E-Scores for (In)Correctness Assessment of Generative Model Outputs

Guneet S. Dhillon*
University of Oxford
guneet.dhillon@stats.ox.ac.uk

Javier González
Microsoft Research
jav.gonzalezh@gmail.com

Teodora Pandeva
Microsoft Research
tpandeva@microsoft.com

Alicia Curth
Microsoft Research
aliciacurth@microsoft.com

Abstract

While generative models, especially large language models (LLMs), are ubiquitous in today's world, principled mechanisms to assess their (in)correctness are limited. Using the conformal prediction framework, previous works construct sets of LLM responses where the probability of including an incorrect response, or error, is capped at a user-defined tolerance level. However, since these methods are based on p-values, they are susceptible to p-hacking, i.e., choosing the tolerance level post-hoc can invalidate the guarantees. We therefore leverage e-values to complement generative model outputs with e-scores as measures of incorrectness. In addition to achieving the guarantees as before, e-scores further provide users with the flexibility of choosing data-dependent tolerance levels while upper bounding size distortion, a post-hoc notion of error. We experimentally demonstrate their efficacy in assessing LLM outputs under different forms of correctness: mathematical factuality and property constraints satisfaction.

1 INTRODUCTION

Generative models, large language models (LLMs) in particular, have gained widespread popularity, with millions of users around the world (OpenAI, 2024a,b,c; Gemini Team, 2025). However, they are susceptible to generating incorrect outputs, or *hallucinations* (Huang et al., 2025), requiring caution in their use. Fig. 1 illustrates an example where an LLM's response contains incorrect steps or sub-responses. Since such correctness labels are unknown at test time, we need a mechanism to assess the (in)correctness of the generated responses.

A recent line of work aims to provide statistical guarantees for such LLM responses (Mohri and Hashimoto, 2024; Cherian et al., 2024; Rubin-Toles et al., 2025). These methods rely on *p-value* based conformal prediction (Shafer and Vovk, 2008; Vovk et al., 2022) to filter the response set such that the probability of including an incorrect response, or error, is capped at a user-defined tolerance level α. Implicitly, this is done by computing a p-value based score for each response, or *p-score*; then, the filtered response set is obtained simply by thresholding the p-scores $\leq \alpha$. Importantly, α must be chosen independently of the data. This begs the question: *what if we want a data-dependent α?*

Recall the example in Fig. 1, and imagine that the scores there are the p-scores in question. If the user pre-sets $\alpha = 0.1$, the responses are filtered for p-scores ≤ 0.1, resulting in the first two sub-responses. However, the user would have obtained the same first two sub-responses if they had pre-set $\alpha = 0.01$ instead. Since a tolerance level of 0.01 conveys a much higher assurance in the responses compared to 0.1, the user would want to update $\alpha = 0.1 \to 0.01$. This necessitates a data-dependent α, also called a *post-hoc* α. Unfortunately, even though such post-hoc α's are commonly used in practice, the guarantees for p-score based methods are invalidated. This is due to the susceptibility of p-values, and hence of p-scores, to *p-hacking* (Carney, 2016).

We propose *e-scores* as *measures of incorrectness*: they are low for correct and high for incorrect responses (depicted in Fig. 1). These scores, based on *e-values*,

*Work done while at Microsoft Research.

