VERISCORE: Evaluating the factuality of verifiable claims in long-form text generation

Yixiao Song[♢♠] **Yekyung Kim**[♢] **Mohit Iyyer**[♢]
[♢] Manning College of Information and Computer Sciences, UMass Amherst
[♠] Department of Linguistics, UMass Amherst
{yixiaosong, yekyungkim, miyyer}@umass.edu

Abstract

Existing metrics for evaluating the factuality of long-form text, such as FACTSCORE (Min et al., 2023) and SAFE (Wei et al., 2024), decompose an input text into "atomic claims" and verify each against a knowledge base like Wikipedia. These metrics are not suitable for most generation tasks because they assume that every claim is *verifiable* (i.e., can plausibly be proven true or false). We address this issue with VERISCORE,[1] a metric for evaluating factuality in diverse long-form generation tasks that contain both verifiable and unverifiable content. VERISCORE can be effectively implemented with either closed or fine-tuned open-weight language models. Human evaluation confirms that VERISCORE's extracted claims are more sensible than those from competing methods across eight different long-form tasks. We use VERISCORE to evaluate generations from 16 different models across multiple long-form tasks and find that while GPT-4o is the best-performing model overall, open-weight models such as Mixtral-8×22 are closing the gap. We show that an LM's VERISCORE on one task (e.g., biography generation) does not necessarily correlate to its VERISCORE on a different task (e.g., long-form QA), highlighting the need for expanding factuality evaluation across tasks with varying fact density.

1 Introduction

Modern approaches for evaluating the factuality of LLM-generated long-form text, such as FACTSCORE (Min et al., 2023) and SAFE (Wei et al., 2024), proceed in three stages: (1) *decomposition* of the text into a list of "atomic" (i.e., short) claims; (2) *retrieval* of relevant evidence for each claim from Wikipedia or Google Search; and (3) *verification* of each claim against the retrieved evidence. These approaches implicitly assume that the input text can be entirely decomposed into *atomic* and *verifiable* claims.

Unfortunately, these assumptions do not always apply to complex generation tasks such as long-form question answering (LFQA) for two reasons. First, outputs for the biography generation task studied in FACTSCORE rarely go beyond introducing entities and events. However, in tasks like LFQA, we observe more complex assertions that cannot be made "atomic" without losing critical context, as in:

> (1) The impeachment of Andrew Johnson set a precedent that impeachment should be reserved for clear cases of serious misconduct rather than political disagreements.

Second, many long-form outputs interleave factual claims with unverifiable content, such as *Betacyanin is like a superhero cape* in Figure 1. Since FACTSCORE and SAFE assume all claims are verifiable, they extract everything from the text (including unverifiable content like examples or hypotheticals), which can unfairly penalize models during the final aggregation process. As such, these metrics are limited to fact-dense and formulaic text (e.g., biographies).

We address these issues by VERISCORE, an automatic metric that assesses models' factuality against Google Search results. VERISCORE's decomposition and verification steps are initially implemented using few-shot prompting. Extensive human studies confirm the quality of both steps. Subsequently, open-weight LLMs are fine-tuned on data generated by GPT-4[2] and GPT-4o to create a cost-efficient and reliable implementation of claim extraction and verification.

Compared to FACTSCORE and SAFE, VERISCORE introduces three key improvements. First, VERISCORE is the first approach

[1] Code and data are available at https://github.com/Yixiao-Song/VeriScore.

[2] Henceforth, unless otherwise specified, GPT-4 refers to `gpt-4-0125-preview` and Claude 3 refers to `claude-3-opus-20240229`.

Findings of the Association for Computational Linguistics: EACL 2024, pages 9447–9474
November 12-16, 2024

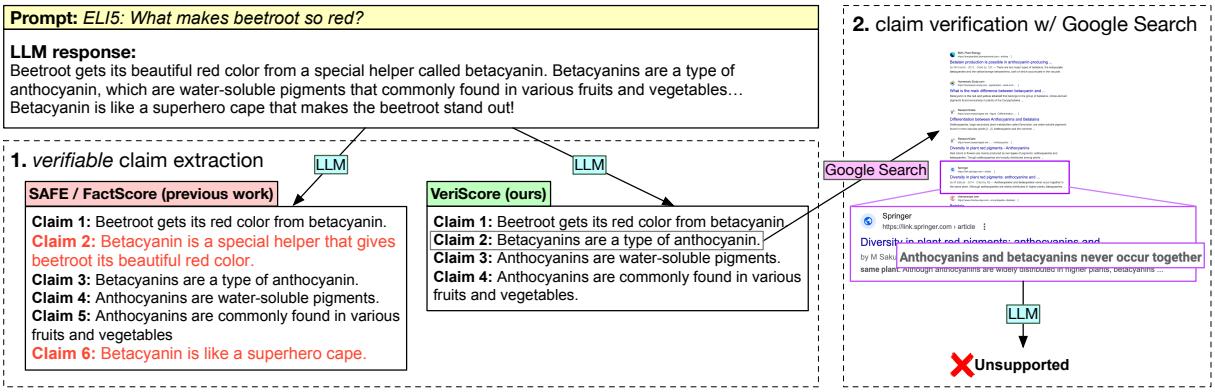

Figure 1: The pipeline of VERISCORE involving claim extraction and claim verification with Google Search. VERISCORE extracts *verifiable* claims. Each claim is used as a search query to retrieve evidence via Google Search, and an LLM then verifies the claim against the search results. We also show SAFE's extracted claims from the same text to highlight its propensity to extract unverifiable claims (Claim 2 and 6); see Section 2.1.1 for more discussion.

that considers inter-sentence context when extracting claims (Section 2.1), removing the need for expensive claim revision steps present in SAFE. Second, VERISCORE only extracts what we term *verifiable claims* (Section 2.1.2), unlike FACTSCORE and SAFE which decompose *everything*. In a human study (Section 3.1), our extraction method is preferred 93% of the time over SAFE's, even in biography generation. Third, unlike SAFE that uses an LLM to iteratively issue search queries for retrieving verification evidence, VERISCORE uses extracted claims per se as search queries (Section 2.2), which is shown to be sufficient for retrieving evidence (Appendix E.3).

To benchmark models with VERISCORE, we gather prompts from eight diverse domains that require long-form responses, ranging from the fact-dense biography generation task of FACTSCORE to the multi-task, open-domain ShareGPT for instruction-following. We evaluate sixteen closed and open-weight LMs with VERISCORE and find that GPT-4o generates the most factually-supported text when averaged across all datasets. Our analyses highlight that (1) multiple tasks (not just biography generation) are needed for comprehensive long-form factuality evaluation because an LLM's factuality varies depending on the task and domain; and (2) verifying complex, lengthy assertions (common in many long-form tasks such as LFQA) against Google Search results can fail due to challenges in retrieving relevant documents from such queries.

2 VERISCORE: an automatic factuality metric

This section details the VERISCORE pipeline, covering claim extraction, evidence retrieval, claim verification, and score calculation.

2.1 Claim extraction

Claim extraction facilitates factuality evaluation by decomposing sentences with potentially multiple independent facts (Min et al., 2023; Tang et al., 2024). We first examine the shortcomings of FACTSCORE and SAFE before developing a new method that focuses on extracting *verifiable* claims.

2.1.1 Issues with claim extraction in FACTSCORE and SAFE

FACTSCORE (Min et al., 2023) extracts *atomic facts*—"short statements that each contains one piece of information". However, their extraction method is optimized for biographies and is inapplicable to other domains. First, it does not resolve pronouns: for example, it extracts "His notable film credits include The Game." from an LLM-generated biography of Lanny Flaherty. Second, it extracts *everything* instead of just verifiable claims, an issue that is inherited by SAFE as in Figure 1.

SAFE (Wei et al., 2024) targets domains beyond biography and adapts FACTSCORE's extraction prompt for a three-step pipeline: (1) claim extraction, (2) claim revision to resolve vague references, and (3) a relevance check to decide if a claim is worth checking. While Wei et al. (2024) proclaim SAFE's superior performance, a closer inspection reveals four issues. First, besides adding a brief task description, SAFE uses FACTSCORE's prompt without changes. Second, the revision and relevance check adds significant processing time and cost.[3] Third, the relevance check unexpectedly re-

[3]Processing 100 claims without parallelization takes 35 minutes using GPT-4. SAFE's prompt templates of claim

moves verifiable claims. Lastly, SAFE's generalizability is questionable given that it is only evaluated on FACTSCORE's biography data. More details of these issues are in Appendix A.

2.1.2 VERISCORE's extraction approach

FACTSCORE and SAFE extract atomic claims with the implicit assumption that all claims are verifiable; unfortunately, this leads to the extraction of unverifiable claims (e.g., Claim 2 and 6 in Figure 1). Achieving atomicity is also hard as exemplified by Example (1). Hence, we aim to extract only *verifiable* claims. Inspired by frameworks of *events* and *states* in linguistics (Maienborn, 2003, 2019), we use the following description as a guideline:

> Verifiable claims describe a single *event* or *state*[4] with all necessary modifiers (e.g., spatial, temporal, or relative clauses) that help denote entities or events in the real world. They should be verifiable against reliable external sources (e.g., Wikipedia). They should exclude personal experiences, subjective opinions, hypotheticals, suggestions, advice, or instructions.

A detailed description can be found in Table 6 in Appendix C. Formally, our claim extraction process produces a set of verifiable claims $C = \{c_1, c_2, ..., c_n\}$.

To address the ambiguous reference issues in FACTSCORE and SAFE, we design few-shot claim extraction prompts given in Appendix C. The prompts extract verifiable claims from a model output sentence by sentence, with the sentences before and after the current sentence being the context. A sliding window is thus formed, formatted as `(context1: 0-3 sentences) <SOS>current sentence<EOS> (context2: 0-1 sentence)`.[5] It is used to guide LLMs to focus on the current sentence while using the context to ensure the claims are self-contained (e.g., pronouns are resolved).[6] Unverifiable content such as advice, fictional stories, or subjective opinions are ignored.

A human evaluation study detailed in Section 3.1 confirms the advantages of our extraction method. It effectively addresses the issue of unresolved referents and eliminates the need of claim revision and removal. Additionally, our method correctly avoids extracting claims from non-factual content.

2.2 Evidence retrieval

As in SAFE, we use Google Search via the Serper API[7] to retrieve evidence. For a claim $c \in C$, we use c as the search query and retrieve the top n search results $E_c = \{e_1, e_2, ..., e_n\}$ ($n \leq 10$). We use the title, snippet, and the link of each search result returned by Serper and combine the results into an evidence list as in Vu et al. (2023).

2.3 Claim verification

Claim verification judges whether a claim c is `supported` or `contradicted` by a corresponding evidence list E_c, or alternatively whether the verification is `inconclusive`. For a claim to be `supported`, everything in the claim need to be supported and no evidence contradicts any part of the claim (e.g., a modifier). For a claim to be contradicted, at least one part of the claim is contradicted by some evidence $e \in E_c$. `Inconclusive` cases can be classified into two types: (1) at least one part of the claim is neither supported nor contradicted with respect to E_c; or (2) at least one part of the claim is both supported and contradicted by different evidences $e \in E_c$. A formal definition of the three scenarios is given in Table 5 of Appendix B. The automatic verifier via prompting is detailed in Section 3.4.

2.4 Score calculation

An ideal generation should have both high factual precision (i.e., low hallucination) and high factual recall (i.e., not be too short or incomplete). We adopt the $F_1@K$ metric from SAFE, which considers both factual precision and recall. K is the minimum number of factual claims a model response must contain to achieve perfect recall. For each tested domain, we set K as the median number of extracted facts among all model responses.

Let $\mathcal{M}$ be a language model to be evaluated and $\mathcal{X}$ be a set of prompts of a given domain. Let $r = \mathcal{M}_x$ be a response of $\mathcal{M}$ to $x \in X$, and let the transitive predicate $\texttt{support}(a, b)$ take a value of either 1 or 0. $S(r) = \sum_{c \in C} \texttt{support}(c, E_c)$ is

revision and relevance check alone, without filling in content, cost about $1.7 per 100 claims (estimated using `https://platform.openai.com/tokenizer`).

[4] Event: change of state, for example, "Jensen Huang founded NVIDIA in 1993 in California, U.S." State: for example, "Westborough is a town in Worcester, MA."

