collapsed to one document. 497 documents.

What the system used. Snippets. A search result gives the model a median 585 characters; an opened page gives a median 4,294 characters, as a line-numbered window of the page. Of the 497 documents Luna cites, 95 (19.1%) were opened; the other 81% are cited from a capped snippet alone. Per report, the share of cited documents opened is 0% in 45 of 65 reports and a median of 71% in the remaining 20 (pooled: 19.1%).

C.2 Docs in Context, with Instructions

System Prompt:
You are a professional researcher producing a structured, citation-backed report. The complete source documents are provided below; they are the ONLY evidence you may use. Ground every factual claim in them and cite by document number, e.g. [doc 3]. Where the documents disagree with each other, say so explicitly rather than choosing a side silently. If the documents do not support a conclusion, say that instead of asserting it from background knowledge.

User Prompt:
SOURCE DOCUMENTS
=== DOCUMENT 1 ===
URL: {{`Document URL`}}
TITLE: {{`Document Title`}}
{{`Document Text`}}
=== DOCUMENT 2 ===
URL: {{`Document URL`}}
TITLE: {{`Document Title`}}
{{`Document Text`}}
...
=== QUESTION ===
{{`Question`}}
Write the report now, citing documents by number.

C.3 Docs in Context, No Epistemic Instructions

System Prompt:
You are a professional researcher producing a structured, citation-backed report. The complete source documents are provided below; they are the ONLY evidence you may use. Ground every factual claim in them and cite by document number, e.g. [doc 3].

User Prompt:
SOURCE DOCUMENTS
=== DOCUMENT 1 ===
URL: {{`Document URL`}}
TITLE: {{`Document Title`}}
{{`Document Text`}}
=== DOCUMENT 2 ===
URL: {{`Document URL`}}
TITLE: {{`Document Title`}}
{{`Document Text`}}
...
=== QUESTION ===
{{`Question`}}
Write the report now, citing documents by number.

C.4 Improved claims under document engagement

D Examples of Smoothing in Model Generated Claims

D.1 Open-source systems

Examples of generations by DR Tulu and Quest are in Table 13.

D.2 Closed-source systems

Table 14 shows examples of claims generated by Luna and Sonnet 5.

Table 12: How smoothing changes when the same model is given the cited documents (Luna, ResearcherBench). Claims are extracted independently per report; each row shows the claim in each arm that reaches the same conclusion, with σ in parentheses. "–" means no claim in that arm reaches the same conclusion.

