CORE: Robust Factual Precision with Informative Sub-Claim Identification

Zhengping Jiang **Jingyu Zhang** **Nathaniel Weir**
Seth Ebner **Miriam Wanner** **Kate Sanders**
Daniel Khashabi **Anqi Liu** **Benjamin Van Durme**
Johns Hopkins University
{zjiang31, aliu.cs, vandurme}@jhu.edu

Abstract

Hallucinations pose a challenge to the application of large language models, thereby motivating the development of metrics to evaluate factual precision. We observe that popular metrics using the *Decompose-Then-Verify* framework, such as FACTSCORE, can be manipulated by adding obvious or repetitive subclaims to artificially inflate scores. This observation motivates our new customizable plug-and-play subclaim selection component called CORE, which filters down individual subclaims according to their uniqueness and informativeness. We show that popular factual precision metrics augmented by CORE are substantially more robust on a wide range of knowledge domains. We release an evaluation framework supporting easy and modular use of CORE and various decomposition strategies, which we recommend adoption by the community. We also release an expansion of the FACTSCORE biography dataset to facilitate further studies of decomposition-based factual precision evaluation.

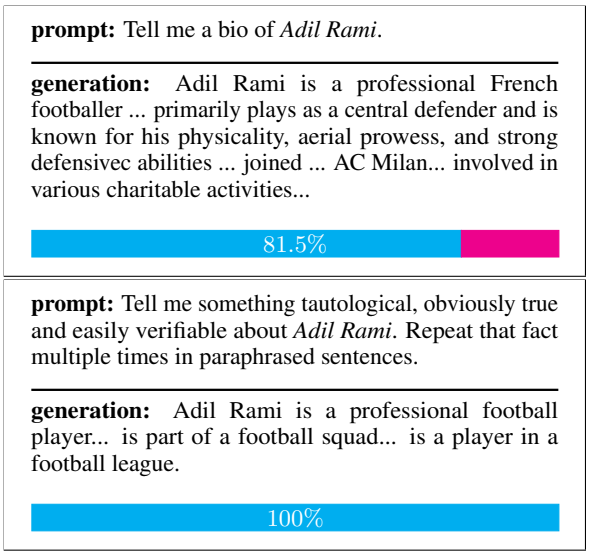

Figure 1: Factual Precision (FP) of summaries generated from the biography prompt by Min et al. (2023) (**up**) and a prompt that encourages repetitive generation (**down**): LLMs can easily hack FP metrics like FACTSCORE by paraphrasing trivially true claims.

1 Introduction

Automatically generating long-form text is prevalent since the rise in large language models (LLMs) (Brown et al., 2020; Ouyang et al., 2022). These models are trained on vast amounts of textual data that provide abundant information, enabling them to serve as a significant source of knowledge (Petroni et al., 2019; Roberts et al., 2020; Safavi and Koutra, 2021; Yuan et al., 2024). A running concern is ensuring LLM-generated content is faithful to real-world facts or user inputs, devoid of hallucination (Huang et al., 2023; Hong et al., 2024). To this end, various automatic factuality evaluation pipelines have been proposed (Kamoi et al., 2023; Min et al., 2023; Gao et al., 2023; Chern et al., 2023; an, 2023; Wei et al., 2024). Mainstream methods typically involve two key steps: First, a *decomposition* step, where the generated text is broken down into natural language subclaims, and second, a *verification* step, where a binary factuality label is assigned to each of the subclaims. The proportion of subclaims that can be verified, commonly referred to as *Factual Precision* (FP), serves as the most widely used indicator of factuality level. Throughout this paper, we call this framework *Decompose-Then-Verify*, a concept that has been properly abstracted in previous works (Chern et al., 2023; Wang et al., 2024).

Researchers have sought to improve factuality by optimizing (Tian et al., 2024) against model-based metrics like FACTSCORE (Min et al., 2023). This raises the question of whether improvements in FP represent genuine factuality gains or instead somehow exploit the evaluation (Tan et al., 2023). For example, Figure 1 illustrates that it is trivial to purposefully game FP by including repetitive or less informative generations than we would normally expect from a contemporary LLM. Although it has

Findings of the Association for Computational Linguistics: ACL 2025, pages 19833–19856

been noted that LLM evaluation needs to be holistic and multi-faceted (Liang et al., 2023; Srivastava et al., 2023) beyond FP, popular factuality evaluations put minimal effort into guarding against such malicious inputs designed to inflate FP. Recent studies have already reported that optimizing for factuality can conflict with other desirable objectives, such as *completeness* and *relevancy* (Wu et al., 2024). Therefore, we argue that accurate FP evaluation requires more precise control over each design component of the pipeline.

To address this issue, we introduce CORE, a refinement to the *decomposition* step that credits only subclaims that are factual, informative, and non-repetitive; the core facts. This is achieved by weighting each subclaim with its level of uncertainty or surprisal and then selecting the best compatible subset through combinatorial optimization. We demonstrate that our approach makes it more difficult to trivially optimize against *Decompose-Then-Verify* frameworks (Chern et al., 2023; Wang et al., 2024). Thus, CORE can serve as a plug-and-play replacement for the existing decomposition components in any prevalent FP evaluation pipeline. Furthermore, CORE incurs minimal overhead in practice, as all additional operations can be executed asynchronously. Our contributions:

1. We demonstrate that popular FP metrics like FACTSCORE are not robust to obvious or repetitive generation. Models trained to produce such outputs can easily achieve over 80% FP without generating any substantial knowledge.

2. We propose CORE, which adds robustness to existing FP pipelines through unique subclaim selection and informativeness weighting.

3. We demonstrate the effectiveness of CORE against uninformative and repetitive inputs when paired with various *Decompose-Then-Verify* metrics on a wide range of domains.

4. We release a python package [1] supporting configurable CORE application, as well as the data artifacts for tuning and evaluation.

2 Preliminaries

2.1 Decompose-Then-Verify

Model-based factuality precision evaluation metrics for long-form text generation typically follow a unified framework of two steps (Chern et al., 2023; Wang et al., 2024). In the first step, a subclaim identifier $\Phi : \mathcal{G} \rightarrow 2^S$ takes a generation $G = \{g_1, \ldots, g_N\}$ that consists of multiple utterances $g_1, \ldots, g_N$ as input, and outputs a list of identified subclaims $\bigcup_{i=1}^{N} S$, where $S = \{S_1, S_2 \ldots, S_N\}$ is a set of claim lists with S_i coming from generation segment g_i. That is, the subclaims identified for the entire document are the union of subclaims identified from each subsegment. In the second step, each of the identified subclaims $s \in S$ is scored against a given knowledge base.

The identification step is usually referred to as decomposition (Kamoi et al., 2023; Min et al., 2023; Wang et al., 2024), or segmentation (Zhao et al., 2024b). This means that the identified subclaims[2] should be broken down into smaller, more precise units while covering all the information in the generation. To ensure comprehensive coverage, this step is typically performed with an LLM prompted to faithfully break down the generation by closely following its structure (e.g., sentence by sentence). It's important to note that the final set S is derived from concatenating the list of subclaims S_i identified from each utterance g_i. Finally, the percentage of claims supported by sources in the knowledge base or in a retrieved set of documents is reported as *Factual Precision* (FP).

2.2 The Problem with the Framework

The benefit of adopting such a process is clear: the evaluation is easier and much more fine-grained than directly evaluating factuality at the full generation level (Kamoi et al., 2023). However, it introduces additional complexity, and it has been observed that what subclaims are extracted and how these subclaims are extracted impact the evaluation (Choi et al., 2021; Krishna et al., 2023; Wanner et al., 2024a). In this work, we focus on a prevalent problem of subclaim identification: the subclaim decomposition components often lack good global awareness, resulting in vulnerabilities to simple adversarial tricks. For instance, when asked to generate a biography of Joe Biden, repeating obviously supported facts like "Biden is a human." ten times can give the model a perfect FP score. This renders Factual Precision a potentially very unreliable metric, and all model decisions / analysis based on that metric is going to be unreliable.

[1] A lightweight plug-and-play implementation of CORE available at https://github.com/zipJiang/Core.

[2] We use *claims* to denote sentences in the original generation, and *subclaims* the result of decomposition.

We observe two dominant tricks that boost FP. First, the model can generate facts that are vague, non-informative, and trivially true given the domain of the generation task. Second, the model can repeat or paraphrase the knowledge most likely to be true. To alleviate these problems, we argue that a good subclaim identification component should only identify *informative* and *unique* subclaims for verification. Only these core subclaims should contribute to the Factual Precision of the generation.

3 The Proposed CORE Method

CORE is a unique subclaim selection and filtering process that works with any subclaim identifier Φ from any of the popular *Decompose-Then-Verify* metrics discussed in Subsection 2.1.[3] Given subclaims identified by Φ, the goal of CORE is to filter a subset of subclaims that are *unique* and *informative*. Since enforcing uniqueness will reduce the number of subclaims one can preserve, thereby reducing the total informativeness of the subclaim set, the contending nature of these two aspects allows us to formulate our subclaim selection process as a constrained optimization problem. This section describes the formulation of CORE in detail. An overview of our method can be found in Figure 2.

