Language Models with Conformal Factuality Guarantees

Christopher Mohri [1] **Tatsunori Hashimoto** [1]

Abstract

Guaranteeing the correctness and factuality of language model (LM) outputs is a major open problem. In this work, we propose *conformal factuality*, a framework that can ensure high probability correctness guarantees for LMs by connecting language modeling and conformal prediction. We observe that the correctness of an LM output is equivalent to an uncertainty quantification problem, where the uncertainty sets are defined as the entailment set of an LM's output. Using this connection, we show that conformal prediction in language models corresponds to a back-off algorithm that provides high probability correctness guarantees by progressively making LM outputs less specific (and expanding the associated uncertainty sets). This approach applies to any black-box LM and requires very few human-annotated samples. Evaluations of our approach on closed book QA (FActScore, NaturalQuestions) and reasoning tasks (MATH) show that our approach can provide 80-90% correctness guarantees while retaining the majority of the LM's original output.

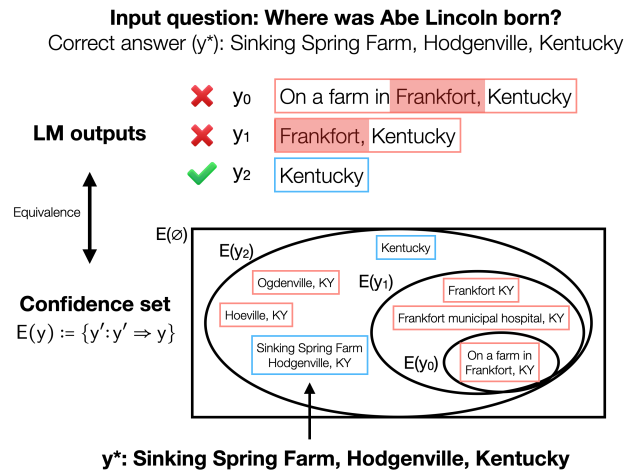

Figure 1. Conformal factuality uses conformal prediction to ensure the correctness of LM outputs. Each potential LM output sequence (top) is associated with an uncertainty set (bottom) that contains every 'more specific' statement that entails it. Conformal prediction provides probabilistic guarantees that these uncertainty sets contain a correct answer (blue), which in turn guarantees the correctness of the associated output.

1. Introduction

Large language models (LLMs) have demonstrated exceptional progress in recent years and are increasingly being adopted in various domains such as search engines and chatbots (Wei et al., 2022; Raffel et al., 2023; Bubeck et al., 2023; Ling et al., 2023). However, their outputs cannot be fully trusted due to their tendency to generate hallucinations and non-factual content (Maynez et al., 2020; Huang et al., 2023a; Ji et al., 2023). This has made the factuality and correctness of LLMs an important and active area of research, with several promising approaches that ground LLMs with knowledge sources (Wang et al., 2023a; Lee et al., 2023; Semnani et al., 2023; Lewis et al., 2021; Du et al., 2023; He et al., 2022) or perform abstention and deferral (Mohri et al., 2023; Mao et al., 2023; Yang et al., 2023a; Cheng et al., 2024).

While the factuality and correctness of language models are improving, precise guarantees are still needed. In many domains such as health (Tang et al., 2023; Thirunavukarasu et al., 2023; Li et al., 2023), law (Huang et al., 2023b; Curran et al., 2023), or robotics (Zeng et al., 2023; Yang et al., 2023b), safely deploying a language model requires outputs to be correct with at least some known, user-specified probability. However, the complexity and opacity of LLMs make it challenging to provide precise performance guarantees.

To enable such high-probability correctness guarantees for black-box LLMs, we take inspiration from conformal prediction (Shafer and Vovk, 2008a; Angelopoulos and Bates, 2022), a framework that performs uncertainty quantification on black-box machine learning systems without strong distributional or modeling assumptions. For any input, this framework produces conformal sets that are guaranteed to

[1]Department of Computer Science, Stanford University. Correspondence to: Christopher Mohri <xmohri@stanford.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

have marginally valid *coverage*, meaning that on average they contain a correct output with any user-specified probability.

While highly successful in regression and classification tasks (Balasubramanian et al., 2014), conformal prediction has had limited success in LLMs for two reasons: the need to score the entire output space makes it intractable, and the resulting confidence sets are so large that they are unusable. Although a few approaches have applied conformal prediction in LLMs for multiple-choice settings (Kumar et al., 2023; Ren et al., 2023), token-level settings (Ravfogel et al., 2023; Ulmer et al., 2024), or approximations (Quach et al., 2023), no existing method can provide exact, conformal guarantees on the open-ended outputs of language models. In our work, we propose a new natural correspondence between conformal sets and an LLM's outputs – this not only resolves the major challenges above, but it will also directly lead to useful LLM-based systems that have correctness guarantees.

The key insight of our work is that each possible LM output defines an associated uncertainty set, where this set is defined as the set of statements that entail the LM's output. Under this definition, there is a direct correspondence between *correctness* and *coverage*, since containing a correct response in the uncertainty set implies that the associated LM's output must also be correct by entailment (Figure 1). Defining the conformal sets *implicitly* via entailment relations makes conformal prediction for language models practical and useful, as we never instantiate uncountably large uncertainty sets and we obtain meaningful and interpretable guarantees directly on model outputs.

Using the correspondence between LM outputs and their associated uncertainty sets, we show how conformal prediction defines a back-off algorithm for ensuring the correctness of LM outputs. This algorithm provides a high-probability correctness guarantee on outputs by producing a chain of output sequences that are increasingly less specific claims and then selects a level of specificity that is correct with high probability, using standard techniques from conformal prediction (Gupta et al., 2022).

Closest to our method is Angelopoulos et al. (2023), which gives an algorithm for classification with hierarchical labels under a conformal risk control framework. Our work also implicitly represents confidence sets but differs in the construction of the set and the actual conformal guarantee. We also note that while the risk control framework has been applied to selecting prompts for LMs (Zollo et al., 2024), our method is exclusively related to LM outputs.

While our approach provides guarantees for *any* black box LM output, we demonstrate the practical utility of our approach by providing correctness guarantees on GPT-4 (OpenAI, 2023) outputs. Our method works by taking the outputs of GPT-4 and repeatedly removing the least certain sub-claims from the output using a GPT-4 prompt. To find the least certain sub-claims, we draw from methods like SelfCheckGPT (Manakul et al., 2023) as well as prompting for uncertainty estimates (Tian et al., 2023). Across closed-book QA and reasoning tasks, we show that conformal factuality enables us to attain any target correctness, and results in usable systems that simultaneously have correctness guarantees far higher than the correctness of the base model (30% → 80% on FactScore, 78% → 93% on NaturalQuestions, and 75% → 95% on MATH) while also retaining the majority of the sub-claims in the output.

We summarize our main contributions below.

- We develop a natural correspondence between conformal prediction and LMs using entailment to define uncertainty sets associated with LM outputs.
- We provide an algorithmic instantiation of conformal factuality by breaking down LM outputs into sub-claims, scoring them, and removing claims according to their uncertainty.
- We demonstrate that our conformal factuality instantiated with GPT-4 can provide high-probability correctness guarantee on closed book QA and reasoning tasks while retaining most of the sub-claims in the outputs.[1]

The rest of this paper is organized as follows. In Section 2, we formulate the desired guarantee and present our algorithm. In Section 3, we present a theoretical analysis of our algorithm. Section 4 describes our implementation, and Section 5 shows the efficacy of our approach on both QA and reasoning tasks.

2. Preliminaries

In the standard language model (LM) generation setting, we receive an input $x \in \mathcal{X}$ and generate an output $y \in \mathcal{Y}$ according to a (potentially stochastic) generator $y = \mathsf{L}(x)$. A key problem is that y may not be fully supported by a ground truth or reference $y^* \in \mathcal{Y}$ (Maynez et al., 2020; Huang et al., 2023a).

Our goal in this work is to provide precise control over the correctness and factuality of an LM's output. As it may be difficult (if not impossible) to provide guarantees on every LM output, our goal will be to provide high-probability guarantees such that for any user-specified probability $\alpha \in (0, 1)$, the LM is correct with probability at least $(1 - \alpha)$

[1] We release our code at `https://github.com/tatsu-lab/conformal-factual-lm`.

over some distribution $\mathbb{P}$. We express this goal as

$$\mathbb{P}(y \text{ is factual and correct}) \geq 1 - \alpha. \tag{1}$$

Throughout this work, we will formalize this correctness constraint in terms of entailments (MacCartney and Manning, 2015) with respect to some reference knowledge y^* where correctness is equivalent to the entailment relation $y^* \Rightarrow y$.[2] Representing factuality and correctness via entailments to a reference is quite general, as we can set y^* to be a broad knowledge base such as 'Wikipedia pages related to x' or even 'all facts accessible via Google' to handle the case where there is no ground truth response for y^*. A key equivalence in our work will be that the factuality constraint $y^* \Rightarrow y$ can be written as a set containment relation $y^* \in \{y' \in \mathcal{Y} : y' \Rightarrow y\}$. We simplify this set using the *entailment operator* $\mathsf{E} : \mathcal{Y} \mapsto 2^{\mathcal{Y}}$ defined by

$$\mathsf{E}(y) := \{y' \in \mathcal{Y} : y' \Rightarrow y\},$$

such that $y^* \in \mathsf{E}(y)$ holds if and only if $y^* \Rightarrow y$.

Using this entailment set, we can now begin to connect our goal (1) to a well-studied statistical inference problem: for exchangeable $(X_i, Y_i^*) \in \mathcal{X} \times \mathcal{Y}, i \in [n+1]$, use $\{(X_i, Y_i^*)\}_{i=1}^n$ to find some uncertainty set $C : \mathcal{X} \mapsto 2^{\mathcal{Y}}$ such that the ground truth Y_{n+1}^* satisfies the following inequality:

$$\mathbb{P}(Y_{n+1}^* \in C(X_{n+1})) \geq 1 - \alpha. \tag{2}$$

The connection between this inference problem and our correctness goal becomes clear if we replace the uncertainty set $C(X_{n+1})$ with the entailment set $\mathsf{E}(\mathsf{L}(X_{n+1}))$ of the LM output $\mathsf{L}(X_{n+1}) \in \mathcal{Y}$. In this case, the inequality (2) would give precisely our correctness bound (1) since the event 'Containing $Y^* \in \mathsf{E}(\mathsf{L}(X))$' is equivalent to the event '$\mathsf{L}(X)$ is correct according to Y^*'. In the subsequent sections, we give formal constructions of our uncertainty sets and introduce conformal prediction techniques for performing the inference problem.

