VERIFASTSCORE: Speeding up long-form factuality evaluation

Rishanth Rajendhran[1] **Amir Zadeh**[2] **Matthew Sarte**[2] **Chuan Li**[2] **Mohit Iyyer**[1]
[1]University of Maryland, College Park [2]Lambda Labs
[1]{rishanth, miyyer}@umd.edu [2]{amirali.zadeh, matt.sarte, c}@lambdal.com

Abstract

Metrics like FACTSCORE and VERISCORE that evaluate long-form factuality operate by decomposing an input response into atomic claims and then individually verifying each claim. While effective and interpretable, these methods incur numerous LLM calls and can take upwards of 100 seconds to evaluate a single response, limiting their practicality in large-scale evaluation and training scenarios. To address this, we propose VERIFASTSCORE, which leverages synthetic data to fine-tune Llama3.1 8B for *simultaneously* extracting and verifying all verifiable claims within a given text based on evidence from Google Search. We show that this task cannot be solved via few-shot prompting with closed LLMs due to its complexity: the model receives ~**4K** tokens of evidence on average and needs to concurrently decompose claims, judge their verifiability, and verify them against noisy evidence. However, our fine-tuned VERIFASTSCORE model demonstrates strong correlation with the original VERISCORE pipeline at both the example level ($r = \mathbf{0.80}$) and system level ($r = \mathbf{0.94}$) while achieving an overall speedup of **6.6**× (9.9× excluding evidence retrieval) over VERISCORE. To facilitate future factuality research, we publicly release our VERIFASTSCORE model and synthetic datasets.[1]

1 Introduction

Modern methods to evaluate the factuality of long-form generation, such as FACTSCORE (Min et al., 2023), SAFE (Wei et al., 2024), and VERISCORE (Song et al., 2024), rely on a pipelined approach that uses LLMs to decompose an input text into "atomic" claims and then verify each claim against retrieved evidence documents. While this pipeline results in interpretable outputs that correlate strongly with human annotations of factuality, it is also *slow* (Liu et al., 2025), requiring many calls to large models to evaluate a text, which limits its usability in both evaluation and training (e.g., as a reward model in RLHF).

In this paper, we propose a significantly faster alternative: a single-pass factuality evaluator that simultaneously performs claim decomposition and verification over a long-form input. Our method, VERIFASTSCORE, is trained to jointly extract all verifiable claims from a response and assess their factuality against retrieved evidence. To train VERIFASTSCORE, we collate sentence-level evidence with claim-level outputs produced by VERISCORE, a high-quality general-domain factuality evaluation metric, into a synthetic dataset used to fine-tune Llama3.1 8B Instruct (Grattafiori et al., 2024).

Each training example contains a response to be evaluated as well as a corresponding evidence context that consists of web search results retrieved by querying individual sentences in the response. VERIFASTSCORE is then asked to output a list of verifiable claims from the response, each labeled as `Supported` or `Unsupported` based on the evidence (see Figure 1). This task is non-trivial: the model must *internally* decontextualize the response, identify atomic claims, resolve coreferences, and assess support against an evidence context spanning up to **4K** tokens of potentially noisy information. Standard few-shot prompting of powerful closed models like GPT-4o struggles with this complexity, achieving a Pearson correlation of only 0.33 with VERISCORE despite high API costs.

Despite these challenges, VERIFASTSCORE matches the quality of VERISCORE closely while offering substantial improvements in both cost and runtime efficiency: it achieves a Pearson correlation of $\mathbf{0.80}$ (Pearson r, p<0.001) with VERISCORE's scores, a **9.9**× modeling speedup and a **6.64**× overall speedup in wall-clock time, reducing both modeling and retrieval latency while maintaining the metric's interpretability. Human

[1]Code and data available at https://github.com/RishanthRajendhran/VeriFastScore.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9234–9259
November 4-9, 2025

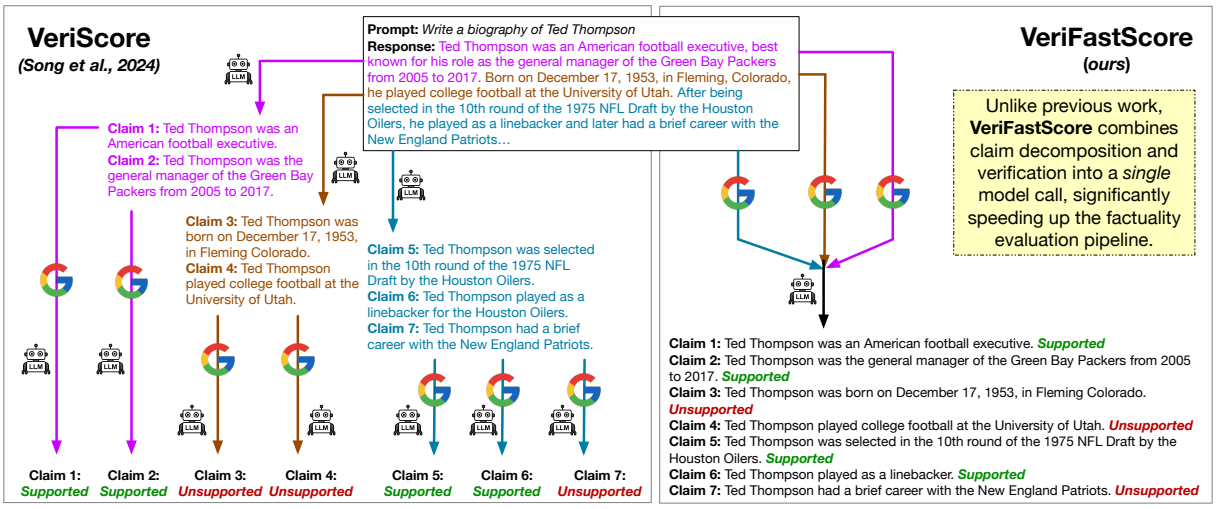

Figure 1: *(left)* The VERISCORE pipeline (Song et al., 2024) involves a separate call to a trained model (or LLM API) to decompose every sentence of the long-form response. More model calls are issued for each decomposed claim to verify them against claim-level evidence retrieved from Google Search. In contrast, our VERIFASTSCORE method retrieves *sentence-level* evidence from Google Search, combines it together with the response into a single long-context prompt, and outputs a list of claims and labels with just a *single call* to a trained model.

evaluation corroborates the high quality of outputs from VERIFASTSCORE. We release VERIFASTSCORE and our synthetic training dataset to support efficient, interpretable factuality evaluation and facilitate its use in alignment settings.

2 Building a faster factuality metric

In this section, we first describe the limitations of existing factuality metrics like VERISCORE, whose separate decomposition, retrieval, and verification stages require several calls to LLMs to complete for a single response. These limitations motivate our design and development of VERIFASTSCORE, which combines all steps in the existing factuality evaluation pipeline into a single model call. To enable such a fast evaluation, we generate synthetic data from VERISCORE and use it to fine-tune a language model on *response-level* decomposition and verification.

2.1 The "decompose and verify" pipeline for factuality evaluation

Recently-proposed metrics for long-form factuality evaluation such as FACTSCORE and SAFE are implemented by the following multi-stage pipeline:

1. **Claim extraction:** decompose the long-form model response into "atomic" (i.e., short) claims
2. **Evidence retrieval:** collect evidence for every claim individually by retrieving relevant documents from a datastore or search engine
3. **Claim verification:** mark each atomic claim as either *supported* or *unsupported* by the evidence collected for that claim

In FACTSCORE and VERISCORE, the first and third stages are implemented by either open-weight or closed LLMs (e.g., GPT-4), while SAFE's decomposition and verification stages are implemented by prompting closed models from OpenAI or Claude.

VERISCORE, our starting point: Unlike FACTSCORE (Min et al., 2023) and SAFE (Wei et al., 2024), which attempt to verify *all* atomic claims in a model's response using external evidence, the decomposition stage of VERISCORE focuses solely on *verifiable* claims—defined as

> Verifiable claims describe a single *event* or *state*[3] 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. (Maienborn, 2003, 2019)

[3]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."

Prompt Source	Split	# Responses	Avg. Claims/Response	Avg. Token Length			Claim Verification Label (%)	
				Response	Evidence	Facts	Supported	Unsupported
VeriScore data	Train	4082	16	298	4381	271	68	32
	Test	1815	16	301	4466	278	68	32
Tulu3 Personas	Train	4860	20	366	3755	334	59	41
	Test	4083	20	370	3714	329	59	41

Table 1: **Overview of the datasets used.** For each data source and split (train/test), we report the total number of model responses, the average number of verifiable claims per response, average token lengths [2] of model responses, consolidated evidence, and consolidated list of verifiable claims extracted by VERISCORE from model responses, and the percentage of claims labeled by VERISCORE as supported and unsupported by retrieved evidence. Labels are more balanced in the dataset collected during this work using prompts from *Tulu3 Personas* dataset.