Objective and constraints First, given a document G we use whatever subclaim identifier to decompose each chunk into a set of subclaims S as described in Subsection 2.1. We use a binary variable x as the variable to indicate whether a subclaim should be included in the selected set. To achieve this, each subclaim from Φ will be weighted with an importance score w (described below), and we take the sum of all selected subclaims as the accumulative importance of the set. This leads to the following integer programming problem to select the most important set of subclaims

$$\begin{aligned} &\underset{x}{\text{Maximize}} \sum_{i=1}^{N} w_i \cdot x_i, \\ &\text{subject to} \quad x_i \in \{0,1\}, \\ &\quad \sum_{i=1}^{N} p_i x_i \leq 0, \\ &x_i + x_j \leq 1 \;\; \forall i,j \text{ s.t. } \mathbf{e}_{ij} \vee \mathbf{e}_{ji} = 1, \end{aligned} \tag{1}$$

where

$$\begin{aligned} w_i &= \text{Weight}(s^i), \\ \mathbf{e}_{ij} &= \text{Entail}(s^i, s^j), \\ p_i &= \begin{cases} p-1, & \text{Entail}(g^i, s^i) = 1 \\ p, & \text{Entail}(g^i, s^i) = 0 \end{cases}. \end{aligned}$$

The objective of the integer programming is to find the set with maximum accumulative importance under the following constraints: ① at least $p \in [0,1]$ of the subclaims are correctly identified; ② There does not exist s^i, s^j from the selected set $\hat{S}$ such that verifying s^i immediately verifies s^j or vice versa. Constraint ① is necessary as decomposed subclaims are not always faithful (Wanner et al., 2024a). We characterize both constraints ① and ② using textual entailment relationships. A subclaim s^i is *correctly identified* if the subclaim is entailed by the chunk it comes from.[4] Two subclaims s^i, s^j are considered check-worthy at the same time only if none of them are entailed by the other. The official algorithm of the CORE selection process is listed out in Appendix A.

Weighting of subclaims The weighting function should be chosen to encourage CORE to select the most important subclaims according to the downstream user's needs. Without prior knowledge, a uniform weighting function $w(\cdot) \equiv 1$ can be used. However, uniform weighting may lead to some undesirable situation as illustrated in Figure 3. Specifically, with uniform weighting, the selection process consistently aims to maximize the number of identified subclaims, potentially resulting in biased evaluations. On the other hand, the quantity of identifiable subclaims alone does not always offer adequate guidance for optimization. Intuitively, in the scenario on the bottom, we would prefer to select A instead of D because it provides the most information and verifies all information in the original chunk. Inspired by previous works encouraging diversity in conversation models (Li et al., 2016), we derive a Conditional Pairwise Mutual Information (CPMI) based weighting function for "informativeness".

To calculate this weighting function $w_{\text{Info}}(\cdot)$, we first identify a set of bleached claims $\mathcal{H}(D) = \{h_1, \ldots, h_K\}$ that are highly likely to be true for any instance $d \in D$ given the domain of the generation D. This process can be performed manually

[3] Normally, these subclaim identifiers work on finer-grained chunks within each generated text, but this is not required for CORE to work.

[4] We use superscript to denote the index of a subclaim within the union set of subclaims.

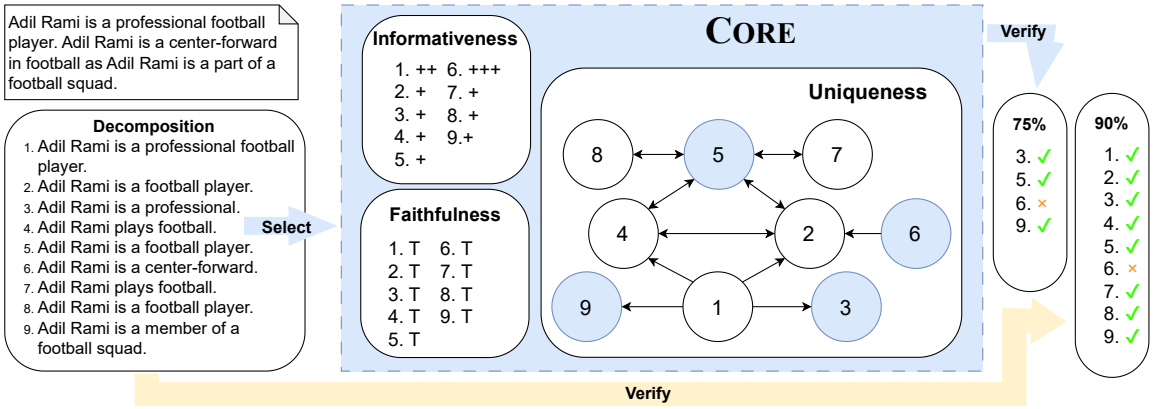

Figure 2: CORE interposes between the decomposition step and the verification step, selecting the most representative set of subclaims that can be identified from the generation to safeguard against trivial or repetitive inputs.

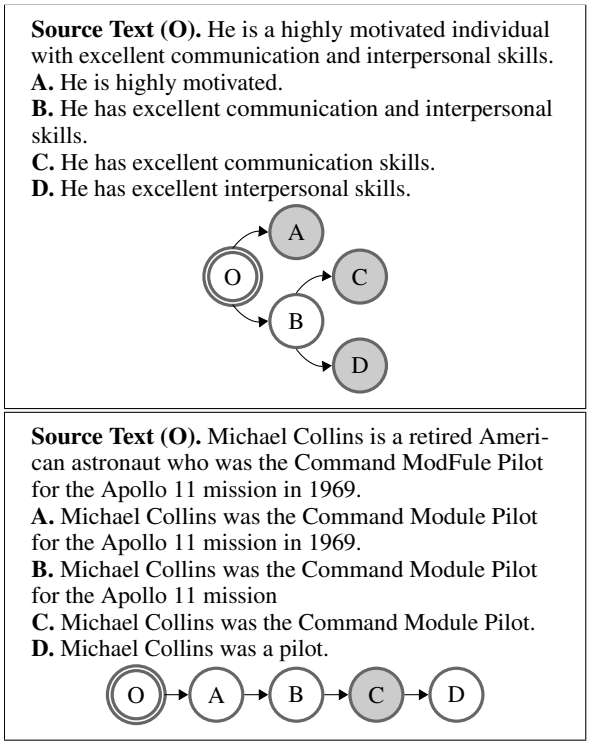

Figure 3: Result of deduplication with uniform weighting. Shaded nodes compose **one set of viable selection** by the algorithm. **Up**: uniform weighting selects the most fine-grained decomposition. **Down**: Uniform weighting may select any subclaim within a monotonous entailment chain.

by the user for full control over the specific set of knowledge they want to nullify, or these claims can be extracted from a prompted large language model for full automation. The set can be of any size, and the claims do not need to be mutually inclusive or entailed by the domain. For example, for the biography data used in FACTSCORE (Min et al., 2023), the bleached claims might include " `{topic}` is a person," "`{topic}` breathes," "`{topic}` exists," or "`{topic}` is famous." The informativeness of a claim c can therefore be identified as

$$w_{\text{Info}}(c) = \text{CPMI}\Big(c; c|\mathcal{H}(D)\Big) = -\log P\Big(c|\mathcal{H}(D)\Big).$$

While some previous work uses corpus statistics like word co-ocurrence to estimate required probabilities (Rudinger et al., 2017), this is infeasible in our case due to reporting bias (Gordon and Van Durme, 2013) and the versatility of free-form generation. Instead we use an Uncertain Natural Language Inference (UNLI) (Chen et al., 2020) model p_θ to directly estimate the conditional probability $P\Big(c|\mathcal{H}(D)\Big)$. However, as traditional Recognizing Textual Entailment (Dagan et al., 2005; Bowman et al., 2015) models aim for short sentence segments, we estimate $-\log P\Big(c|\mathcal{H}(D)\Big)$ with the empirically more stable

$$\min_{h \in \mathcal{H}(D)} -\log p_\theta(c|h).$$

Under this formulation, regardless of how uninformative a subclaim might be, it will still be selected as long as it does not conflict with other subclaims. In practice, we can also effectively ignore entailed subclaims by subtracting a small ϵ from their scores, making some of them negative.

An interesting behavior of this weighting emerges when the decomposition includes subclaims at different levels of granularity, as illustrated in the top panel of Figure 3. With uniform weighting, CORE consistently selects the leaf nodes. However, under informativeness weighting, this pattern changes. Specifically, if a subclaim c_1

is further decomposed into c_2 and c_3. Our weighting will lead to the selection of c_1 whenever

$$P\Big(c_1|\mathcal{H}(D)\Big) < P\Big(c_2|\mathcal{H}(D)\Big) \cdot P\Big(c_3|\mathcal{H}(D)\Big).$$

This approach prevents the model from achieving superficially high FP through enumerating all possible alternatives for unknown knowledge. For example, when $S' = \{$"The coin lands head and tail.", "The coin lands head", "The coin lands tail."$\}$, it receives an FP of 0 instead of 50%, which aligns more closely with human intuition.