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

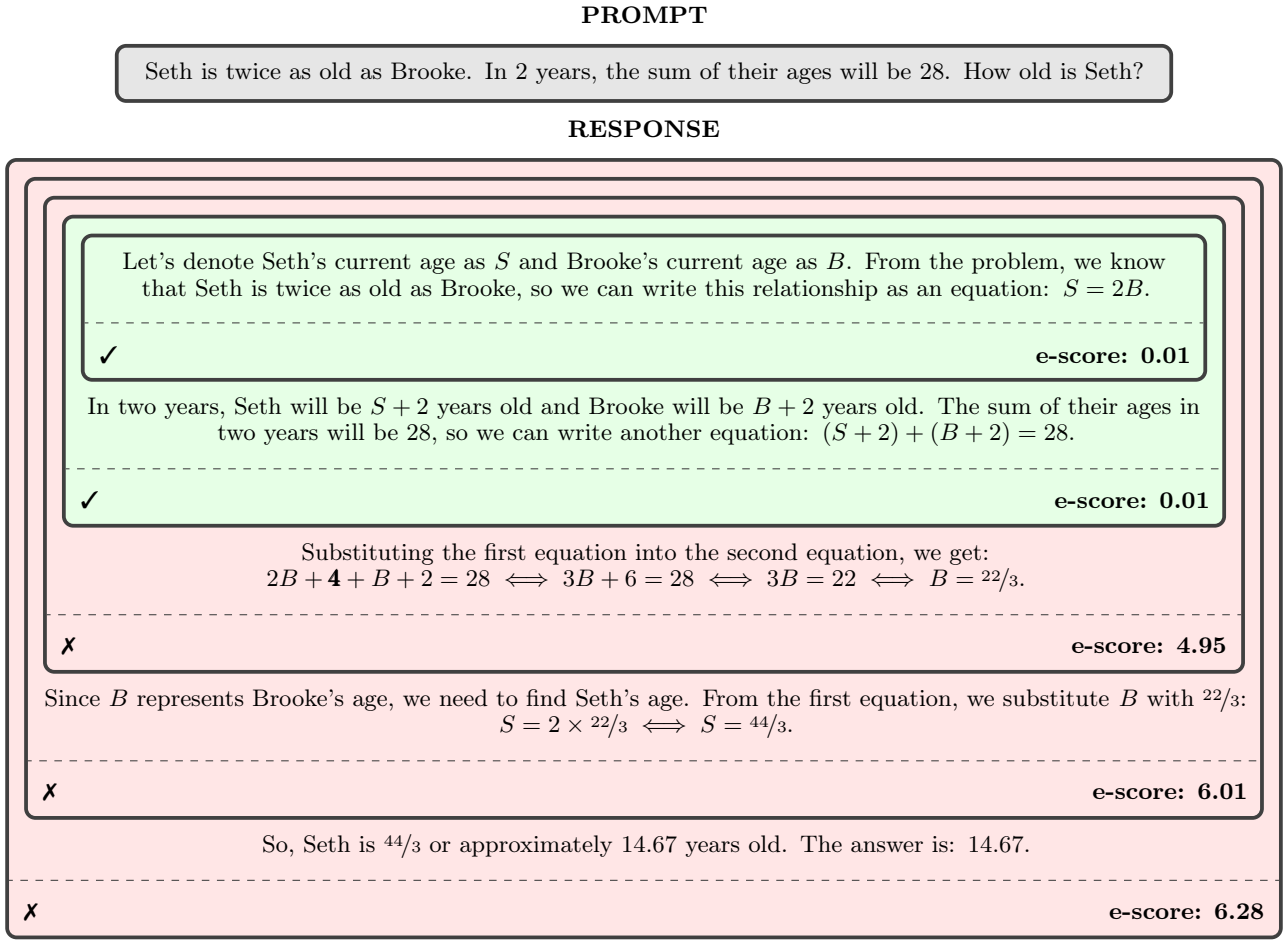


Figure 1: **E-scores example for mathematical factuality.** This is an example prompt and response from the ProcessBench benchmark (Zheng et al., 2025) (cf. Section 5.1). The LLM's response is made up of 5 sub-responses, each a step in the mathematical reasoning (starting from the inner and ending on the outer block). The checks/crosses on the bottom left and the green/red colour of each block represent whether the response is correct/incorrect up till that point (which is not known at test-time). On manual inspection, we bold part of the third sub-response causing the incorrectness, cascading to subsequent sub-responses. The e-scores on the bottom right of each block are our proposed measures of incorrectness: low for correct and high for incorrect responses.

provide statistical guarantees on a post-hoc notion of error called *size distortion* (Koning, 2024): the distortion between observing an error and the user's post-hoc tolerance level. The non-post-hoc error guarantee mentioned earlier arises as a special case of this generalization. Furthermore, we show that our statistical guarantees remain valid for any generative model and for a super-set of the response sets considered by Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025). Altogether, this provides avenues for more diverse applications and use-cases. We experimentally demonstrate the efficacy of our e-scores in the assessment of LLM responses under two settings: (i) mathematical factuality, for sound mathematical reasoning, and (ii) property constraints satisfaction, to ensure responses satisfy certain desirable properties.

We summarize the contributions of our work as follows.

- We study the problem of achieving statistical guarantees for a post-hoc notion of error, namely size distortion, for generative model outputs. This generalizes the non-post-hoc guarantees studied before for LLM responses (Mohri and Hashimoto, 2024; Cherian et al., 2024; Rubin-Toles et al., 2025).
- We propose e-scores, based on e-values, as measures of incorrectness. Our theoretical results show that e-scores achieve the post-hoc statistical guarantees mentioned above. In doing so, e-scores provide users the flexibility of choosing data-dependent tolerance levels. Furthermore, we corroborate our theory with experimental results.