Background on split conformal prediction. Split conformal prediction (Shafer and Vovk, 2008a; Gupta et al., 2022) provides standard tools by which we can construct $C(\cdot)$ that satisfy the constraint in inequality (2). Our work follows the standard split conformal prediction approach in (Gupta et al., 2022), where one constructs a sequence of nested sets and uses exchangeable calibration data to pick a nested set that is sufficiently large to fulfill the inequality in (2). Formally, for a threshold set $\mathcal{T} \subseteq \mathbb{R}$ and each input $x \in \mathcal{X}$, let $\{\mathcal{F}_t(x)\}_{t \in \mathcal{T}}$ denote a sequence of output sets following the *nested property*, meaning that $\mathcal{F}_t(x) \subseteq \mathcal{F}_{t'}(x)$ for $t \leq t'$. Consider the score

$$r(x, y) := \inf\{t \in \mathcal{T} : y \in \mathcal{F}_t(x)\}. \tag{3}$$

This can be thought of as the minimum *safe* threshold where $y \in \mathcal{F}_t(x)$ for every $t > r(x, y)$. Split conformal prediction then sets the final confidence set as

$$C(x) = \mathcal{F}_{\hat{q}_\alpha}(x),$$

where $\hat{q}_\alpha$ is the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th quantile of the scores $\{r(X_i, Y_i)\}_{i=1}^n$. This implementation satisfies the constraint in inequality (2) (see Proposition 1 of (Gupta et al., 2022) for a proof). We now show how to generalize $\mathcal{F}_t$ for language models in a way that also leads to factuality and correctness guarantee on LM outputs.

Application to the language setting. Recall that the correctness of an LM output y is equivalent to the event $y^* \in \mathsf{E}(y)$, and we seek to find some y that makes this event hold with probability at least $1 - \alpha$. To do this, we construct sequences of outputs $\{y_t\}_{t \in \mathcal{T}}$ which induce sequences of associated conformal sets $\{\mathsf{E}(y_t)\}_{t \in \mathcal{T}}$ on which we can apply conformal prediction methods. While these sets could be nested like in Gupta et al. (2022) (implying that the associated outputs y_t become strictly 'more generic' as in Figure 1), this constraint can be hard to enforce for a language model, and we show that our main guarantees do not require nestedness.

In conformal prediction, the key objects are the sequence of conformal sets $\{\mathcal{F}_t\}_{t \in \mathcal{T}}$ and the score $r(x, y)$. We will define these two quantities for the LM setting below.

For the conformal sets, we will define these sets using the entailment operator E as $\mathcal{F}_t(x) = \mathsf{E}(\mathsf{F}_t(x, \mathsf{L}(x)))$ where $\mathsf{F}_t : \mathcal{X} \times \mathcal{Y} \mapsto \mathcal{Y}$ is a 'back off' function and the threshold $t \in \mathcal{T} \subseteq \mathbb{R}$ controls how much $\mathsf{F}_t(x, y_0)$ 'backs off' from the base output y_0 by removing (unreliable) claims. We call F_t *sound* if it satisfies the property that $\mathsf{F}_{\sup \mathcal{T}}(x, y_0) = \varnothing$, where $\varnothing$ represents some output sequence that abstains from making any claim. For notational clarity, we will omit the second argument whenever there is only one relevant language model $\mathsf{L}(x)$ that can generate y_0. In this case, we use the shorthand $\mathsf{F}_t(x) := \mathsf{F}_t(x, \mathsf{L}(x))$.

For the score function, we can redefine the score in (3) as

$$r(x, y^*) := \inf\{t \in \mathcal{T} : \forall j \geq t, y^* \in \mathsf{E}(\mathsf{F}_j(x))\}. \tag{4}$$

This matches the original score with one minor modification where we take the minimum *strictly safe* threshold—we consider a threshold *strictly safe* if any threshold greater than or equal to this one is safe.[3] For the example in Figure 1,

[2]Entailment can be ambiguous, and our work provides guarantees for any definition of entailment (such as entailment as judged by domain experts, crowd workers, or even an automated fact checker), as long as the user has access to a binary entailment oracle. Our guarantees only require that $\forall y \in \mathcal{Y}, y \Rightarrow \varnothing$.

[3]The key difference with respect to (Gupta et al., 2022) is that in the definition of r we write: $\forall j \geq t$. This is implicitly encoded in their definition due to the *nested property* of their set predictors, but as we do not require nestedness, we instead explicitly modify our thresholds to be strictly safe.

Algorithm 1 α-conformal-factuality algorithm

Inputs: base LM $\mathsf{L}\colon \mathcal{X} \mapsto \mathcal{Y}$, confidence α, calibration data $\{X_i, Y_i^*\}_{i=1}^n$, and back-off mechanism $\{\mathsf{F}_t\}_{t\in\mathcal{T}}$
for $i \leftarrow 1$ **to** n **do**
$\quad r(X_i, Y_i^*) \leftarrow \inf\{t \in \mathcal{T} \colon \forall j \geq t, y^* \in \mathsf{E}(\mathsf{F}_j(x, \mathsf{L}(x)))\}$
end
$\hat{q}_\alpha \leftarrow \frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th quantile of the scores $\{r(X_i, Y_i^*)\}_{i=1}^n$
Output: conformally factual $\bar{\mathsf{L}}(x) \coloneqq \mathsf{F}_{\hat{q}_\alpha}(x, \mathsf{L}(x))$.

if we add $y_3 = \varnothing$ and define $\mathsf{F}_t(x) \coloneqq y_t$, we would have the minimum strictly safe threshold $r(x, y^*) = 2$.

With these two components in hand, we can directly apply the split conformal prediction method to obtain an LM with our desired correctness guarantees in inequality 1. Formally, we say that a model $\bar{\mathsf{L}}$ is *α-conformally factual* if for exchangeable $(X_i, Y_i^*) \in \mathcal{X} \times \mathcal{Y}, i \in [n+1]$ and $\{(X_i, Y_i^*)\}_{i=1}^n$ used to construct $\bar{\mathsf{L}}$, the reference output Y_{n+1}^* satisfies the following inequality:

$$\mathbb{P}(Y_{n+1}^* \in \mathsf{E}(\bar{\mathsf{L}}(X_{n+1}))) \geq 1 - \alpha.$$

We present our algorithm for achieving α-conformal-factuality in Algorithm 1, which is a procedure that takes in a base LM $\mathsf{L}\colon \mathcal{X} \mapsto \mathcal{Y}$, target error rate α, back-off mechanism F_t and a calibration dataset, and produces a new LM $\bar{\mathsf{L}}$ that is α-conformally factual. In the following sections, we prove high-probability factuality guarantees for Algorithm 1 (Section 3), and provide our implementation of F_t (Section 4) with experiments on several datasets (Section 5).

3. Theoretical analysis

In this section, we present a theoretical analysis of Algorithm 1, giving upper and lower bounds matching those of standard split conformal prediction and providing a guarantee of the form in inequality (1). The only difference with respect to the analysis and proof of standard split conformal prediction is the new score function.

Theorem 3.1. *(adapted from Shafer and Vovk (2008a)) Let $\{X_i, Y_i^*\}_{i=1}^{n+1}$ be exchangeable, F_t be sound, and $\hat{q}_\alpha$ be defined as the $\frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th quantile of the scores $\{r(X_i, Y_i^*)\}_{i=1}^n$, which we assume to be distinct without loss of generality. Then, for $\alpha \in \left[\frac{1}{n+1}, 1\right]$, the following lower bound holds:*

$$\mathbb{P}(Y_{n+1}^* \in \mathsf{E}(\mathsf{F}_{\hat{q}_\alpha}(X_{n+1}))) \geq 1 - \alpha.$$

If $\mathsf{E}(\mathsf{F}_t(\cdot))$ follows the nested property, then the following upper bound holds:

$$1 - \alpha + \frac{1}{n+1} \geq \mathbb{P}(Y_{n+1}^* \in \mathsf{E}(\mathsf{F}_{\hat{q}_\alpha}(X_{n+1}))).$$

Proof. Let $r_i = r(X_i, Y_i^*)$ for $i \in [n]$ and $r_{\text{test}} = r(X_{n+1}, Y_{n+1}^*)$. These scores are all well-defined because $\mathsf{F}_{\sup \mathcal{T}} = \varnothing$ and $\forall y \in \mathcal{Y}, y \Rightarrow \varnothing$. Without loss of generality, we can assume that the scores are sorted $r_1 < r_2 < \ldots < r_n$. In that case, $\hat{q}_\alpha = r_{\lceil (1-\alpha)(n+1) \rceil}$ when $\alpha \geq \frac{1}{n+1}$. We note that by exchangeability,

$$\mathbb{P}(r_{\text{test}} \leq r_{\lceil (1-\alpha)(n+1) \rceil}) = \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1 - \alpha.$$

We now observe the relationship between the following two events:

$$\{r_{\text{test}} \leq \hat{q}_\alpha\} \text{ implies } \{Y_{n+1}^* \Rightarrow \mathsf{F}_{\hat{q}_\alpha}(X_{n+1})\},$$

because if $r_{\text{test}} \leq \hat{q}_\alpha$, then $\hat{q}_\alpha$ is a safe threshold. This completes the proof of the lower bound. Now, since for any $a \in \mathbb{R}, \lceil a \rceil \leq a + 1$, we obtain the upper bound:

$$\begin{aligned}\mathbb{P}(r_{\text{test}} \leq r_{\lceil (1-\alpha)(n+1) \rceil}) &= \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \\ &\leq \frac{(1-\alpha)(n+1) + 1}{n+1} \\ &= 1 - \alpha + \frac{1}{n+1}.\end{aligned}$$

If $\mathsf{E}(\mathsf{F}_t(\cdot))$ follows the *nested property*, we now observe the equality of two events:

$$\{r_{\text{test}} \leq \hat{q}_\alpha\} = \{Y_{n+1}^* \Rightarrow \mathsf{F}_{\hat{q}_\alpha}(X_{n+1})\},$$

as $\hat{q}_\alpha$ being a safe threshold now implies that it is larger than or equal to the minimum strictly safe threshold. This completes the proof of the upper bound. □

Thus, Algorithm 1 achieves α-conformal factuality for any user-specified correctness target with $\alpha \in \left[\frac{1}{n+1}, 1\right]$, along with an upper bound when the nested property holds. Remarkably, we can guarantee that output sequences in $\mathsf{F}_{\hat{q}_\alpha}(\cdot)$ are factual with high probability over exchangeable sequences. While we can always obtain this guarantee, it does not necessarily imply that we can retain the usefulness of the LM outputs—the threshold $\hat{q}_\alpha$ may be so large that they are uninformative or even empty. In the next section, we provide an implementation of F_t that aims to keep this threshold small.

4. Implementation of F_t via sub-claims

Our guarantees hold with any sound F_t, but ideally, F_t should first remove unreliable parts of an output sequence as the threshold t increases. We now construct an empirically effective instantiation of F_t that makes use of the following observation: the LM often confidently knows that some subparts of its answer are correct, so it often suffices to remove the 'uncertain' subparts to balance correctness and

$$x = \texttt{Who was Abe Lincoln?}$$

$$\mathsf{F}_1(x) = \texttt{Abraham Lincoln, } \underbrace{\texttt{born in Idaho}}_{(1,\ \text{score:}1.5)}\texttt{, } \underbrace{\texttt{was the 16th President of the United States}}_{(2,\ \text{score:}3.8)}\texttt{. } \underbrace{\texttt{He is best known for leading the country through the Civil War.}}_{(3,\ \text{score:}2.2)}$$

$$\mathsf{F}_2(x) = \underbrace{\texttt{Abraham Lincoln was the 16th President of the United States}}_{(2,\ \text{score:}3.8)}\texttt{. } \underbrace{\texttt{He is best known for leading the country through the Civil War.}}_{(3,\ \text{score:}2.2)}$$

$$\mathsf{F}_3(x) = \underbrace{\texttt{Abraham Lincoln was the 16th President of the United States}}_{(2,\ \text{score:}3.8)}\texttt{.}$$

$$\mathsf{F}_4(x) = \varnothing$$

Figure 2. Example $\{\mathsf{F}_t(x)\}_{t\in\mathcal{T}}$ via sub-claims. Here we identified three sub-claims corresponding to (1) Abe Lincoln's birthplace, (2) his notable job, and (3) what he was best known for.

usefulness. We start by defining our implementation of F_t inspired by this idea, and then we analyze the simple procedure that it admits for computing scores r.