[5] The sentence number in `context1` and 2 depends on how many sentences proceed and follow the current sentence, with the maximum sentence number being 3 and 1 respectively.

[6] For QA tasks, we always prepend the question to the sliding window. For non-QA tasks, we prepend the first sentence of a paragraph to the sliding window if the paragraph is longer than five sentences to mitigate lack-of-context issues.

[7] `https://serper.dev/`

the number of supported claims of r. $P(r) = \frac{S(r)}{|C|}$ and $R(r) = min(\frac{S(r)}{K}, 1)$ are precision and recall. VERISCORE of $\mathcal{M}$ is the average of the responses' $F1@K$ within each domain, defined as:

$$F_1@K(r) = \begin{cases} \frac{2P(r)R_K(r)}{P(r)+R_K(r)} & \text{if } S(r) > 0 \\ 0 & \text{if } S(r) = 0 \end{cases}$$

$$\text{VERISCORE} = \frac{1}{|X|} \sum_{x \in X} F_1@K(\mathcal{M}_x)$$

3 Validation of VERISCORE's claim extraction and verification

SAFE and FACTSCORE use closed LLMs for claim extraction and verification. Following them, as shown in Figure 2, we first develop VERISCORE's extraction and verification by prompting closed LLMs, whose effectiveness is verified by human evaluations. To mitigate the high cost of closed LLMs, we develop a free alternative in Section 4.1 by fine-tuning open-weight LLMs on data from GPT-4 and GPT-4o.

3.1 Human evaluation on claim extraction

To verify our extraction method's efficacy, we conducted a pairwise comparison of claims extracted by VERISCORE and SAFE with three human raters. They were asked to choose the claim lists with least unverifiable content. Half of the claims were extracted by GPT-4 and the other half by Claude 3. Our method outperforms SAFE regardless of the model used. Because GPT-4 was preferred more often than Claude 3 with our prompts, we use GPT-4 as the claim extractor in Section 4.

Setup: We extracted claims from 15 randomly sampled long-form texts from the eight datasets in Table 1 (Usage = HE), using both SAFE's method and ours. The datasets were selected to have a range of verifiable factual content. For time and cost efficiency, we only used SAFE's fact extraction and revision steps (see Appendix A and Footnote 3). To ensure the comparison was independent of the model used, we used GPT-4, paired with SAFE's and our methods, to extract claims from half of the data points and Claude 3 for the other half.[8] For each text, the annotators were asked to choose the claim list that better covered the verifiable content in the text—the one with the most verifiable and the least/no unverifiable content. The annotators are also asked to indicate whether it was hard to choose and briefly justify their choice. Data preparation and annotation details are in Appendix D. In total, we collected 360 data points.

[8] All claim extractions are done on April 3rd and 4th, 2024.

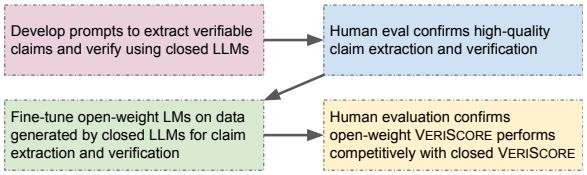

Figure 2: The development of the open-weight VERISCORE. Details in Section 3 and Section 4.

Results: The three annotators fully agreed on 99 out of 120 annotated data points, resulting in a Fleiss $\kappa = 0.7662$ (substantial agreement, Landis and Koch, 1977). Of the 360 annotated items, claims extracted by SAFE were preferred only 26 times, with 19 of those preferences being marginal. The annotator preferences across the data domains are detailed in Figure 3. Notably, our approach is significantly favored even on biography generation.

Annotator comments on SAFE's claims: The annotators identified three major issues with SAFE's extraction pipeline. First, it indiscriminately extracts *everything*, such as subjective content (2a) and personal experience (2b).

(2) a. I am 1000% better.
b. My grandpa assembled a TV.

Second, SAFE overly decomposes texts, causing meaning overlaps between claims as in (3) which can disproportionately affect the final score.

(3) Longwood House is a place.
Longwood House is a Napoleonic Museum.
Longwood House is one of the best Napoleonic Museums.
Longwood House is one of the best Napoleonic Museums in the world.

Third, SAFE often extracts trivial (4a) or vague claims (4b) that do not need to or cannot be verified.

(4) a. 3.2 is a number.
b. *All My Sons* has key themes.

3.2 VERISCORE's claim extractor only extracts verifiable claims

To further support that our claim extraction method with GPT-4 extracts only verifiable claims, we applied it to LLM-generated responses to 200 prompts from each domain in Table 1 (Usage = Dev) and calculated the average ratio of verifiable

Name	Description	Usage	VerRatio	Source
Scruples	Community judgements on real-life anecdotes from r/AmItheAsshole from November 2018 to April 2019	HE	—	Lourie et al. (2021)
CommonCrawl	A corpus of raw web page data, metadata extracts, and text extracts	HE	—	CommonCrawl
wikitext-103	Wikipedia articles	HE	—	Merity et al. (2016)
WritingPrompts [WP]	Story premises and stories written by online users on r/WritingPrompts	HE/Dev	0.03	Fan et al. (2018)
ShareGPT [S.GPT]	User-shared conversations (prompts and responses) with ChatGPT on `ShareGPT.com`	HE/Dev	0.92	Chiang et al. (2023)
ELI5	Questions and layperson-friendly answers posted on r/explainlikeimfive	HE/Dev	1.71	Scraped by Xu et al. (2023)
AskHistorians [AskH]	Questions and answers on history topics posted on r/AskHistorians	HE/Dev	1.90	Same as above
Biography[Bio]	Biography text generated by PerplexityAI, InstructGPT, and ChatGPT	HE/Dev	2.08	Min et al. (2023)
LongFact [LF]	A prompt set of 38 topics generated by GPT-4; each topic has prompts about object & concept; we randomly sampled 5 object and 5 concept prompts from 10 topics	Dev	2.24	Wei et al. (2024)
FreshQA	A dynamic QA benchmark whose answers can change w.r.t. updated world knowledge; we randomly sampled 200 questions with a true premise from the never- and fast-changing categories in the test set of the April 1st version	Dev	1.00	Vu et al. (2023)
FreshBooks [FBs]	We collected 20 non-fictional books that are published in 2023 and 2024. Ten paragraphs are taken from each book. LLMs are prompted to generate a continuation given a paragraph	Dev	2.31	Current paper; Details in Table 13

Table 1: Datasets used in the human evaluation of claim extraction (Usage = HE) in Section 3.1 and in the VERISCORE development (Usage = Dev) in Section 3 and 4.1. Short name of each dataset is in square brackets. VerRatio column presents the ratio of verifiable claims to sentences per domain of GPT-4 generated responses.

claims to sentences per response, shown in the VerRatio column of Table 1. We observe significant and intuitive differences in this ratio across domains: fact-seeking domains (e.g., FreshBooks) have a higher density of verifiable claims, while WritingPrompts' creative story outputs contain almost no verifiable claims with a ratio of 0.03, despite containing the longest responses.[9] This level of variation shows that our method effectively discriminates verifiable and unverifiable content.

3.3 Human evaluation on claim verification

We conducted a human study where three annotators verified claims given search results. The study has three purposes: (1) to understand the feasibility of the task, (2) to see the distribution of the labels in Table 5, and (3) to later judge automatic verifiers by their agreement with human annotations.

Setup: We sampled 320 GPT-4-extracted claims from model responses to the prompts from the datasets (Usage = Dev) in Table 1. Evidence was retrieved as described in Section 2.2. The <claim, evidence list> pairs were split into subsets of 50 for agreement analysis and three subsets of 90, with each annotator doing one. The annotators evaluated each claim on two levels: (1) evidence level: assess if each search result supports, contradicts, or is inconclusive for the claim; (2) claim level: whether the claim is `supported`, `contradicted`, or `inconclusive` given all the evidence.

Human agreement: The agreement result shows that the verification task is well-defined and feasible. Of the 50 triple-annotated items, 82% had complete agreement among the annotators, and 14% had two annotators in consensus. The Fleiss κ is 0.7316 (substantial agreement). An analysis of annotator disagreements is provided in Appendix E.

Reasons for being inconclusive: Among the 41 fully agreed items, 15 are `inconclusive`. There are two reasons. First, a claim is too general to be verified (e.g., "A systematic review on sex differences in the reinforcing effects of nicotine was published in Nicotine & Tobacco Research in 2019." without specifying *which* systematic review it was.) Second, there is no direct mentioning of a part of the claim or no evidence verifies the connection between the parts of a claim (see Table 8). Overall, no triple-annotated item is marked as `inconclusive` for the reason that there are both supporting and contradicting search results.

Only over half of the claims are `supported`. We analyzed the distribution of the claim level labels of all annotated items. For the triple-annotated data, we use the majority vote, if there is one, as the final label. Otherwise, the label is `inconclusive`. Results in Table 2 show that only 55% of the claims are `supported`. As discussed later in Section 4.3, the low supported rate showcases that open-domain claim verification is beyond identifying exact or related terms but requires extensive reasoning to verify the connection between parts of a claim.

Top search results are more informative. We consider a search result informative if it is marked as supporting or contradicting a claim. We analyzed the frequency with which search results were deemed informative. The first five search results show higher utility, with over 30% being useful—

[9] It has on average 34 sentences per response.

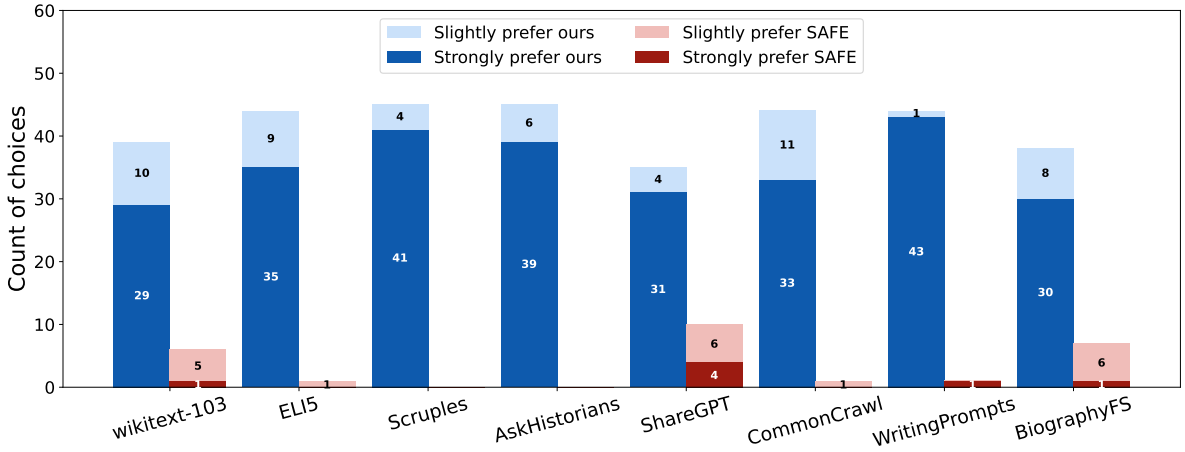

Figure 3: Results of the pair-wise performance comparison between our one-step extraction prompts and SAFE's two-step extraction pipeline. The claims extracted by our prompts are overwhelmingly preferred by the annotators across all eight domains. The dark area in each bar indicates that the annotators strongly preferred one choice over the other. The light area represents slight preference. The numbers are aggregated over three annotators.

Label	Count	%
Claim supported	176	55%
Claim contradicted	9	2.8%
Inconclusive (a)	128	40%
Inconclusive (b)	7	2.2%

Table 2: The distribution of the four labels (Table 5) that can happen in the claim verification step.

the highest being the first search result at 35.6%. For search results six to nine, their usefulness percentages range from 27.0% to 29.2%. The utility of the last search result drops to only 13.3%.

3.4 Automatic verifier via prompting

To find the best performing LLM on the claim verification task, we tested Mixtral-8×22-Instruct-v0.1, Claude 3, GPT-4, and GPT-4o on the human annotated verification data using the binary classification prompt in Table 12 in Appendix G.[10] We calculated the precision, recall, and F_1 on all items as well as separately on the `supported` and `unsupported` items. Results in Table 11 show that GPT-4o aligns the best with the human performance. Hence, we use GPT-4o data for fine-tuning an open claim verifier.