Like SAFE, VERISCORE also relies on Google Search to retrieve evidence in the second stage of the pipeline. We use VERISCORE as a starting point due to its increased domain generality compared to prior long-form factuality metrics.

2.2 Issues with VeriScore

VERISCORE evaluates factuality using a multi-stage pipeline that proceeds sentence-by-sentence through a model response. For each sentence, it extracts verifiable claims using a claim extraction model, where the full response is included in the input to ensure proper decontextualization. Each extracted claim is then used as a query to retrieve evidence via Google Search. Finally, a claim verification model checks the factuality of each claim against the retrieved evidence.

While this careful decomposition and step-wise verification ensures precision, it incurs significant computational cost. On average, a single response of 14 sentences leads to 23 extracted claims (see Table 6). Each response thus results in:

- 14 calls to the claim extraction model (one per sentence),
- 23 Google searches (one per claim),
- 23 calls to the claim verification model.

This totals roughly 60 model or API calls per response. Processing a single response takes approximately 100 seconds, making the method impractical for real-time or large-scale evaluation (Liu et al., 2025; Song et al., 2024).

2.3 VERIFASTSCORE's Integrated Approach

Given advances in instruction following and long-context reasoning in modern LLMs, we investigate whether it is possible to perform claim decomposition and verification in tandem using a single model pass. A central challenge in such an approach is the order of operations: traditional pipelines perform decomposition before retrieval in order to resolve coreferences, identify verifiability, and construct precise search queries. By bypassing decomposition, we must ensure that evidence retrieval remains effective despite potential ambiguity or context dependence in the original response.

Evidence Retrieval Without Claim Decomposition: To retrieve evidence without prior knowledge of claims, we treat each sentence in the model response as a search query. We use the SERPER API[4], as in VERISCORE, to obtain the top 10 search results per sentence. Snippets from all queries are concatenated to form a consolidated evidence context. This method leverages redundancy in web search and the semantic breadth of full-sentence queries to compensate for the lack of explicitly formulated claims.

Simultaneous Claim Decomposition and Verification: With the consolidated evidence and the full model response as inputs, VERIFASTSCORE performs both claim extraction and verification in a single forward pass. This removes the need for intermediate steps and per-claim processing, offering a significantly more efficient pipeline.

Despite the lack of decomposition at the retrieval stage, we find that VERIFASTSCORE maintains strong performance. On average, each response contains ∼23 verifiable claims and is associated with ∼186 evidence snippets, totaling over 4k tokens. These are processed together, with VERIFASTSCORE identifying and verifying each claim directly against the evidence context. As shown in

[4] serper.dev

Table 8 and Table 5, VERIFASTSCORE achieves high factual precision and recall with respect to VERISCORE, while reducing inference time per response by over 6×.

Score calculation: To score a model's response, we adapt the metrics used in VERISCORE. Factual precision of a model response r is defined as:

$$P(r) = \frac{S(r)}{|C|}$$

where C is the set of all verifiable claims in r and $S(r)$ denotes the number of verifiable claims in r that are supported by the consolidated evidence E.

Furthermore, to prevent models from gaming the evaluation, it is important to also consider the number of verifiable claims produced by the model in its response. Thus, we also compute factual recall as follows:

$$R(r) = min(1, \frac{|C|}{K})$$

where K is the median number of verifiable claims across model responses.

We compute $F_1@K(r)$:

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

The final score for a model $\mathcal{M}$ across a dataset X is:

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

2.4 Generating Synthetic Training Data for VERIFASTSCORE

Data Sources and Motivation: To train VERIFASTSCORE, we began with data collected for human evaluation during the development of VERISCORE (see *VeriScore data* in Table 1) consisting of prompts sourced from LongFact (Wei et al., 2024), AskHistorian, ELI5 (Xu et al., 2023), ShareGPT (Chiang et al., 2023), FreshQA (Vu et al., 2023) etc. Early experiments with subsets of this data of different sizes ranging between 0.5K and 5K revealed a clear trend of improved model performance with increasing training data size. Motivated by this, we aimed to scale up training via a synthetic dataset generated using the VERISCORE pipeline. Specifically, we targeted prompts likely to elicit factual claims, enabling us to generate high-quality supervision for claim decomposition and verification.

Section	Content (see Table 10 for full text)
SYSTEM	`You are trying to verify factuality by extracting atomic, verifiable claims. . . Output: <fact_i>: {Supported, Unsupported}. If none, output "No verifiable claim."`
RESPONSE	`. . . In the summer of 1215, a pivotal moment in the course of Western political and legal history unfolded in England. . . .`
EVIDENCE	`The Magna Carta is the charter of English liberties granted by King John on June 15, 1215,. . .`
FACTS	`1. The Magna Carta is considered a pivotal moment in Western political and legal history:` **`Supported`** `2. . . .`

Table 2: Heavily-truncated example of the **prompt format** used to fine-tune VERIFASTSCORE for response-level factuality evaluation. The model receives a SYSTEM prompt, RESPONSE, and EVIDENCE, and it is asked to generate a list of FACTS that consists of verifiable claims and their labels.

Prompt Selection and Dataset Choice: We selected prompts from the *Tulu3 Personas* dataset (Lambert et al., 2025), which contains diverse and high-quality instructions across domains intended to test model's instruction-following ability and adaptation to user intent. This dataset was chosen because it aligns well with our goal of generating model responses that contain factual, verifiable content. We used GPT-4o-mini to identify a subset of prompts likely to elicit factual claims[5]. A manual inspection of 200 filtered prompts showed that over 90% did indeed lead to factually dense responses.

Data Collection Pipeline: From the filtered prompts, we generated responses using a selection of LLMs: GPT-4o (OpenAI et al., 2024), Llama3.1 8B, Gemma2 7B (Team et al., 2024), Mistral-v2 7B (Jiang et al., 2023), and Qwen2 7B (Yang et al., 2024). We collected a total of about 9K prompt-response pairs, ensuring diversity across model architectures and outputs. Each response was pro-

[4]Token lengths are estimated using Llama 3.1 Instruct tokenizer.

[5]More details in A.3

Evidence Granularity	Model	Precision ↑	Recall ↑	Claim Accuracy (%)			Correlation b/w factuality scores ↑
				Correct ↑	Incorrect ↓	Missing ↓	
claim-level	GPT-4o few-shot ICL	0.30	0.31	18.1	**7.2**	73.2	0.28
	VERIFASTSCORE	**0.87**	**0.90**	**71.6**	15.5	**12.3**	**0.87**
sentence-level	GPT-4o few-shot ICL	0.38	0.34	19.3	**8.2**	71.2	0.33
	VERIFASTSCORE	**0.83**	**0.86**	**66**	17.6	**15**	**0.80**

Table 3: **Performance metrics across test splits:** VERIFASTSCORE produces claims and verification labels that are largely consistent with VERISCORE, achieving a strong corelation of **0.80** even with sentence-level evidence, while significantly outperforming the GPT-4o few-shot prompting baseline. *Evidence Granularity* denotes whether evidence was retrieved using full sentences or claims: claim-level includes *VeriScore data* and *Tulu3 Personas* data while sentence-level includes *Tulu3 Personas* data only. *Precision* and *Recall* measure overlap between claims identified by VERIFASTSCORE and VERISCORE. Under *Claim Accuracy*, *Correct*, *Incorrect (Label)*, and *Missing (Claim)* capture agreement, label mismatches, and omissions relative to VERISCORE. A small fraction of instances were marked *Erroneous* (not shown) due to ambiguities in automatic evaluation (see A.2). *Correlation* reports Pearson r between factuality scores (see Eq. 1) assigned by VERIFASTSCORE and VERISCORE. For all Pearson's correlation scores reported in this paper, p<0.001, unless stated otherwise.

cessed through the VERISCORE pipeline to decompose it into verifiable claims and assign verification labels based on evidence retrieved using those claims.

Discrepancy between evidence used during training vs. test-time: A key difference between VERISCORE and VERIFASTSCORE is the granularity at which evidence for claim verification is retrieved. As illustrated in Figure 1, VERISCORE first extracts verifiable claims from the model response and then uses them as search queries to retrieve evidence for claim verification. This is in contrast to the setting in VERIFASTSCORE where evidence for claim verification is collected even before claim decomposition. This is achieved by using the sentences in the model response to collect search results which are then consolidated together for verification.

To train and evaluate VERIFASTSCORE model, we opted to keep the claim-level evidence collected by VERISCORE during the synthetic data generation pipeline outlined earlier for two reasons:

- it allowed us to reuse the unmodified and validated VERISCORE pipeline to ensure high-quality data, and
- it promotes robustness by training the model to verify claims against arbitrary but relevant evidence sets.

Additionally, we also collected sentence-level evidence for the Tulu3-personas subset of our training and test data, in-line with the true setting of VERIFASTSCORE. We evaluate the model's ability to generalize to test-time evidence conditions in Section Section 3.