Our implementation identifies unreliable parts of an output sequence by decomposing it into sub-claims. Let $\mathsf{L}: \mathcal{X} \to \mathcal{Y}$ be a mapping derived from a language model, $\mathsf{S}: \mathcal{Y} \mapsto 2^{\mathcal{Y}}$ be a function that separates an output sequence into sub-claims, and $\mathsf{M}: 2^{\mathcal{Y}} \mapsto \mathcal{Y}$ be a function that merges sub-claims into a single sequence and satisfies $\mathsf{M}(\varnothing) = \varnothing$. Let $s: 2^{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$ be a sub-claim scoring function, where a larger score is meant to denote a larger probability of a sub-claim being factual. Intuitively, we merge a set of extracted sub-claims that were scored at least t. We implement F_t as follows:

$$\mathsf{F}_t(x) = \mathsf{M}(\{c \in (\mathsf{S} \circ \mathsf{L})(x): s((\mathsf{S} \circ \mathsf{L})(x), c) \geq t\}). \quad (5)$$

This implementation is *sound*, as no sub-claims are accepted for large enough t. We provide an example in Figure 2 to show how a sequence of F_t might look. Note that under the assumptions of Theorem 3.1, applying Algorithm 1 with F_t implemeted as in (5) leads to α-conformal factuality. We will also see that the upper bound of Theorem 3.1 holds under a simple assumption on M.

One additional advantage of implementing F_t via subclaims is that it can substantially reduce annotation effort. Normally, to compute the infimum in the definition of r, one has to evaluate entailment across all outputs $\{\mathsf{F}_t(\cdot)\}_{t\in\mathcal{T}}$. However, we now show that this can be done much more cheaply by only evaluating the entailment of the sub-claims once and computing an infimum over the sub-claims. This trick of computing entailments on sub-claims preserves all our guarantees under the natural assumption that the merger function M that does not add or remove any sub-claims, thus preserving the entailment relations between the sub-claims and the merged output.

Algorithm 2 Score computation with Assumption 4.1

Inputs: base LM $\mathsf{L}: \mathcal{X} \mapsto \mathcal{Y}$, input-reference pair (x, y^*), sub-claim separator S, sub-claim scoring function s

$\{c_i\}_{i=1}^m \leftarrow (\mathsf{S} \circ \mathsf{L})(x)$ `// generate sub-claims`

for $i \leftarrow 1$ **to** m **do**

  $s_i \leftarrow s((\mathsf{S} \circ \mathsf{L})(x), c_i)$ `// score sub-claim`

  $a_i \leftarrow \mathbb{1}_{y^* \Rightarrow c_i}$ `// annotate sub-claim`

end

`// find largest score assigned to a`

`// non-entailed sub-claim`

Sort $\{(c_i, s_i, a_i)\}_{i=1}^m$ in decreasing order of s_i

$k \leftarrow 1$

while $a_k = 1$ *and* $k \leq m$ **do**

  $k \leftarrow k + 1$

end

Output: score $r := s_k$ if $k \neq m + 1$ else $\inf \mathcal{T}$.

Assumption 4.1. For any $y^* \in \mathcal{Y}$ and $\{c_i\}_{i=1}^n \in 2^{\mathcal{Y}}$,

$$\{y^* \Rightarrow \mathsf{M}(\{c_i\}_{i=1}^n)\} \iff \{\forall i \in [n], y^* \Rightarrow c_i\},$$

which could equivalently be written as:

$$\mathsf{E}(\mathsf{M}(\{c_i\}_{i=1}^n)) = \bigcap_{i=1}^n \mathsf{E}(c_i).$$

In this case, the r admits a simpler form.

Proposition 4.2. *For $x \in \mathcal{X}$, let the sub-claims accepted by F_t be denoted by $A_t(x) := \{c \in (\mathsf{S} \circ \mathsf{L})(x): s((\mathsf{S} \circ \mathsf{L})(x), c) \geq t\} \in 2^{\mathcal{Y}}$. Under Assumption 4.1, $r(x, y^*)$ can be computed as*

$$r(x, y^*) = \inf\{t \in \mathcal{T}: \forall j \geq t, \forall c \in A_j(x), y^* \Rightarrow c\}.$$

Proof. We first observe the following equivalence:

$$\begin{aligned} \{y^* \in \mathsf{E}(\mathsf{F}_j(x))\} &\iff \{y^* \Rightarrow \mathsf{F}_j(x)\} \\ &\iff \{y^* \Rightarrow \mathsf{M}(A_j(x))\} \\ &\iff \{\forall c \in A_j(x), y^* \Rightarrow c\}. \end{aligned}$$

Using the definition of r, we can write

$$\begin{aligned} r(x, y^*) :&= \inf\{t \in \mathcal{T}: \forall j \geq t, y^* \in \mathsf{E}(\mathsf{F}_j(x))\} \\ &= \inf\{t \in \mathcal{T}: \forall j \geq t, \forall c \in A_j(x), y^* \Rightarrow c\}, \end{aligned}$$

which completes the proof. □

Thus, if Assumption 4.1 holds, one only has to call the entailment oracle on the sub-claims appearing in the original output, which can be significantly cheaper than calling the entailment oracle on the merger of every possible set of accepted sub-claims. We detail this procedure in Algorithm 2,

Algorithm 3 Score computation without Asssumption 4.1

Inputs: base LM $\mathsf{L}: \mathcal{X} \mapsto \mathcal{Y}$, input-reference pair (x, y^*), merger M, sub-claim separator S, sub-claim scoring function s

$\{c_i\}_{i=1}^m \leftarrow (\mathsf{S} \circ \mathsf{L})(x)$ `// generate sub-claims`

for $i \leftarrow 1$ **to** m **do**

  $s_i \leftarrow s((\mathsf{S} \circ \mathsf{L})(x), c_i)$ `// score sub-claim`

end

`// annotate merged sets of sub-claims`

Sort $\{(c_i, s_i)\}_{i=1}^m$ in decreasing order of s_i

$k \leftarrow 1$

while $\mathbb{1}_{y^* \Rightarrow \mathsf{M}(\{c_i\}_{i=1}^k)} = 1$ *and* $k \le m$ **do**

  $k \leftarrow k + 1$

end

Output: score $r := s_k$ if $k \neq m + 1$ else $\inf \mathcal{T}$.

Algorithm 4 Inference with F_t via sub-claims

Inputs: input x, base LM $\mathsf{L}: \mathcal{X} \mapsto \mathcal{Y}$, sub-claim separator S, merger M, sub-claim scoring function s, threshold t

$\{c_i\}_{i=1}^m \leftarrow (\mathsf{S} \circ \mathsf{L})(x)$ `// generate sub-claims`

for $i \leftarrow 1$ **to** m **do**

  $s_i \leftarrow s((\mathsf{S} \circ \mathsf{L})(x), c_i)$ `// score sub-claim`

end

$A \leftarrow \{c_i \in \{c_i\}_{i=1}^m : s_i \ge t\}$ `// filter sub-claims`

Output: $\mathsf{M}(A)$ `// merge`

as well as the procedure for computing the score r without Assumption 4.1 in Algorithm 3.

We also note that Assumption 4.1 also gives us the upper bound of Theorem 3.1 since F_t follows the nested property—we have $\mathsf{E}(\mathsf{F}_t(x)) = \mathsf{E}(\mathsf{M}(A_t(x))) = \bigcap_{c \in A_t(x)} \mathsf{E}(c)$ and an intersection of sets becomes larger as sets are removed.

Finally, in Algorithm 4 we present the steps for running inference once a threshold on scores r has been computed.

4.1. Partial entailment

Finally, we note that our framework can be extended to provide guarantees for partial correctness. Instead of guaranteeing full factuality, one may want to guarantee that $a \in [0, 1]$ fraction of the accepted sub-claims are factual. To achieve this, we can modify the definition of r to allow for partially-entailed sets of sub-claims. Let $\mathcal{T}_{y^*}: 2^{\mathcal{Y}} \mapsto [0, 1]$ denote an operator indicating the entailed fraction of a set of sub-claims. That is, $\mathcal{T}_{y^*}(\{s_i\}_{i=1}^m) := \frac{1}{m} \sum_{i=1}^m \mathbb{1}_{y^* \Rightarrow s_i}$. Define a new score with acceptable entailment level $a \in [0, 1]$:

$$r_a(x, y^*) := \inf\{t \in \mathcal{T} : \forall j \ge t, \mathcal{T}_{y^*}(A_t(x)) \ge a\}.$$

Note that $r_1(\cdot) = r(\cdot)$ as defined in (4). Replacing r in Algorithm 1 with r_a leads to our partial factuality algorithm, which we outline in Appendix D for completeness. Then, we obtain a result similar to the lower bound of Theorem 3.1.

Corollary 4.3. *(Partial entailment via sub-claims) Let the assumptions of Theorem 3.1 hold, but with $\hat{q}_\alpha$ as the $\frac{\lceil (n+1)(1-a) \rceil}{n}$ quantile of the scores $r_a(X_i, Y_i^*)$ for $a \in [0, 1]$ and F_t implemented as in (5). Then, the following lower bound holds:*

$$\mathbb{P}(\mathcal{T}_{Y^*_{n+1}}(A_{\hat{q}_\alpha}(X_{n+1})) \ge a) \ge 1 - \alpha.$$

Proof. The proof is identical to lower bound of Theorem 3.1, but now we note the following relationship between two events (where $r_{\text{a, test}} := r_a(X_{n+1}, Y^*_{n+1})$):

$$\{r_{\text{a, test}} \le \hat{q}_\alpha\} \text{ implies } \{\mathcal{T}_{Y^*_{n+1}}(A_{\hat{q}_\alpha}(X_{n+1})) \ge a\}.$$

This is not an equivalence, because we could have a safe threshold $\hat{q}_\alpha$ that is less than $r_{\text{a, test}}$. That is, $\mathcal{T}_{y^*}(A_t(x))$ is not monotonically increasing with t since one could remove sub-claims that are entailed and fall below a. □

Note that this result is with respect to the sub-claims themselves, rather than the merged sub-claims, but simple approaches such as providing the sub-claims alongside the output can enable users to verify the correctness of M.[4] Empirically, we found gains from allowing partial factuality to be small, and cover empirical evaluations of this class of approaches in Appendix D.