[10]We also experimented with a ternary classification prompt but the LMs' performance was worse. See Appendix G.

4 Using VERISCORE to benchmark LM factuality

In this section, we use VERISCORE to benchmark 16 LMs on 6 long-form fact-seeking domains.[11] We first introduce our fine-tuned claim extraction and verification models that are used for our large-scale study. The results of our study highlights the gap between closed and open-weight LMs, where GPT-4o achieving significantly higher VERISCORE than any open LM. We also note tasks whose VERISCORE does not correlate well, and conclude with qualitative analysis revealing limitations of VERISCORE's verification step.

4.1 An open-weight VERISCORE pipeline

To facilitate affordable factuality evaluation, we fine-tuned open LMs for a deterministic and cost-efficient VERISCORE pipeline. We use the few-shot prompting pipeline developed in Section 3 to generate 13403 training data.[12] We experimented with Llama3-8B-Instruct and Mistral-7B-Instruct-v0.2 (henceforth Llama3 and Mistral) as the base models (see Appendix G for details of fine-tuning). The benchmark experiments are then performed with the best performing fine-tuned models. The fine-tuned VERISCORE saves considerable money, making the evaluation process more accessible.[13]

[11]The results on FreshQA and WritingPrompts are reported in Appendix J because the former mostly requires short answers and the latter is not fact-seeking.

[12]Each data point consists of a claim, search results, and a label.

[13]Evaluating 400 GPT-4o generations in the domains in Table 16 using our prompting method cost $1,038 USD.

For claim extraction, the fine-tuned Mistral on GPT-4 data achieves the most competitive performance. The model sees the whole prompt and model response and extract claims sentence by sentence. In a quality comparison of 300 pairs of Mistral and GPT-4 extracted claims in Appendix F, the exact match rate is 43.7% and RougeL is 0.801. For claim verification, a verifier should be equally adept at identifying valid claims as well as recognizing unsupported claims. The fine-tuned Llama3 on GPT-4o data performs the best on the human annotated data in Section 3.3, achieving $F1 = 0.841$ (see Table 11). Details of the fine-tuning process and quality analysis are in Appendix F and G.

4.2 Data domains and studied LMs

VERISCORE aims to operate on a wide range of domains. We prompt 16 LMs using prompts from the datasets in Table 1 and benchmark their factuality. The datasets include prompts that require various degree of factual content, from highly fact-dense (e.g., AskHistorians and ELI5) to moderately factual (e.g., ShareGPT). We also collect a dataset FreshBooks that consists of 10 paragraphs from each of 20 non-fictional books in Table 13 published between 2023 and 2024. Models are required to generate a continuation of the paragraphs.

The three largest model families are the GPT, Claude 3, and Mistral/Mixtral models. We also evaluate LMs of various sizes—Qwen1.5-1.8B-Chat, Gemma-2B-it, OLMo-7B-Instruct, Vicuna-7B-v1.5, and DBRX Instruct (132B). Details of the models are in Table 14. To generate model responses for evaluation, the default model hyperparameters were used. The maximum token length was set to 1024. We used 50 prompts per domain.[14]

4.3 VERISCORE results

The factuality performance of the 16 LMs on VERISCORE is reported in Table 3. We tune K in $F_1@K$ for each domain, which is the median number of verifiable claims extracted from each response in each domain from all models. From the results, we observe the following:

Closed models are more factual. Overall, the GPT models performs better than the Claude 3 models.[15] DBRX-Instruct and the Mixtral models performs competitively to some versions of the GPT and Claude 3 models. The smaller models fall behind on VERISCORE, with Gemma-2b-it performing the worse across all domains. The overall trend underscores the correlation of model size and VERISCORE in long-form fact-seeking outputs.

Dataset K	LF (32)	Bio (26)	ELI5 (21)	AskH (21)	FBs (24)	S.GPT (11)	Avg.
Gemma-2B-it	60.7	4.6	28.8	17.8	25.1	27.6	27.4
Mist-7B-Inst-v0.1	57.6	20.3	42.2	36.5	39.8	41.2	39.6
Vicuna-7B-v1.5	63.4	23.0	51.3	39.7	39.0	43.6	43.3
Qwen1.5-1.8B-Chat	70.3	14.1	57.9	45.2	52.6	49.2	48.2
OLMo-7B-Inst	73.4	19.4	58.8	43.2	53.7	49.4	49.6
Mist-7B-Inst-v0.2	72.0	30.0	58.8	41.2	52.4	**54.8**	51.5
Mixt-8x7B-Inst-v0.1	77.3	42.5	61.9	50.7	57.4	51.5	56.9
DBRX-Inst	75.9	46.5	61.9	49.5	60.2	48.9	57.2
Mixt-8x22B-Inst-v0.1	78.0	47.6	64.9	51.1	58.0	51.4	58.5
GPT3.5-turbo-1106	64.7	38.1	42.8	40.8	32.5	42.1	43.5
Claude-3-Haiku	79.4	37.1	58.7	43.5	49.5	44.7	52.2
Claude-3-Sonnet	80.7	37.6	56.2	40.7	59.3	51.7	54.4
GPT3.5-turbo-0613	77.6	45.9	62.9	51.8	49.0	48.6	56.0
Claude-3-Opus	83.6	52.7	63.4	49.8	66.4	51.6	61.2
GPT4-0125-preview	85.9	56.4	70.7	56.6	69.7	53.5	65.5
GPT-4o	**86.7**	**56.7**	**71.7**	**61.4**	**70.9**	51.5	**66.5**
Kendall's τ w/ Avg.	0.78	0.83	0.82	0.73	0.73	0.56	1.00

Table 3: VERISCORE on 50 responses per LM per dataset (FreshQA and WritingPrompts in Table 16). K is in brackets. Dataset full names are in Table 1. Precision and recall are in Table 16. All correlations are statistically significant. GPT-4o is the best closed LLM and Mixt-8x22B-Inst-v0.1 the best open LLM. Kendall's τ measures correlation between models' performance on individual datasets and their average scores, indicating how well each domain aligns with the overall trends.

Multiple generation tasks are needed for a comprehensive factuality evaluation. The Kendall's τ correlations between LMs' performance in domains in Figure 4 indicate that LMs' VERISCORE on two fact-seeking domains (e.g., ELI5 and Biography) do not necessarily correlate well. This suggests that LMs exhibit varying strengths in different domains, highlighting the need for diverse tasks to comprehensively assess LMs' factuality.

$F1@K$ favors longer outputs. $F_1@K$ (Wei et al., 2024) considers both factual precision and recall, which improves on measuring factual precision alone (Min et al., 2023). A model must generate at least K supported claims per response to achieve perfect recall. However, for domains that do not require long generations, models that generate short to-the-point outputs will be penalized if other models generate lengthy outputs with auxhurts the recall. On average, GPT-3.5-1106 generates 8.56 sentences per response and GPT-3.5-0613 generates 15.95.

[14]The instruction "Generate a continuation of the following text. The continuation should be objective and factual" is prepended to the FreshBooks paragraphs.

[15]GPT-3.5-turbo-1106 is an exception because the model generates shorter responses than GPT-3.5-turbo-0613, which

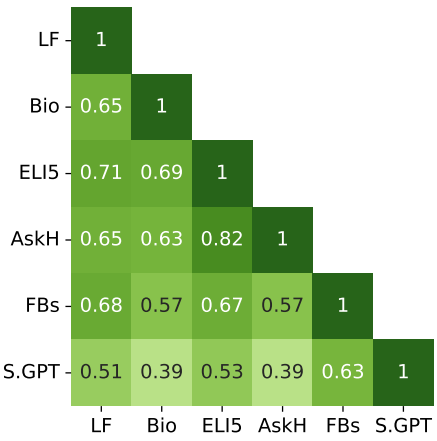

Figure 4: Kendall's τ correlations of LMs' performance between domains. All correlations are statistically significant. Models' VERISCORE on dissimilar tasks do not necessarily have high correlation, highlighting the need of using different tasks to assess LMs' factuality.

iliary information (e.g., FreshQA in Appendix J). It is debatable whether longer responses should always be preferred. They provide more details but, as Min et al. (2023) shows, later facts in long responses tend to be less accurate.

4.4 Qualitative analysis

This subsection examines VERISCORE's performance, highlighting the limitations of decompositional factuality evaluation for generations that are not entity-centric and not formulaic. Two issues are identified: (1) not all claims can be short, and long claims are harder to verify, and (2) search results may be insufficient as expertise or extensive logical reasoning is often needed for verification.

4.4.1 Claim complexity increases outside of entity-centric tasks

Shorter, self-contained claims are desired because they help locate factual errors and are easy to be verified as employed by FACTSCORE for biography. However, claims extracted from other fact-seeking generations are often long.[16] While some long claims could be split at conjunctions like *and* or *or*, this does not significantly shorten claims with inherently long core content, as seen in (5).

(5) Travelers *and* crusaders during the medieval period depended on established infrastructure to secure clean *and* consistent sources of water.

Occasionally, shorter claims can be extracted from a longer one, as the bracketed content in (6). However, verifying the shorter claims does not mean the longer one is verified because of *solidified*.

(6) [Chuck Norris's victory in the 1968 World Full-Contact Karate Championships] *solidified* [his reputation as one of the best martial artists in the world].

For these reasons, long and complex claims are likely to be marked as `inconclusive`.

4.4.2 Google Search may be insufficient for complex claims

To understand what types of claims are supported and unsupported by Google Search snippets, we examine 80 claims from GPT-4o generated ELI5 and FreshBooks responses, along with their search and verification results. Half of these claims were verified as supported and the other half were not.[17]

The supported claims resemble encyclopedic writing. The content is supported by search results via semantic or string match, as in (7).

(7) Indigenous women in Australia were not fully enfranchised until much later.[18]

The unsupported claims do not have direct contradicting evidence. They are unsupported because there is no direct mention of (parts of) the claims or the connection between the parts of the claims. Example (8) is judged as `unsupported` because there is no mention of the Meiji era and stuffed tigers occurring together in the search results.[19] Snippets do not offer enough background for such reasoning. Expertise or more sophisticated search is needed to verify/falsify such claims.

(8) Japanese people encountered tigers in the form of stuffed animals before the Meiji era.

Some unsupported claims require extensive supporting evidence. This happens the most often in FreshBooks when a claim encapsulates aspects like someone's achievements or historical movements, as in (9-10). Such content might not be directly

[16] To confirm this, we randomly sampled 200 claims from Min et al. (2023)'s model extracted claims and 200 claims extracted by GPT-4 from GPT-4 generated ELI5 responses. On average, FACTSCORE's claims have 7 words, with the longest one having 18 words. In contrast, the ELI5 claims on average have 12 words, with the longest one having 25 words.

[17] More `unsupported` claim examples are in Appendix I.

[18] Search snippet: In Australia, Indigenous women were not enfranchised until 1962, six decades after non-Indigenous women were able to vote. (link)

[19] After extensive search, we did not find any supporting or contradicting evidence to the claim.

mentioned in search results but need to be inferred from a large body of documents.

(9) Marshall's leadership and strategic acumen ensured the maneuver was carried out flawlessly during a field maneuver in the Philippines.

(10) Germany is maintaining its competitive edge in a rapidly changing global landscape.

In sum, with the current system, it is hard to decide whether an `unsupported` claim is hallucinated because it is beyond what reasoning over search snippets can achieve. This indicates the need to move beyond semantic or string matching for verification as they fail to uncover possible hallucination.

5 Related work

Our work builds on prior research in claim verification and long-form factuality evaluation. Users rely on the accuracy of LLM-generated content, yet LLMs often produce unreliable information (Maynez et al., 2020; Xu et al., 2023; Huang et al., 2023; Rawte et al., 2023). Research has thus focused on enhancing factual precision (Lin et al., 2024) and identifying inaccuracies.

Factual error detection: Prior research targets error detection in individual sentences (Mihaylova et al., 2019; Wadden et al., 2020; Shaar et al., 2022). FEVER (Thorne et al., 2018) features synthesized incorrect sentences from Wikipedia. FEVEROUS (Aly et al., 2021) and AVERITEC (Schlichtkrull et al., 2023) build on FEVER but remained limited to sentence-level facts. At the paragraph level, Li et al. (2023) test LLMs' detection of synthesized factual errors but do not locate the errors.