Evidence Overlap and Label Robustness: To assess whether our training setup, consisting largely of claim-level evidence, is compatible with test-time inference, we analyzed the overlap between evidence gathered using these two different approaches. We found that 17% of claim-level retrieved URLs are also present in the sentence-level evidence, and about 60% of claims have at least one matching URL in the sentence-level evidence.[6]

Empirical proof of soundness: Despite this modest overlap, we observed a strong correlation between verification labels and model-level factuality scores produced by VERIFASTSCORE and VERISCORE (discussed in Section 3), suggesting that the model generalizes well across evidence types. This empirical agreement indicates that the discrepancy in evidence provenance has limited practical impact, and supports the validity of using claim-level supervision for training.

2.5 Evaluation Metrics

Since we motivate VERIFASTSCORE as a speedy replacement to the multi-stage pipeline of VERISCORE, we want its outputs to be closely aligned with outputs from VERISCORE. We evaluate VERIFASTSCORE using three main metrics:

- **Claim Precision:** Computed as the proportion of claims produced by VERIFASTSCORE that are also produced by VERISCORE,

[6]Note that we only count exact URL matches; semantic overlap is likely higher.

- **Claim Recall:** Computed as the proportion of claims produced by VERISCORE that are also produced by VERIFASTSCORE,
- **Claim Accuracy:**
  - **Correct:** Percentage of claims produced by both VERIFASTSCORE and VERISCORE with the same verification label,
  - **Incorrect Label:** Percentage of claims produced by both VERIFASTSCORE and VERISCORE with the differing verification label,
  - **Missing Claim:** Percentage of claims produced by VERISCORE but not by VERIFASTSCORE (analogous to Claim recall),
- **Pearson's Correlation:** Correlation coefficient between factuality scores of model responses produced by VERIFASTSCORE and by VERISCORE.

Automatic Evaluation: During early experiments, we found that when using exact match metrics, comparing model-extracted claims and verification labels directly to reference outputs, even minor surface-level differences (e.g., paraphrasing, slight reordering, or omission of non-essential modifiers) led to false negatives, thereby underestimating the model's actual performance (see Appendix Table 8). For example, VERIFASTSCORE achieved an exact match claim accuracy of 23.7% on the test set, versus 71.6% with paraphrase-aware evaluation.[7]

While exact match metrics were retained during training for their efficiency and stability, we adopted a more nuanced automatic evaluation for the test set. Specifically, we used GPT-4o-mini as a judge to compare the model outputs against gold reference annotations and assess semantic overlap. This helped capture more accurate performance trends. The evaluation setup and prompts are detailed in Appendix A.2.

2.6 Training the VERIFASTSCORE Model

We train VERIFASTSCORE using a two-stage fine-tuning procedure on synthetic data generated via the VERISCORE pipeline. The base model is Llama3.1 8B Instruct, trained first on claim-level evidence and then on a mixture of claim- and sentence-level evidence to better match the test-time setting. This mixed-evidence setup improves robustness and mitigates the observed performance drop when relying solely on sentence-level evidence (see Table 9).

All training examples consist of a model-generated response and retrieved evidence, with supervision derived from VERISCORE's annotated claim lists and verification labels. Details of the multi-stage training setup are provided in Appendix A.1.

3 Results

In this section, we describe the test set performance and computational cost of the trained VERIFASTSCORE model and compare it with VERISCORE.

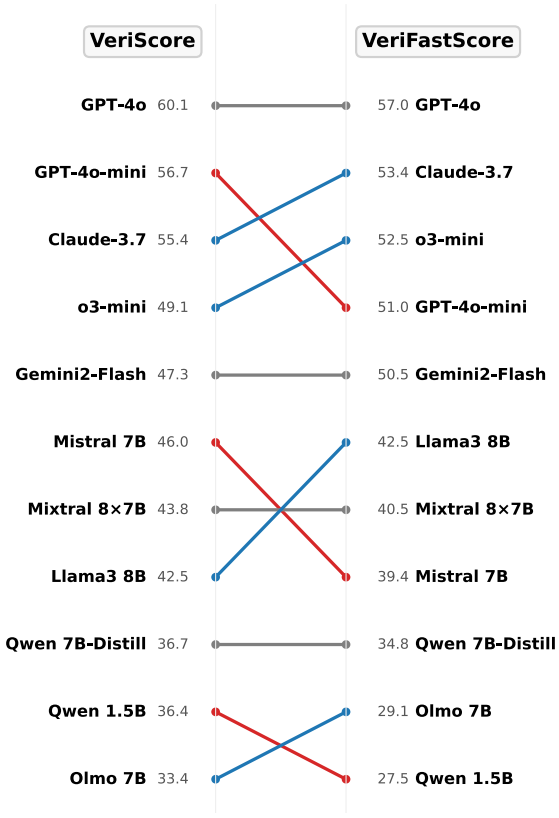

Figure 2: Model rankings produced by VERISCORE and VERIFASTSCORE using a random subset of 100 prompts from the test split of Tulu3-Personas data. The rankings produced by both models are fairly consistent with each other despite some differences in order. There is a clear separation between open-weight and closed-weight models.

[7] We re-report Table 8 with exact match metrics in Appendix.

Example (truncated)	Remarks
Response … The "chimney" and "riverbank" serve as powerful metaphors — the chimney suggesting an unreachable goal while the riverbank signifies his current position … **VERIFASTSCORE**: (Unsupported) The main character in "Araby" imagines climbing a chimney to reach his desire. **VERIFASTSCORE**: (Unsupported) The "riverbank" in "Araby" signifies the main character's current position. VERISCORE: (Unsupported) The chimney symbolizes an unreachable goal. VERISCORE: (Unsupported) The riverbank symbolizes the protagonist's current position.	This example illustrates the benefit of response-level extraction in VERIFASTSCORE over the sentence-focused VERISCORE; VERIFASTSCORE better de-contextualises each claim.
Response … Navigating a rainforest can pose several challenges … unpredictable weather which can lead to sudden rainstorms or flooding … **VERIFASTSCORE**: (Supported) Unpredictable weather in a rainforest can lead to sudden rainstorms. VERISCORE: (Unsupported) Navigating a rainforest can lead to sudden rainstorms.	In this example, VERIFASTSCORE exhibits precise semantic alignment between extracted claim and model response while VERISCORE does not.
Response … JSON-formatted letter quoting Admiral Chester W. Nimitz: "Courage is not the absence of fear …" … **VERIFASTSCORE**: (Unsupported) Admiral Chester W. Nimitz said, "Courage is not the absence of fear, but rather the judgment that something else is more important than fear." VERISCORE: No verifiable claim.	VERIFASTSCORE remains robust even when facts are buried in noisy, non-standard contexts (here, a JSON letter).
Response … The fast-paced editing, over-the-top physical comedy, and clever sound effects have become hallmarks of animation, influencing everything from anime to Pixar films … **VERIFASTSCORE**: (Unsupported) Over-the-top physical comedy is a hallmark of animation. VERISCORE: (Supported) Classic cartoons influenced the use of over-the-top physical comedy in animation.	**Spurious inference via clause-conflation:** VERIFASTSCORE sometimes glues fragments into an unsupported cause–effect claim.

Table 4: **Qualitative analysis:** VERIFASTSCORE exhibits generalization and robustness at test time due to diverse training, but still inherits some undesirable behavior from VERISCORE.

3.1 Main results

High correlation between outputs from VERIFASTSCORE and those from VERISCORE: In Table 8, we observe that VERIFASTSCORE outputs have both a high factual precision and factual recall with outputs from VERISCORE as ground truth irrespective of the provenance of the evidence. There is a drop in claim accuracy when using sentence-level evidence which manifests itself as an increase in incorrect verification label and in missing claims. The higher percentage of missing claims can possibly be mitigated by more training on sentence-level evidence.

GPT-4o Baseline: For a baseline, we prompt GPT-4o with three few-shot demonstrations. Despite the high API cost (roughly **$300** for our evaluation), GPT-4o is about **50% less accurate** than VERIFASTSCORE which can even be run locally. A closer inspection revealed that GPT-4o failed to extract any verifiable claim from about 29% of the instances. These results underscore the complexities involved in simultaneous claim extraction and verification.

3.2 Speedup

	Average time spent (s)	
	Evidence retrieval	**Modeling**
VERISCORE	21.4	83.0
VERIFASTSCORE	13.5	9.5
Speedup	1.9x	9.9x

Table 5: **Execution time:** VERIFASTSCORE spends significantly lesser time than VERISCORE in the modeling stages i.e. extraction and verification stages. Overall, VERIFASTSCORE is 6.6x faster than VERISCORE[8].