5. Experiments

While we provide conformal factuality guarantees on LM outputs derived from Algorithm 1, we still need to verify that they are indeed factual and useful. To verify these two items, we apply our algorithm to standard question-answering and reasoning tasks where correctness and factuality guarantees would be useful. We first explicitly describe our experimental set-up, and then present results across 3 datasets.

5.1. Experimental set-up

Here we instantiate all the pieces necessary to implement F_t via sub-claims and describe our datasets.

5.1.1. Models

The definition of F_t via sub-claims in (5) depends on multiple language models: a base mapping from input sequences to output sequences L, a sub-claim separator S, and a merger function M. As a proof of concept, we implement each of

[4] We leave this result in terms of the sub-claims since there is no agreed-upon sub-claim separator function and it would not make sense to say that an output sequence is a-fraction factual without one.

these by using GPT-4. L is implemented by using GPT-4 directly, and S and M by using GPT-4 with prompts to separate and merge (which we present in Appendix C).

5.1.2. Sub-claim scoring functions

Recall that any sub-claim scoring function $s(\cdot)$ introduced in Section 4 leads the guarantees of Theorem 3.1. Below, we provide the definitions of several natural ones that are used in our experiments. For some input and ground truth pair $(x, y^*) \in \mathcal{X} \times \mathcal{Y}$, let the ordered set of n extracted sub-claims be $\{c_i\}_{i=1}^n = (\mathsf{S} \circ \mathsf{L})(x) \in 2^{\mathcal{Y}}$, sorted by where they appear in $\mathsf{L}(x)$. Below we define scoring functions for a particular subclaim $c_j \in \{c_i\}_{i=1}^n$ (and assign a score of $-\infty$ to any sub-claim not in the set $\{c_i\}_{i=1}^n$).

We first define two baselines.

Random scoring. This method assigns random scores to sub-claims and is defined as:

$$s_r(\{c_i\}_{i=1}^n, c_j) = X_j, \text{ where } X_j \sim \mathcal{N}(0, 1).$$

Ordinal scoring. This method assigns scores corresponding to the order a sub-claim appeared in an output sequence, and is defined as:

$$s_b(\{c_i\}_{i=1}^n, c_j) = n - j.$$

The following two scoring functions use an LLM like GPT-4 and are ones to consider using in practice.

GPT-4 confidence scoring. Motivated by Tian et al. (2023); Guan et al. (2023), this method directly asks GPT-4 for a confidence score. We present our prompt in Appendix C.

Frequency scoring. Motivated by self-consistency approaches (Wang et al., 2023b; Manakul et al., 2023), this method first samples 5 alternate output sequences with temperature 1.0 and then counts (with GPT-4) the number of times a sub-claim appeared in the alternate output sequences. The prompt used to implement this method appears in Appendix C as well.[5]

Finally, to provide an upper bound on performance, we include the following oracle scoring method.

Oracle scoring. This method assigns scores corresponding to true entailment. Of course, this is not possible to use without knowledge of y^*, and is not efficient when entailment is meant to be checked by a human. It is defined as:

$$s_o(\{c_i\}_{i=1}^n, c_j) = \mathbb{1}_{y^* \Rightarrow c_j}.$$

Since our guarantees require no ties among the scores r, we tie-break using $\mathcal{N}(0, 0.001)$ noise, and we ensure the noise terms are consistent across sub-claim scoring functions.

[5] In our experiments, we break the ties among these scores using the GPT-4 confidence score.

5.1.3. Datasets and Annotation

We study 3 datasets covering a range of tasks that require correctness. Below, we describe both the datasets we build on and the additional factuality annotations we collect.

FActScore (Min et al., 2023). FActScore is a common factuality evaluation for open-ended generation, which works by breaking a generation down into atomic facts and then evaluating them with a given knowledge source. We use the people entities from their biography generation dataset, but we generate our own sub-claims using S for consistency.

Natural Questions (NQ) (Kwiatkowski et al., 2019). NQ evaluates factuality in open-ended question answering through real queries to the Google search engine. We use questions from the simplified training dataset and allow the model to respond as a long-form response.

MATH (Hendrycks et al., 2021). This is a dataset of math word problems, and we use it to show that our correctness framework can also be applied to reasoning tasks. Answers to reasoning tasks typically involve a sequence of steps, and when we associate these steps with sub-claims, we can immediately apply our framework to return only the correct steps and abstain from the rest. While we observe cohesive selected steps in our experiments, here one may also apply heuristics such as only having the option to return the first k steps.

We select the first 50 inputs from each dataset and manually annotate the sub-claims produced by S applied to GPT-4's outputs. When examining earlier works on factuality evaluation with crowd-workers, we found extensive errors and thus chose to annotate the data ourselves using a 4-way label (Factual, Subjective, Unverifiable, and False), where factuality judgments were verified using Google. We considered Factual and Subjective as entailed, and others as not entailed. All factuality annotations were done before running our experiments.

5.2. Results

In this section, we first verify that Algorithm 1 indeed achieves the factuality guarantees of Theorem 3.1, and then we assess the utility of our outputs both quantitatively and qualitatively.

5.2.1. Empirical factuality

Our main result, Theorem 3.1, states that we should attain roughly $1 - \alpha$ factuality. To check that this happens in practice, we randomly split our datasets into 25 calibration examples and 25 test examples 1000 times, fitting a threshold on the calibration set and measuring the empirical factuality on the test set. The threshold is computed as in Algorithm 2 using our annotated sub-claims, and for empiri-

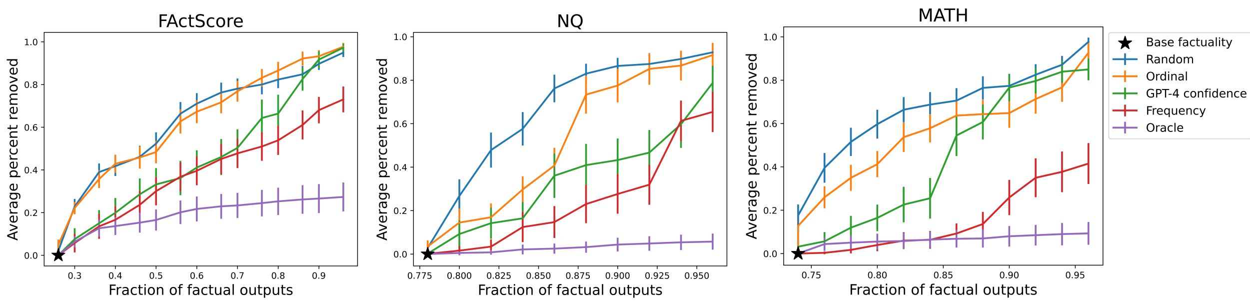

Figure 3. Factuality vs. percent of sub-claims removed across all datasets. Frequency scoring (red) can lead to significant (20 – 50%) gains in correctness while retaining the majority of claims when compared to the base GPT-4 model (star). The tick marks correspond to different values of target α, and the standard deviations represent standard error.

cal factuality, we consider what fraction of the modified LM outputs from Algorithm 4 on the test set are factual (again using our annotated sub-claims). Here F_t is implemented via sub-claims with frequency scoring.

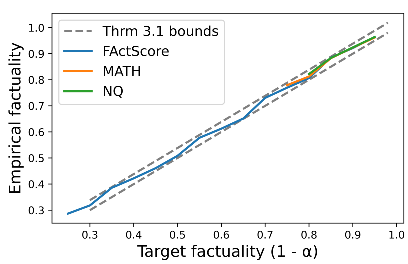

Figure 4. Target vs. empirical factuality. Each solid line starts at the base factuality of GPT-4 on the associated dataset. NQ and MATH overlap on the top right.

We plot the results in Figure 4, and defer similar plots with the remaining scoring function to Appendix G. These results show very tight control over factuality, which Theorem 3.1 guarantees marginally over both the draw of a calibration set and test point. Going beyond standard conformal guarantees, if we require additional high probability guarantees over the calibration set, we find that there is additional variation in the empirical factuality with standard deviation ~0.09, which we expect to decrease to zero as the calibration set grows large.

5.2.2. Utility

The utility of the outputs crucially depends on the quality of the sub-claim scoring function: if a scoring function removes many correct sub-claims before the incorrect ones, the resulting output may be uninformative. We now evaluate our scoring functions to identify whether any of them achieve acceptable utility at various factuality levels.

We start by plotting the percent of sub-claims that are removed under a leave-out-out evaluation and varying the confidence level (Figure 3).

First, the star on the bottom left of each plot represents the base performance of GPT-4, which has relatively low correctness across all tasks. For FActScore, this is very low and the large majority of outputs had at least one sub-claim that we deemed non-factual. For the remaining two datasets, the base performance of GPT-4 is higher at around 75%, but still too low for many high-stakes settings.

Second, Figure 3 suggests that our two scoring functions derived from an LM — frequency scoring and GPT-4 confidence scoring — provide reasonable tradeoffs between utility and factuality. For frequency scoring, we observe particularly promising performance: factuality on FActScore can be increased from ~25% to ~80% while keeping around half of the original sub-claims, on NQ we can increase factuality by ~15% by only removing about one-quarter of sub-claims, and on MATH we can similarly increase factuality by ~15% but by only removing ~10% of sub-claims.

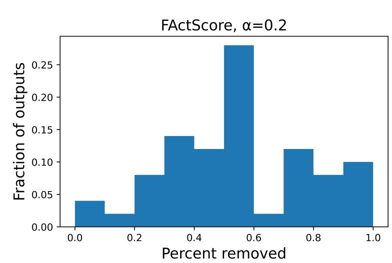

Figure 5. Histogram of percent of sub-claims removed for frequency scoring and α = 0.2 on FActScore.