Long-form factuality evaluation: Detecting factual errors in a long-form text at once is hard (Li et al., 2023). Decomposing a piece of long-form text into shorter sentences or search queries for factuality evaluation is commonly implemented in previous works (Kamoi et al., 2023; Gao et al., 2023; Wang et al., 2023; Min et al., 2023; Chern et al., 2023; Wanner et al., 2024; Wei et al., 2024; Guan et al., 2024; Chen et al., 2024). The decomposition helps locate factual errors and offers a fine-grained estimate of models' factuality (Min et al., 2023). It also help with identifying facts that do not pertain to the same entities (Chiang and Lee, 2024). Factuality evaluation often requires world knowledge, which can be achieved by employing retrieval (Ram et al., 2023; Vu et al., 2023). It is commonly used in factual error detection (Min et al., 2023; Wei et al., 2024; Thorne et al., 2018) and helps evaluation by providing up-to-date knowledge. The overall evaluation pipeline helps generate post hoc citations and iteratively improve model generations' factuality, which improves models' trustworthiness model (Huang and Chang, 2024; Ye et al., 2024).

6 Conclusion and future work

We propose VERISCORE, a factuality metric that focuses exclusively on *verifiable* claims. Human evaluations validate that VERISCORE is more effective than existing metrics for diverse long-form generation tasks that contain both verifiable and unverifiable content. We open-source both a closed- and open-weight implementation of VERISCORE, with the latter's performance approaching that of the former. Finally, we notice that complex claims (e.g., not entity-centric or formulaic) are challenging to verify against search results. We hope that future work will improve on this aspect to develop more robust factuality metrics.

Limitations

We acknowledge further limitations of the current work and the decompositional approach below.

First, formally defining verifiable claims poses a significant challenge. Although our definition advances beyond the concept of *atomic facts* (Min et al., 2023), it remains a working definition rather than a formal one. For instance, consider the sentence (6): it is problematic to determine whether it describes a single state of "solidifying" or encompasses one event, "Chuck Norris's victory in ... Championships," along with two states, "solidifying" and "as one of the best ... in the world,". We hope future studies can improve on this.

Second, the decomposition method is slow. With one RTX8000 GPU, it takes about 4 hours to extract claims from 400 GPT-4o responses without parallelization. The reason is that VERISCORE uses a sliding window to scan through a model response. In our experiments, each response on average has 40 sentences (20 sentences on average if excluding WritingPrompts responses). For verification, it takes about two hours to verify 10k claims. Future work can aim for a claim extractor that works without a sliding window to speed up the claim extraction.

Third, for model responses that are extremely infactual (e.g., WritingPrompts in Appendix J), our claim extractor might still extract a small amount

of unverifiable claims. However, we contend that the percentage of factual content in creative writing in response to fictional premises is less concerned than in the fact-seeking domains. Hence, we do not consider this as a major concern of VERISCORE.

Fourth, VERISCORE aims to extract verifiable claims and verify them, but does not judge if the extracted claims pertains to the topic of generation queries. This introduces an intriguing dimension to factuality evaluation, further explored by Chiang and Lee (2024) in the context of biographies of ambiguous entities.

Fifth, in the current work, we did not search exhaustively for the best hyperparameters for fine-tuning the open-source claim extractor and verifier. It is possible that, after searching, a better performance can be achieved. However, it is resource-intensive and time-consuming.

Ethics Statement

Our project aimed to minimize the computational cost by using LoRA (Hu et al., 2022) for efficient model fine-tuning. For the annotation work, an IRB review was exempted. By signing a data conset, each annotators agreed on the annotated data being used for scientific research and published. No personally identifiable information was collected. We paid annotators $18 per hour. Additional bonus were paid for reasonable extra time spent.

Acknowledgement

We extend our special gratitude to Kalpesh Krishna, who extensively discussed the project details with us and offered invaluable insights. We extend gratitude to the Upwork annotators for their hard work, and to members from the UMass NLP lab for their feedback. This project was partially supported by awards IIS-2202506, IIS-2046248, and IIS-2312949 from the National Science Foundation (NSF).

References

Rami Aly, Zhijiang Guo, Michael Sejr Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. 2021. The fact extraction and VERification over unstructured and structured information (FEVEROUS) shared task. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 1–13, Dominican Republic. Association for Computational Linguistics.

Anthropic. 2023. Model Card: Claude 3. Technical report, Anthropic. Accessed: 2024-03-25.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Jifan Chen, Grace Kim, Aniruddh Sriram, Greg Durrett, and Eunsol Choi. 2024. Complex claim verification with evidence retrieved in the wild. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3569–3587, Mexico City, Mexico. Association for Computational Linguistics.

I-Chun Chern, Steffi Chern, Shiqi Chen, Weizhe Yuan, Kehua Feng, Chunting Zhou, Junxian He, Graham Neubig, and Pengfei Liu. 2023. Factool: Factuality detection in generative ai – a tool augmented framework for multi-task and multi-domain scenarios. *Preprint*, arXiv:2307.13528.

Cheng-Han Chiang and Hung-yi Lee. 2024. Merging facts, crafting fallacies: Evaluating the contradictory nature of aggregated factual claims in long-form generations. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2734–2751, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Databricks. 2024. Introducing DBRX: A New State-of-the-Art Open LLM.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. RARR: Researching and revising what language models say, using language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 16477–16508, Toronto, Canada. Association for Computational Linguistics.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Jian Guan, Jesse Dodge, David Wadden, Minlie Huang, and Hao Peng. 2024. Language models hallucinate, but may excel at fact verification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1090–1111, Mexico City, Mexico. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Jie Huang and Kevin Chang. 2024. Citation: A key to building responsible and accountable large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 464–473, Mexico City, Mexico. Association for Computational Linguistics.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L'elio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *ArXiv*, abs/2401.04088.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore. Association for Computational Linguistics.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Preprint*, arXiv:2405.01525.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. *Preprint*, arXiv:2008.09094.

Claudia Maienborn. 2003. *Event-internal modifiers: Semantic underspecification and conceptual interpretation*, pages 475–510. De Gruyter Mouton, Berlin, Boston.

Claudia Maienborn. 2019. Events and States. In *The Oxford Handbook of Event Structure*. Oxford University Press.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2016. Pointer sentinel mixture models. *Preprint*, arXiv:1609.07843.

Tsvetomila Mihaylova, Georgi Karadzhov, Pepa Atanasova, Ramy Baly, Mitra Mohtarami, and Preslav Nakov. 2019. SemEval-2019 task 8: Fact checking in community question answering forums. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 860–869, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100, Singapore. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report. *CoRR*, abs/2303.08774.

OpenAI. 2024. Model release blog: GPT-4o. Technical report, OpenAI. Accessed: 2024-05-23.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. 2023. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331.

Vipula Rawte, Amit Sheth, and Amitava Das. 2023. A survey of hallucination in large foundation models. *Preprint*, arXiv:2309.05922.

Michael Sejr Schlichtkrull, Zhijiang Guo, and Andreas Vlachos. 2023. AVeritec: A dataset for real-world claim verification with evidence from the web. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Shaden Shaar, Firoj Alam, Giovanni Da San Martino, and Preslav Nakov. 2022. The role of context in detecting previously fact-checked claims. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1619–1631, Seattle, United States. Association for Computational Linguistics.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. *Preprint*, arXiv:2404.10774.

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

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Tu Vu, Mohit Iyyer, Xuezhi Wang, Noah Constant, Jerry Wei, Jason Wei, Chris Tar, Yun-Hsuan Sung, Denny Zhou, Quoc Le, and Thang Luong. 2023. Freshllms: Refreshing large language models with search engine augmentation. *Preprint*, arXiv:2310.03214.

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

Yuxia Wang, Revanth Gangi Reddy, Zain Muhammad Mujahid, Arnav Arora, Aleksandr Rubashevskii, Jiahui Geng, Osama Mohammed Afzal, Liangming Pan, Nadav Borenstein, Aditya Pillai, et al. 2023. Factcheck-GPT: End-to-end fine-grained document-level fact-checking and correction of LLM output. *arXiv preprint arXiv:2311.09000*.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. *Preprint*, arXiv:2403.11903.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, et al. 2024. Long-form factuality in large language models. *arXiv preprint arXiv:2403.18802*.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3225–3245, Toronto, Canada. Association for Computational Linguistics.

Xi Ye, Ruoxi Sun, Sercan Arik, and Tomas Pfister. 2024. Effective large language model adaptation for improved grounding and citation generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6237–6251, Mexico City, Mexico. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

A Weaknesses in FACTSCORE and SAFE

In Section 2.1.1, we pointed out four major issues in SAFE's claim extraction pipeline. Details of the issues are provided in this appendix section.

First, for claim extraction, aside from prepending a brief task description to FACTSCORE's prompt, SAFE does not make other modifications. Consequently, the prompt only focuses on biography.

Second, SAFE's extraction pipleine is multi-step. Because FACTSCORE extracts claims by sentence without context, it cannot resolve references. This limitation is not an issue for FACTSCORE because each claim is verified against one predefined Wikipedia article. However, SAFE uses Google Search and thus must resolve all vague references. SAFE addresses this by deploying claim revision which prompts a language model once for each claim to revise vague references. Following that, a language model reviews each claim again to decide whether they are worth checking. The entire pipeline adds significant processing time and cost.

Third, the relevance check step negatively impacts evaluation. Wei et al. (2024) justifies this step with an example in their Figure 1—when asked about the Eiffel Tower, a model generates *The Nile River is in Egypt*. First of all, such behaviour is not observed in our experiments. Second, we applied SAFE's extraction pipeline to five texts and examined which claims were removed. It turns out that 11% of 211 claims were removed, of which 58% were actually relevant. The remaining 42% were either tautologies or not claims and should not have been extracted.[20] Table 4 provides an example of SAFE removing a relevant claim.

Fourth, there is no guarantee that SAFE works across domains. Despite being applied to 38 fact-seeking topics, SAFE's performance is only evaluated on FACTSCORE's biography data. Among the 38 topics, SAFE is solely applied to model outputs that responses to object-related prompts. Six topics mostly contain biography questions (i.e., *Who is*).[21] Some topics (e.g., sports) contain only *who*, *what*, and *can you tell me about* questions, making them fact-dense and entity-centric. A human study in Section 3.1 confirms that SAFE falls short in less entity-centric domains.

[20] For example, *Castello Maniace is Castello Maniace.* is a tautology; *As always, there is some disclaimer.* is not a verifiable claim.

B Formal definition of claim verification

Formally, our claim extraction process produces a set of verifiable claims $C = \{c_1, c_2, ..., c_n\}$ from a model response r where c consists of meaningful parts p such that $c = \{p_1, p_2, ..., p_n\}$. Each p does not have to be a full proposition. For example, if c = "Jensen Huang founded NVIDIA in 1993 in California, U.S.", p_1 = *Jensen Huang founded NVIDIA*, p_2 = *in 1993*, and p_3 = *in California, U.S.*

In Section 2.3, we described the definition of the four possible scenarios that can happen when verifying a claim with respect to evidence. We give a formal definition of such scenarios in Table 5.

C Claim extraction prompts

We developed two claim extraction prompts: one for question-answering (QA) type of input data, and the other for non-QA data. For evaluating model outputs, the QA prompt is generally applicable with the prompt being the question. The non-QA prompt is used for cases where neither a question nor a prompt is available.

What is common in the two prompts is a sliding window for claim extraction. Each window has the format `(context1 = 0-3 sentence) <SOS>Sentence to be focused on<EOS> (context2 = 0-1 sentence)`. The goal is to extract claims from the sentences marked by `SOS` and `EOS` while using the information in `context1` and `context2` to make the claims self-contained.

What is different in the two prompts is that for non-QA-type of inputs, we always prepend the first sentence of a paragraph to `context1` if the paragraph is longer than five sentences; for QA-type of inputs, we always prepend the question to `context1`. This is based on the observation that, when answering a question or when an answer

[21] The six topics are: celebrities, jurisprudence, mathematics, medicine, philosophy, and sociology.