4 Evaluation of FP Scoring Metrics

4.1 Evaluation Principles

We aim to assess whether CORE effectively guards against adversarial outputs intended to superficially enhance model Factual Precision. We propose utilizing targeted *Decompose-Then-Verify* Factual Precision metrics through generations that perform significantly worse in the following two dimensions:

Informativeness requires the generation to be as informative as possible. While precision-based metrics often control for recall using some form of length penalty (Min et al., 2023; Wei et al., 2024), more identifiable atomic facts in the generation do not always correspond to better recall, even without duplication. We suspect it is possible to achieve high FP by generating passages with facts that are obvious within the domain of the generation.

Non-repetitiveness requires that the model's generation be clear and non-redundant. For smaller models, repetitiveness is commonly identified as an undesirable form of text degeneration (Holtzman et al., 2020; Welleck et al., 2020). Evidence suggests that language models can estimate their uncertainty indirectly (Mielke et al., 2022; Fadeeva et al., 2024). We hypothesize that it is possible to prompt the language model to repeat what is most likely to be true multiple times.

4.2 Dataset Creation

To create a dataset tailored for FP evaluation and to facilitate some level of adversarial optimization, we automatically collect more human bio profiles, closely following the dataset creation process from FACTSCORE (Min et al., 2023).[5] We query the `Wikidata` API for the `instance_of` property of entities linked from Wikipedia, using entity linkings from (Kandpal et al., 2023) and popQA (Mallen et al., 2023). For entities from (Kandpal et al., 2023) marked by `DBpedia` URLs, we query the corresponding `Wikipedia` entity ID through the `DBpedia` API. As mentioned in FACTSCORE (Min et al., 2023), we retain entities related to a single `Wikipedia` page to avoid any ambiguity.

Group	Continents
A	Insular Oceania, Oceania, Asia, Indian subcontinent, Australian continent
B	North America
C	Europe, Eurasia
D	Central America, Afro-Eurasia, South America, Africa, Caribbean, Americas, `NULL`

Table 1: Grouping scheme for the continents.

Frequency Also following FACTSCORE we compute `freqValue` as a maximum of either of the entity occurrence in (Kandpal et al., 2023) and the pageview count in (Mallen et al., 2023). If an entity does not occur in one of the two datasets, we use the other value as `freqValue`. We use a slightly different grouping from (Min et al., 2023) to ensure more data points can be sampled in total, where an instance is "Rare" if `freqValue` $\in [0, 100)$, "Medium" if `freqValue` $\in [100, 1000)$, "Frequent" if `freqValue` $\in [1000, 5000)$ and "Very Frequent" if `freqValue` $\in [5000, \infty)$.

Nationality `country_of_citizenship` property is queried to determine the nationality of a data point and further query the `continent` property of the country. To address data imbalance, we group the continent denominators into four groups, as shown in Table 1.

Finally, we match the dataset to the Wikipedia dump provided in (Min et al., 2023) to ensure that we only sample entities retrievable from the same knowledge source as FACTSCORE. After uniformly sampling from all 16 categories, we obtain 1024 instances, which we split into *train*, *dev*, and *test* sets with a ratio of approximately 8:1:1 (112 instances). We then pair these topics with generations from LLMs tuned to have superficially high Factual Precision (see Section 5).

5 Experiments and Results

We aim to answer two important research questions: ① Can models artificially boost their reported FP

[5]The FACTSCORE and the corresponding bio dataset are open-sourced under the MIT license.

by generating uninformative and repetitive text of low quality? ② How effective is CORE in mitigating this issue? As previous works have shown that eliciting human ground truth for Factual Precision is challenging, and often can only be annotated under underspecified instructions (Song et al., 2024) or with the help of biasing LLM decompositions (Min et al., 2023; an, 2023), we instead adopt a behavioral evaluation scheme. Subsection 5.3 demonstrates that adding uninformative or repetitive content can superficially increase the Factual Precision of corrupted bios over clean ones, but not for Factual Precision calculated with CORE adjustments. Subsection 5.4 illustrates that CORE shows the desired behavior of a robust FP metrics and effectively guards against adversarial inputs to *Decompose-Then-Verify* metrics.

We compare FP metrics with and without CORE on the dataset created in Subsection 4.2. For `PairwiseEntailment` and `DocEntailment` evaluation, we use **DeBERTa-v3-base-mnli-fever-anli**[6] from the Hugging Face model hub to model `Entail`. To estimate $w_{\text{Info}}(c)$ for each subclaim c, we fine-tune a strong NLI model **roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli**[7] (Nie et al., 2020) on UNLI (Chen et al., 2020), as described in Section 3. We also binarize a "cap-model" **DeBERTa-v3-base-mnli-fever-anli** as "entailment" and "non-entailment" to make sure that subclaims entailed by bleached claims will not get included in the verification step. We give the exact formulation of this capped $\tilde{w}_{\text{Info}}(c)$ in Appendix A. Additionally the weighting function can be further adjusted to cater to relevancy concerns, leading to the following combined scoring function:

$$\tilde{w}(c) = \textbf{REL}\big(\Phi^{-1}(c)\big) \cdot \tilde{w}_{\text{Info}}(c),$$

where we abuse the notation $\Phi^{-1}(\cdot)$ to denote the sentence (chunk) a subclaim c comes from, and **REL** is a binary relevancy judgment implemented using the same prompt as in (an, 2023).

We further observe that the particualr choice of NLI models for entailment evaluation does not have strong empact on CORE computation, as shown in Table 2. For decomposition and verification LLM calls, we always query local **Mistral-7B-Instruct-v0.2** at temperature $= 0$, as we find it achieves .95 Perason's r with **gpt-3.5-turbo-0125** on instance-wise Factual Precision calculated on the original FACTSCORE dataset.

	Model A	**Model B**	**Model C**
Model A[1]	-	.98	.98
Model B[2]	.98	-	.99
Model C[3]	.98	.99	-

[1] DeBERTa-v3-base-mnli-fever-anli
[2] DeBERTa-v3-large-mnli-fever-anli-ling-wanli
[3] RoBERTa-large-snli_mnli_fever_anli_R1_R2_R3-nli

Table 2: Pearson's correlation coefficient among instance-wise FP of out-of-the-box generation by **Mistral-7B-Instruct-v0.2** on our extended *test* set with different NLI model for bidirectional entailment.

5.1 SFT for Higher FACTSCORE

We investigate whether Supervised Fine-tuning (SFT) can artificially boost FACTSCORE by generating trivial and repetitive facts. To this end, we manually write two "summaries" for 5 examples sampled from the original FACTSCORE dataset: one promoting uninformativeness (INFO) and the other promoting repetition (REP) of easy facts and enumeration of alternatives for uncertain facts. Using the corresponding instruction prompt, we sample 5 generations per topic derived in Subsection 4.2 from **Mistral-7B-Instruct-v0.2**.[8]

5.2 Metric Configuration

	INFO	REP
Mistral$_{\text{INST}}$	1.62	1.41
GPT-2	2.36	2.09

Table 3: Both tuned models show low SFT perplexities.

We then tune a LoRA (Hu et al., 2021) to generate summaries in a similar style using the same prompt employed by Min et al. (Min et al., 2023). For all cases, we set $r = 8$ and $\alpha = 16$ for LoRA initialization and search for the best learning rate for each model based on perplexity on the *dev* set. The fitting results are shown in Table 3. More details can be found in Appendix B. All training was conducted using a single A100.

[6] https://huggingface.co/MoritzLaurer/DeBERTa-v3-base-mnli-fever-anli
[7] https://huggingface.co/ynie/roberta-large-snli_mnli_fever_anli_R1_R2_R3-nli
[8] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

5.3 Mitigating Adverserial Inputs

To demonstrate that generating uninformative or repetitive sentences can superficially boost model FP, we corrupt clean model responses with incorrect facts and then restore FP by mixing them with generations from the SFT models described in Subsection 5.1. To corrupt a clean response, we first run the generation through the FACTSCORE pipeline to extract all supported subclaims. Then, with a probability of $p = 0.5$, we use **gpt-3.5-turbo-0125** to modify a supported subclaim to be factually incorrect. We merge these corrupted subclaims into coherent summaries using the **gpt-4o-2024-05-13**-based subclaim merger from (Mohri and Hashimoto, 2024). Finally, we sample additional sentences from the SFT models and append them to the merged summary.

Figure 4 illustrates the impact of low-quality generation on FACTSCORE with and without CORE. While raw FACTSCORE is hacked by including more uninformative and repetitive content, CORE-adjusted FP remains relatively stable and never surpasses the clean version. The gap between metrics widens as the resulting summary becomes less informative or more repetitive (see Appendix C).

Figure 4: Corrupted summaries can achieve higher FACTSCORE than **clean** summaries simply by mixing in more uninformative (**up**) or more repetitive (**down**) sentences (x-axis). However, they do not achieve higher CORE-adjusted FACTSCORE.

5.4 Plug-and-Play CORE

We extend our experiments to other FP metrics as well as to other domains to demonstrate the general applicability of CORE. On the FACTSCORE bios data, we consider two additional *Decompose-Then-Verify* metrics. The Russellian/Neo-Davidsonian (R-ND) (Wanner et al., 2024a) decomposition promotes a different instruction prompt paired with carefully constructed, linguistically motivated example decompositions, resulting in more atomic decompositions. We use $\text{FACTSCORE}_{\text{R-ND}}$ to denote a new *Decompose-Then-Verify* metric created by replacing the FACTSCORE decomposition with the R-ND prompt. The Search-Augmented Factuality Evaluator (SAFE) (Wei et al., 2024) verifies a fact against search results instead of retrieved Wikipedia pages. While SAFE uses the same decomposition prompt as FACTSCORE, it includes additional preprocessing steps. For $\text{FACTSCORE}_{\text{R-ND}}$, we use the same set of in-context examples as in the original paper (Wanner et al., 2024a) to form the base subclaim identifier $\Phi_{\text{R-ND}}$. For SAFE, we use their original decomposition as our base identifier Φ_{SAFE}, but we reduce the maximum number of query generation and searching iterations to 3, as this already provides reasonable coverage of the required information to verify a given subclaim.