Taking a closer look at individual points on the plots of Figure 3, we show that the behavior of our method is far from simply abstaining on difficult examples. A method that does this would accept or reject entire outputs, instead of removing a varying fraction of sub-claims. Analyzing the α = 0.2 model for FActScore in Figure 5, we find that most examples have ~50% of their claims removed, leading

Original	Conformal Factuality
Zamfir Ralli-Arbore (1848-1933) was a Romanian political activist and historian from Bessarabia, who spent much of his life in exile. As a member of the National Liberal Party, he campaigned for the union of his native region with the Kingdom of Romania, and was a prominent opponent of Russian and Soviet policies. He was also a noted historian, specializing in the history of the Moldavia and Wallachia during the Middle Ages.	Zamfir Ralli-Arbore, born in 1848, was a Romanian political activist from Bessarabia. He passed away in 1933.
John Estes may refer to: 1. Sleepy John Estes (1899-1977), American blues guitarist, songwriter and vocalist. 2. John Adam Estes (1919-2001), American politician who served as a member of the Texas House of Representatives. 3. John Estes (American football) (born 1986), American football center. 4. John C. Estes, American screenwriter known for his work on films like "Mean Creek" and "The Details". Without more specific information, it's hard to provide details on the correct John Estes.	Sleepy John Estes was an American blues guitarist, songwriter and vocalist. He was born in 1899 and died in 1977.
Michael Valpy is a Canadian journalist, author, and lecturer. He was born on July 30, 1942 in Toronto, Ontario. He is best known for his work as a senior writer for The Globe and Mail, a leading Canadian newspaper, where he covered both national and international news. Valpy has also written several books on Canadian politics and history, and has taught journalism at several universities. He is a three-time nominee for the Gover nor General's Award, Canada's top literary honor.	Michael Valpy is a Canadian journalist, author, and lecturer. He is best known for his work as a senior writer for The Globe and Mail, a leading Canadian newspaper. In his career, Valpy covered both national and international news for the newspaper. In addition to his journalism career, he has written several books on Canadian politics and history. Valpy has also taught journalism at several universities.
We have $i^5 = i^4 \cdot i = i$, $i^{-25} = i^{-24} \cdot i^{-1} = i$, and $i^{45} = i^{44} \cdot i = i$. Therefore, $i^5 + i^{-25} + i^{45} = \boxed{3i}$	The given steps are:$i^5 = i^4 \cdot i = i\backslash n\backslash n i^{45} = i^{44} \cdot i = i\backslash n\backslash n$Using these steps, we can simplify the given expression:$i^5 + i^{-25} + i^{45} = i + i^{-25} + i = i + i^{-25} + i$

Table 1. Comparison of original output by GPT-4 versus our modified output from conformal factuality with frequency scoring. Incorrect claims are highlighted in red. The first three examples are from FActScore and use $\alpha = 0.2$. The last sample from MATH, where the question asked for the evaluation of the expression $i^5 + i^{-25} + i^{45}$, uses $\alpha = 0.1$.

to a less detailed but still informative responses in most cases. The standard deviations for the percent of sub-claims removed can be made very low (at the cost of slightly worse means) with *ranking-based* scoring functions, which we define and present results for in Appendix E.

Finally, we complement our quantitative analysis with samples generated by the conformal factuality model with frequency scoring. Our main dataset of interest is FActScore, as GPT-4 hallucinates aggressively on this task. We thus choose the $\alpha = 0.2$ threshold and present examples in Table 1, with further examples in Appendix F. In all cases shown here, we find that the original GPT-4 output contained a significant fraction of falsehoods, which was successfully removed with conformal factuality.

In addition to the FActScore outputs, we also show an example from MATH in Table 1 (with $\alpha = 0.1$) which shows how our method can be successfully applied to reasoning tasks. Here, the math problem required evaluating three expressions and then taking their sum, but GPT-4 was not able to correctly evaluate one of the expressions and thus provided an incorrect final answer. The conformal factuality model identified and removed the incorrect claim, leading to a partial proof that was presented to the user with a single remaining step left for the user.

6. Conclusion

We described conformal factuality, a framework that connects conformal prediction and language modeling, and importantly leads to a practical algorithm for obtaining conformal factuality guarantees on LM outputs. We gave a natural implementation with sub-claim scoring functions and showed that we can indeed get factual and useful outputs on both question-answering and reasoning datasets. Our work still has limitations, including the guarantees being restricted to a prescripted distribution P, and the fact that our bounds are marginal with respect to the draw of calibration and test data. But our work is the first step to enabling the application of many sophisticated conformal prediction algorithms in improving the outputs of language models – including those that address challenges such as distribution shifts.

Acknowledgements

We thank Zitong Yang for comments related to an earlier version of this manuscript, as well as Neil Band and John Duchi for helpful discussions. TH was supported by a gift by Open Philantropy, the Tianqiao and Chrissy Chen Foundation, and a grant by Samsung GRO.

Impact Statement

Establishing factuality guarantees for Large Language Models (LLMs) using conformal prediction holds the potential to impact society. This could lead to more trustworthy LLMs, offering increased reliability in areas where accuracy is vital, such as health, law, and journalism. Additionally, it may help mitigate biases in AI systems and make their decision-making processes more explainable. However, there are potential challenges, including the risk of over-reliance on these models, and the ethical considerations surrounding LLMs that are perceived as highly factual. Overall, this work carries significant promise for improving AI capabilities, but responsible development and implementation are crucial to ensure the benefits outweigh any potential risks.

References

Angelopoulos, A. N. and Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification.

Angelopoulos, A. N., Bates, S., Fisch, A., Lei, L., and Schuster, T. (2023). Conformal risk control.

Balasubramanian, V., Ho, S.-S., and Vovk, V. (2014). *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes.

Barber, R. F., Candes, E. J., Ramdas, A., and Tibshirani, R. J. (2023). Conformal prediction beyond exchangeability.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4.

Cheng, Q., Sun, T., Liu, X., Zhang, W., Yin, Z., Li, S., Li, L., Chen, K., and Qiu, X. (2024). Can ai assistants know what they don't know?

Curran, S., Lansley, S., and Bethell, O. (2023). Hallucination is the last thing you need.

Das, R., Dhuliawala, S., Zaheer, M., and McCallum, A. (2019). Multi-step retriever-reader interaction for scalable open-domain question answering. In *International Conference on Learning Representations (ICLR)*.

Ding, T., Angelopoulos, A. N., Bates, S., Jordan, M. I., and Tibshirani, R. J. (2023). Class-conditional conformal prediction with many classes.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. (2023). Improving factuality and reasoning in language models through multiagent debate.

Einbinder, B.-S., Romano, Y., Sesia, M., and Zhou, Y. (2022). Training uncertainty-aware classifiers with conformalized deep learning.

Gibbs, I., Cherian, J. J., and Candès, E. J. (2023). Conformal prediction with conditional guarantees.

Guan, J., Dodge, J., Wadden, D., Huang, M., and Peng, H. (2023). Language models hallucinate, but may excel at fact verification.

Gupta, C., Kuchibhotla, A. K., and Ramdas, A. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496.

He, H., Zhang, H., and Roth, D. (2022). Rethinking with retrieval: Faithful large language model inference.

Hendrycks, D., Burns, C., Kadavath, S., Arora, A., Basart, S., Tang, E., Song, D., and Steinhardt, J. (2021). Measuring mathematical problem solving with the math dataset.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., and Liu, T. (2023a). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Huang, Q., Tao, M., Zhang, C., An, Z., Jiang, C., Chen, Z., Wu, Z., and Feng, Y. (2023b). Lawyer llama technical report.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Kang, D. and Hashimoto, T. B. (2020). Improved natural language generation via loss truncation. In *Association for Computational Linguistics (ACL)*, pages 718–731, Online. Association for Computational Linguistics.

Karpukhin, V., Oğuz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and tau Yih, W. (2020). Dense passage retrieval for open-domain question answering.

Kuhn, L., Gal, Y., and Farquhar, S. (2023). Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *ArXiv*, abs/2302.09664.

Kumar, B., Lu, C., Gupta, G., Palepu, A., Bellamy, D., Raskar, R., and Beam, A. (2023). Conformal prediction with large language models for multi-choice question answering.

Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., Toutanova, K., Jones, L., Kelcey, M., Chang, M.-W., Dai, A. M., Uszkoreit, J., Le, Q., and Petrov, S. (2019). Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Lee, N., Ping, W., Xu, P., Patwary, M., Fung, P., Shoeybi, M., and Catanzaro, B. (2023). Factuality enhanced language models for open-ended text generation.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., Riedel, S., and Kiela, D. (2021). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S., and Zhang, Y. (2023). Chatdoctor: A medical chat model fine-tuned on a large language model meta-ai (llama) using medical domain knowledge.

Ling, C., Zhao, X., Lu, J., Deng, C., Zheng, C., Wang, J., Chowdhury, T., Li, Y., Cui, H., Zhang, X., Zhao, T., Panalkar, A., Cheng, W., Wang, H., Liu, Y., Chen, Z., Chen, H., White, C., Gu, Q., Pei, J., and Zhao, L. (2023). Domain specialization as the key to make large language models disruptive: A comprehensive survey.

MacCartney, B. and Manning, C. D. (2015). Natural logic and natural language inference. In *Computing Meaning: Volume 4*, pages 129–147. Springer.

Manakul, P., Liusie, A., and Gales, M. J. F. (2023). Self-checkgpt: Zero-resource black-box hallucination detection for generative large language models.

Mao, A., Mohri, C., Mohri, M., and Zhong, Y. (2023). Two-stage learning to defer with multiple experts. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). On faithfulness and factuality in abstractive summarization.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-t., Koh, P. W., Iyyer, M., Zettlemoyer, L., and Hajishirzi, H. (2023). FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *EMNLP*.

Mohri, C., Andor, D., Choi, E., Collins, M., Mao, A., and Zhong, Y. (2023). Learning to reject with a fixed predictor: Application to decontextualization.

OpenAI (2023). Gpt-4 technical report.

Quach, V., Fisch, A., Schuster, T., Yala, A., Sohn, J. H., Jaakkola, T. S., and Barzilay, R. (2023). Conformal language modeling.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

Ravfogel, S., Goldberg, Y., and Goldberger, J. (2023). Conformal nucleus sampling.

Ren, A. Z., Dixit, A., Bodrova, A., Singh, S., Tu, S., Brown, N., Xu, P., Takayama, L., Xia, F., Varley, J., Xu, Z., Sadigh, D., Zeng, A., and Majumdar, A. (2023). Robots that ask for help: Uncertainty alignment for large language model planners.

Semnani, S., Yao, V., Zhang, H., and Lam, M. (2023). WikiChat: Stopping the hallucination of large language model chatbots by few-shot grounding on Wikipedia. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP*

2023, pages 2387–2413, Singapore. Association for Computational Linguistics.

Shafer, G. and Vovk, V. (2008a). A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3).

Shafer, G. and Vovk, V. (2008b). A tutorial on conformal prediction. *Journal of Machine Learning Research (JMLR)*, 9:371–421.

Shi, W., Han, X., Lewis, M., Tsvetkov, Y., Zettlemoyer, L., and Yih, S. (2023). Trusting your evidence: Hallucinate less with context-aware decoding. *ArXiv*, abs/2305.14739.

Tang, X., Cohan, A., and Gerstein, M. (2023). Aligning factual consistency for clinical studies summarization through reinforcement learning. In Naumann, T., Ben Abacha, A., Bethard, S., Roberts, K., and Rumshisky, A., editors, *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 48–58, Toronto, Canada. Association for Computational Linguistics.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. (2023). Large language models in medicine.

Tian, K., Mitchell, E., Zhou, A., Sharma, A., Rafailov, R., Yao, H., Finn, C., and Manning, C. D. (2023). Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback.

Tibshirani, R. J., Barber, R. F., Candes, E. J., and Ramdas, A. (2020). Conformal prediction under covariate shift.

Ulmer, D., Zerva, C., and Martins, A. F. T. (2024). Non-exchangeable conformal language generation with nearest neighbors.

Vovk, V. (2012). Conditional validity of inductive conformal predictors.

Wang, C., Liu, X., Yue, Y., Tang, X., Zhang, T., Jiayang, C., Yao, Y., Gao, W., Hu, X., Qi, Z., Wang, Y., Yang, L., Wang, J., Xie, X., Zhang, Z., and Zhang, Y. (2023a). Survey on factuality in large language models: Knowledge, retrieval and domain-specificity.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., and Zhou, D. (2023b). Self-consistency improves chain of thought reasoning in language models.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., and Fedus, W. (2022). Emergent abilities of large language models.