SAFE's relevance check identifies relevant claim as irrelevant.
Question: At their peak, what did the insides of the most beautifully decorated castles look like? Today, castles seem to just be giant fortresses but I would like to know how they looked when they were fully furnished. How were they decorated? What treasures were stored there? Are there a few castles that were especially beautiful? **Human response**: It is quite a broad subject because castles varied quite a lot depending on location, time of construction and wealth of the constructor; u/valkine talked about Caenarfon Castle (link) specifically in another question (link) is a part of the inside of Castello Maniance in Siracusa, Italy. It was built from 1232 to 1239 during a large castle-construction effort by Emperor Frederick II. I do find it particularly beautiful but this doesn't really say much about what other castles looked like. **Extracted claim**: Castello Maniance in Siracusa, Italy was built from 1232 to 1239. **Authors' note**: Although the human answer does not answer all parts of the question, the content that is deemed as irrelevant by SAFE is actually pointing to a castle that is relevant to answering the question.

Table 4: An example illustrating SAFE's relevance assessment does not work as expected.

Scenario	Description
Claim supported	$\forall p \in c.[\exists e \in E_c.\texttt{support}(e,p) \wedge \neg\exists e \in E_c.\texttt{contradict}(e,p)]$
Claim contradicted	$\exists p \in c.[\neg\exists e \in E_c.\texttt{support}(e,p) \wedge \exists e \in E_c.\texttt{contradict}(e,p)]$
Inconclusive (a)	$\exists p \in c.[\neg\exists e \in E_c.\texttt{support}(e,p) \wedge \neg\exists e \in E_c.\texttt{contradict}(e,p)]$
Inconclusive (b)	$\exists p \in c.[\exists e \in E_c.\texttt{support}(e,p) \wedge \exists e \in E_c.\texttt{contradict}(e,p)]$

Table 5: Four scenarios that can happen in the claim verification step. $\texttt{support}(a,b)$ and $\texttt{contradict}(a,b)$ are two transitive predicates such that $\neg\texttt{support}(a,b) \neq \texttt{contradict}(a,b)$ and $\neg\texttt{contradict}(a,b) \neq \texttt{support}(a,b)$.

gets long, people might take the information in the question or previous sentences for granted and do not refer to an entity using its full name. Adding the question or the first sentence of a paragraph into `context1` can help a model better recover time, location, and person references in a claim.

The prompts are given in Table 6 and Table 7

D Human evaluation of claim extractions by our prompt and SAFE

To verify the effectiveness of our proposed claim extraction method, we conducted a human evaluation of pair-wise comparison between claims extracted by our prompts and SAFE's. We hired three experienced data annotators on Upwork[22].

To prepare the data for evaluation, we sampled 15 data points from each dataset in Table 1. We used the first four datasets as non-QA datasets and the others as QA datasets. Each data point was truncated to 300 white-space-separated words at the sentence boundary. Each annotator was asked to annotate the same set of 120 sampled data.

The evaluation was conducted on the open-source data labeling platform Label Studio (Tkachenko et al., 2020-2022). The task interface is given in Figure 5. Before the task begins, each annotator needs to read through the instructions of the task.[23] We estimated the annotation task to take approximately two hours to complete. Therefore, each annotator was compensated at a rate of $15 per hour.

Figure 3 depicts the human preference in each domain of data in Table 1. Among the 360 annotated data points, the claims extracted by SAFE are only preferred 26 times by the three annotators in total, among which 19 were chosen hesitatingly, as indicated by the light red color in Figure 3.

E Human study on claim verification

This appendix section provides supplementary details to Section 3.3.

E.1 Detailed examples of human verification

In this section, we provide detailed examples from our human study on the claim verification task in Section 3.3. Table 8 presents the examples of the annotation items whose claim was labeled as `inconclusive` by all annotators.

E.2 Reason of disagreement in human verification

There are 9 items in the human study in Section 3.3 on which the annotators did not reach a full agree-

[22] https://www.upwork.com/

[23] The instructions are on Google slides.

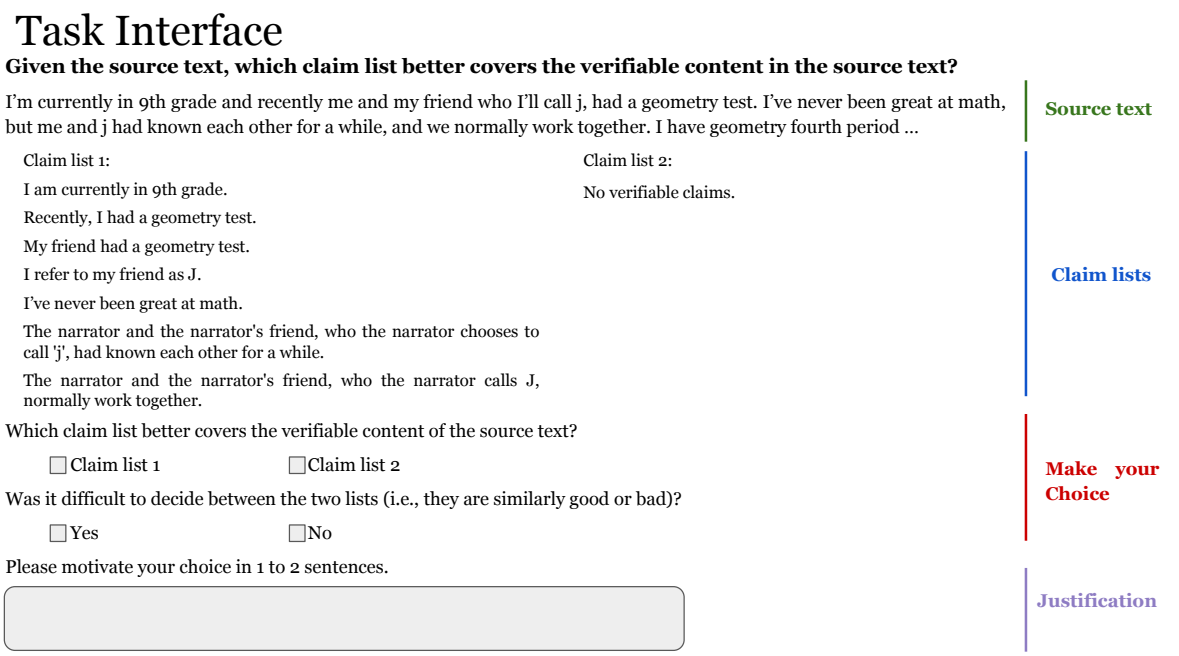

Figure 5: The interface design of the human evaluation described in Section 3.1 and Appendix D. The interface consists of four parts. Source text: the text from which claims are extracted. Claim lists: Two claim lists extracted by our prompt and SAFE respectively. The order of the two lists are randomized. Decisions: annotators indicate here which claim list is better and whether it is hard to choose between the two. Justification: annotators should briefly explain why they choose one list over the other.

ment. After inspecting these items, we conclude 4 sources of disagreement, listed in Table 9 with examples. First, an annotator made a mistake (e.g., misread a name). Second, there is disagreement in the interpretation of the claim and evidence (e.g., *can* in the claim vs. *could* in the evidence or an ambiguous referent). Third, the claim is complex and long, hence, is hard to verify. Fourth, the evidence indirectly supports the claim which means intermediate reasoning process is needed. one annotator have overlooked the connection between the claim and the search result.

E.3 Verifying/falsifying inconclusive claims is hard

In Section 3.3, we presented the distribution of verification labels in Table 2. As many as 42.2% of the annotated items are labeled as `inconclusive` by our annotators. In order to understand whether the inconclusive cases can be verified/falsified by checking the full web page of the returned search results, we randomly picked 15 inconclusive cases and manually verified them. Results show that two claims are not specified enough to be verified, for example, (11).

(11) A group of archaeologists unearthed a cache of Roman weaponry near the ancient ruins of the Colosseum on a sweltering summer afternoon.

Only one claim in (12) can be verified the full web page. The search result snippets do not mention that Angular is maintained by Google but this is confirmed by a notice at the end of the web page.

(12) Google Angular supports dependency injection.
angular.io/guide/architecture-services

For the remaning 12 cases, we used Google search to find more evidence but only the claim in (13a) was weakly contradicted by a popular science article, which states the content in (13b). If "As a battery discharges" is describing the status change of a battery from not discharging to discharging, the article implies that the chemical reactions become active but not slowing down. However, if the claim is understood as "as a battery continues to discharge", the article does not mention anything about chemical reactions slowing down gradually.

(13) a. As a battery discharges, the chemical reactions inside the battery slow down.

b. When a battery is discharged, chemical reactions within the battery cells facilitate the movement of electrons from the negative terminal (anode) to the positive terminal (cathode) [...]. (link)

F Details of fine-tuning open-source models for claim extraction

In this section, we specify the details of fine-tuning open-source models for the claim extraction task.

Data We used two types of data for GPT-4 to extract claims, which are used to fine-tuned open-source models. The first type of data is the existing open-source data in Table 1. We sampled 100 data points from Scruples and 200 from the other datasets. The reason of sampling less data points from Scruples is that the majority of them is invariably subjective and yields the "No verifiable claim." output. Hence, they are not very helpful in teaching an open-source model how to extract claims from factual texts.

The second data type is our newly generated model responses. For this, we sampled 63 prompts from Biography and 80 from the other datasets listed in Table 1 (Usage = Dev). To generate the responses, we prompted the first 12 LLMs in Table 14 with their default hyperparameters and the maximum token requirement was set to 1000. The LLMs are chosen in a way that we have both close- and open-source models as well as models in the same family with different sizes or versions.

After collecting and generating all the model long-form responses, we decompose them into claims with GPT-4 using the prompts in Appendix C. The temperature was set to 0.

We formed the fine-tuning data in the following way. As the input, we used the prompt (if there is one) and the response text with one sentence being marked with <SOS> and <EOS>. The output is the claims extracted from the marked sentence.

To get the final set of fine-tuning data, we randomly removed 80% of the data points whose marked sentence is shorter than 10 characters. These short marked sentences are usually the numbering of numbered lists. We also randomly dropped 50% data whose output is "No verifiable claim." In total, we got 99592 input-output pairs, among which 902 pairs have a short marked sentence, and 31819 pairs have "No verifiable claim." as the output. We took 95%, 4%, and 1% of the dataset as the training, validation, and test splits.

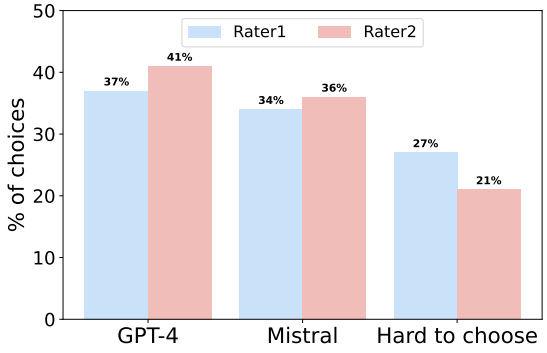

Figure 6: Results of comparing GPT-4 and fine-tuned Mistral-7B-Instruct-v0.2 on the claim extraction task. Numbers are in percentage. The Cohen's κ between the two annotators is 0.4320. Mistral achieves a competitive result compared to GPT-4.

Fine-tuning We chose Llama3-8B-Instruct and Mistral-7B-Instruct-v0.2 as the base models (henceforth Llama3 and Mistral). Both were fine-tuned via Unsloth[24] for two epochs using LoRA (Hu et al., 2022). Checkpoints were saved at each epoch and tested on the test set. We used string-based metrics for evaluation and selected Mistral fine-tuned for two epochs as the best checkpoint.[25] We further evaluated this checkpoint manually.

Manual quality comparison To understand how good the performance of the fine-tuned model is compared to GPT-4, the first two authors did a pairwise comparison between the outputs from the two models on 300 test data points. After removing the data whose GPT-4 and Mistral outputs match exactly, there were 169 data points left for manual evaluation. Each annotator annotated the same data and were asked to choose which output was better or whether it was hard to choose between the two.

Figure 6 shows the percentage of each annotator's choices. The two authors fully agreed in 106 data points, achieving a Cohen's $\kappa = 0.4320$ (moderate agreement, Landis and Koch, 1977). Given that the quality of both models' outputs is close with GPT-4 being slightly better, such a moderate agreement is expected.