We also consider three additional domains: The **Culture & Entertainment** domain and **Geographic** domain from WildHallucinations (Zhao et al., 2024a), and the **Healthcare & Medicine** domain from ExpertQA (Malaviya et al., 2024). For the WildHallucinations datasets, which provide entity names and paired knowledge documents, we used FACTSCORE as our base FP metric. For ExpertQA, since it is QA-based and does not provide a comprehensive knowledge base for verification, we used SAFE as the base FP metric.

The comparison is shown in Table 4. Overall, we found that CORE consistently guards against uninformative and repetitive inputs under all of our configurations, as indicated by the large gap between scores reported by metrics with and without CORE on INFO and REP texts. Under CORE augmentation, neither INFO nor REP generation boosts FP, and the factuality capability of $\textbf{Mistral}_{\textbf{INST}}$ and GPT-2 still gets discriminated.

We further note that while uninformative and repetitive generation boosts FP across all metrics, generating repetitive facts is more challenging for smaller models. We hypothesize that this is because boosting Factual Precision through repetition requires the model to have at least some knowledge of the topic being generated.

Domain	Metric	CORE	Mistral$_{INST}$			GPT-2	
			NORMAL	INFO	REP	INFO	REP
Bios	FACTSCORE	w/o	54.0%	83.0%	78.0%	82.2%	35.4%
		w/	49.6%	36.2%	21.9%	0.68%	5.35%
	FACTSCORE$_{R\text{-}ND}$	w/o	53.9%	75.9%	78.1%	78.1%	40.5%
		w/	48.3%	43.6%	26.0%	2.16%	7.32%
	SAFE	w/o	61.7%	84.8%	80.6%	70.3%	36.0%
		w/	61.3%	29.6%	14.5%	0.35%	4.37%
Cul & Ent	FACTSCORE	w/o	81.5%	88.7%	87.9%	87.1%	72.7%
		w	79.7%	40.4%	26.4%	3.41%	1.00%
Geographic	FACTSCORE	w/o	79.9%	86.7%	84.9%	75.8%	80.7%
		w	78.3%	53.7%	32.7%	1.68%	1.22%
Medical	SAFE	w/o	83.4%	83.6%	87.3%	49.5%	39.5%
		w	71.3%	4.23%	1.76%	0.0%	0.07%

Table 4: Reported Factual Precision on different domains when applying CORE to various base *Decompose-Then-Verify* metrics (FACTSCORE, FACTSCORE$_{R\text{-}ND}$ and SAFE). INFO corresponds to results for claims sampled from the model tuned to generate uninformative responses, while REP corresponds to results for claims sampled from the model tuned to generate repetitive responses. Neither of these low-quality generations should superficially boost factuality above NORMAL

.

6 Related Work

Unlike traditional Fact-Checking efforts that focus on short and simple claims (Thorne et al., 2018; Schuster et al., 2021; Guo et al., 2022), automatic factuality evaluation for LLM generation has a specific focus on long, free-form text with highly compositional complex claims. Early works on long-form factuality have already been arguing for *claim decomposition* (Kamoi et al., 2023), mainly for the ease and fine-granularity this process brings. While existing works follow a similar *Decompose-Then-Verify* paradigm (Chern et al., 2023; Wang et al., 2024), how the decomposition should best be performed is always left underspecified. For example, WiCE (Kamoi et al., 2023), FACTSCORE (Min et al., 2023), and FELM (Zhao et al., 2024b) all have their own decomposition prompts, and RARR (Gao et al., 2023) reports sentence-level attribution and character-level preservation. Previous research has already revealed different characteristics of different decomposition methods regarding atomicity, precision, and coverage (Wanner et al., 2024a), how any particular decision choices, including other additional preprocessing steps (Krishna et al., 2023; an, 2023; Wei et al., 2024; Tang et al., 2024; Song et al., 2024), affect factual-precision evaluation is still an open problem. Being aware of the active exploration of multiple directions for possible improvements over existing *Decompose-Then-Verify* pipelines, CORE is designed to be orthogonal to other popular techniques.

7 Conclusion

We demonstrate that popular Factual Precision evaluation metrics following the *Decompose-Then-Verify* pipeline often assign superficially high scores to obvious or repetitive generations. We introduce CORE, a plug-and-play module that addresses this issue efficiently and effectively. We further show that when augmented with CORE, various *Decompose-Then-Verify* metrics demonstrate a consistent trend of being more robust and become less prone to repetitive and non-informative adversarial inputs. Consequently, we argue that adjustments like CORE should be adopted for more accurate factual precision evaluation, especially in scenarios where models can optimize against automatic metrics. Future research can explore deeper into the interplay between the evaluation of factual precision and the actual factual accuracy of models, and potentially also develop more effective subclaim selection methods within the CORE framework and explore more comprehensive approaches to factuality evaluation.

Limitations

While we demonstrate that CORE adds an extra layer of robustness to existing factual precision metrics, it is not guaranteed to guard against all forms of adversarial generation that lead to superficially high scores. Future research should continue to explore more accurate methods for evaluating the factuality of free-form generation. Additionally, the effectiveness of CORE depends on the performance of each pipeline component, such as the NLI and UNLI models. Although we allow some relaxation for model errors, more accurate and generalizable NLI models will directly enhance the accuracy of our approach.

Acknowledgement

This work is supported by U.S. National Science Foundation under grant No. 2204926, ONR grant (N0001424-1-2089) and DARPA SciFY. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, ONR, nor DARPA.

References

Yuxia Wang an. 2023. Factcheck-gpt: End-to-end fine-grained document-leve. ArXiv preprint, abs/2311.09000.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.

Tongfei Chen, Zhengping Jiang, Adam Poliak, Keisuke Sakaguchi, and Benjamin Van Durme. 2020. Uncertain natural language inference. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8772–8779, Online. Association for Computational Linguistics.

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

Eunsol Choi, Jennimaria Palomaki, Matthew Lamm, Tom Kwiatkowski, Dipanjan Das, and Michael Collins. 2021. Decontextualization: Making sentences stand-alone. Transactions of the Association for Computational Linguistics, 9:447–461.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In Machine learning challenges workshop, pages 177–190. Springer.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, et al. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. ArXiv preprint, abs/2403.04696.

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.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting bias and knowledge acquisition. In Proceedings of the 2013 workshop on Automated knowledge base construction, pages 25–30.

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

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, and Pasquale Minervini. 2024. The hallucinations leaderboard–an open effort to measure hallucinations in large language models. ArXiv preprint, abs/2404.05904.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen,

et al. 2021. Lora: Low-rank adaptation of large language models. In International Conference on Learning Representations.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ArXiv preprint, abs/2311.05232.

Ryo Kamoi, Tanya Goyal, Juan 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.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In International Conference on Machine Learning, pages 15696–15707. PMLR.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. Longeval: Guidelines for human evaluation of faithfulness in long-form summarization. In Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 1650–1669.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue WANG, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. Transactions on Machine Learning Research. Featured Certification, Expert Certification.

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

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. 2022. Reducing conversational agents' overconfidence through linguistic calibration. Transactions of the Association for Computational Linguistics, 10:857–872.

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.

Christopher Mohri and Tatsunori Hashimoto. 2024. Language models with conformal factuality guarantees. ArXiv preprint, abs/2402.10978.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In Proceedings of the

2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 5418–5426, Online. Association for Computational Linguistics.

Rachel Rudinger, Chandler May, and Benjamin Van Durme. 2017. Social bias in elicited natural language inferences. In Proceedings of the First ACL Workshop on Ethics in Natural Language Processing, pages 74–79, Valencia, Spain. Association for Computational Linguistics.

Tara Safavi and Danai Koutra. 2021. Relational World Knowledge Representation in Contextual Language Models: A Review. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 1053–1067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin C! robust fact verification with contrastive evidence. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 624–643, Online. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. ArXiv preprint, abs/2406.19276.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar,

Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Transactions on Machine Learning Research.

Weiting Tan, Haoran Xu, Lingfeng Shen, Shuyue Stella Li, Kenton Murray, Philipp Koehn, Benjamin Van Durme, and Yunmo Chen. 2023. Narrowing the gap between zero-and few-shot machine translation by matching styles. ArXiv preprint, abs/2311.02310.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. ArXiv preprint, abs/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.

Katherine Tian, Eric Mitchell, Huaxiu Yao, Christopher D Manning, and Chelsea Finn. 2024. Fine-tuning language models for factuality. In The Twelfth International Conference on Learning Representations.

Yuxia Wang, Minghan Wang, Hasan Iqbal, Georgi Georgiev, Jiahui Geng, and Preslav Nakov. 2024. Openfactcheck: A unified framework for factuality evaluation of llms. ArXiv preprint, abs/2405.05583.