Yang, Y., Chern, E., Qiu, X., Neubig, G., and Liu, P. (2023a). Alignment for honesty.

Yang, Z., Raman, S. S., Shah, A., and Tellex, S. (2023b). Plug in the safety chip: Enforcing constraints for llm-driven robot agents.

Zeng, F., Gan, W., Wang, Y., Liu, N., and Yu, P. S. (2023). Large language models for robotics: A survey.

Zollo, T. P., Morrill, T., Deng, Z., Snell, J. C., Pitassi, T., and Zemel, R. (2024). Prompt risk control: A rigorous framework for responsible deployment of large language models.

A. Related Work

Our work relates to a growing literature on factuality for LLMs, conformal prediction, and conformal prediction for language generation.

LM factuality. The factuality of language models is a major concern and a topic of significant research interest. We refer our reader to the several surveys covering research on reducing, detecting, and evaluating hallucinations for a broader view (Ji et al., 2023; Huang et al., 2023a; Wang et al., 2023a) and mention a few relevant works below.

One class of works aims to improve LLMs' access to knowledge, via mechanisms such as retrieval augmentation (Lewis et al., 2021; Das et al., 2019; Karpukhin et al., 2020) while another seeks to suppress LLMs' tendency to fabricate facts by adjusting the training method (Kang and Hashimoto, 2020), ensembling (Kuhn et al., 2023), or modifying the decoder (Shi et al., 2023; Yang et al., 2023a). Our work falls in the latter group, and shares some similarities to methods such as semantic uncertainty in that it is a decoding-time method, but has different goals in that we aim for precise, probabilistic guarantees of correctness.

Our work also relates to a line of work on uncertainty quantification for LLMs, as we must be able to identify and remove unreliable sub-claims from an LLM's output. The implementation of conformal factuality in this work uses a method closely related to SelfCheckGPT (Manakul et al., 2023) as well as prompting the LM for uncertainty estimates (Tian et al., 2023). These works are complementary to ours, as our contribution is a meta-algorithm that uses these uncertainty estimators to return LM outputs that have conformal factuality guarantees. We expect future developments in this line of work to improve the performance of our methods as our algorithms work with any uncertainty quantification method.

Conformal prediction. Conformal prediction is a statistical technique for constructing confidence sets with precise, marginal coverage guarantees without strong distributional or modeling assumptions (Shafer and Vovk, 2008b; Angelopoulos and Bates, 2022; Balasubramanian et al., 2014; Barber et al., 2023). This approach has been successful in providing confidence sets for black-box models such as deep neural networks (Einbinder et al., 2022; Balasubramanian et al., 2014) but its application to language models has been limited.

In language models, there are three major families of applications of conformal prediction to language models: token-based approaches construct confidence sets on *individual tokens* (Ravfogel et al., 2023; Ulmer et al., 2024). This constrains the prediction space and makes it possible to apply standard conformal prediction techniques, but coverage guarantees over tokens cannot be converted into correctness guarantees for sequences. Multiple-choice reduction approaches reduce the prediction space of the LM in constrained domains like question answering. These approaches provide coverage guarantees over the output (e.g. the confidence sets contain the answer with probability $1 - \alpha$) but can only be applied in highly constrained QA-style domains (Kumar et al., 2023). This class of approaches is also related to selecting prompts from a set while controlling their risk (Zollo et al., 2024). Finally, recent work has attempted to instantiate conformal uncertainty sets directly on the space of sequences (Quach et al., 2023) but the space of all sequences is intractably large, and this necessitates approximations to the true confidence set. Most importantly, all of these approaches return *sets* of tokens and sequences, which are difficult to interpret and act upon. In contrast, our conformal factuality approach returns a *single natural language sequence* (which implicitly represents a confidence set) and thus can be used directly to improve the output of LMs.

Finally, closest to our work is Angelopoulos et al. (2023), which gives a conformal prediction algorithm that can be applied to classification problems with hierarchical labels. In that setting, their approach can return an intermediate node in the tree which implicitly represents a confidence set consisting of all the leaves of this subtree. While our work is similar in that we implicitly represent confidence sets for conformal prediction, our work differs in the construction of the set (via entailments), representation of the hierarchy (we do not enumerate its edges), the setting and implementation (language models) and the actual conformal guarantee (correctness of an output rather than risk control).

B. Limitations

Our work depends on standard split conformal prediction, which has important limitations in practice. First, the coverage guarantee is not conditional, meaning that we do not have a guarantee on the conformal set associated with every input, but instead, we have coverage on average over inputs. This means that when the calibration set is defined across multiple tasks or users, the factuality of any one domain or user may differ significantly from the target factuality level. Second, the conformal guarantee is also marginal over the draw of the calibration set. This means that in settings where one has a small, fixed calibration dataset and repeatedly uses the same threshold, the coverage associated with this threshold may deviate from the target coverage. Lastly, in real-world scenarios where distributions change, the threshold computed on past calibration data can fail to maintain the desired coverage. In our case, this means that the factuality guarantee of our language model may be lost if, for example, the distribution of inputs were to change drastically.

While these drawbacks of conformal prediction exist, we believe this work is a step in the right direction toward guaranteeing the factuality of language models. Several works build on the framework of split conformal prediction to tackle the challenges mentioned, and the connections established in this work could enable the use of those approaches (Gibbs et al., 2023; Ding et al., 2023; Barber et al., 2023; Tibshirani et al., 2020; Vovk, 2012). Moreover, we observe promising experimental results that show how we can effectively remove hallucinations from language model outputs, and we expect this to improve with better uncertainty quantification methods.

C. Prompts used in experiments

We use prompts to implement both the sub-claim separator S and merger function M (defined in Section 4). The only other prompt we use is for frequency scoring. For convenience, we perform GPT-4 confidence scoring alongside sub-claim separation. All of these prompts appear in Table 2. We use the `'gpt-4'` endpoint, set `max_tokens` to 1000 and `temperature` to 0.0. We used GPT-4 with these prompts between December 15 and January 15.

Separator (for all datasets)/GPT-4 confidence scoring

```
Please breakdown the following input into a set of small, independent claims (make sure not to add any
information), and return the output as a jsonl, where each line is subclaim:[CLAIM], gpt-score:[CONF]. The
confidence score [CONF] should represent your confidence in the claim, where a 1 is obvious facts and results
like 'The earth is round' and '1+1=2'.  A 0 is for claims that are very obscure or difficult for anyone to know,
like the birthdays of non-notable people.  If the input is short, it is fine to only return 1 claim.  The input is:
```

Merger (for FActScore)

```
You will get an instruction and a set of facts that are true.  Construct an answer using ONLY the facts
provided, and try to use all facts as long as its possible.  If no facts are given, reply to the instruction
incorporating the fact that you dont know enough to fully respond.  \n\nThe facts:\n {claim_string}\n\nThe
instruction:\n{prompt}
```

Merger (for NQ)

```
You will get a natural question and parts of an answer, which you are to merge into coherent prose.  Make sure to
include all the parts in the answer.  There may be parts that are seemingly unrelated to the others, but DO NOT
add additional information or reasoning to merge them.  \n\nThe parts:\n{claim_string}\n\nThe question:\n{prompt}.
Remember, DO NOT add any additional information or commentary, just combine the parts.
```

Merger (for MATH)

```
"You will get a math problem and a set of steps that are true.  Construct an answer using ONLY the steps provided.
Make sure to include all the steps in the answer, and do not add any additional steps or reasoning.  These
steps may not fully solve the problem, but merging them could assist someone in solving the problem.  \n\nnThe
steps:\n{claim_string}\n\nThe math problem:\n{prompt}.  Remember, do not do any additional reasoning, just combine
the given steps.
```

Frequency scoring

```
You will get a list of claims and piece of text.  For each claim, score whether the text supports, contradicts,
or is unrelated to the claim.  Directly return a jsonl, where each line is {"id":[CLAIM_ID], "score":[SCORE]}.
Directly return the jsonl with no explanation or other formatting.  For the [SCORE], return 1 for supports, -1 for
contradicts, and 0 for unrelated.  The claims are:\n{claim_string}\n\nThe text is:\n{output}
```

Table 2. Prompts for sub-claim separator S and merger function M, as well as LM-based sub-claim scoring functions (GPT-4 confidence and frequency). The prompt for frequency scoring is used to evaluate 5 alternate output sequences generated by GPT-4 with temperature 1.0 .

Algorithm 5 α-conformal-partial-factuality algorithm

Inputs: base LM $\mathsf{L}: \mathcal{X} \mapsto \mathcal{Y}$, confidence α, calibration data $\{X_i, Y_i^*\}_{i=1}^n$, and back-off mechanism $\{\mathsf{F}_t\}_{t \in \mathcal{T}}$
for $i \leftarrow 1$ **to** n **do**
  $r_a(X_i, Y_i^*) \leftarrow \inf\{t \in \mathcal{T}: \forall j \geq t, \mathcal{T}_{y^*}(A_t(x)) \geq a\}$
end
$\hat{q}_\alpha \leftarrow \frac{\lceil (n+1)(1-\alpha) \rceil}{n}$th quantile of the scores $\{r_a(X_i, Y_i^*)\}_{i=1}^n$
Output: conformally factual $\bar{\mathsf{L}}(x) := \mathsf{F}_{\hat{q}_\alpha}(x, \mathsf{L}(x))$.

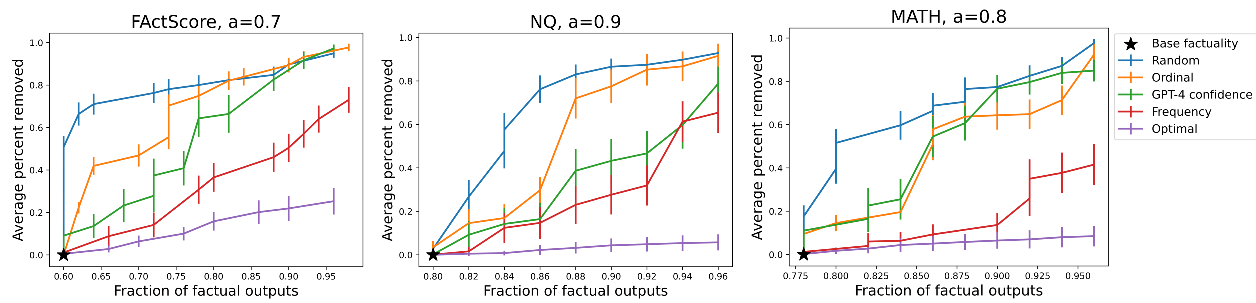

Figure 6. Factuality vs. percent of sub-claims removed across all datasets for partial factuality setting. The tick marks correspond to different values of target α, and the standard deviations represent standard error.

D. Partial factuality

Here we give experimental results for the partial entailment/factuality setting of Section 4.1, which relaxes the standard formulation with an acceptable fraction of entailed sub-claims $a \in [0, 1]$. Our partial factuality algorithm is explicitly defined in Algorithm 5. The only difference with respect to Algorithm 1 is that we use r_a instead of r. We choose $a = 0.7$ for FActScore, $a = 0.9$ for NQ, and $a = 0.8$ for MATH, and plot factuality vs. percent of sub-claims removed in Figure 6. While the condition to be considered factual is relaxed, we observe a lower percentage of sub-claims removed.