We measured the execution time for both the VERIFASTSCORE and VERISCORE pipelines. By retrieving evidence at a coarser granularity, VERIFASTSCORE achieves a 39% reduction in API costs and cuts retrieval time by almost half. Additionally, VERISCORE takes significantly longer due to its sequential claim extraction and verification processes. VERIFASTSCORE is approximately 10 times faster in extracting and verifying claims without sacri-

[8]Estimated from 200 instances

ficing performance (see Table 8). Overall, VERIFASTSCORE achieves a 6.64x speedup compared to VERISCORE. Longer model responses further increase speedup, as VERISCORE's sentence-by-sentence processing is particularly slow and expensive, in-part due to the larger number of SERPER queries VERISCORE makes compared to VERIFASTSCORE (see Table 6).

	Retrieval time (%)	Average SERPER queries	Cost estimate[9] ($)
VERISCORE	20.5	23	$0.01725
VERIFASTSCORE	58.8	14	$0.0105

Table 6: **Cost estimates:** VERIFASTSCORE makes lesser search queries than VERISCORE thereby reducing cost[8].

3.3 Ranking factuality of models with VERIFASTSCORE

We randomly sample 100 prompts from the Tulu3-personas test set to obtain responses from 12 different LLMs:

- **Closed-weight:** GPT-4o, GPT-4o-mini, o3-mini (OpenAI, 2025), Claude-3.7-Sonnet (Anthropic, 2025), Gemini-2.0-Flash (Google, 2025)
- **Open-weight:** Llama3.1 8B, Mistral v0.3 7B, Mixtral v0.1 8x7B, Qwen2.5 1.5B (Qwen et al., 2025), Deepseek-R1-distill Qwen 7B (DeepSeek-AI et al., 2025), Olmo2 7B (OLMo et al., 2025)

As shown in Figure 2, factuality scores produced by VERIFASTSCORE correlate strongly with those from VERISCORE (Pearson $r = 0.94$, p=$1.1e^{-5}$), indicating alignment despite differing approaches. Given that VERIFASTSCORE operates over a single model pass and avoids per-claim retrieval, it delivers these rankings with significantly reduced latency and cost (see Table 5, 6).

4 Runtime Analysis

To more precisely quantify the efficiency difference, we implemented a parallelized version of VERISCORE and compared it directly with VERIFASTSCORE. We distributed sentence-level claim decomposition and claim-level verification across **8 GPUs** on a H100 node.

[9]Estimated using the tier priced at $0.75/1000 queries

Parallelized VERISCORE Implementation: For a 100–instance subset of the test set, this setup reduced the *claim-decomposition* time to an average of **59.7s** (compared to **253s** for the standard single-GPU implementation), with a total modeling time (decomposition + verification) of approximately **235s**.

Comparison with Unparallelized VERIFASTSCORE: By contrast, VERIFASTSCORE processes the same 100 instances in only **29s on a *single* GPU** on the same H100 node, while performing both claim extraction and verification in one pass.

Limits of Parallelization: Although parallelization yields moderate speedups, VERISCORE remains fundamentally constrained by its *sequential pipeline*: (1) claims must first be extracted sentence-by-sentence, and (2) each extracted claim must then be verified individually. These dependencies create repeated tokenization and input-formatting costs, along with inter-process communication overhead, all of which limit scalability and make efficient batching difficult. In contrast, VERIFASTSCORE integrates decomposition and verification into a single forward pass over the entire response and consolidated evidence, eliminating intermediate stages and achieving consistent, end-to-end efficiency gains.

5 Qualitative Analysis

To better understand VERIFASTSCORE's behavior, the first author conducted a small-scale human evaluation of 401 (response, evidence, claim, label) tuples, drawn from 21 test instances in the *Tulu3 Personas* dataset. The evaluation focused on three core aspects of VERIFASTSCORE's output: claim extraction, claim verifiability, and verification accuracy.

Claim grounding and hallucination: All 401 extracted claims were grounded in the original model response, with no instances of hallucinated or fabricated claims. This is a crucial property for a factuality metric, particularly in downstream use cases like hallucination detection and RLHF training where spurious claim hallucination could mislead supervision.

Extraction coverage and omissions: VERIFASTSCORE was generally effective at identifying claims, especially in responses rich with fac-

tual content. However, it showed reduced reliability in responses that interweave factual and non-factual content, or contain few factual claims overall. Claim omission was observed in about 6 out of the 21 evaluated responses, though 4 of these cases involved fewer than four missing claims. Notably, in 3 of the omission cases, VERIFASTSCORE extracted claims only from the final few sentences of the response, missing earlier factual content—suggesting a positional bias or early truncation.

Unverifiable and ill-formed claims: About 42 out of the 401 extracted claims were judged to be unverifiable due to missing key details (e.g., named entities, dates) in the original model response. In rare cases, VERIFASTSCORE also extracted claims from inherently unverifiable sentences, such as advice or vague generalizations. Ill-formed claims, those missing critical contextual elements, were rare (around 4 claims out of the total 401 claims), but they did alter the meaning of the original sentence in notable ways.

Verification accuracy: VERIFASTSCORE was largely accurate in labeling claims against the provided evidence, with an estimated verification error rate of 91.5%. Among the 34 incorrect verification labels, 25 of them were false positives: instances where VERIFASTSCORE incorrectly marked a claim as supported. These errors often arose when partial matches or topical overlaps misled the model, particularly in cases where evidence spanned multiple unrelated contexts or referenced similar entities in different time periods or locations.

Evidence quality and noise: Finally, we observed that the retrieved evidence was often noisy, especially in cases where the original response contained few or vague factual claims. This is likely due to the sentence-level retrieval strategy used by VERIFASTSCORE, which may issue underspecified queries and produce loosely related results. These limitations in evidence quality can contribute both to omission in extraction and errors in verification.

6 Related work

Factuality evaluation via decomposition and verification FACTSCORE computes factual precision over atomic claims (Min et al., 2023), while SAFE uses GPT-4 for claim-wise verification (Wei et al., 2024). VERISCORE extracts only verifiable claims and uses Google Search to verify them (Song et al., 2024). Other systems like FACTOOL (Chern et al., 2023), RARR (Gao et al., 2023), and COVE (Elazar et al., 2021) operate similarly albeit with document-level or retrieval-augmented reasoning.

Decomposition Quality and Claim Granularity. Wanner et al. (2024) introduce DECOMPSCORE to quantify decomposition consistency. Chen et al. (2023) and Yue et al. (2024) report failures tied to vague or anaphoric claims. Hu et al. (2025) catalog common decomposition errors and tradeoffs. VERISCORE explicitly avoids unverifiable content like opinions or suggestions (Song et al., 2024).

End-to-End and Lightweight Approaches. MINICHECK avoids decomposition entirely, using a small model to classify sentence-level factuality (Tang et al., 2024). LLM-OASIS introduces a synthetic benchmark for scalable factuality evaluation (Scirè et al., 2025), while FACTCHECK-GPT trains an end-to-end verifier integrating retrieval and judgment (Wang et al., 2024).

7 Conclusion

We present VERIFASTSCORE, a model that unifies claim decomposition and verification into a single model pass. Trained on synthetic data derived from VERISCORE, our method processes long-form outputs with consolidated evidence and produces fine-grained factuality judgments.

Despite the complexity of the task, VERIFASTSCORE closely matches the accuracy of VERISCORE (Pearson $r = 0.80$), while achieving a 6.6× speedup in evaluation time. It also outperforms strong few-shot prompting baselines such as GPT-4o, and provides interpretable, claim-level outputs suitable for evaluation and alignment.

We release the model and data to support scalable, efficient factuality assessment and to facilitate further research on long-form factuality and reward modeling.

8 Limitations

Latency remains a bottleneck. While VERIFASTSCORE offers substantial speedups over multi-stage pipelines like VERISCORE, it still requires full-sequence generation conditioned on long evidence and response contexts. This makes it less

suited for real-time or low-latency feedback settings such as RLHF. Although we explored a scoring-head variant (Appendix A.3) to reduce inference cost, it struggled to match the accuracy and reliability of the generative approach.

Reliance on synthetic supervision. VERIFASTSCORE is trained on synthetic labels produced by VERISCORE, which, while effective, is not free of noise or bias. Errors in claim decomposition, incomplete evidence, or incorrect verifiability judgments in the teacher outputs may propagate into the learned model, potentially reinforcing systemic weaknesses in downstream predictions.

Retrieval noise and specificity. Our retrieval strategy issues Google search queries for each sentence in the model response, aggregating the results into a shared evidence context. While this aligns with the single-pass structure of VERIFASTSCORE, it can introduce irrelevant or insufficient evidence, especially for vague or underspecified sentences. More adaptive or agentic retrieval, allowing the model to detect failures and issue improved follow-up queries, may yield higher evidence quality.

Lack of explicit rationales. VERIFASTSCORE produces fine-grained factuality labels for claims but does not provide natural language justifications. Although rationale generation adds computational cost, it could improve interpretability and trustworthiness, especially in sensitive domains. Training with rationales, even if only used during finetuning (e.g., as in Qwen3's "reasoning mode" (Yang et al., 2025)), may offer benefits without incurring runtime overhead at inference time. Initial experiments on training a model to reason first using GRPO showed promise.