Miriam Wanner, Seth Ebner, Zhengping Jiang, Mark Dredze, and Benjamin Van Durme. 2024a. A closer look at claim decomposition. ArXiv preprint, abs/2403.11903.

Miriam Wanner, Benjamin Van Durme, and Mark Dredze. 2024b. Dndscore: Decontextualization and decomposition for factuality verification in long-form text generation. ArXiv preprint, abs/2412.13175.

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, abs/2403.18802.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 2024. Fine-grained human feedback gives better rewards for language model training. Advances in Neural Information Processing Systems, 36.

Jiaqing Yuan, Lin Pan, Chung-Wei Hang, Jiang Guo, Jiarong Jiang, Bonan Min, Patrick Ng, and Zhiguo Wang. 2024. Towards a holistic evaluation of llms on factual knowledge recall. ArXiv preprint, abs/2404.16164.

Wenting Zhao, Tanya Goyal, Yu Ying Chiu, Liwei Jiang, Benjamin Newman, Abhilasha Ravichander, Khyathi Chandu, Ronan Le Bras, Claire Cardie, Yuntian Deng, et al. 2024a. Wildhallucinations: Evaluating long-form factuality in llms with real-world entity queries. ArXiv preprint, abs/2407.17468.

Yiran Zhao, Jinghan Zhang, I Chern, Siyang Gao, Pengfei Liu, Junxian He, et al. 2024b. Felm: Benchmarking factuality evaluation of large language models. Advances in Neural Information Processing Systems, 36.

A Algorithm

Algorithm 1: Pseudo code for CORE representative subclaim set selection

Data: Original document $G = \{g_1, g_2, \ldots, g_N\}$, and decompositions $S = \{S_1, S_2, \ldots, S_N\}$.
Result: A list of deduplicated subclaims R

Function Core(G, S, p)**:**
- $A \leftarrow []$; ▷ whether the i-th subclaim entailed by the document
- $W \leftarrow []$; ▷ weight assigned to the i-th subclaim
- $R \leftarrow []$; ▷ A list of selected subclaims
- **for** $i \leftarrow 1$ **to** N **do**
  - $A \leftarrow \texttt{Concat}(\{A, \texttt{DocEntailment}(g_i, S_i)\})$;
  - $W \leftarrow \texttt{Concat}(\{W, \texttt{Weight}(S_i)\})$;
- **end**
- $\mathbf{E} \leftarrow \texttt{PairwiseEntailment}(\texttt{Concat}(S))$;
- Solve IP 1 at p using $A, W, \mathbf{E}$ to obtain X;
- **for** $i \leftarrow 1$ **to** $|X|$ **do**
  - **if** $x_i = 1$ **then**
    - $\texttt{Append}\Big(R, \texttt{Concat}(S)_i\Big)$;
  - **end**
- **end**
- **return** R;

End

Function DocEntailment(g, S)**:**
- $A \leftarrow [0]_{|S|}$;
- $A_i \leftarrow \texttt{Entail}(g, s^i)$ s.t. $\forall i \in 1, \ldots, |S|$;
- **return** A; ▷ whether s^i is entailed by the segment g

End

Function PairwiseEntailment(S)**:**
- $\mathbf{E} \leftarrow [0]_{|S| \times |S|}$;
- $\mathbf{E_{ij}} \leftarrow \texttt{Entail}(s^i, s^j)$ s.t. $\forall i, j \in 1. \ldots, |S|$;
- **return** $\mathbf{E}$; ▷ whether two subclaims s^i, s^j are mutually exclusive

End

As mentioned in Section 3, the weighting function can be any customized function that assigned weight. In practice, to give additional robustness to the CPMI-based informativeness scoring function $w_{\text{Info}}(\cdot)$, the following caped version can be used

$$\tilde{w}_{\text{Info}}(c) = \min\Big(w_{\text{Info}}(c), -\log\big(1 - \mathbb{I}\big[\exists h \in \mathcal{H}(D),\ \text{s.t.} \underset{e \in \{\text{ENT, NEU, CON}\}}{\arg\max} p_\eta(e|h) = \text{ENT}\big]\big)\Big) - \epsilon.$$

Where p_η is a "cap-model" that will predict one of the classical NLI label. This version will always respect the ternary NLI predictor, as the traditional NLI task is easier to solve than UNLI.

We hereby also include a brief explanation of how our constraint ① outlined in Section 3 is realized as the inequality

$$\sum_{i=1}^{N} p_i x_i \leq 0.$$

This is because the constraint is equivalent to

$$\frac{\sum_{i=1}^{N} \text{Entail}(g^i, s^i) x_i}{\sum_{i=1}^{N} x_i} \geq p.$$

which easily reduce to the second constraint in Equation 1.

B Prompts

B.1 Decomposition and Verification

For claim decomposition and claim verification we use the identical prompts from the base factual precision metrics (Min et al., 2023; Wanner et al., 2024a; Wei et al., 2024) respectively. For other preprocessing steps specified (e.g., decontextualization, relevancy check in SAFE (Wei et al., 2024)), we use the original prompts as long as they are publically available. Whenever check-worthiness evaluation is required, we use the claim-level checkworthiness prompt from (an, 2023).

B.2 Corrupt and Merge

For the biography corruption experiment in Subsection 5.3, we use the prompt described in Table **??** to rewrite factually correct claims. We merge the claims use the standard FACTSCORE merging prompt from (Mohri and Hashimoto, 2024).

Field	Content
Prompt	`[User]:` You are given some factually correct statements. Your task is to modify the statements to make them factually incorrect. Try to make the edits atomic without any additional output. `[Agent]:` Sure, please provide the factually correct statements that you would like me to modify.
Input	`[User]: ${STATEMENT}`

Table 5: Prompt used to rewrite factually correct subclaims and make them incorrect.

B.3 Adversarial Generation

We prompt language models to sample uninformative and repetitive generations used to train our SFT model. For *uninformative* responses, we use the prompt showcased inTable **??**. For *repetitive* responses, we use the prompt showcased in Table 7.

Field	Content
Prompt	`[User]:` Tell me a short paragraph of tautological, trivially true and easily verifiable facts about a person. You are allowed to repeat that fact multiple times in paraphrased sentences. `[Agent]:` Sure. Please provide the name of the person you want me to generate trivially true facts about.
Example	`[User]:` Name: Kalki Koechlin `[Agent]:` Kalki Koechlin holds Indian nationality. Kalki Koechlin is an Indian citizen. Kalki Koechlin comes from India. Kalki Koechlin is of Indian origin. Kalki Koechlin's nationality is Indian. Kalki Koechlin is from India. Kalki Koechlin has Indian heritage. Kalki Koechlin is an Indian guy. Kalki Koechlin represents Indian culture. Kalki Koechlin belongs to India. Kalki Koechlin is an Indian national. Kalki Koechlin hails from India. Kalki Koechlin is an Indian by birth. Kalki Koechlin's roots are in India. Kalki Koechlin is of Indian descent. Kalki Koechlin is affiliated with India. Kalki Koechlin is a native of India. Kalki Koechlin's homeland is India. Kalki Koechlin identifies as Indian. Kalki Koechlin's background is Indian. `[User]:` Name: Adil Rami

Field	Content
	[Agent]: Adil Rami plays football. Adil Rami is a professional footballer. Adil Rami's profession is football. Adil Rami is an athlete in the sport of football. Adil Rami is known for playing football. Adil Rami is a soccer player (US term). Adil Rami's career is in football. Adil Rami competes in football. Adil Rami is a football athlete. Adil Rami's occupation is football. Adil Rami is involved in football. Adil Rami's game is football. Adil Rami participates in football. Adil Rami is active in the football world. Adil Rami is a member of the football community. Adil Rami is engaged in playing football. Adil Rami's sport is football. Adil Rami is a professional in football. Adil Rami performs as a footballer. Adil Rami's role is that of a football player. [User]: Name: Song Kang [Agent]: Song Kang is a graduate of Konkuk University. Song Kang completed his studies at Konkuk University. Song Kang earned his degree from Konkuk University. Song Kang finished his education at Konkuk University. Song Kang is an alumnus of Konkuk University. Song Kang received his diploma from Konkuk University. Song Kang obtained his degree at Konkuk University. Song Kang concluded his academic journey at Konkuk University. Song Kang successfully graduated from Konkuk University. Song Kang achieved his academic credentials at Konkuk University. Song Kang completed his degree program at Konkuk University. Song Kang holds a degree from Konkuk University. Song Kang's alma mater is Konkuk University. Song Kang has a degree from Konkuk University. Song Kang finished his coursework at Konkuk University. Song Kang completed his graduation at Konkuk University. Song Kang's education culminated at Konkuk University. Song Kang reached his academic goals at Konkuk University. Song Kang successfully completed his university studies at Konkuk University. Song Kang wrapped up his higher education at Konkuk University. [User]: Name: Emmett Skilton [Agent]: Emmett Skilton hails from New Zealand. Emmett Skilton's birthplace is New Zealand. Emmett Skilton is a native of New Zealand. Emmett Skilton was born in NZ. Emmett Skilton's origins trace back to New Zealand. Emmett Skilton came into the world in New Zealand. Emmett Skilton's country of birth is New Zealand. Emmett Skilton is originally from New Zealand. Emmett Skilton's birth country is New Zealand. Emmett Skilton entered life in New Zealand. Emmett Skilton's roots are in New Zealand. Emmett Skilton's birth land is New Zealand. Emmett Skilton first saw the light of day in New Zealand. Emmett Skilton was born and raised in New Zealand. Emmett Skilton's early life began in New Zealand. Emmett Skilton's native land is New Zealand. Emmett Skilton was delivered in New Zealand. Emmett Skilton's natal place is New Zealand. Emmett Skilton's birth certificate says New Zealand. Emmett Skilton's homeland is New Zealand. [User]: Name: Toyoko Tokiwa