Pattern	**Question**	**WebSearch (σ)**	**Docs in context (σ)**	**Docs + instructions (σ)**
A result from one setup stated without its conditions	Analyze the differences between the LIMO and S1 papers: objectives, methodology, key findings, contributions.	LIMO appears stronger in broad benchmark performance, particularly on GPQA Diamond and MATH500. (4)	LIMO reports higher scores than *original* s1 on AIME24, MATH500 and GPQA, but the studies differ in evaluation details (greedy decoding vs. temperature-0) … the comparison does not by itself establish a controlled head-to-head superiority claim. (1)	LIMO is the more explanatory and curation-focused paper … [doc 1]. (3)
A result from one setup stated without its conditions	How can we detect hallucinations in LLMs using their internal states, and what advantages does this offer over external methods?	Lower latency and computational cost: internal detectors can operate in the same forward pass and use a lightweight probe. MIND targets real-time detection, and HIDE reports substantially lower computation time than multi-pass methods. (2)	–	Its principal advantages over external or output-only detection are earlier intervention, richer process-level information, reduced reliance on external knowledge bases, *lower latency for single-pass methods*, and improved interpretability. (1)
Separately supported fragments combined into one demonstrated conclusion	Could transformer architectures be reimagined to process multimodal inputs (video/audio/-text) as efficiently as text?	The more promising principle is: use modality-specific perception, adaptive semantic tokenization, sparse cross-modal attention, persistent latent memory, and on-demand access to raw evidence. (4)	The evidence suggests that multimodal Transformers *may* need to be redesigned around asymmetric processing, learned compression, latent bottlenecks, adaptive tokenization, and selective cross-modal interaction. (3)	The likely path … is to redesign the architecture so that all modalities are converted into compact, queryable representations … Whether that bottleneck can remain small enough … *remains unresolved by the supplied evidence*. (1)
Separately supported fragments combined into one demonstrated conclusion	How might federated learning combined with model distillation overcome technical and legal barriers in domains like healthcare?	The most credible approach is therefore a layered architecture: local training + distillation + secure aggregation + differential privacy + strong governance and clinical validation. (3)	The strongest approach is therefore layered: local data retention, carefully selected distillation, secure aggregation, differential privacy where appropriate … and clinical validation. (3)	The strongest approach is therefore layered: [same list] … Such an architecture can substantially lower both technical and legal risk, *but the supplied documents do not support claiming that it eliminates either category of barrier*. (1)
Conclusive or superlative wording beyond what the sources state	What is the evolution path of multimodal models from early visual representations to current multimodal LLMs, and what are the key breakthroughs?	Language as a *universal* interface for vision: natural language became both a source of supervision and a way to query visual representations. (3)	CLIP's central breakthrough was language-grounded open-vocabulary recognition … The approach also exposed important limitations: performance depended on prompt wording, and CLIP struggled with counting, fine-grained recognition, and out-of-distribution images [doc 11]. (1)	CLIP and ALIGN demonstrated that web-scale paired text can train general visual representations, support retrieval, and produce zero-shot classifiers through natural language [docs 11, 13]. (1)
Conclusive or superlative wording beyond what the sources state	(same question)	BLIP-2 made the connection more parameter-efficient. This was an *important engineering breakthrough* because it showed that multimodal capability could be obtained without end-to-end retraining of every parameter. (3)	–	Flamingo and BLIP-2 showed that strong pretrained vision and language models can be connected through relatively small trainable modules rather than fully retrained end-to-end [docs 15–16]. (1)

Table 13: High-severity claims from the open-source systems, with the SMOOTHSCORE judge's verdict.

Pattern	System	Question	Claim (score)	Judge verdict
Unsupported composition	DR Tulu	*DRB-II ex16:* I am researching the challenges faced by SRAM operating at low voltages, and I hope you can help me survey and compile a technical report on mainstream solutions. The report needs to include the following ...	Alternate Cells: 7T adds explicit precharge–equalize sequence, 8T virtual-ground/zero-precharge, 9T conflict-free write with equalized leakage, 10T quasi-differential read, 11T DDPS for ultra-low leakage, 12T gating and two-stage read. (4)	The claim over-compresses heterogeneous literature into a fixed one-to-one mapping and misattributes key mechanisms (notably 7T and 11T), while several other pairings are only variant-specific. It requires substantial revision to a qualified, example-based mapping rather than categorical assignments.
Unsupported composition	DR Tulu	*CS2-Expert ex6:* Describe the Sunflower conjecture, progress towards a solution, and related problems.	Decision trees, DNF/branching programs, and communication: Variants of the sunflower conjecture directly constrain the size/width/depth of Boolean formulas and decision trees, with tight results (e.g., $\tilde{\Omega}(k^2)$) for unambiguous DNF; these bounds, proved via explicit constructions and sunflower arguments, carry over to communication protocols and algebraic complexity analogs (e.g., matrix multiplication exponent connections surveyed by Alon). (4)	The claim over-compresses heterogeneous results: it correctly cites tight unambiguous DNF bounds and connections to communication, and separately, sunflower-variant tools and matrix-multiplication connections, but it incorrectly attributes the tight DNF bound to sunflower arguments, overstates direct constraints on decision trees/branching programs, and overstates transfer to algebraic complexity. It needs substantial narrowing and reattribution.
Premature closure	QUEST	*CS2-Expert ex11:* Survey recent papers on the use of LLMs for 5G network orchestration	Validates that LLM-powered prompt engineering outperforms static ML methods in highly variable O-RAN environments. (5)	The claim overstates what the evidence shows. The record supports that LLM prompting can beat simple heuristics and be useful in hybrid schemes, but it does not validate outperformance over static ML in highly variable O-RAN environments.
Premature closure	QUEST	*ResearcherBench ex3:* In multimodal pre-training, the current mainstream paradigms are based on image tokens and stable diffusion. Analyzing the latest advancements in these two technical approaches, with reference ...	SOTA Across Multiple Benchmarks: DDT-based MLLMs outperform all previously reported models in zero-shot/non-zero-shot evaluation for image-text matching, captioning, question answering (VQA, GQA), object detection, and metric scores (GenEval 0.66, NOCAPs 124.2, EVR L1 14.5, CVS 85.5%). Even compared to diffusion specialists trained on similar data, DDT shows clear superiority in compositional edits and counterfactual interpolation, owing to its recursive, attribute-rich tokens that mimic natural language structure. (4)	The claim needs substantial revision: while the quoted metric values are accurate and DDT-based MLLMs are competitive, the evidence shows they do not outperform all previously reported models on several listed benchmarks and provides no support for object detection.