Noise in evaluation signals. We use GPT-4o-mini to assess label correctness and semantic equivalence, which enables scalable evaluation but introduces potential noise, particularly for nuanced paraphrase detection. This may affect metric estimates such as precision and recall, and obscure more subtle performance differences across models.

Language and configuration limitations. Our experiments focus primarily on English-language outputs. Although some non-English generations were present, we did not conduct language-specific analysis. Additionally, we only minimally modified prompts from VERISCORE and did not exhaustively tune training configurations. Future work could explore multilingual robustness and systematically investigate the effects of prompt and hyperparameter choices.

Ethics Statement

This work involves training and evaluating factuality metrics for long-form language generation. All model outputs analyzed in this study were generated by publicly available language models. The human evaluation was conducted by the first author and involved no sensitive or private user data. No personally identifiable information (PII) was collected, and no crowdworkers or third-party annotators were employed. The synthetic data used to train our model was generated automatically using existing metrics and publicly available datasets. We will release our model and training data upon publication to support reproducibility and future research.

Acknowledgments

We extend our special gratitude to Yixiao Song for sharing data and details about VERISCORE with us. We also extend gratitude to members from the NGRAM lab at UMD for their feedback. This project was partially supported by awards IIS-2046248, IIS-2312949, and IIS-2202506 from the National Science Foundation (NSF).

References

Anthropic. 2025. Claude 3.7 sonnet system card.

Shiqi Chen, Yiran Zhao, Jinghan Zhang, I-Chun Chern, Siyang Gao, Pengfei Liu, and Junxian He. 2023. Felm: Benchmarking factuality evaluation of large language models. *Preprint*, arXiv:2310.00741.

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.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang,

Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, and Chengda Lu... 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Preprint*, arXiv:2102.01017.

Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. 2023. Rarr: Researching and revising what language models say, using language models. *Preprint*, arXiv:2210.08726.

Google. 2025. Gemini 2.0 flash. `https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash`. Accessed: 2025-05-16.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, and Anthony Hartshorn... 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qisheng Hu, Quanyu Long, and Wenya Wang. 2025. Decomposition dilemmas: Does claim decomposition boost or burden fact-checking performance? *Preprint*, arXiv:2411.02400.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, Noah A. Smith, Yizhong Wang, Pradeep Dasigi, and Hannaneh Hajishirzi. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Siyi Liu, Kishaloy Halder, Zheng Qi, Wei Xiao, Nikolaos Pappas, Phu Mon Htut, Neha Anna John, Yassine Benajiba, and Dan Roth. 2025. Towards long context hallucination detection. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7827–7835, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *Preprint*, arXiv:2305.14251.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 olmo 2 furious. *Preprint*, arXiv:2501.00656.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and Alex Renzin... 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Openai o3-mini system card. `https://cdn.openai.com/o3-mini-system-card-feb10.pdf`. Accessed: 2025-05-16.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alessandro Scirè, Andrei Stefan Bejgu, Simone Tedeschi, Karim Ghonim, Federico Martelli, and Roberto Navigli. 2025. Truth or mirage? towards end-to-end factuality evaluation with llm-oasis. *Preprint*, arXiv:2411.19655.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable

claims in long-form text generation. *Preprint*, arXiv:2406.19276.

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

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, and Léonard Hussenot ... 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

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.

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

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024. A closer look at claim decomposition. In *Proceedings of the 13th Joint Conference on Lexical and Computational Semantics (*SEM 2024)*, pages 153–175, Mexico City, Mexico. Association for Computational Linguistics.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, Cosmo Du, and Quoc V. Le. 2024. Long-form factuality in large language models. *Preprint*, arXiv:2403.18802.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. *Preprint*, arXiv:2305.18201.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Zhenrui Yue, Huimin Zeng, Lanyu Shang, Yifan Liu, Yang Zhang, and Dong Wang. 2024. Retrieval augmented fact verification by synthesizing contrastive arguments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10331–10343, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Training VERIFASTSCORE

We follow a two-stage pipeline to train our VERIFASTSCORE model.

- In the **first stage**, we finetune Llama3.1 8B on the combined train split of *VeriScore data* and *Tulu3 Personas* data using claim-level evidence.
- In the **second stage**, we further finetune the best model from stage 1 on the same dataset but with a mixture (60%:40%) of claim-level and sentence-level evidence.

Motivation for multi-stage training: In our synthetic data, ground-truth claim verification labels were obtained from VERISCORE using claim-level evidence. To ensure correctness of labels and model's reliance on evidence provided in the input during generation of output labels, the first stage trains VERIFASTSCORE model using consolidated claim-level evidence.

Drop in accuracy with sentence-level evidence: As seen in Table 8, there was a ∼20% drop in claim accuracy when using sentence-level evidence, which is our true test setting, instead of claim-level evidence. This is possibly due to differences in quality of retrieved search results and length of consolidated evidence depending on

the granularity at which evidence was collected. To mitigate this, the model was further trained on a mixture of claim- and sentence-level evidence in the second stage of training.

Compute details: We finetune our models for 10 epochs on 4 A100 GPUs for about 1-2 day(s). We evaluate the model after every epoch on our validation split and save the best model based on claim accuracy. Test evaluations of the model were performed on a single GH200 GPU node.

A.2 Automatic Evaluation

To evaluate the factuality labels produced by VERIFASTSCORE, we perform automatic evaluation using GPT-4o-mini as an entailment judge. Given an extracted claim and a set of reference claims, the model is prompted to determine whether the claim can be inferred from the reference set. If supported, the model also identifies a minimal subset of reference claims that together entail the extracted claim. More concretely, GPT-4o-mini was prompted to judge factual consistency of each predicted claim against gold claims, and vice versa. Factual precision is computed from predicted claims given the gold claims as the reference set, while factual recall is computed from gold claims given the predicted claims as the reference set. This two-way setup accounts for granularity mismatches (e.g., one composite claim vs. multiple atomic claims), which precludes one-to-one alignment while also ensuring that factual consistency is assessed in a targeted, evidence-grounded manner while minimizing overreliance on superficial textual overlap.

For scoring purposes, we use the original VERISCORE verification labels assigned to the retrieved subset of justifying reference claims. If any of the claims in this minimal set are labeled as Unsupported by VERISCORE, the extracted claim is marked as Unsupported for evaluation. This setup enables us to scale the evaluation of thousands of (claim, evidence) pairs with approximate but consistent supervision.

A.3 VERIFASTSCORER model

To take the cost-efficiency one-step further, we trained a model with a scoring head to directly produce the factuality score given the model response and consolidated sentence-level evidence without going through the whole process of claim decomposition and verification. We envision such a model being used as a reward model in RLHF setups to train models to produce more factual responses.

Model architecture: We experimented with starting from the base model as well as from the best model from our multi-stage training pipeline. We explored two different architecture for the scoring head:

- Linear output layer on top of the last hidden state of the last token in the input prompt
- Linear projection layer on top of a learnable attention layer over the last hidden state of all tokens in the input prompt followed by a linear output layer

To avoid overfitting, we added a dropout layer in both of these configurations and ablated by freezing the language model parameters. We also experimented with a multi-task learning setup with losses computed with both the scoring head output and the decomposition and verification outputs from the language modeling head.

Abysmal generalization: Our best model achieved 75% score accuracy on the train split, 55% on the validation split and 45% on the test split. Due to poor correlation with ground-truth scores, this avenue was abandoned. We leave it for future work to explore more complex scoring head architectures and training configurations.

Table 7: **Automatic Evaluation:** Prompt used for automatic evaluation with GPT-4o-mini. Given an extracted claim and a list of reference claims, the task is to judge whether the extracted claim can be inferred from the provided list of claims. If that is true, the model also needs extract a minimal subset of claims from the provided list of claims which entails the extracted claim.

You are a helpful assistant tasked with judging the factuality of a statement given some evidence by retrieving a minimal set of facts in the evidence that is necessary and sufficient to justify your judgement of the factuality of the statement.
A fact is justifying your judgement if and only if the fact needs to be true for your judgement to be true.
Below are the definitions of the two categories:
Supported: A fact is supported by the evidence if everything in the fact is supported and nothing is contradicted by the evidence. Evidence may contain other information not fully related to the fact.
Unsupported: A fact is unsupported by the evidence if something in the fact is either contradicted by some information in the evidence or cannot be verified from the evidence.
Input format:
Evidence
1. <fact 1 here>
2. <fact 2 here>
...
n. <fact n here>
Statement
<statement here>
Response format:
Thoughts
<your thoughts here>
Justifying Facts
<justifying fact number 1>, <justifying fact number 2>, ..., <justifying fact number k>
Judgement
Supported / Unsupported
Guidelines:
Only mention fact numbers as a comma separated list under justifying facts and not the entire text of the claims.
Do not include facts that are not strictly required to be true to justify your judgement of the factuality of the statement.
If there are no justifying facts, return None.
Claims
1. Colleen Hoover's *Without Merit* book is a work by Colleen Hoover.
2. Several titles by Colleen Hoover have become *New York Times* bestsellers.
3. Colleen Hoover's *Heart Bones* book is a work by Colleen Hoover.
4. Colleen Hoover's *Slammed* series includes 3 books.
5. Colleen Hoover has published 24 books as of February 2023.
6. Colleen Hoover's books include series.
7. Colleen Hoover's *All Your Perfects* book is a work by Colleen Hoover.
8. Colleen Hoover's *Reminders of Him* book is a work by Colleen Hoover.
9. Colleen Hoover's *Hopeless* series includes 2 books.