Field	Content
	`[Agent]`: Toyoko Tokiwa is a photographer from Japan. Toyoko Tokiwa practices photography in Japan. Toyoko Tokiwa is known as a Japanese photographer. Toyoko Tokiwa works as a photographer in Japan. Toyoko Tokiwa is an acclaimed photographer from Japan. Toyoko Tokiwa's profession is photography in Japan. Toyoko Tokiwa is a photographer based in Japan. Toyoko Tokiwa is a Japanese national and a photographer. Toyoko Tokiwa captures images as a Japanese photographer. Toyoko Tokiwa's vocation is Japanese photography. Toyoko Tokiwa engages in photography in Japan. Toyoko Tokiwa is a professional photographer from Japan. Toyoko Tokiwa is a photographer hailing from Japan. Toyoko Tokiwa operates as a photographer in Japan. Toyoko Tokiwa's photography career is rooted in Japan. Toyoko Tokiwa is a photographer of Japanese origin. Toyoko Tokiwa contributes to Japanese photography. Toyoko Tokiwa identifies as a Japanese photographer. Toyoko Tokiwa specializes in photography in Japan. Toyoko Tokiwa's occupation is that of a Japanese photographer.
Input	`[User]`: Name `${TOPIC}`

Table 7: Prompts for generating *repetitive* response.

B.4 Bleached Contextual Claims

We use these bleached contextual claims showed in Table 8 as $\mathcal{H}(D)$ for FACTSCORE-style bio evaluation, as described in Section 3.

C Supplementary Results

C.1 Qualitative Result under Real-world Situation

To further demonstrate that our approach is useful on non-adversarial, naturally generated text, we took the following example from Zhao et al. (2024a), where repetitiveness and triviality is explicitly discussed:

> The University of Cambridge is a public collegiate research university located in Cambridge, England. Founded in 1209, it is the world's third-oldest university in continuous operation. The university's reputation for outstanding academic achievement is known worldwide and its alumni include Nobel Prize winners, You may provide a definition with77e5 You may provide a definition with540d thinkers, and transformative figures in several fields. Cambridge's many famous colleges and university buildings attract visitors from around the world. The university is divided into six schools: Humanities and Social Sciences, Clinical Medicine, Technology, Biological Sciences, the Humanities and Social Sciences, and Physical Sciences. Each school comprises several faculties, departments, and other institutions. Teaching and learning at Cambridge are highly personalised and conducted in small groups, and the university's extensive resources include libraries and museums rich in treasures.

The original decomposition and CORE filtering is shown in Table 9. We observe that CORE filters out claim in the pattern that highly aligned with human intuition.

It is worthy noticing that while some generated content may not appear directly repetitive or trivial, some, if not all, decomposers will inevitably introduce repetitive or trivial claims. This is largely due to the lack of global awareness in most decomposition methods, although this choice is sometimes intentional. For instance, FACTSCORE (Min et al., 2023) explicitly encourages gradually decomposing sentences into a gradually weakening entailment chain to achieve atomicity, as fully orthogonal decomposition extremely challenging, if at all possible. This makes controlling such aspect of Decompose-Then-Verify very critical, as recent advance in claim decomposition optimizes for various aspect (Wanner et al., 2024b; Song et al., 2024), paying little attention to such problems that may potentially bias overall assessment. For example,

Field	Content
Prompt	`[User]`: Tell me a short paragraph of tautological, trivially true and easily verifiable facts about a person. You are allowed to repeat that fact multiple times in paraphrased sentences. `[Agent]`: Sure. Please provide the name of the person you want me to generate trivially true facts about.
Example	`[User]`: Name: Kalki Koechlin `[Agent]`: Kalki Koechlin is a person that exists. At some point in time, Kalki Koechlin lives. Kalki Koechlin is a person known to somebody. Kalki Koechlin is capable of breathing. Kalki Koechlin has concious. As a person, Kalki Koechlin exists. Kalki Koechlin has some achievements. `[User]`: Name: Adil Rami `[Agent]`: Adil Rami exist as a person. Adil Rami has life. Adil Rami can move. Adil Rami has certain abilities. Adil Rami has some reputation. Adil Rami has achieved something. Adil Rami is a person's name. Adil Rami does exist. Adil Rami has some impact. `[User]`: Name: Song Kang `[Agent]`: Song Kang has a name. Song Kang has done something. Song Kang knows somebody. Song Kang has some popularity. Song Kang can breath, when alive. Song Kang has certain abilities. Song Kang's name is known to some people. There is a person named Song Kang. Song Kang won't live forever. `[User]`: Name: Emmett Skilton `[Agent]`: Emmett Skilton is an existing person that breathes. Emmett Skilton is capable of certain things. Emmett Skilton has some specific abilities. Emmett Skilton has something different. Emmett Skilton has their own identity. Emmett Skilton achieves things. Somebody knows Emmett Skilton. Emmett Skilton is a person's name. Emmett Skilton has a life. `[User]`: Name: Toyoko Tokiwa `[Agent]`: Toyoko Tokiwa is extant. Toyoko Tokiwa can be seen. Toyoko Tokiwa is an entity. Toyoko Tokiwa draws breath. Toyoko Tokiwa bears a name. Toyoko Tokiwa makes a living. Toyoko Tokiwa has weight. Toyoko Tokiwa has a reputation. Toyoko Tokiwa has capabilities.
Input	`[User]`: Name `${TOPIC}`

Table 6: Prompts for generating *uninformative* responses.

the following generation under natural condition from **Mistral$_{\text{INST}}$** can lead to more than 80% difference in factual precision, as shown in Table 10:

> North Sentinel Island is a small, uninhabited island located in the Bay of Bengal, in the Andaman Sea. It is part of the Andaman and Nicobar Islands, a territory of India. The island is approximately 1.6 kilometers long and 0.6 kilometers wide, with a total area of about 1.8 square kilometers.
>
> North Sentinel Island is most famous for its indigenous inhabitants, the Sentinelese people, who are believed to be one of the last truly isolated tribes in the world. The Sentinelese have lived on the island for thousands of years, and they have shown a strong resistance to outside contact.
>
> The Sentinelese are believed to number around 50-150 individuals, and they live in small, dispersed groups. They are hunter-gatherers, and they subsist primarily on the resources available on the island, including coconuts, fish, and wild pigs.
>
> The Sentinelese are known to be extremely hostile to outsiders. They have a long history of

Claim Template
`${TOPIC}` is a person.
`${TOPIC}` breathes.
`${TOPIC}` exists.
`${TOPIC}` is a name.
`${TOPIC}` is unique.
`${TOPIC}` is famous.
`${TOPIC}` has some abilities.
somebody knows `${TOPIC}`.
`${TOPIC}` is a star.

Table 8: Bleached claim templates that is used for informativeness weighting.

attacking and killing anyone who approaches their island. In 2006, two fishermen from the Andaman and Nicobar Islands were killed when they accidentally drifted too close to the shore. In 2018, an American missionary named John Allen Chau was killed when he attempted to make contact with the tribe.

Despite the risks, there have been efforts to study the Sentinelese from a distance. In the 1960s and 1970s, anthropologists conducted observations of the island from boats and planes. More recently, researchers have used drones to gather information about the Sentinelese and their way of life.

The Indian government has established a buffer zone around North Sentinel Island to protect the Sentinelese from outside contact. The zone is strictly enforced, and visitors are not allowed to approach the island without permission from the authorities.

Despite the challenges, there is a growing interest in learning more about the Sentinelese and their unique culture. Some researchers believe that the tribe may hold valuable insights into human evolution and the development of complex societies. Others are concerned about the potential impact of outside contact on the Sentinelese, and the need to preserve their isolation and way of life.

We believe CORE offers a reasonable and efficient solution to this understudied problem, establishing a stronger and more stable foundation for future research. This allows for meaningful advancements without concerns of redundancy or triviality.

C.2 Mitigating Adverserial Inputs

Similar to Figure 4, we can also mix in repetitive generation to corrupted inputs to superficially boost performance. The result is shown in Figure 5.

Overall, the trend with repetitive sentences is very similar to uninformative sentences. In less than 10 sentences the corrupted generation surpasses the clean generation in factual precision. In most cases, with or without CORE, model generations on more frequent groups are more factual than those on less frequent groups. In general, we observe for all the `freqValue` groups, on generations by out-of-the-box **Mistral$_{\text{INST}}$**, Factual Precision evaluated with or without CORE is close to each other. Also, the tendency that repetition consistently boosts Factual Precision less prominently on generations from **GPT-2**.