Quality analysis GPT-4 and the fine-tuned Mistral perform similarly. In many cases, their outputs are identical with certain phrases being relocated in the sentences. However, there are cases where Mistral lost to GPT-4 because it misses small words and

[24] https://unsloth.ai/

[25] The string-based metrics and the scores are the following: exact match = 0.4317, Rouge1 = 0.8243, Rouge2 = 0.7576, RougeL = 0.8009, and CHRF++ = 74.6686.

makes the extracted claims less specific. Occasionally, Mistral puts multiple pieces of information into one claim while GPT-4 breaks the information down into multiple claims. Concrete examples can be seen in Table 10. Besides these issues, in certain cases, Mistral does not refrain itself to the marked sentence in an input but also extracts claims that come after the span.

Overall, our fine-tuned Mistral model performs comparable to GPT-4. As an open-source model, it is also cost-efficient.

G Details of fine-tuning open-source models for claim verification

In this section, we offer details of fine-tuning the open-source models for the claim verification task, in addition to Section 4.1.

In order to select the best model for generating fine-tune data, we tested Mixtral-8×22-Instruct-v0.1, Claude-3-Opus, GPT-4, and GPT-4o with the few-shot prompt in Table 12 on the 320 human annotated verification data in Section 3.3. We tried two types of verification task, one with `supported` and `unsupported` as the labels and one with `supported`, `contradicted`, and `inconclusive` as the labels. For a fair comparison between the model performance on the binary and ternary task, we convert the `contradicted` and `inconclusive` labels in the ternary task to `unsupported`. We then calculated the F1 scores on all 320 items as well as on `supported` and `unsupported` items separately. The results are in Table 11. Overall, GPT-4o with ternary labels has the most balanced performance on the `supported` and `unsupported` items. Hence, we generated data from GPT-4o for fine-tuning open-source models and converted the ternary labels to binary ones.

Data We sampled 10 prompts from the datasets listed in Table 1 (Usage = Dev) and prompted the LLMs in Table 14 with their default hyperparameters and the maximum token requirement was set to 1024. The model responses are decomposed into claims with GPT-4 using the prompts in Appendix C. The temperature was set to 0. Serper is then used to retrieve search results as described in Section 2.2. As the prompt, we use the binary prompt template in Table 12 without the few-shot examples. From the generated data, we randomly sampled 13403 data points, among which 9996 has the `supported` label and 3407 has the `unsupported` label. We split the dataset into 85%, 3%, and 12% as the training, validation, and test splits.

Fine-tuning Similar to the fine-tuned claim extractor, we fine-tuned Llama3-8B-Instruct and Mistral-7B-Instruct-v0.2 via Unsloth for 5 epochs. Because the number of `supported` and `unsupported` items in the training dataset are imbalanced, we triplicated the `unsupported` data points. Checkpoints were saved at each epoch and tested on both the test dataset and the 320 human annotated set. The Llama3-8B-Instruct model fine-tuned for one epoch achieves the most balanced performance on `supported` and `unsupported` data points on the test and human data. Hence, we use this checkpoint for further experiments.

H Details of data domains and studied LLMs

This section gives the details of the datasets and LLMs in Section 4.2. Table 1 lists the datasets that are used for developing, test, and benchmark models on VERISCORE. Table 13 further expands the name and details of the FreshBooks dataset. For developing VERISCORE, we used the model generations from the first 12 models in Table 14. For benchmarking models on VERISCORE, we used all 16 models in Table 14.

I `Unsupported` cases in VERISCORE outputs

This appendix section provides more examples of the `unsupported` claims in Table 15 in complementary to Section 4.4.2.

J VERISCORE on WritingPrompts and FreshQA

In Section 4, we presented VERISCORE of 16 models on 6 domains of long-form model generation. In this section, we focus solely on model responses in the FreshQA and WritingPrompts datasets. As shown in Table 16, both domains yield very few verifiable facts. The median number for verifiable claims (K) in FreshQA is four because the questions in general do not require long-form answers. WritingPrompts requires long-form generations but they are conditioned on fictional premises. Hence, the generations contain very few verifiable claims, resulting in $K = 1$.

For the generations in WritingPromtps, the results in Table 16 show that the VERISCORE of

the models is very low, matching the expectation that there are few supported verifiable claims in a creative writing task. We examine the extracted claims from two models—Gemma-2b-it and GPT-4o. It turns out that the majority of the claims (82.61% for Gemma and 71.15% for GPT-4o) has the potential to be verified/falsified. The results of the models on WritingPrompts prove that the VERISCORE pipeline works effectively on fictional content although it occasionally extracts unverifiable claims.

For the generations in FreshQA, the results show that the Claude 3 models perform the best. However, upon careful examination of the outputs of Claude 3 Haiku and GPT-4o, we notice that GPT-4o has a higher percentage of supported claim among all the extracted claims (74.70%) compared to Claude 3 Haiku (72.71%). The fact that GPT-4o generates shorter outputs than Claude 3 Haiku contributes to the lower final VERISCORE of GPT-4o. On average, Claude 3 Haiku has 5.5 claims per response, higher than $K = 4$, but GPT-4o has only 3.39. After reading the claims extracted from both models in the FreshQA domain, we notice that GPT-4o tends to generate short and to-the-point answers while Claude 3 Haiku tends to generate longer answers, offering more auxiliary information. Within the Claude 3 model family, Claude 3 Haiku generates longer outputs than the other two and has more responses with 4 or more supported claims, resulting in a higher VERISCORE.

The results of FreshQA shows that, for a fair comparison, when calculating VERISCORE, the length of model responses should be taken into account. This can be done by forcing all the models to generate a similar length of outputs. However, this will not result in a setting of how end-users would use language models. Forcing models to generate longer responses than necessary can also elicit more infactual content, as noticed by Min et al. (2023) that later content in model responses tends to be less factual.

K Detailed results of models' VERISCORE

In this appendix section, we provide the breakdown of VERISCORE of the 16 models on all 8 domains in Table 16.

Prompt for Extracting Verifiable Claims from Non-Question-Answering Type of Inputs

You are trying to verify how factual a piece of text is. To do so, you need to break down a sentence and extract as many fine-grained facts mentioned in the sentence as possible. Each of these fine-grained facts should be verifiable against reliable external world knowledge (e.g., via Wikipedia). Any story, personal experiences, hypotheticals (e.g., "would be" or subjunctive), subjective statements (e.g., opinions), suggestions, advice, instructions, and other such content should not be included in the list. Biographical, historical, scientific, and other such texts are not personal experiences or stories. You should extract verifiable facts from them. Each fact should also be describing either one single event (e.g., "Nvidia is founded in 1993 in Sunnyvale, California, U.S.") or single state (e.g., " [redacted] has existed for 161 years.") with necessary time and location information. Quotations should be extracted verbatim with the source when available. Listed references should be ignored.

Extract fine-grained facts from the sentence marked between <SOS> and <EOS>. You should focus on the named entities and numbers in the sentence and extract relevant information from the sentence. Other sentences are only context for you to recover pronouns, definite phrases (e.g., "the victims" or "the pope"), and so on. Each fact should be understandable on its own and require no additional context. This means that all entities must be referred to by name but not pronoun. Use the name of entities rather than definite noun phrases (e.g., 'the teacher') whenever possible. If a definite noun phrase is used, be sure to add modifiers (e.g., a embedded clause, a prepositional phrase, etc.). Each fact must be situated within relevant temporal and location whenever needed. Keep each fact to one sentence with zero or at most one embedded clause. You do not need to justify what you extract.

If there is no verifiable fact in the sentence, please write "No verifiable claim."
Here are some examples:

Text: The sweet potato or sweetpotato (Ipomoea batatas) is a dicotyledonous plant that belongs to the bindweed or morning glory family, Convolvulaceae. <SOS>Its large, starchy, sweet-tasting tuberous roots are used as a root vegetable.<EOS> The young shoots and leaves are sometimes eaten as greens.
Sentence to be focused on: Its large, starchy, sweet-tasting tuberous roots are used as a root vegetable.
Facts:
- Sweet potatoes' roots are large.
- Sweet potatoes' roots are starchy.
- Sweet potatoes' roots are sweet-tasting.
- Sweet potatoes' roots are tuberous.
- Sweet potatoes' roots are used as a root vegetable.

Text: Garnett had spent well over a decade with the Minnesota Timberwolves, and while he stayed loyal to that team, he found little success there. <SOS>When he said "you can't get your youth back," he meant it - because from a human standpoint, had he been able to apply his talents somewhere else, NBA history might have been different.<EOS>
Sentence to be focused on: When he said "you can't get your youth back," he meant it - because from a human standpoint, had he been able to apply his talents somewhere else, NBA history might have been different.
Facts:
- Kevin Garnett said "you can't get your youth back."

Text: I (27f) and my fiance "Leo" (27m) decided to let my FSIL "Maya" (32f) stay at our house because she needed space from her husband due to some relationship struggles they're having. Leo and I had gotten wedding cake samples from an expensive bakery specializing in wedding cakes. We planned to test them along with Maya after we finished up some other wedding plans yesterday. <SOS>However, when I came home from work to see Leo yelling at Maya, the box the samples came in wide open on the living room table, and Maya arguing with him.<EOS> I asked what was happening, and Leo angrily told me that while we were both at work, Maya had some friends over and they ended up eating almost all of our cake samples.
Sentence to be focused on: However, when I came home from work to see Leo yelling at Maya, the box the samples came in wide open on the living room table, and Maya arguing with him.
Facts:
No verifiable claim.

... <Total of 13 Examples> ...
Extract *verifiable atomic* facts.

Text: `{sliding window}`
Sentence to be focused on: `{sentence}`
Facts:

Table 6: Claim extraction prompt for non-question-answering type of inputs. The sliding window follows the template `(context1 = 0-3 sentence) <SOS>Sentence to be focused on<EOS> (context2 = 0-1 sentence)`. If the paragraph from which the sentence is taken is longer than five sentences, the first sentence of the paragraph is always prepended before `context1`. Marked out content will be uncovered after the review process.

Prompt for Extracting Verifiable Claims from Question-Answering Type of Inputs

You are trying to verify how factual a response to a question or request is. To do so, you need to break down a sentence and extract as many fine-grained facts mentioned in the response. Each of these fine-grained facts should be verifiable against reliable external world knowledge (e.g., via Wikipedia). Any story, personal experiences, hypotheticals (e.g.,"would be" or subjunctive), subjective statements (e.g., opinions), suggestions, advice, instructions, and other such content should not be included in the list. Biographical, historical, scientific, and other such texts are not personal experiences or stories. You should extract verifiable facts from them. Each fact should also be describing either one single event (e.g., "Nvidia is founded in 1993 in Sunnyvale, California, U.S.") or single state (e.g., " [redacted] has existed for 161 years.") with necessary time and location information. Quotations should be extracted verbatim with the source when available. Listed references should be ignored.
Extract fine-grained facts from the sentence between <SOS> and <EOS>. You should focus on the named entities and numbers in the sentence and extract relevant information from the sentence. Do not extract claims from the question. The question and other sentences are only context for you to recover pronouns, definite phrases (e.g., "the victims" or "the pope"), and so on. Each fact should be understandable on its own and require no additional context. This means that you need to always related the extracted claims to the question. This also means that all entities must be referred to by name but not pronoun. Use the name of entities rather than definite noun phrases (e.g., 'the teacher') whenever possible. If a definite noun phrase is used, be sure to add modifiers (e.g., a embedded clause, a prepositional phrase, etc.). Each fact must be situated within relevant temporal and location whenever needed. Keep each fact to one sentence with zero or at most one embedded clause. You do not need to justify what you extract.
If there is no verifiable fact in the sentence, please write "No verifiable claim."
Here are some examples:

Question: What NASA programs would support our college in starting a robotics program?
Response: NASA has several programs that can support colleges in starting a robotics program. Here are a few:
<SOS>1. NASA Robotics Alliance Project (RAP): This program provides educational resources and support for robotics teams, including college-level teams, that are participating in NASA robotics competitions.<EOS>
2. NASA Minority University Research and Education Project (MUREP): This program provides funding and resources for colleges and universities with a significant minority student population to develop research and education programs in STEM fields, including robotics.
3. NASA's Robotics Education Project: This project provides robotics education materials and resources for educators, including college-level educators, to use in their classrooms.
4. NASA's Space Technology Mission Directorate (STMD): This directorate funds research and development in advanced technologies, including robotics, that can support NASA's mission to explore space.
Sentence to be focused on: 1. NASA Robotics Alliance Project (RAP): This program provides educational resources and support for robotics teams, including college-level teams, that are participating in NASA robotics competitions.
Facts:
- NASA has a program called NASA Robotics Alliance Project (RAP).
- NASA Robotics Alliance Project provides educational resources for robotics teams.
- NASA Robotics Alliance Project provides supports for robotics teams.
- NASA Robotics Alliance Project provides supports for college-level teams that are participating in NASA robotics competitions.