Automatic Evaluation: *Continued on next page*

10. Colleen Hoover's books have gained immense popularity in recent years.
11. Colleen Hoover is a prolific writer in the New Adult genre.
12. Colleen Hoover's *Layla* book is a work by Colleen Hoover.
13. Colleen Hoover's *It Starts with Us* book is a sequel to *It Ends with Us*.
14. Colleen Hoover's *It Ends with Us* book is a work by Colleen Hoover.
15. Colleen Hoover is a prolific writer in the romance genre.
16. Colleen Hoover's books include standalone novels.
17. Colleen Hoover's books include novellas.
Statement
Colleen Hoover has published a book titled "Hopeless" as part of the "Hopeless" series.

Data Source, Evidence Granularity	Model	Precision ↑	Recall ↑	Claim Accuracy (%)			Correlation b/w factuality scores ↑
				Correct ↑	Incorrect ↓	Missing ↓	
VERISCORE & Tulu3 Personas, claim-level	GPT-4o few-shot ICL	0.10	0.04	3.54	**0.82**	95.64	0.28
	VERIFASTSCORE-*Stage-1*	**0.06**	**0.29**	**24.64**	4.53	**70.84**	0.85
	VERIFASTSCORE	0.05	0.28	23.69	4.35	71.96	**0.87**
Tulu3 Personas, sentence-level	GPT-4o few-shot ICL	**0.15**	0.03	2.61	**0.57**	96.82	0.33
	VERIFASTSCORE-*Stage-1*	0.06	0.19	15.33	3.43	81.24	0.75
	VERIFASTSCORE	0.05	**0.23**	**18.62**	4.18	**77.26**	**0.80**

Table 8: **Exact match metrics for different test splits.** VERIFASTSCORE-*Stage-1* is the best model from Stage 1 training with claim-level evidence. *Evidence Granularity* indicates whether provided evidence was collected using claims or sentences in the model responses as search queries. *Precision* measures how many claims produced by VERIFASTSCORE is also produced by VERISCORE. *Recall* measures how many claims produced by VERISCOREis also produced by VERIFASTSCORE. Under *Claim accuracy*, *Correct* measures the accuracy of the verification label produced by VERIFASTSCORE with label produced by VERISCORE as the ground truth. *Incorrect (Label)* measures how many claims produced by both VERIFASTSCORE and VERISCORE but with differing labels. *Missing (Claim)* measures how many claims produced by VERISCORE but not by VERIFASTSCORE. *Correlation b/w factuality scores* is the Pearson's correlation between factuality scores computed by VERIFASTSCORE and VERISCORE for model responses.

Data Source, Evidence Granularity	Model	Precision ↑	Recall ↑	Claim Accuracy (%)			Correlation b/w factuality scores ↑
				Correct ↑	Incorrect ↓	Missing ↓	
VeriScore data & *Tulu3 Personas*, claim-level	GPT-4o few-shot ICL	0.30	0.31	18.1	**7.2**	73.2	0.28
	VERIFASTSCORE-*Stage-1*	**0.88**	0.87	63.0	21.3	15.1	0.85
	VERIFASTSCORE	0.87	**0.90**	**71.6**	15.5	**12.3**	**0.87**
Tulu3 Personas, sentence-level	GPT-4o few-shot ICL	0.38	0.34	19.3	**8.2**	71.2	0.33
	VERIFASTSCORE-*Stage-1*	0.72	0.69	50.8	20.8	27.7	0.75
	VERIFASTSCORE	**0.83**	**0.86**	**66**	17.6	**15**	**0.80**

Table 9: **Automatic performance metrics across test splits.** VERIFASTSCORE-*Stage-1* refers to the best model trained with claim-level evidence. *Evidence Granularity* denotes whether evidence was retrieved using full sentences or claims. *Precision* and *Recall* measure overlap between claims identified by VERIFASTSCORE and VERISCORE. Under *Claim Accuracy*, *Correct*, *Incorrect (Label)*, and *Missing (Claim)* capture agreement, label mismatches, and omissions relative to VERISCORE. A small fraction of instances were marked *Erroneous* (not shown) due to ambiguities in automatic evaluation (see A.2). *Correlation* reports Pearson r between factuality scores (see Eq. 1) assigned by VERIFASTSCORE and VERISCORE. For all Pearson's correlation scores reported in this paper, p<0.001, unless stated otherwise.

Table 10: **VERIFASTSCORE prompt format:** Sample input and output prompt

You are trying to verify how factual a response is by extracting fine-grained, verifiable claims.
Each claim must describe one single event or one single state (for example, "Nvidia was founded in 1993 in Sunnyvale, California, U.S.") in one sentence with at most one embedded clause.
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 by 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., an embedded clause or a prepositional phrase).
Each fact must be situated within relevant temporal and location details whenever needed.

All necessary specific details-including entities, dates, and locations—must be explicitly named, and verify here means that every detail of a claim is directly confirmed by the provided evidence.
The verification process involves cross-checking each detail against the evidence; a detail is considered verified if it is clearly confirmed by the evidence.

Avoid extracting stories, personal experiences, hypotheticals (e.g., those using "would be" or the subjunctive mood), subjective opinions, suggestions, advice, instructions, or similarly non-factual content; however, biographical, historical, scientific, and similar texts are acceptable.
Also, ignore any listed references.

For each extracted claim, classify it as follows:

***Supported**: Every detail of the claim (including entities, dates, and locations) is directly confirmed by the provided evidence with no contradictions.*
***Unsupported**: One or more details of the claim are either missing from or contradicted by the provided evidence, even though the claim remains verifiable using external sources.*

You do not need to justify what you extract.

Output format:
<fact 1>: <your judgment of fact 1>
<fact 2>: <your judgment of fact 2>
...
<fact n>: <your judgment of fact n>

If no verifiable claim can be extracted, simply output "No verifiable claim."
Response:
Historical Context and Significance: In the summer of 1215, a pivotal moment in the course of Western political and legal history unfolded in England. Amidst mounting discontent between King John and his barons, a document was drafted in Runnymede, a meadow along the River Thames. This document, known as the Magna Carta, was the first significant attempt to limit the power of a monarch by law, and it marked a crucial step towards modern representative democracy.

The Heart of the Matter: The Magna Carta's Provisions: The Magna Carta, officially known as the "Great Charter", was a compact between King John and his barons, sealing a series of agreements that aimed to secure political rights and liberties. Some of the most influential clauses protected church rights, granted trial by jury, limited the amount of unjust taxes, and established the principle that no free man could be imprisoned or exiled without a fair trial. These provisions laid the foundation for the rule of law, constitutional government, and individual rights that are widely upheld today.

VERIFASTSCORE prompt format: *Continued on next page*

Lasting Impact and Legacy: The Magna Carta has left an indelible mark on the annals of history and continues to serve as a symbol of liberty and the rule of law in the modern world. Although it has undergone numerous reissues and modifications throughout the centuries, the original 1215 charter remains a significant historical artifact. It has been celebrated as a precursor to the Bill of Rights in the United States, the Universal Declaration of Human Rights, and many other international charters and constitutions. More than eight centuries after its inception, the Magna Carta's enduring legacy continues to inspire efforts towards justice, freedom, and governance worldwide.

Evidence
The Magna Carta is the charter of English liberties granted by King John on June 15, 1215, under threat of civil war.
The Magna Carta (or Great Charter) was written in Latin and was effectively the first written constitution in European history.
This significant moment, the first time a ruling monarch had been forcibly persuaded to renounce a great deal of his authority, took place at Runnymede, a ...
It promised the protection of church rights, protection for the barons from illegal imprisonment, access to swift and impartial justice, and limitations on ...
Eight hundred years ago today, King John of England sealed the Magna Carta, a groundbreaking legal document that served as the foundation for ...
The Magna Carta, officially granted by King John of England on 15 June 1215, stands as one of the most influential and pivotal documents in ...
On June 15, 1215, in a field at Runnymede, King John affixed his seal to Magna Carta. Confronted by 40 rebellious barons, he consented to their ...
The Department of Defense undertook preparation ofa history ofthe 11. September 2001 attack on the Pentagon at the suggestion of Brig. Gen. John S. Brown, ...
hours This course examines the causes and consequences of globalization. Issues are examined from a changing historical context of economy, politics, and ...