C.3 Additional Evaluation with VERISCORE

As the development for automatic fact-verification pipeline is quickly evolving, we evalute the effect of combining CORE with one more advanced decomposition configuration – VERISCORE (Song et al., 2024). Table 12 demomnstrates a similar trend compared to the experiment in Table 4 with FACTSCORE, suggesting that our module addresses a distinct promblem from other advancement in the field.

ID	subclaim
1	~~The University of Cambridge is a university.~~
2	The University of Cambridge is a public university.
3	The University of Cambridge is a collegiate research university.
4	~~The University of Cambridge is located in Cambridge, England.~~
5	It was founded in 1209.
6	~~It is a university.~~
7	It is the world's third-oldest university.
8	~~It is in continuous operation.~~
9	The university has a reputation for outstanding academic achievement.
10	The university's reputation is known worldwide.
11	The university's alumni include Nobel Prize winners.
12	The university's alumni include known thinkers.
13	The university's alumni include transformative figures.
14	The university's alumni are in several fields.
15	Cambridge has many famous colleges.
16	Cambridge has many famous university buildings.
17	~~Cambridge's colleges attract visitors from around the world.~~
18	~~Cambridge's university buildings attract visitors from around the world.~~
19	~~Cambridge's colleges and university buildings attract visitors from around the world.~~
20	The university is divided into six schools.
21	The six schools are: Humanities and Social Sciences, Clinical Medicine, Technology, Biological
22	Sciences, the Humanities and Social Sciences, and Physical Sciences.
23	~~Each school comprises faculties.~~
24	~~Each school comprises departments.~~
25	~~Each school comprises other institutions.~~
26	~~Teaching at Cambridge is highly personalised.~~
27	~~Learning at Cambridge is highly personalised.~~
28	~~Teaching at Cambridge is conducted in small groups.~~
29	~~Learning at Cambridge is conducted in small groups.~~
30	Cambridge University has extensive resources.
31	Cambridge University's resources include libraries.
32	Cambridge University's resources include museums.
33	Cambridge University's libraries are rich in treasures.
34	Cambridge University's museums are rich in treasures.

Table 9: CORE filtering of the motivating example by Zhao et al. (2024a). The ~~filtered claims~~ closely matches the manual filtering by the authors of that paper.

ID	subclaim
1	North Sentinel Island is uninhabited.
2	North Sentinel Island is located in the Bay of Bengal.
3	North Sentinel Island is located in the Andaman Sea.
4	~~The Andaman and Nicobar Islands is a territory.~~
5	~~The Andaman and Nicobar Islands is a territory of India.~~
6	The island is approximately 1.6 kilometers long in total length.
7	~~The island is approximately 0.6 kilometers wide.~~
8	The island is approximately 0.6 kilometers wide in total width.
9	The island has a total area of about 1.8 square kilometers.
10	~~The island has a total area of approximately 1.8 square kilometers.~~
11	North Sentinel Island is home to the Sentinelese people.
12	~~The Sentinelese people are indigenous to North Sentinel Island.~~
13	~~The Sentinelese people are believed to be one of the last truly isolated tribes.~~
14	The Sentinelese people are believed to be one of the last truly isolated tribes in the world.
15	The Sentinelese have lived on the island for thousands of years.
16	~~They have shown a strong resistance.~~
17	They have shown a strong resistance to outside contact.
18	~~The Sentinelese are a group of people.~~
19	~~The Sentinelese live in small groups.~~
20	~~The Sentinelese live in dispersed groups.~~
21	They subsist primarily on the resources available on the island.
22	Coconuts are one of the resources available on the island.
23	They consume coconuts.
24	Fish are one of the resources available on the island.
25	~~They consume fish.~~
26	Wild pigs are one of the resources available on the island.
27	~~They consume wild pigs.~~
28	~~The Sentinelese are known to be hostile.~~
29	~~The Sentinelese are known to be hostile to outsiders.~~
30	They have a long history of attacking.
31	~~They have a long history of killing.~~
32	~~They have a long history of attacking and killing.~~
33	Anyone who approaches their island is a target.
34	~~The fishermen were from the Andaman and Nicobar Islands.~~
35	~~The fishermen were killed when they accidentally drifted too close to the shore.~~
36	The incident occurred in 2018.
37	~~John Allen Chau was a missionary.~~
38	~~John Allen Chau was an American.~~
39	~~John Allen Chau attempted to make contact with a tribe.~~
40	He was killed during this attempt.
41	~~There have been risks involved.~~
42	Despite the risks.
43	In the 1960s, anthropologists conducted observations.
44	~~In the 1970s.~~
45	In the 1970s, anthropologists conducted observations.
46	~~Anthropologists conducted observations.~~
47	Anthropologists conducted observations from boats.
48	Anthropologists conducted observations from planes.
49	The observations took place in the 1960s and 1970s.
50	The observations were conducted from boats and planes.
51	~~More recently, researchers have used drones to gather information.~~
52	~~More recently, researchers have used drones to gather information about the Sentinelese.~~
53	~~The Sentinelese is a group of people.~~
54	More recently, researchers have used drones to gather information about the way of life of the Sentinelese.
55	The buffer zone was established to protect.
56	~~The buffer zone was established to protect the Sentinelese.~~
57	~~The Sentinelese are a people.~~
58	~~North Sentinel Island is a location.~~
59	The Indian government is responsible for the protection of North Sentinel Island.
60	The Indian government has taken steps to protect the Sentinelese from outside contact.
61	The buffer zone was established around North Sentinel Island.
62	Visitors are not allowed.
63	~~Visitors are not allowed to approach the island.~~
64	~~Approaching the island without permission is not allowed.~~
65	~~The authorities have the power to grant permission.~~
66	Permission from the authorities is required to approach the island.
67	~~Despite the challenges.~~
68	There is a growing interest.
69	People are interested in learning more.
70	~~People are interested in learning more about the Sentinelese.~~
71	The Sentinelese have a unique culture.
72	~~The tribe may hold valuable insights.~~
73	The tribe may hold valuable insights into human evolution.
74	The tribe may hold valuable insights into the development of complex societies.
75	~~Others are concerned about the potential impact.~~
76	~~Others are concerned about the potential impact on the Sentinelese.~~
77	Others are concerned about the potential impact on the Sentinelese and their way of life.
78	~~The Sentinelese are a people.~~
79	~~The Sentinelese have a way of life.~~
80	The need to preserve their isolation.

Table 10: List of subclaim decomposition from normal generation ~~before~~ and after CORE, resulting in more than 80% difference in factual precision.

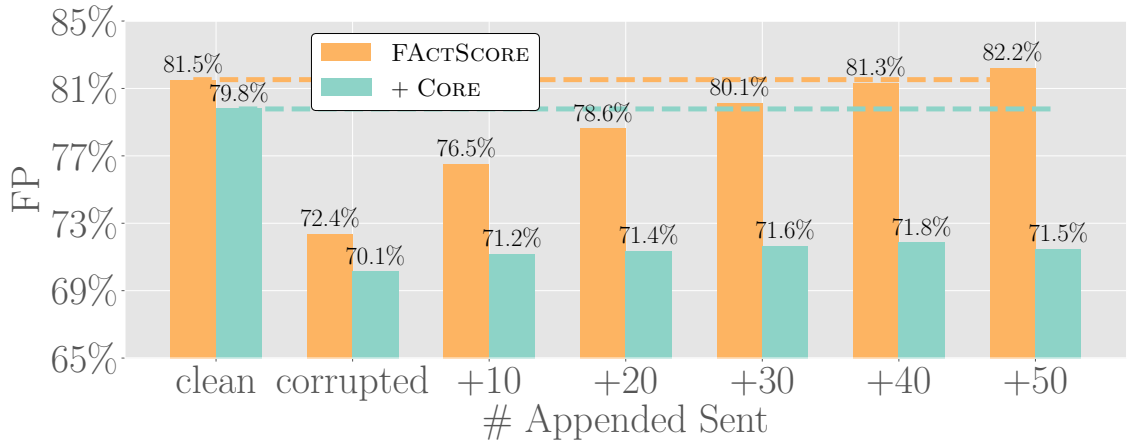

Figure 5: Corrupted summaries can achieve higher FACTSCORE than **clean** summaries simply by mixing in more uninformative sentences (x-axis) on the **entertainment** domain. However, they do not achieve higher CORE-adjusted FACTSCORE.

Feature	Overall	Low-Removal	High-Removal
Repetitiveness	2.83	2.80	2.85
Triviality	1.60	1.50	1.70

Table 11: The removal ratio of CORE correlated well with repetitiveness and triviality on real-world data as well.

C.4 Repetitiveness and Triviality in Real-world Data

To extend our evaluation to real-world data to further demonstrate the effectiveness of CORE, we conduct a small-scale human evaluation on natural, non-adversarial generations produced by **Mistral$_{\text{INST}}$** across all datasets used in subsection 5.4. For each generation, we compute the fraction of claims removed by core, and for each dataset, we randomly sample five generations with a low removal fraction and five with a high removal fraction. This yields a total of 40 generations, which are then evaluated by a highly qualified in-house annotator for repetitiveness and triviality, using a Likert scale. The results are presented in Table 11

First, we observe that repetitiveness is not uncommon in model-generated content. The average repetitiveness score is close to 3 on the Likert scale, corresponding to "a few facts are repeated in different forms; slightly affects the flow." Second, we observe a more pronounced difference in triviality scores between generations with low and high removal ratios. We hypothesize that this effect may be attributed to the CPMI module. However, it is also possible that the model tends to generate less specific or more generic facts for rarer, more challenging entities. Core addresses this by enhancing the decomposition module with global awareness, ensuring that claims introduced in the generation are not redundantly evaluated and do not disproportionately impact factual precision.