Question: How do trees know when to stop growing?
Thanks everyone i learned a lot more about trees.(:
Response: <SOS>Ah yes, tomatoes, this is a big problem with tomato plants.<EOS>
Sentence to be focused on: Ah yes, tomatoes, this is a big problem with tomato plants.
Facts:
No verifiable claim.

. . . <Total of 10 Examples> . . .
Extract *verifiable atomic* facts.

`{sliding window}`
Sentence to be focused on: `{sentence}`
Facts:

Table 7: Claim extraction prompt for question-answering type of inputs. The sliding window consists of the question and `(context1 = 0-3 sentence) <SOS>Sentence to be focused on<EOS> (context2 = 0-1 sentence)`. Marked out content will be uncovered after the review process.

The claim is too general to be verified.

Example Claim 1:
A wooden spoon creates a small gap between the pot and the spoon.
Example Claim 2:
Martha had passed away.
Example Claim 3:
A systematic review on sex differences in the reinforcing effects of nicotine was published in the journal Nicotine & Tobacco Research in 2019.

A part of the claim is not mentioned in the evidence, or no evidence confirms the relationship between the parts in a claim—Inconclusive (a) Case

Reasoning:
(1) Only the search result 7 hints that there might be a Persuasive Technology Lab at Stanford University.
(2) No evidence mentions this lab aims to create positive behavior change.
Claim
The intention of Stanford University's Persuasive Technology Lab was to create positive behavior change .
Evidence
Search result 1
Title: Behavior Design Lab - Stanford University
Content: Behavior Design is a new approach to understanding human behavior and how to design for behavior change. Based on the work of Dr. BJ Fogg , Behavior Design ...
Link: https://behaviordesign.stanford.edu/
Search result 2
Title: About Us - Behavior Design Lab - Stanford University
Content: Our lab's overall mission is this: Teach good people how human behavior works so they can create solutions that effectively increase health, boost happiness, ...
Link: https://behaviordesign.stanford.edu/about-us
Search result 3
Title: Building Habits: The Key to Lasting Behavior Change
Content: "Habits are easier to form than most people think," he says, "If you do it in the right way." As the founder and director of Stanford's Behavior ...
Link: https://www.gsb.stanford.edu/insights/building-habits-key-lasting-behavior-change
Search result 4
Title: The Ethical Use of Persuasive Technology - Behavior Design Lab
Content: While our research has moved on from persuasive technology to focus on designing for healthy behavior change, we believe it is important to continue to ...
Link: https://behaviordesign.stanford.edu/ethical-use-persuasive-technology
Search result 5
Title: Fiddling With Human Behavior - WIRED
Content: Researchers at Stanford are studying technology designed to persuade people to change the way they think or act.
Link: https://www.wired.com/2000/03/fiddling-with-human-behavior/
Search result 6
Title: BJ Fogg - Behavior Design Lab - Stanford University
Content: BJ wrote a seminal book, Persuasive Technology: Using Computers to Change What We Think and Do , about how computers can be designed to influence attitudes ...
Link: https://behaviordesign.stanford.edu/people/bj-fogg
Search result 7
Title: How Stanford Profits Off Addiction
Content: Back in 1998, one of Stanford's eccentric social scientists, B.J. Fogg, founded the Persuasive Technology Lab to research how tech products ...
Link: https://stanfordreview.org/how-stanford-profits-tech-addiction-social-media/
Search result 8
Title: Tech companies use "persuasive design" to get us hooked ... - Vox
Content: Big tech now employs mental health experts to use persuasive technology, a new field of research that looks at how computers can change the way ...
Link: https://www.vox.com/2018/8/8/17664580/persuasive-technology-psychology
Search result 9
Title: Stanford Behavior Design Lab - Wikipedia
Content: The Stanford Behavior Design Lab is a research organization advancing behavior change methods and models based at Stanford University. Founded in 1998 and ...
Link: https://en.wikipedia.org/wiki/Stanford_Behavior_Design_Lab
Search result 10
Title: How to create new good habits, according to Stanford ... - Quartz
Content: To create a real lifelong habit, the focus should be on training your brain to succeed at a small adjustments, then gaining confidence from that ...
Link: https://qz.com/877795/how-to-create-new-good-habits-according-to-stanford-psychologist-b-j-fogg

Table 8: Examples of the annotaton items whose claim was labeled as `inconclusive` by all three annotators.

Annotator mistake

Explanation: Two annotators chose the `inconclusive` label for this claim but one chose `supported` based on one search result as given below. This is an annotator mistake because the name Luis Guillermo Rivera is not mentioned in the evidence.
Claim: Luis Guillermo Rivera has written literary criticism.
Evidence:
Search result 5
Title: I Write with Words That Have Shadow but Don't Shelter
Content: Born in Tumeremo, Bolívar, in 1933, Venezuelan writer Guillermo Sucre is also an essayist, translator, literary critic, and educator. A ...
Link: https://www.worldliteraturetoday.org/blog/poetry/i-write-words-have-shadow-dont-shelter-guillermo-sucre

Ambiguity in the interpretation of the claim and evidence

Explanation: Two annotators chosen the `supported` label but one chose `inconclusive`. The annotator's comment states:"I chose inconclusive as the claim is 'can be chosen' and all the results are that they potentially 'could' be chosen, not that they actually CAN."
Claim: Traits that can be chosen include eye color, hair color, intelligence, and athletic ability.
Evidence:
Search result 2
Title: [PDF] Sex Selection, Genetic Analysis, and Designer Babies
Content: In theory, parents could also select embryos on the basis of eye color, hair color, or any other genetic trait.
Link: https://med.nyu.edu/departments-institutes/population-health/divisions-sections-centers/medical-ethics/sites/default/files/medical-ethics-sex-selection-genetic-analysis.pdf

The claim contains an unclear referent.

Explanation: The annotators did not agree on this item at all. It is probably because there are multiple people with the name Jessica Barboza, making it hard to make a decision.
Claim: Jessica Barboza was born in São Paulo, Brazil.
Evidence:
Search result 1
Title: Jessica Barboza - Wikipedia
Content: Jessica Barboza. Born. Jessica Cristina Barboza Schmidt. (1987-08-14) 14 August 1987 (age 36). Maracaibo, Zulia, Venezuela. Height, 1.79 m (5 ft 10+1⁄2 in).
Link: https://en.wikipedia.org/wiki/Jessica_Barboza
Search result 3
Title: Jessica Barboza - Age, Family, Bio | Famous Birthdays
Content: Style blogger and makeup guru known for her Peace and Vogue blog and YouTube channel. The blossoming beauty maven has gained a following of more than 550,000 ...
Link: https://www.famousbirthdays.com/people/jessica-barboza.html
Search result 5
Title: Jessica Barboza - Facebook
Content: Jessica Barboza ; Lives in [redacted] ; From São Paulo, Brazil ; In a relationship with [redacted].
Link: https://www.facebook.com/[redacted]

The claim is hard to verify because it is long and complex.

Explanation: The annotators did not agree on this item at all. The claim contains multiple parts that are correlated to each other. However, it is also hard to further break the claim down to smaller claims.
Claim: Chuck Norris's victory in the 1968 World Full-Contact Karate Championships solidified his reputation as one of the best martial artists in the world.

Intermediate reasoning process is needed because the evidence might indirectly supports the claim.

Explanation: Two annotators chose the `supported` label and one chose `inconclusive`. It is possible to interprete he "safety objectives" in search result 8 as it includes "public health".
Claim: The disposal of radioactive waste is aimed at ensuring public health.
Evidence:
Search result 8
Title: PART 61—LICENSING REQUIREMENTS FOR LAND DISPOSAL ...
Content: (1) Disposal of radioactive waste in near-surface disposal facilities has the following safety objectives: protection of the general population from releases of ...
Link: https://www.nrc.gov/reading-rm/doc-collections/cfr/part061/full-text.html

Table 9: Example of the annotation items on which the annotators did not fully agree with each other. We mark out the private sensitive content.

Quality description	**Model outputs** (G = GPT-4, M = Mistral)
Almost identical with certain phrases being relocated	**G** Rubbing alcohol works by a different mechanism than antibiotics. **M** Rubbing alcohol works by a mechanism different than antibiotics.
Mistral misses small words, hence the extracted claims are less specific	**G** Summarily laying off workers can have devastating impacts on individuals **M** Laying off workers can have devastating impacts on individuals.
Mistral puts multiple pieces of information into one claim while GPT-4 breaks the information down into multiple claims	**G** (1) The regimental flags used during Napoleon's invasion of Russia by Bavaria were similar to their design. (2) The regimental flags for Bavaria during Napoleon's invasion of Russia had a green wreath added around the eagle. **M** Bavarian regimental flags used during Napoleon's invasion of Russia in 1812 featured a green wreath around the eagle.

Table 10: Quality analysis of the outputs generated by prompted GPT-4 and fine-tuned Mistral-7B-Instruct-v0.2. The table lists three common observations made in the pairwise comparison task.

Model	Label #	Overall F1	S. F1	U. F1
Mixtral	2	0.817	0.817	0.686
Mixtral	3	0.807	0.807	0.680
Claude 3	2	0.839	0.839	**0.758**
Claude 3	3	0.826	0.826	0.681
GPT-4	2	0.829	0.829	0.696
GPT-4	3	0.812	0.812	0.639
GPT-4o	2	0.813	0.813	0.649
GPT-4o	3	**0.841**	**0.841**	0.731

Table 11: The results of testing four prompted LLM on the claim verification task. S. = `supported` and U. = `unsupported`. Mixtral stand for Mixtral-8x22B-Instruct-v0.1. Claude 3 stands for Claude 3 Opus. GPT-4o achieves the highest overall F1 and F1 on the items with `supported` as the label. Although its `unsupported` F1 is not the highest, it is not far below the highest, which is Claude 3 with 0.758.

Prompt for verifying claims with three labels

You need to judge whether a claim is supported or contradicted by Google search results, or whether there is no enough information to make the judgement (i.e., inconclusive). When doing the task, take into consideration whether the link of the search result is of a trustworthy source. Mark your answer with ### signs.

Below are the definitions of the three categories:
Supported: A claim is supported by the search results if everything in the claim is supported and nothing is contradicted by the search results. There can be some search results that are not fully related to the claim.
Contradicted: A claim is contradicted by the search results if something in the claim is contradicted by some search results. There should be no search result that supports the same part.
Inconclusive: A claim is inconclusive based on the search results if:
- a part of a claim cannot be verified by the search results,
- a part of a claim is supported and contradicted by different pieces of evidence,
- the entity/person mentioned in the claim has no clear referent (e.g., "the approach", "Emily", "a book").

Here are some examples:

Claim: Vikings used their longships to transport livestock.
Search result 1
Title: How did the Vikings transport animals on their ships? - Quora
Content: The Vikings transported horses overseas in boats very similar to Viking longships, but with flat flooring built within the hulls, which allowed ...
Link: https://www.quora.com/How-did-the-Vikings-transport-animals-on-their-ships
Your decision: ###Contradicted.###
<Other search results omitted for the sake of space>

<nine such examples>

Your task:
Claim: {claim to be verified}
{search results}
Your decision:

Prompt for verifying claims with two labels

Everything being the same but the definitions of the labels are changed as below. The decisions in the few-shot exmaples are converted to `supported` and `unsupported` accordingly (i.e., `contradicted` and `inconclusive` become `unsupported`).

Supported: A claim is supported by the search results if everything in the claim is supported and nothing is contradicted by the search results. There can be some search results that are not fully related to the claim.
Unsupported: If a claim is not supported by the search results, mark it as unsupported.