Table 14: High-severity claims from the closed-source systems, with the SMOOTHSCORE judge's verdict.

Pattern	System	Question	Claim (score)	Judge verdict
Scope expansion	Sonnet 5	*ResearcherBench ex49:* From a technical perspective, how to understand the similarities and differences between Reinforcement Learning (RL) algorithms and Supervised Fine-Tuning (SFT) in Large Language Models (LLMs), as well as …	RL, especially when trained with an outcome-based reward, generalizes across both rule-based textual and visual variants, while SFT, in contrast, tends to memorize training data and struggles to generalize out-of-distribution scenarios. (4)	The claim overstates a mixed and conditional landscape. While there is solid evidence that outcome-based RL can improve OOD generalization in both textual and visual tasks and that SFT often overfits, there is also substantial, direct evidence that SFT can generalize and sometimes surpass RL, and that RL's gains depend on reward design, initialization, and compute. The conclusion needs rebalancing and explicit conditions.
Premature closure	Luna	*ResearcherBench ex3:* In multimodal pretraining, the current mainstream paradigms are based on image tokens and stable diffusion. Analyzing the latest advancements in these two technical approaches, with reference …	For multimodal pretraining as a route toward a single general-purpose model, discrete image-token modeling appears more promising than a Stable-Diffusion-centered architecture. It fits naturally with the scaling machinery of language models: images, text, and video can be serialized into one sequence and trained with a common next-token objective. The strongest evidence comes from Emu3, Janus-Pro, Chameleon, and the rapidly improving tokenizers TiTok/FlexTok. (4)	The claim needs substantial revision. The evidence supports discrete-token, next-token training as a viable and rapidly improving path, but not that it is more promising overall than diffusion-centered approaches; multiple studies show diffusion/flow or hybrid models scaling better and being more compute-efficient, and the 'natural fit' is tempered by instability and modality-specific shortcomings.
Unsupported composition	Gemini	*CS2-Expert ex2:* What are some of the methods of detecting GPS location spoofing attack in Connected and Autonomous Vehicles?	At the mechanical layer, the tight fusion of GNSS data with highly resilient IMU kinematics and LiDAR map-matching ensures that even if a spoofed signal perfectly replicates the cryptographic and spatial signatures of a satellite, it fundamentally cannot alter the physical reality of the road ahead. (4)	The claim needs substantial revision. The evidence supports that GNSS–IMU–LiDAR/map fusion can greatly harden localization and often detect/mitigate spoofing, but it does not ensure immunity—even against sophisticated, spatially and cryptographically convincing attacks—and has documented failure modes and delays.
Scope expansion	Gemini	*ResearcherBench ex4:* Please analyze the differences between the LIMO and S1 these two papers. Provide a detailed comparison, considering aspects such as their research objectives, methodologies, key findings, and overall contributions.	LIMO and s1 definitively prove that reasoning is not created during the post-training phase; it is merely unlocked. This validates an extension of the superficial alignment hypothesis into the realm of complex cognition: pre-training builds the cognitive engine, and a mere 800 to 1,000 pristine examples act as the ignition key. (4)	The claim overstates what the evidence supports. A faithful conclusion is mixed and conditional: LIMO and s1 provide evidence consistent with the hypothesis that, in sufficiently pretrained models, small, carefully curated datasets (sometimes plus inference-time control) can elicit strong reasoning—but they do not definitively prove that post-training cannot create reasoning, nor that 800–1,000 pristine examples universally suffice or that this extends across all complex cognition.