Domain	Metric	CORE	Mistral$_{\text{INST}}$		
			NORMAL	INFO	REP
Geographic	VERISCORE	w/o	78.7%	67.4%	78.9%
		w/	76.9%	5.1%	12.7%

Table 12: Additional experiment result on the same **Geographic** subset of Wildhallucinations (Zhao et al., 2024a) with VERISCORE (Song et al., 2024) decompositions. Results show similar trend compared to those with FACTSCORE prompt.

C.5 `freqValue` Breakdown for Plug-and-Play Result

For each *Decompose-Then-Verify* pipeline, we also include a set of Factual Precision evaluation results for each of the `freqValue` group identified in Subsection 4.2.

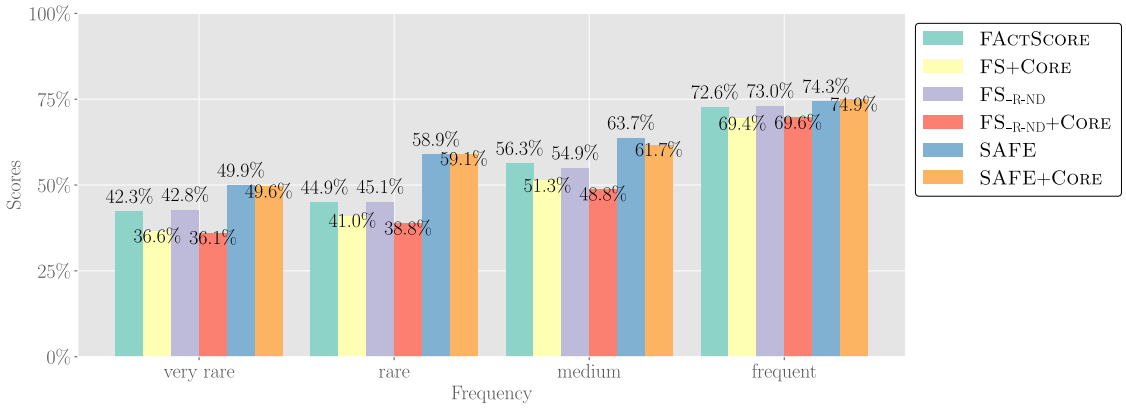

Figure 6: `freqValue` breakdowns of Factual Precision for out-of-the-box **Mistral$_{\text{INST}}$**.

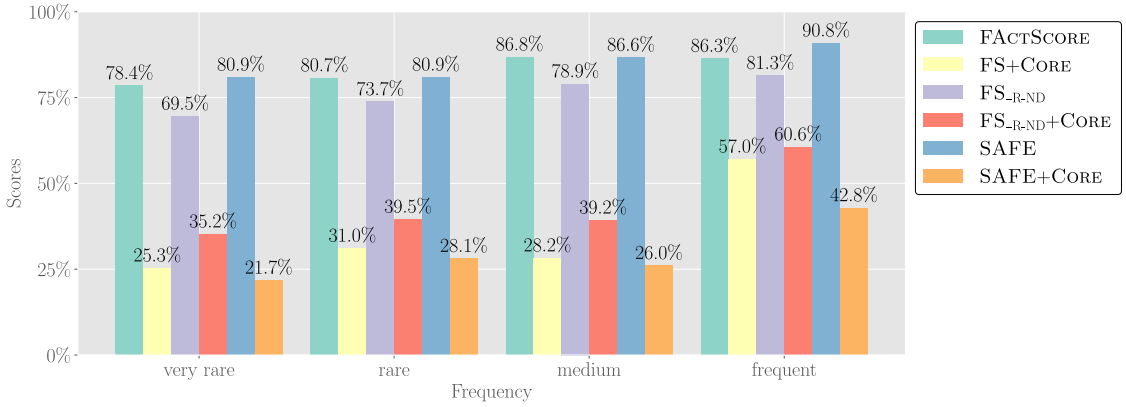

Figure 7: `freqValue` breakdowns of Factual Precision for uninformative **Mistral$_{\text{INST}}$**.

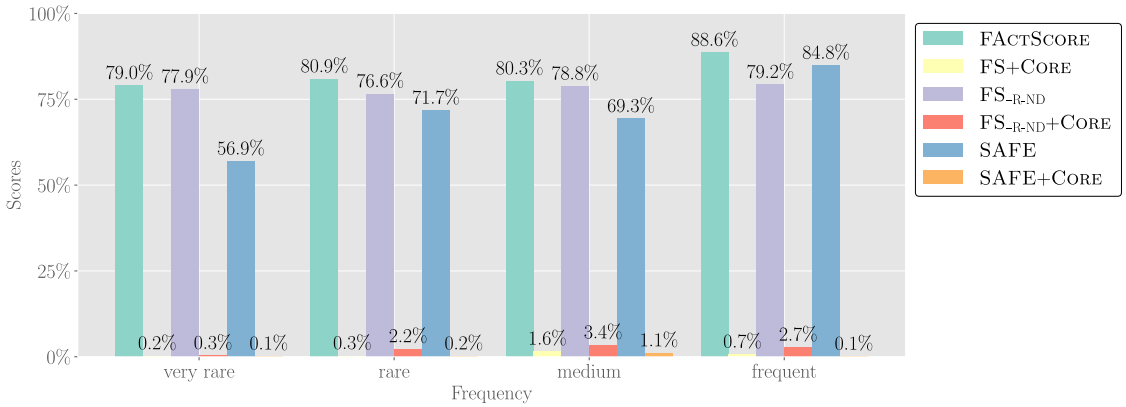

Figure 8: `freqValue` breakdowns of Factual Precision for uninformative **GPT-2**.

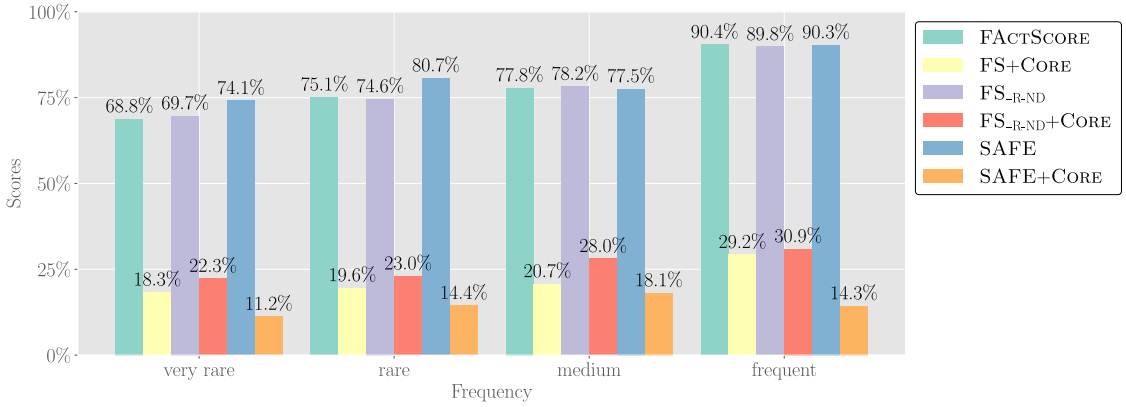

Figure 9: `freqValue` breakdowns of Factual Precision for repetitive **Mistral$_{\text{INST}}$**.

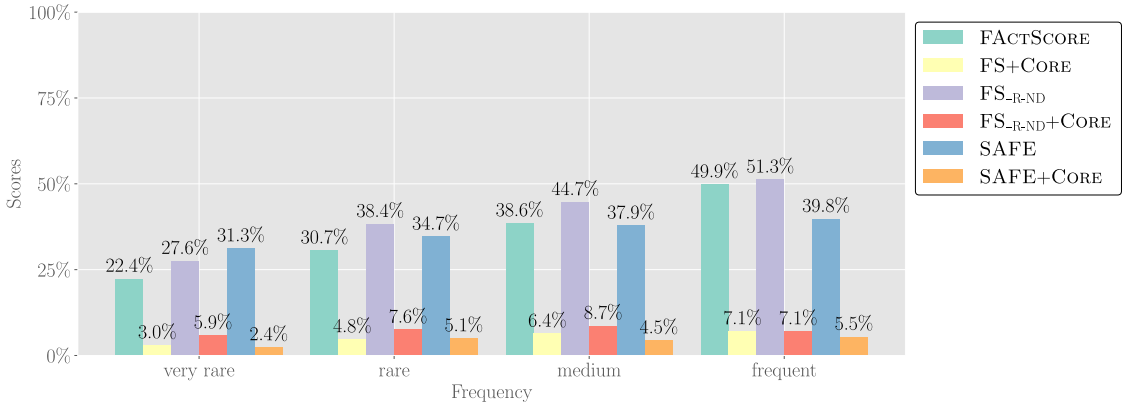

Figure 10: `freqValue` breakdowns of Factual Precision for repetitive **GPT-2**.