E. Ranking-based scoring functions

Any sub-claim scoring function can be transformed into a *ranking-based* scoring function. These aim to remove a fixed percentage of facts, which may be desirable in cases where modifying the outputs in roughly the same way is desirable. Here, for a sub-claim scoring function $s: 2^{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{R}$, we define a ranking function $f_s: 2^{\mathcal{Y}} \times \mathcal{Y} \mapsto \mathbb{N}$, where $f(\{c_i\}_{i=1}^n, c_j)$ returns the rank of $s(\{c_i\}_{i=1}^n, c_j)$ among the set $\{s(\{c_i\}_{i=1}^n, c_k)\}_{k=1}^n$ in increasing order. Then, for each of sub-claim scoring functions s defined above, we apply the score:

$$\frac{f_s(\{c_i\}_{i=1}^n, c_j)}{n}.$$

We divide by n to account for varying-length output sequences. In Figure 7, we plot factuality vs. percent of sub-claims removed for the ranking-based versions of all our sub-claim scoring functions. These do remove a higher percent of sub-claims, but the standard deviations are smaller.

F. More conformal factuality output examples

Here we give more conformally factual output examples, supplementing those in Table 1. We give examples for FActScore in Table 3, NQ in Table 4, and MATH in Table 5.

G. Empirical factuality for all scoring functions

To complete the study of empirical factuality in Section 5.2.1, we repeat the same procedure for the remaining sub-claim scoring functions. We present the results in Figure 4.

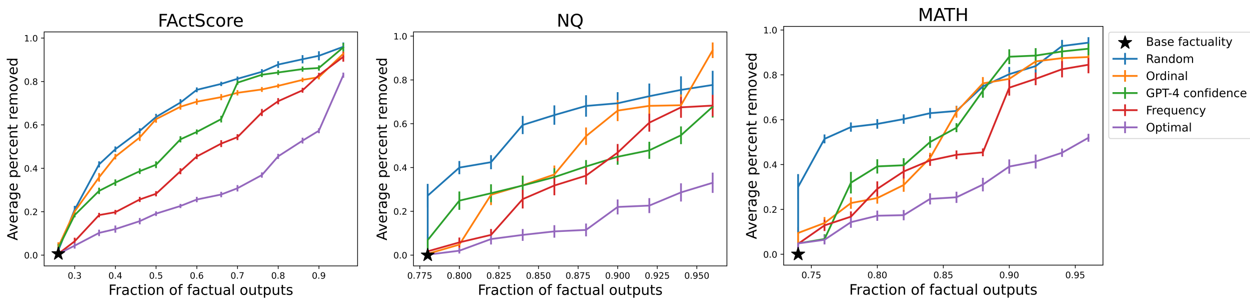

Figure 7. Factuality vs. percent of sub-claims removed across all datasets with *ranking-based* scoring functions. The tick marks correspond to different values of target α, and the standard deviations represent standard error.

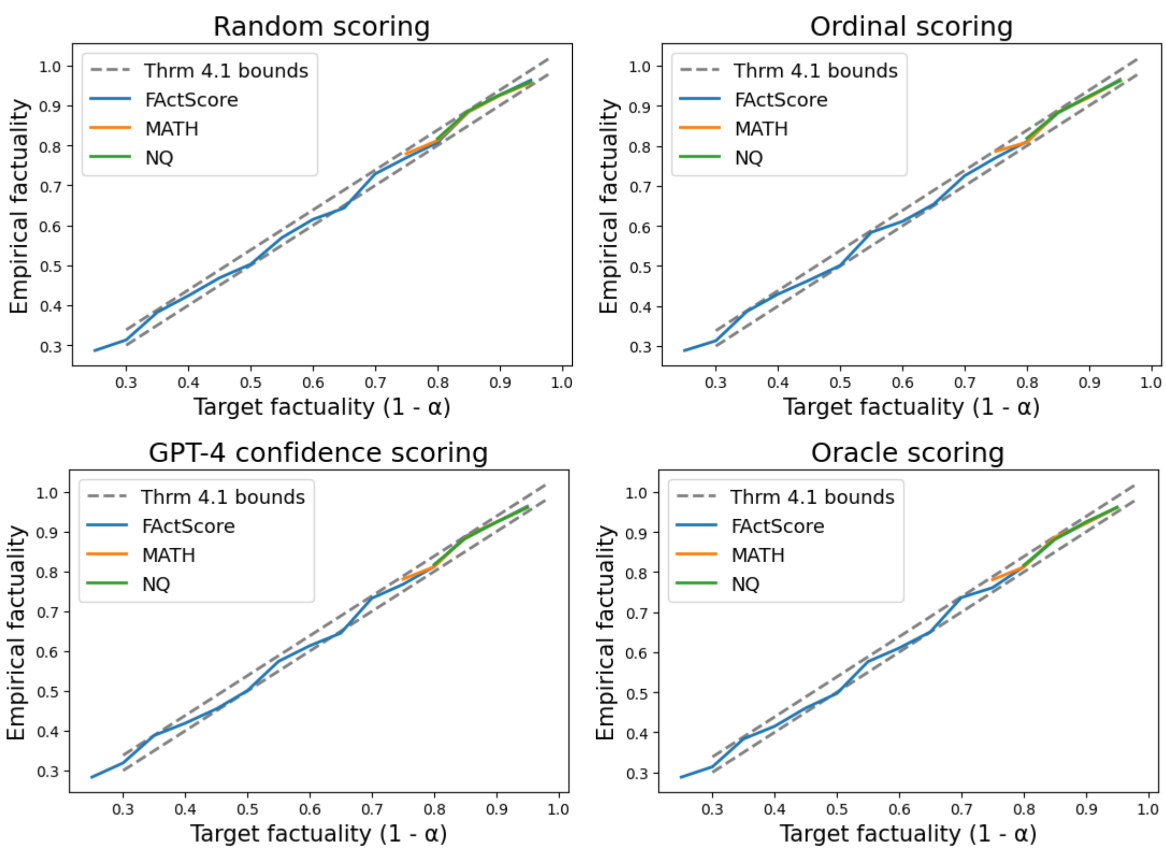

Figure 8. Empirical factuality plots for all scoring functions.

Original	Conformal Factuality
Luis Guillermo Rivera is a professional athlete from Colombia who specializes in high jump. He was born on March 20, 1989 in Apartado, Colombia. Rivera has represented Colombia in several international competitions, including the World Championships in Athletics and the Pan American Games. He won the gold medal in the high jump at the 2013 South American Championships in Athletics. His personal best in the high jump is 2.31 meters, which he achieved in 2013.	I'm uncertain about the specific details, and I recommend using a search engine like Google for more accurate and up-to-date information.
Roxana Diaz is a Venezuelan athlete who specializes in the 400 meters hurdles. She represented her country in various international competitions, including the World Championships and the Pan American Games. Diaz is known for her speed, agility, and determination on the track. She has won several medals throughout her career, making her one of the most successful athletes in her country.	Roxana Diaz is an athlete who has represented Venezuela in various international competitions. She has also participated in the World Championships.
Sharad Kumar is an Indian para-athlete who competes in high jump events. He was born on March 1, 1992, in Bihar, India. Kumar was diagnosed with polio at a young age, which affected his left leg. Despite this, he pursued his passion for sports and began training as a high jumper.\n\nKumar has represented India in several international competitions and has won numerous medals. He won a gold medal at the 2014 Asian Para Games in Incheon, South Korea, and a silver medal at the 2018 Asian Para Games in Jakarta, Indonesia. He also won a bronze medal at the 2017 World Para Athletics Championships in London.\nnIn 2021, Kumar won a bronze medal in the men's high jump T42 category at the Tokyo Paralympics, marking his first Paralympic medal. His personal best in high jump is 1.90 meters, which he achieved at the 2017 World Para Athletics Championships.	Sharad Kumar is an Indian para-athlete who was born on March 1, 1992, in Bihar, India. He competes in high jump events and despite his condition, he pursued his passion for sports and began training as a high jumper. Sharad Kumar has represented India in several international competitions and has won numerous medals in his career. Most notably, in 2021, he won a bronze medal in the men's high jump T42 category at the Tokyo Paralympics.
Katsunosuke Hori is a Japanese actor and voice actor. He was born on November 28, 1941 in Tokyo, Japan. He is best known for his work in anime and video games, including roles in "Neon Genesis Evangelion", "Ghost in the Shell: Stand Alone Complex", and "Final Fantasy VII: Advent Children". He is also known for his dubbing roles in foreign films, such as the voice of Gandalf in the Japanese version of "The Lord of the Rings" trilogy.	Katsunosuke Hori is a Japanese actor and voice actor. He is known for his work in anime and video games.
Ra Jong-yil is a South Korean diplomat and academic. He served as the South Korean ambassador to the United Kingdom and Japan. He also served as the deputy director of the National Intelligence Service. As an academic, he has taught at various universities and written several books on North Korea and international relations.	Ra Jong-yil is a South Korean diplomat and academic who has served as the South Korean ambassador to both the United Kingdom and Japan. In addition to his diplomatic work, he has written several books on North Korea and international relations.
Luis N. Rivera-PagŎ0e1n is a Puerto Rican theologian and author. He was the first Latino to be tenured as a professor at Princeton Theological Seminary, where he taught for many years. His work focuses on the intersection of religion, colonialism, and social justice, particularly in the context of Latin America and the Caribbean. He has written several books, including "A Violent Evangelism: The Political and Religious Conquest of the Americas" and "Essays from the Margins".	Luis N. Rivera-PagŎ0e1n is a Puerto Rican theologian and author who has written several books. His work primarily focuses on the intersection of religion, colonialism, and social justice. He has spent many years teaching at Princeton Theological Seminary."
Jesse Foppert is a former professional baseball player from the United States. He was born on July 10, 1980 in Reading, Pennsylvania. Foppert was a pitcher and played in Major League Baseball (MLB) for the San Francisco Giants and the Seattle Mariners. He was considered one of the top pitching prospects in baseball in 2003, but his career was derailed by injuries. After retiring from professional baseball, Foppert became a coach.	Jesse Foppert is a former professional baseball player from the United States, born on July 10, 1980 in Reading, Pennsylvania. He was a pitcher in Major League Baseball (MLB), playing for the San Francisco Giants and the Seattle Mariners. Unfortunately, his career was derailed by injuries. After retiring from professional baseball, Foppert became a coach.
Zelma Wilson (1926-1996) was an American architect known for her work in the field of sustainable design. She was one of the first women to establish her own architectural firm in California and was a pioneer in the use of solar energy in architecture. Wilson's designs often incorporated natural materials and were designed to blend in with their surroundings. Some of her most notable projects include the Ojai City Hall and the Ventura County Government Center. She was also a dedicated advocate for women in architecture and served as a mentor to many young architects.	Zelma Wilson was an American architect.

Table 3. FActScore with $\alpha = 0.2$ and frequency scoring: comparison of original output by GPT-4 versus our modified output from conformal factuality. Incorrect claims are highlighted in red.