- We show that our guarantees extend to any generative model and to a super-set of the response sets considered by Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025).

We begin by formulating our problem in Section 2, and discuss related work in Section 3. We define our proposed e-scores in Section 4. Sections 5 and 6 include experimental results and theoretical analyses, respectively. We finish with concluding remarks in Section 7.

2 PROBLEM FORMULATION

We are interested in providing guarantees pertaining to the (in)correctness of generative model outputs. We describe: (i) the generative model outputs we consider, (ii) their (in)correctness, (iii) the desired post-hoc guarantees, and (iv) practical examples of post-hoc use. While we use LLMs to provide concrete examples, our setup could be instantiated with any generative model.

2.1 Generative Model Outputs

We define the prompt space $\mathcal{X}$, a sub-space of language. Given a prompt $x \in \mathcal{X}$, an LLM π generates a response,

$$g_\pi(x) = \mathbf{y} = (y_1, y_2, \ldots) \sim \pi(\cdot|x),$$

an ordered set of sub-responses that collectively answer the prompt. These sub-responses could be sentences, steps in chain-of-thought reasoning (Wei et al., 2022), etc., akin to the example in Fig. 1. While natural for auto-regressive models, we do not assume any particular dependency structure. Note that singular responses of length $|\mathbf{y}| = 1$ are a special case. We define the sub-response space $\mathcal{Y}$, also a sub-space of language. Then, each sub-response $y_i \in \mathcal{Y}$ and the response $\mathbf{y} \in \cup_{i \geq 1} \mathcal{Y}^i$.

We could evaluate $\mathbf{y}$ alone, but we consider a larger set of responses. Notice that a *single* generated response can form *multiple* responses: the partial responses $\mathbf{y}_{\leq i} = (y_1, \ldots, y_i)$, for $i = 1, \ldots, |\mathbf{y}|$. In doing so, the user could find partial but correct responses, similar to the example in Fig. 1. We define such a response set,

$$\mathbb{Y}(g_\pi(x)) = \{\mathbf{y}_{\leq i} : \mathbf{y} = g_\pi(x), i = 1, \ldots, |\mathbf{y}|\}. \quad (1)$$

Note that this includes the full generated response itself $g_\pi(x) \in \mathbb{Y}(g_\pi(x))$. We will generalize this response set to a larger one in Section 6.1. However, we continue to use the definition in Eq. (1) as it is suited for the experimental benchmarks we consider (cf. Section 5).

2.2 Oracle (In)Correctness

There are different notions of *correctness* that are of interest to us. Factuality is a pertinent one, to verify whether responses are based on facts (Mohri and Hashimoto, 2024; Cherian et al., 2024; Rubin-Toles et al., 2025). Another is property constraints satisfaction, to ensure responses have certain desirable properties (Dhillon et al., 2025). To adapt to such use-cases, we define (in)correctness *with respect to an oracle* o,

$$o(x, \mathbf{y}) = \begin{cases} 1, & \mathbf{y} \text{ is correct as a response to } x \\ 0, & \mathbf{y} \text{ is incorrect as a response to } x \end{cases},$$

determining the correctness of $\mathbf{y}$ as a response to the prompt x. Further, we define the labeled response set,

$$\mathbb{O}(x, g_\pi(x)) = \{(\mathbf{y}, o(x, \mathbf{y})) : \mathbf{y} \in \mathbb{Y}(g_\pi(x))\}.$$

2.3 Desideratum for Statistical Guarantees

Given an LLM π and a prompt $x \in \mathcal{X}$, we construct its (unlabeled) response set $\mathbb{Y}(g_\pi(x))$. Our goal then is to assess the (in)correctness of each response in this set, i.e., we want to reason about the unknown oracle labels $o(x, \mathbf{y})$ for every response $\mathbf{y} \in \mathbb{Y}(g_\pi(x))$. We do so by complementing each response with a non-negative score $s(x, \mathbf{y}) \in \mathbb{R}_{\geq 0}$ as a *measure of incorrectness*: low for correct responses and high for incorrect responses. Consequently, we will provide the scored response set,

$$\mathbb{S}(x, g_\pi(x)) = \{(\mathbf{y}, s(x, \mathbf{y})) : \mathbf{y} \in \mathbb{Y}(g_\pi(x))\},$$

to facilitate the user in deciding which responses to include (and use) or not. In particular, they could decide to filter the scored response set at some $\alpha \in \mathbb{R}_{\geq 0}$,

$$\mathbb{S}_\alpha(x, g_\pi(x)) = \{(\mathbf{y}, v) \in \mathbb{S}(x, g_\pi(x)) : v \leq \alpha\}.$$