... along the south bank of the Thames River, in a meadow called Runnymede. King John affixed his seal to the Magna Carta on June 15, 1215. The document then ...
On June 15, 1215, John met the barons at Runnymede on the Thames and set his seal to the Articles of the Barons, which after minor revision was ...
The result was the Magna Carta, which was signed on June 15, 1215, at Runnymede, a meadow by the River Thames. Key Provisions of the Magna Carta. While the ...
They accrued at Runnymede, a meadow alongside the banks of the River Thames, on a fateful day in June. They demanded that King John renowned their rights and ...
Magna Carta originated in 1215 as a peace treaty between King John and a group of rebellious barons · The original document was written in Latin ...
A deal was brokered between King John and the barons who agreed to meet on 10th June 1215 at Runnymede, a large flat expanse of meadow near Egham, ideally ...
In May the barons took London, and withdrew their homage and fealty. In the middle of June, 1215, on a meadow, Runnymede, along the River Thames the rebellious ...
McKechnie's introduction to Magna Carta: A Commentary on the Great Charter of King John, with an Historical Introduction, by William Sharp McKechnie.
The story culminates at Runnymede, a meadow on the banks of the River Thames. ... by King John in 1215, was a document of unprecedented ... its origins in the ...

The Magna Carta is the charter of English liberties granted by King John on June 15, 1215, under threat of civil war.

VERIFASTSCORE prompt format: *Continued on next page*

Magna Carta was issued in June 1215 and was the first document to put into writing the principle that the king and his government was not above the law.
In the early 17th century, Magna Carta became increasingly important as a political document in arguments over the authority of the English monarchy.
June 15, 2015 marks the 800th anniversary of the Magna Carta, a pioneering legal document that served as the foundation for American democracy and individual ...
The document, often referred to simply as Magna Carta, did indeed innumerate several various rights and liberties, as well as limitations on the King's powers ...
Magna Carta is significant because it is a statement of law that applied to the kings as well as to his subjects.
The Magna Carta was the first step of the English Monarchy giving up some power and allowing the Judicial Branches have more power. In theory it ...
The Magna Carta established that the king and all individuals were subject to the law, codified the principles of due process and trial by jury.
Magna Carta marked an important step, in the process by which England became a nation; but that step was neither the first nor yet the final one.2. In ...

The Magna Carta, Latin for "Great Charter," is one of the most famous documents in history. It was originally issued in 1215 during the reign of ...
A law made by the king in one national assembly might be repealed by the king in another; whereas the Great Charter was intended by the barons to be ...
MCMXIV. "vo. Page 7. MAGNA CARTA. A COMMENTARY ON THE GREAT. CHARTER OF KING JOHN. WITH AN. HISTORICAL INTRODUCTION. BY. WILLIAM SHARP McKECHNIE ... Great Charter ...
Magna Carta: A Commentary on the Great Charter of King John, with an Historical Introduction, by William Sharp McKechnie (Glasgow: Maclehose, 1914). Copyright.
Full text of "Magna carta : a commentary on the Great Charter of King John with an historical introduction" ... Great Charter was intended by the barons to be ...
a Commentary on the Great Charter of King John (2nd edn., Glasgow, 1914). ... have all their rights and liberties according to the great charter of England'.
During his final year of public service, Coke forced a reconsideration of the Great Charter's overall meaning and promise. Would the liberties of the.
But minding and mining the Magna Carta became a fun task, not just to learn about the Great Charter, but to glean from his notes or his citations the major ...
called "the statute r :lied the Great Charter of the Liberties of England." The simple fact that the language of Maps Carta bestowed its benefits on "free ...
Indeed, it was named the 'Great Charter', several years after the initial 1215 settlement, not in recognition of its importance, but because it ...

It promised the protection of church rights, protection from illegal imprisonment, access to swift justice, and, most importantly, limitations on taxation and ...
Of enduring importance to people appealing to the charter over the last 800 years are the famous clauses 39 and 40: "No free man shall be seized, imprisoned, ...
The two most famous clauses; establishing the right of all to be judged by their equals, and outlawing imprisonment of free men without a trial, were clauses 39 ...
Among these are the principle of no taxation without representation and the right to a fair trial under law.
Clause 12 prevented kings from imposing taxes 'without common counsel'. The principle – that taxation must be by consent – became fixed in English politics.

VERIFASTSCORE prompt format: *Continued on next page*

The English church shall be free and shall have its rights intact and its liberties uninfringed upon. And thus we will that it be observed.
... No free man shall be seized or imprisoned, or stripped of his rights or possessions, or outlawed or exiled, or deprived of his standing in any other way, nor.
+ (39) No free man shall be seized or imprisoned, or stripped of his rights or possessions, or outlawed or exiled, or deprived of his standing in any way, nor ...
These guaranteed the Church the freedom to handle its business without interference from the throne; that no free man could be imprisoned or outlawed except by ...
One of the most important, and often quoted, provisions stated that "no freeman shall be seized, imprisoned, dispossessed, outlawed, or exiled, or in any way ...

The Federalist # 78 states further that, if any law passed by Congress conflicts with the Constitution, "the Constitution ought to be preferred to the statute, ...
The Bill of Rights is a founding documents written by James Madison. It makes up the first ten amendments to the Constitution including freedom of speech ...
Article III of the Constitution establishes the federal judiciary. Article III, Section I states that "The judicial Power of the United States, shall be vested ...
The Bill of Rights became a document that defends not only majorities of the people against an overreaching federal government but also minorities against ...
"[A] bill of rights is what the people are entitled to against every government on earth, general or particular, and what no just government ...
The higher law, reciprocity and mutuality of obligations, written charters of rights, the right to be consulted on policy and to grant or refuse one's consent, ...
The. Supreme Court upheld the law, ruling that Congress has the power to enact laws that directly affect the acts of individuals, thereby making ...
The Universal Declaration of Human Rights is generally agreed to be the foundation of international human rights law. Adopted in 1948, the UDHR has inspired ...
Article III of the Constitution establishes and empowers the judicial branch of the national government.
Today, the rule of law is often linked to efforts to promote protection of human rights worldwide.

Magna Carta was the result of the Angevin king's disastrous foreign policy and overzealous financial administration.
Magna Carta's clauses provided the basis for important principles in English law developed in the fourteenth through to the seventeenth century.
Magna Carta was written by a group of 13th-century barons to protect their rights and property against a tyrannical king.
Though the king never meant to keep his promises, Magna Carta survived. Down through the centuries, it has been a symbol of opposition to arbitrary government.
The Magna Carta is the charter of English liberties granted by King John on June 15, 1215, under threat of civil war.
The Magna Carta (or Great Charter) was written in Latin and was effectively the first written constitution in European history.
The Magna Carta, the medieval English historic legal document that is seen as the origin of many modern-day legal rights and constitutional principles.
It is a story that runs 800 years forward and is still unfolding. It is the story of our rule of law tradition and of how our American system of government is ...
exploration of how historical events have left an indelible mark on present-day national political systems, employing a comparative analysis of the British ...

VERIFASTSCORE prompt format: *Continued on next page*

None of the original 1215 Magna Carta is currently in force since it has been repealed; however, four clauses of the original charter are enshrined in the 1297 ...
On the back of this charter issued in the name of King Æthelbald of Mercia in 736, it is still possible to see the impressions from where the document was ...
The king and the rebel barons negotiated a peace settlement in June 1215. The king agreed to accept the terms of Magna Carta, which is dated 15 June 1215.
David Carpenter considers the huge significance of the 13th-century document that asserted a fundamental principle – the rule of law.
That any copies of the 1215 Magna Carta survive is even more remarkable given the number of times it was reissued—most copyholders would ...
Magna Carta, charter of English liberties granted by King John on June 15, 1215, under threat of civil war and reissued, with alterations, in 1216, 1217, and ...
Such an evaluation will not prove that everything commonly said in praise of the 1215 Charter is true. Some of it is myth. But it is not all myth. Many of the ...
Representing a would-be peace treaty between the king and rebellious nobles, the 1215 Charter did not survive its year of issue. Pope Innocent ...
The Magna Carta is a document written in 1215 that limited the power of the English king. It remains important today as a symbol of protest against ...
This version is arguably just as significant as the 1215 charter in that it remained materially unchanged through various reissues and confirmations until in ...

The UDHR is widely recognized as having inspired, and paved the way for, the adoption of more than seventy human rights treaties, applied today on a permanent ...
It directly inspired the development of international human rights law, and was the first step in the formulation of the International Bill of Human Rights, ...
The Universal Declaration of Human Rights, which was adopted by the UN General Assembly on 10 December 1948, was the result of the experience of the Second ...
These three documents, known collectively as the Charters of Freedom, have secured the rights of the American people for more than two and a quarter centuries.
The Magna Carta established the rule of law and the idea that all citizens, including those in power, should be fairly and equally ruled by the law.
This exhibit celebrates the leadership of Eleanor Roosevelt in writing the Universal Declaration of Human Rights as we mark the 70 th anniversary of its ...
The Universal Declaration of Human Rights has been the centrepiece of the modern international law of human rights for more than sixty years. If anything ...
Everyone is entitled to all the rights and freedoms set forth in this Declaration, without distinction of any kind, such as race, colour, sex, language, ...
Known today as the Cyrus Cylinder, this ancient record has now been recognized as the world's first charter of human rights. It is translated into all six ...
The Universal Declaration of Human Rights describes human rights of all people around the world. D. The Bill of Rights to the U.S. Constitution ...