Question	Original	Conformal Factuality
when is the last episode of season 8 of the walking dead	The last episode of season 8 of The Walking Dead aired on April 15, 2018.	The Walking Dead, which has at least 8 seasons, aired its last episode of season 8 on April 15, 2018.
in greek mythology who was the goddess of spring growth	The goddess of spring growth in Greek mythology was Persephone. She was also the queen of the underworld as she was abducted by Hades, the god of the underworld. Her return to the surface each year was believed to symbolize the rebirth of plants and the arrival of spring.	In Greek mythology, the goddess of spring growth was Persephone. Interestingly, Persephone also held the title of the queen of the underworld.
until we meet again filipino soap episode 10	"Until We Meet Again" is a popular Filipino soap opera. However, I'm sorry, but I can't provide a summary for episode 10 as the details for specific episodes of this show are not readily available online. It's best to watch the series through a legitimate streaming service or on the network it originally aired on.	I'm uncertain about the specific details, and I recommend using a search engine like Google for more accurate and up-to-date information.
what is the name of spain's most famous soccer team	Real Madrid and FC Barcelona are the most famous soccer teams in Spain.	Real Madrid and FC Barcelona are both soccer teams in Spain and are considered to be two of the most famous soccer teams in the country.
who are uncle owen and aunt beru related to	Uncle Owen and Aunt Beru are related to Luke Skywalker. They are his uncle and aunt, who raised him on the planet Tatooine in the Star Wars series. Uncle Owen is also the stepbrother of Anakin Skywalker, who later becomes Darth Vader.	Uncle Owen and Aunt Beru are related to Luke Skywalker, as they are his uncle and aunt. They are characters in the Star Wars series, along with Luke. Uncle Owen is also the stepbrother of Anakin Skywalker, who later becomes Darth Vader in the series. Uncle Owen and Aunt Beru raised Luke Skywalker.
where was donovan mitchell picked in the draft	Donovan Mitchell was picked 13th overall in the 2017 NBA Draft.	Donovan Mitchell, who is a person, was picked 13th overall in the 2017 NBA Draft.
where did the beatles final live performance take place	The Beatles' final live performance took place on the rooftop of the Apple Corps building in London, England on January 30, 1969.	The Beatles' final live performance took place on January 30, 1969, on a rooftop. This rooftop was of the Apple Corps building, which is located in London, England.
when does maggie come on grey's anatomy	Maggie Pierce, played by Kelly McCreary, first appears in Grey's Anatomy in the last episode of Season 10, which aired on May 15, 2014. However, her character is formally introduced in the first episode of Season 11, which aired on September 25, 2014.	Maggie Pierce, a character in Grey's Anatomy, first appears in the series on the last episode of Season 10. This episode aired on May 15, 2014.
who are the actors in fast and furious	1. Vin Diesel as Dominic Toretto\n 2. Paul Walker as Brian O'Conner\n 3. Michelle Rodriguez as Letty Ortiz\n 4. Jordana Brewster as Mia Toretto\n 5. Tyrese Gibson as Roman Pearce\n 6. Ludacris as Tej Parker\n 7. Dwayne Johnson as Luke Hobbs\n 8. Jason Statham as Deckard Shaw\n 9. Sung Kang as Han Lue\n 10. Gal Gadot as Gisele Yashar\n 11. Elsa Pataky as Elena Neves\n 12. Nathalie Emmanuel as Ramsey\n 13. John Cena as Jakob Toretto\n 14. Charlize Theron as Cipher\n 15. Lucas Black as Sean Boswell\n 16. Bow Wow as Twinkie (Fast and Furious: Tokyo Drift)\n 17. Eva Mendes as Monica Fuentes (2 Fast 2 Furious and Fast Five)\n 18. Kurt Russell as Mr. Nobody\n 19. Helen Mirren as Magdalene Shaw\n 20. Scott Eastwood as Little Nobody (The Fate of the Furious and F9) \n Please note that this list includes main characters from the entire Fast & Furious franchise, not just the first movie.	The Fast & Furious franchise features a star-studded cast. Vin Diesel plays the character Dominic Toretto, while Paul Walker is known for his role as Brian O'Conner. Michelle Rodriguez portrays Letty Ortiz and Jordana Brewster plays Mia Toretto. Tyrese Gibson takes on the role of Roman Pearce, and Ludacris is cast as Tej Parker. Dwayne Johnson is recognized as Luke Hobbs, and Jason Statham plays Deckard Shaw. Sung Kang is cast as Han Lue, and Gal Gadot plays Gisele Yashar. Elsa Pataky portrays Elena Neves, and Nathalie Emmanuel is known for her role as Ramsey. John Cena plays Jakob Toretto, and Charlize Theron is cast as Cipher. This list includes main characters from the entire Fast & Furious franchise, not just the first movie.

Table 4. NQ with $\alpha = 0.1$ and frequency scoring: comparison of original output by GPT-4 versus our modified output from conformal factuality. Incorrect claims are highlighted in red.

Question	Original	Conformal Factuality
How many vertical asymptotes does the graph of $y = \frac{2}{x^2+x-6}$ have?	The graph of $y = \frac{2}{x^2+x-6}$ has vertical asymptotes where the denominator is equal to zero. The denominator x^2+x-6 factors as $(x-2)(x+3)$, so the vertical asymptotes are at $x = 2$ and $x = -3$. Therefore, the graph has $\boxed{2}$ vertical asymptotes.	The graph of $y = \frac{2}{x^2+x-6}$ has vertical asymptotes where the denominator is equal to zero. The denominator $x^2 + x - 6$ factors as $(x - 2)(x + 3)$. Therefore, the vertical asymptotes are at $x = 2$ and $x = -3$. So, the graph has 2 vertical asymptotes.
What is the positive difference between 120% of 30 and 130% of 20?	120% of 30 is $1.2(30) = 36$ and 130% of 20 is $1.3(20) = 26$. The positive difference between these two numbers is $36 - 26 = \boxed{10}$.	First, find 120% of 30 by multiplying 1.2 by 30 to get 36. Then, find 130% of 20 by multiplying 1.3 by 20 to get 26. Finally, find the positive difference between 36 and 26 which is 10.
Find the sum of all integers that satisfy these conditions: $\|x\|+1>7$ and $\|x+1\|\le 7$.	The first inequality, $\|x\|+1>7$, is equivalent to $\|x\|>6$, which means $x>6$ or $x<-6$. The second inequality, $\|x+1\|\le 7$, is equivalent to $-7\le x+1\le 7$, or $-8 \le x \le 6$. The integers that satisfy both inequalities are -7 and 7, so their sum is $\boxed{0}$.	The first inequality, $\|x\|+1>7$, is equivalent to $\|x\|>6$. This means $x>6$ or $x<-6$. The second inequality, $\|x+1\|\le 7$, is equivalent to $-7\le x+1\le 7$. This can be rewritten as $-8 \le x \le 6$. The sum of the integers -7 and 7 is 0.
What is the number of units in the distance from the origin to the point $(-8,6)$ in a coordinate system?	The distance from the origin to a point (x,y) in a coordinate system is given by the formula $\sqrt{x^2+y^2}$. Substituting $x = -8$ and $y = 6$ gives $\sqrt{(-8)^2+6^2} = \sqrt{64+36} = \sqrt{100} = \boxed{10}$ units.	The distance from the origin to a point (x,y) in a coordinate system is given by the formula $\sqrt{x^2+y^2}$. If we substitute $x = -8$ and $y = 6$ into the formula, we get $\sqrt{(-8)^2+6^2}$. This calculation simplifies to $\sqrt{64+36}$, which further simplifies to $\sqrt{100}$. Therefore, the distance from the origin to the point $(-8,6)$ is 10 units.
If A, B and C are positive integers such that $\frac{A\sqrt{B}}{C} = \frac{9}{2\sqrt{3}}$, what is the value of $A+B+C$ given that A and C have no common prime factors, and B has no perfect-square factors other than 1?	We can simplify the right side of the equation to get $\frac{A\sqrt{B}}{C} = \frac{9}{2\sqrt{3}} = \frac{9\sqrt{3}}{6} = \frac{3\sqrt{3}}{2}$. Setting this equal to $\frac{A\sqrt{B}}{C}$, we get $A=3$, $B=3$, and $C=2$, so $A+B+C = \boxed{8}$.	First, the right side of the equation is simplified to get $\frac{A\sqrt{B}}{C} = \frac{9}{2\sqrt{3}} = \frac{9\sqrt{3}}{6} = \frac{3\sqrt{3}}{2}$. Setting this equal to $\frac{A\sqrt{B}}{C}$, we find that $A = 3$, $B = 3$, and $C = 2$. Therefore, the sum of A, B, and C is 8.
Ten treeks weigh as much as three squigs and one goolee. Two treeks and one goolee are equal in weight to one squig. The combined weight of how many treeks equals the weight of one squig?	From the first statement, we can write the equation 10T = 3S + G (where T represents the weight of a treek, S represents the weight of a squig, and G represents the weight of a goolee). From the second statement, we can write the equation 2T + G = S. Substituting the second equation into the first, we get 10T = 3(2T + G) + G, which simplifies to 10T = 6T + 3G. Subtracting 6T from both sides, we get 4T = 3G. Substituting this into the second equation, we get 2T + 4T/3 = S, which simplifies to 10T/3 = S. Therefore, the combined weight of 10/3, or approximately 3.33 treeks, equals the weight of one squig.	First, we can represent the weight of a treek as T, the weight of a squig as S, and the weight of a goolee as G. From the first statement, we can derive the equation 10T = 3S + G. From the second statement, we can derive the equation 2T + G = S. We can then substitute the second equation into the first, resulting in the equation 10T = 3(2T + G) + G. This equation represents the relationship between the weights of treeks, squigs, and goolees.
Evaluate the expression $a^3 \cdot a^2$ if $a=5$.	We have $a^3 \cdot a^2 = a^{3+2} = a^5$. Therefore, if $a = 5$, then $a^3 \cdot a^2 = 5^5 = \boxed{3125}$.	First, we know that $a^3 \cdot a^2$ is equivalent to a^{3+2}. Then, a^{3+2} is equivalent to a^5. Given that $a=5$, we can substitute this value into the expression, so $a^3 \cdot a^2$ is equivalent to 5^5. Finally, we know that 5^5 equals 3125. Therefore, if $a = 5$, the expression $a^3 \cdot a^2$ evaluates to 3125.
Let $f(y) = y^4-3y^3+y-3$ and $g(y) = y^3+7y^2-2$. Find $f(y) + g(y)$. Write your answer as a polynomial with terms of decreasing degree.	We have $f(y)+g(y) = (y^4-3y^3+y-3)+(y^3+7y^2-2) = \boxed{y^4-2y^3+7y^2+y-5}$.	The function f(y) is represented by the equation y^4-3y^3+y-3 and the function g(y) is represented by the equation y^3+7y^2-2. The sum of these two functions, f(y) + g(y), equals $y^4-2y^3+7y^2+y-5$. Therefore, f(y) + g(y) = $y^4-2y^3+7y^2+y-5$.

Table 5. MATH with $\alpha = 0.1$ and frequency scoring: comparison of original output by GPT-4 versus our modified output from conformal factuality. Incorrect claims are highlighted in red.