Table 12: Claim verification prompt for Mixtral-8×22B-Instruct-v0.1, GPT-4, and GPT-4o. For Claude 3 Opus, the order of the claim, search results, and the decision is rearranged. Otherwise, the model does not always output a decision marked by ###. The rearranged order is search results, the claim, a short task description, and the decision. The short task description is "Task: Given the search results above, is the claim {supported, contradicted, or inconclusive}/{supported or unsupported}? Mark your decision with ### signs." The set of labels in the curly brackets depends on whether the verificaton task is binary or trinary.

Book name	Author/editor/translator	Publication date
Blunt Instruments: Recognizing Racist Cultural Infrastructure in Memorials, Museums, and Patriotic Practices	Kristin Ann Hass	January, 2023
Every Living Thing: The Great and Deadly Race to Know All Life	ason Roberts	April, 2024
It's OK to Be Angry About Capitalism	Bernie Sanders, John Nichols	2024
Out of the Darkness: The Germans, 1942-2022	Frank Trentmann	February, 2024
Takeover: Hitler's Final Rise to Power	Timothy W. Ryback	March, 2024
The Exhausted of the Earth: Politics in a Burning World	Ajay Singh Chaudhary	February, 2023
The Making of a Leader: The Formative Years of George C. Marshall	Josiah Bunting III	March, 2024
They Were Here Before Us: Stories from the First Million Years	Eyal Halfon, Ran Barkai	March, 2024
The Green Power of Socialism: Wood, Forest, and the Making of Soviet Industrially Embedded Ecology	Elena Kochetkova	February, 2024
A Brief History of Feminism	Patu, Antje Schrupp, Sophie Lewis	April, 2024
Handbook of Formal Analysis and Verification in Cryptography	Sedat Akleylek, Besik Dundua	September, 2023
Handbook on Renewable Energy and Green Technology	S. Pugalendhi, J. Gitanjali, R. Shalini, P. Subramanian	February, 2024
The Handbook of Sex Differences Volume I Basic Biology	Lee Ellis, Craig T. Palmer, Rosemary Hopcroft, Anthony W. Hoskin	September, 2023
The Oxford Handbook of Thomas More's Utopia	Cathy Shrank, Phil Withington	February, 2024
The Routledge Handbook of Commodification	Elodie Bertrand, Vida Panitch	December, 2023
The Routledge Handbook of Green Finance	Othmar M. Lehner, Theresia Harrer, Hanna Silvola, Olaf Weber	November, 2023
The Routledge Handbook of Language and Religion	Stephen Pihlaja, Helen Ringrow	December, 2023
The Routledge Handbook of Language and Youth Culture	Bente A. Svendsen, Rickard Jonsson	December, 2023
Clinical Handbook of Nephrology	Robert S. Brown MD	August, 2023
Handbook of Face Recognition: The Deep Neural Network Approach	Stan Z. Li, Anil K. Jain, Jiankang Deng	2024

Table 13: Twenty newly published non-fictional books. We took ten paragraphs from each book and used them to prompt LLMs to generate continuations. The selected paragraphs are all located at the beginning of a chapter/section.

Model Name	Release	Reference
GPT-3.5-turbo	2023.06	
GPT-3.5-turbo	2023.11	OpenAI (2023)
GPT-4	2024.01	
Claude-3-Haiku	2024.03	
Claude-3-Sonnet	2024.02	Anthropic (2023)
Claude-3-Opus	2024.02	
Mist-7B-Inst-v0.1	2023.09	
Mist-7B-Inst-v0.2	2023.12	
Mixt-8×7B-Inst-v0.1	2023.12	Jiang et al. (2023, 2024)
Mixt-8×22B-Inst-v0.1	2024.04	
OLMo-7B-Inst	2024.01	Groeneveld et al. (2024)
DBRX Inst (132B)	2024.03	Databricks (2024)
Qwen1.5-1.8B-Chat	2023.11	Bai et al. (2023)
Gemma-2B-it	2024.04	Gemma Team et al. (2024)
Vicuna-7B-v1.5	2023.12	Zheng et al. (2023)
GPT-4o	2024.05	OpenAI (2024)

Table 14: Sixteen models that are tested in the current work. Inst, Mist, and Mixt stands for instruction, Mistral, and Mixtral. The numbers of total parameters of each model are given in brackets if provided by the model providers and not in the model names. The first 12 models are also used to generate model responses for fine-tuning claim extraction and verification models.

Claim that is unsupported because it uses a different term as its search results

Claim: In long trains, additional locomotives can be placed along the train to help distribute the pulling force more evenly. (ELI5)

Search result

Title: Nuts & Bolts: Why is there an engine in the middle of that train?

Content: By placing DPUs* throughout the train rather than just at the rear—thus distributing power more evenly—railroads were able to enhance a train's ...

Link: https://gorail.org/infrastructure/nuts-bolts-why-is-there-an-engine-in-the-middle-of-that-train

* DPU stands for Distributed Power Unit, a locomotive set.

Unsupported claims that have reasonable content

- Spreading the load across multiple axles reduces the stress on individual components in trains. (ELI5)
- Detailed interviews with patients and their contacts help to establish timelines. (ELI5)

Unsupported claims that are too vague to be verified

- Missiles travel through the Earth's atmosphere for most or all of their flight. [Not clear which type of missiles] (ELI5)
- The forests of Siberia and the Far East were crucial for meeting the demand for wood and wood products for export. [Not clear whose demand it is] (FreshBooks)

Unsupported claims that are too broad to be verified

- Marshall's leadership and strategic acumen ensured the maneuver was carried out flawlessly during a field maneuver in the Philippines. (FreshBooks)
- Germany is maintaining its competitive edge in a rapidly changing global landscape. (FreshBooks)
- The initiatives of the General German Women's Union helped to lay the groundwork for future advancements in women's rights in Germany. (FreshBooks)

General unsupported claims—no supporting evidence

- The movement of the sun or stars could be compared with the rate of flow in water clocks. (ELI5)
- Driving from the middle of a car complicates interactions with road design elements. (ELI5)
- The patronage of William Warham reflects the broader trend of Renaissance humanism gaining foothold in England. (FreshBooks)

Table 15: Types and reasons of claims being `unsupported`.

Model	LongFact (32)				Biography (26)				ELI5 (21)				AskHist (21)				FreshBooks (24)				ShareGPT (11)				FreshQA (4)				WP (1)			
	L	P	R	F	L	P	R	F	L	P	R	F	L	P	R	F	L	P	R	F	L	P	R	F	L	P	R	F	L	P	R	F
Gemma-2b-it	20.4	67.2	61.4	60.7	11.8	4.3	5.1	4.6	8.7	38.6	27.2	28.8	6.6	28.2	17.2	17.8	5.0	43.1	20.0	25.1	15.4	28.8	32.7	27.6	1.4	8.7	4.0	4.8	23.8	3.2	6.1	3.7
Qwen1.5-1.8B-Chat	22.1	64.4	78.2	70.3	16.2	11.4	18.5	14.1	**21.2**	49.1	74.0	57.9	16.6	37.7	59.2	45.2	19.4	45.1	64.8	52.6	24.1	42.7	**64.7**	49.2	**7.1**	35.9	68.5	43.9	29.5	4.1	10.0	5.5
Vicuna-7b-v1.5	12.1	79.3	57.0	63.4	8.8	26.3	21.9	23.0	8.3	59.8	47.5	51.3	8.7	43.3	39.7	39.7	7.1	50.2	35.0	39.0	15.1	46.6	45.8	43.6	2.6	42.5	29.5	30.9	20.1	2.0	2.0	2.0
OLMo-7B-Inst	21.2	72.3	77.3	73.4	**16.8**	17.4	22.7	19.4	18.3	55.0	65.2	58.8	16.7	41.5	50.5	43.2	**24.1**	49.9	60.6	53.7	25.5	48.0	60.5	49.4	3.6	46.1	56.5	49.2	31.2	2.0	2.0	2.0
DBRX-Inst	15.6	85.0	72.3	75.9	13.2	45.4	48.6	46.5	11.5	69.0	59.3	61.9	13.6	49.7	52.2	49.5	13.1	58.9	62.6	60.2	18.0	49.0	54.0	48.9	4.3	66.1	69.5	64.9	26.4	**12.1**	**18.0**	**13.3**
Mist-7B-Inst-v0.1	10.3	76.5	50.6	57.6	10.3	22.1	19.6	20.3	7.0	59.2	35.2	42.2	8.0	45.1	33.3	36.5	7.1	53.3	34.7	39.8	16.3	46.6	43.3	41.2	1.8	46.8	24.0	29.0	21.2	6.4	10.0	7.3
Mist-7B-Inst-v0.2	16.0	83.2	68.4	72.0	11.9	29.6	31.2	30.0	11.3	63.9	58.7	58.8	14.5	43.9	45.4	41.2	10.7	54.6	52.4	52.4	19.1	**54.0**	61.1	**54.8**	4.1	57.9	55.5	54.2	29.8	2.6	6.0	3.4
Mixt-8x7B-Inst-v0.1	17.7	84.4	75.1	77.3	11.1	42.2	44.0	42.5	11.6	68.0	60.7	61.9	13.8	50.9	54.3	50.7	10.4	59.8	56.8	57.4	17.7	53.0	56.4	51.5	3.4	59.8	52.5	53.1	30.7	5.3	8.0	5.8
Mixt-8x22B-Inst-v0.1	17.6	86.7	76.2	78.0	12.6	47.4	49.2	47.6	12.4	**69.1**	64.6	64.9	13.3	52.5	54.0	51.1	12.4	60.3	58.3	58.0	19.8	50.4	57.3	51.4	3.3	65.6	58.0	59.6	30.4	5.4	8.0	6.1
Claude-3-haiku	17.4	85.9	76.4	79.4	8.0	42.9	35.4	37.1	14.8	60.5	60.5	58.7	13.8	42.8	46.3	43.5	9.0	53.9	47.3	49.5	18.0	44.8	51.1	44.7	5.5	**75.3**	**89.0**	**77.9**	24.6	4.0	4.0	4.0
Claude-3-sonnet	19.4	85.8	78.4	80.7	7.9	44.2	36.0	37.6	13.7	58.9	56.3	56.2	13.1	39.6	43.6	40.7	10.6	59.6	60.1	59.3	16.2	50.4	58.5	51.7	4.9	71.3	86.5	76.2	25.6	2.1	4.0	2.2
Claude-3-opus	21.4	88.3	81.9	83.6	10.5	51.8	54.4	52.7	14.3	66.6	63.3	63.4	14.8	49.2	52.9	49.8	12.2	63.3	70.5	66.4	19.0	50.9	56.5	51.6	5.4	72.2	81.0	72.2	24.7	1.3	4.0	1.9
GPT-3.5-turbo-0613	16.9	87.7	75.1	77.6	14.5	43.9	49.2	45.9	13.1	67.8	63.2	62.9	12.9	52.9	53.6	51.8	11.7	50.3	51.3	49.0	19.1	51.1	52.5	48.6	1.7	68.4	38.5	44.9	37.2	4.3	6.0	4.9
GPT-3.5-turbo-1106	10.2	**90.8**	56.3	64.7	5.3	**54.1**	30.6	38.1	6.4	61.9	35.3	42.8	7.8	51.8	37.0	40.8	4.3	51.5	24.8	32.5	13.3	47.9	43.1	42.1	1.5	65.7	36.5	44.0	24.2	4.0	8.3	5.0
GPT-4-0125-preview	20.6	84.3	89.2	85.9	13.0	52.2	**63.8**	56.4	20.1	63.7	**83.0**	70.7	18.8	47.8	72.6	56.6	12.8	64.6	**77.0**	69.7	23.6	51.5	60.2	53.5	2.9	67.8	59.5	60.3	33.3	3.1	6.0	3.7
GPT-4o	**25.8**	85.4	**89.8**	**86.7**	13.1	53.5	61.5	**56.7**	20.4	67.1	79.8	**71.7**	**22.6**	**54.2**	**77.9**	**61.4**	11.6	**68.8**	75.2	**70.9**	**28.2**	49.3	56.5	51.5	2.6	67.9	55.0	58.2	**48.6**	5.3	8.0	6.2

Table 16: Details of VERISCORE of 16 models on all 8 domains. The maximum values for each metric in every category are highlighted in bold. L = average sentence count per response; P = average response precision; R = average response recall; F = VERISCORE. AskHist = AskHistorians; WP = WritingPrompts.