In 1215, when King John confirmed Magna Carta with his seal, he was acknowledging the now firmly embedded concept that no man–not even the king ...
As we celebrate 800 years of Magna Carta, it's timely to reflect on the bloodshed and tyranny behind its creation and the reasons for its ...
June 15, 2015 marks the 800th anniversary of the Magna Carta, a pioneering legal document that served as the foundation for American democracy and individual ...

VERIFASTSCORE prompt format: *Continued on next page*

The story of Magna Carta began in 1215, but it continues eight centuries forward and is still unfolding. It is the story of modern constitutional government and ...
Magna Carta established a number of very important political principles that framed and informed our discussions about human rights in the centuries that ...
The Magna Carta is the charter of English liberties granted by King John on June 15, 1215, under threat of civil war.
A Magna Carta moment. After eight centuries the revered document of liberty still grips the political imagination, says David Hayes in London.
Magna Carta was not made law; it was more of a working document setting out how the country would be run.

Facts
The Magna Carta serves as a symbol of the rule of law.: Supported
The Magna Carta is celebrated as a precursor to the Bill of Rights in the United States.: Supported
The Magna Carta is officially known as the "Great Charter".: Unsupported
The Magna Carta's legacy continues to inspire efforts towards freedom worldwide.: Unsupported
The Magna Carta's legacy continues to inspire efforts towards justice worldwide.: Supported
The provisions of the Magna Carta laid the foundation for constitutional government.: Supported
Individual rights are widely upheld today.: Unsupported
The Magna Carta has influenced many international charters and constitutions.: Supported
The Magna Carta protected church rights.: Supported
The Magna Carta established the principle that no free man could be imprisoned or exiled without a fair trial.: Supported
The provisions of the Magna Carta laid the foundation for individual rights.: Supported
The Magna Carta was the first significant attempt to limit the power of a monarch by law.: Supported
The Magna Carta has left an indelible mark on the annals of history.: Supported
The Magna Carta's legacy continues to inspire efforts towards governance worldwide.: Supported
The provisions of the Magna Carta laid the foundation for the rule of law.: Supported
The Magna Carta is celebrated as a precursor to the Universal Declaration of Human Rights.: Supported
The drafting of the Magna Carta was amidst mounting discontent between King John and his barons.: Supported
The Magna Carta was a compact between King John and his barons.: Supported
Constitutional government is widely upheld today.: Unsupported
The Magna Carta is considered a pivotal moment in Western political and legal history.: Supported
The Magna Carta has influenced the Universal Declaration of Human Rights.: Supported
The Magna Carta took place in England.: Supported
The Magna Carta aimed to secure political rights and liberties.: Supported
The Magna Carta granted trial by jury.: Supported
The Magna Carta was incepted more than eight centuries ago.: Unsupported
The drafting of the Magna Carta took place in the summer of 1215.: Unsupported
The Magna Carta was drafted in the summer of 1215.: Unsupported
The Magna Carta serves as a symbol of liberty.: Supported
The Magna Carta has influenced the Bill of Rights in the United States.: Supported
The rule of law is widely upheld today.: Unsupported
The original 1215 charter of the Magna Carta remains a significant historical artifact.: Unsupported
The Magna Carta was drafted in Runnymede, a meadow along the River Thames.: Supported

VERIFASTSCORE prompt format: *Continued on next page*

The Magna Carta marked a crucial step towards modern representative democracy.: Supported The Magna Carta limited the amount of unjust taxes.: Unsupported The Magna Carta is celebrated as a precursor to many international charters and constitutions.: Supported The original 1215 charter of the Magna Carta has undergone numerous reissues and modifications throughout the centuries.: Supported
VERIFASTSCORE prompt format

Table 11: **Automatic Filtering:** Prompt used for automatic filtering of Tulu3-personas prompts with GPT-4o-mini.

You are a helpful assistant. Given a prompt, you are to judge whether a response to the prompt would contain fine-grained factually verifiable claims. 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 is considered factually verifiable. Biographical, historical, scientific, and other such texts are not personal experiences or stories.

Here are some examples of responses and some fine-grained factually verifiable claims present in those responses:

Text:
The sweet potato or sweetpotato (Ipomoea batatas) is a dicotyledonous plant that belongs to the bindweed or morning glory family, Convolvulaceae. Its large, starchy, sweet-tasting tuberous roots are used as a root vegetable. 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:
After the success of the David in 1504, Michelangelo's work consisted almost entirely of vast projects. He was attracted to these ambitious tasks while at the same time rejecting the use of assistants, so that most of these projects were impractical and remained unfinished.
Sentence to be focused on:
After the success of the David in 1504, Michelangelo's work consisted almost entirely of vast projects.
Facts:

- *Michelangelo achieved the success of the David in 1504.*
- *After 1504, Michelangelo's work consisted almost entirely of vast projects.*

Text:
After the success of the David in 1504, Michelangelo's work consisted almost entirely of vast projects. He was attracted to these ambitious tasks while at the same time rejecting the use of assistants, so that most of these projects were impractical and remained unfinished. In 1504 he agreed to paint a huge fresco for the Sala del Gran Consiglio of the Florence city hall to form a pair with another just begun by Leonardo da Vinci. Both murals recorded military victories by the city (Michelangelo's was the Battle of Cascina), but each also gave testimony to the special skills of the city's much vaunted artists.
Sentence to be focused on:

Automatic Filtering: *Continued on next page*

In 1504 he agreed to paint a huge fresco for the Sala del Gran Consiglio of the Florence city hall to form a pair with another just begun by Leonardo da Vinci.
Facts:

- *In 1504, Michelangelo agreed to paint a huge fresco for the Sala del Gran Consiglio of the Florence city hall.*
- *Around 1504, Leonardo da Vinci just began with a mural for the Florence city hall.*

Text:
After the success of the David in 1504, Michelangelo's work consisted almost entirely of vast projects. He was attracted to these ambitious tasks while at the same time rejecting the use of assistants, so that most of these projects were impractical and remained unfinished. In 1504 he agreed to paint a huge fresco for the Sala del Gran Consiglio of the Florence city hall to form a pair with another just begun by Leonardo da Vinci. Both murals recorded military victories by the city (Michelangelo's was the Battle of Cascina), but each also gave testimony to the special skills of the city's much vaunted artists. Leonardo's design shows galloping horses, Michelangelo's active nudes—soldiers stop swimming and climb out of a river to answer an alarm.
Sentence to be focused on:
Both murals recorded military victories by the city (Michelangelo's was the Battle of Cascina), but each also gave testimony to the special skills of the city's much vaunted artists.
Facts:

- *Michelangelo's murals for the Florence city hall recorded military victories by the city.*
- *Leonardo da Vinci's murals for the Florence city hall recorded military victories by the city.*
- *Michelangelo's mural for the Florence city hall was the Battle of Cascina.*

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. 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. 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.

Text:
I was a catholic school kid, educated by nuns and somehow on a spring day in 1972, I was called down to the principal's office by Sister Mary Roberts, who informed me that I had gained admission to Stuyvesant High School. I was excited to be freshman in one of New York City's elite public schools but soon came to realize that my catholic school education did not provide the groundwork for abstract concepts like science and algebra. My parochial education in Science at St. Joseph's was essentially "God made it, what else do you need to know?"

Automatic Filtering: *Continued on next page*

Sentence to be focused on:
I was excited to be freshman in one of New York City's elite public schools but soon came to realize that my catholic school education did not provide the groundwork for abstract concepts like science and algebra.
Facts:

- *Stuyvesant High School is in New York City.*
- *Stuyvesant High School is an elite high school.*
- *Stuyvesant High School is a public school.*
- *In 1972, St. Joseph's catholic school education did not provide the groundwork for abstract concepts like science and algebra.*

Text:
Ãcariya Mun related the story of a dhutanga monk (ascetic monk) who inadvertently went to stay in a forest located next to a charnel ground. He arrived on foot at a certain village late one afternoon and, being unfamiliar with the area, asked the villagers where he could find a wooded area suitable for meditation. They pointed to a tract of forest, claiming it was suitable, but neglected to tell him that it was situated right on the edge of a charnel ground. They then guided him to the forest, where he passed the first night peacefully. On the following day he saw the villagers pass by carrying a corpse, which they soon cremated only a short distance from where he was staying.
Sentence to be focused on:
They then guided him to the forest, where he passed the first night peacefully.
Facts:
No verifiable claim.

Input format:
Prompt
<prompt here>

Output format:
Thoughts
<your thoughts here>

Judgement
Yes/No