Since we want to avoid incorrect responses, we treat the inclusion of any incorrect response in the filtered set $\mathbb{S}_\alpha(x, g_\pi(x))$ as an *error at* α. Then, a possible desideratum for our measures of incorrectness is to upper bound the probability of error at α by α itself,

$$\mathbf{P}\{\text{error at } \alpha\} = \mathbf{P}\left\{\begin{array}{r} \exists (\mathbf{Y}, \cdot) \in \mathbb{S}_\alpha(X, g_\pi(X)) \\ \text{s.t. } o(X, \mathbf{Y}) = 0 \end{array}\right\} \leq \alpha. \quad (2)$$

In doing so, α represents the user's tolerance level. This is considered by Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025), who use p-value based conformal prediction (Shafer and Vovk, 2008; Vovk et al., 2022) to achieve this. Note that the above requirement assumes that α is determined independently of the data (the prompt, the responses, and the scores). However, in practice, users would want to use a data-dependent tolerance level α, as highlighted by the scenario in Section 1. On realizing that they would obtain the same filtered set if they pre-set $\alpha = 0.01$ instead of 0.1 in the example in Fig. 1, the user wants to update $\alpha = 0.1 \to 0.01$, conveying higher

assurance in the responses with a smaller tolerance level. This necessitates a data-dependent α, or a *post-hoc* α.

Specifically, we want to enable the user to choose α after observing the prompt x and the scored response set $\mathbb{S}(x, g_\pi(x))$. Since α is now a random variable, Eq. (2) cannot be applied directly. We therefore generalize our desideratum instead to a post-hoc notion of error at α,

$$\mathbf{E}\left[\frac{\mathbb{1}\left\{\text{error at } \alpha\left(X, \mathbb{S}\left(X, g_\pi\left(X\right)\right)\right)\right\}}{\alpha\left(X, \mathbb{S}\left(X, g_\pi\left(X\right)\right)\right)}\right] = \mathbf{E}\left[\frac{\mathbb{1}\left\{\begin{array}{r}\exists(\mathbf{Y}, \cdot) \in \mathbb{S}_{\alpha(X, \mathbb{S}(X, g_\pi(X)))}\left(X, g_\pi\left(X\right)\right) \\ \text{s.t. } o\left(X, \mathbf{Y}\right) = 0\end{array}\right\}}{\alpha\left(X, \mathbb{S}\left(X, g_\pi\left(X\right)\right)\right)}\right] \leq 1. \tag{3}$$

The ratio of observing an error at α and α itself is expected to be at most 1. The ratio captures the distortion between observing an error and the user's tolerance level; furthermore, the bound ensures that the expected distortion is controlled. Koning (2024) uses such an expected distortion for hypothesis testing, calling it *size distortion* (with size referring to an error). This generalizes Eq. (2), recovered when α is a fixed pre-set value. Thus, Eq. (3) will be our new desideratum.

Role of the Calibration Data To aid us in our desideratum, we are given labeled calibration data that is assumed to be exchangeable with the test data, which is similar to Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025). This includes n calibration prompts $x^i \in \mathcal{X}$, for $i = 1, \ldots, n$, each with their corresponding labeled response set $\mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)$. Now, given the test prompt $x^{n+1} \in \mathcal{X}$, we will complement each test response $\mathbf{y}^{n+1} \in \mathbb{Y}\left(g_\pi\left(x^{n+1}\right)\right)$ with a non-negative test score $s\left(x^{n+1}, \mathbf{y}^{n+1}\right) \in \mathbb{R}_{\geq 0}$ as a measure of incorrectness, which can additionally depend on the given calibration data. For simplicity and compactness, we will suppress the notation for the dependence of the scores on the calibration data; note that the requirement in Eq. (3) (and in Eq. (2)) is now *marginal over both the test and the calibration data.*

2.4 Use of Post-Hoc α's

We provide two concrete examples of post-hoc α strategies. A user could choose either one of these or any other strategy; in return, we will aim to satisfy Eq. (3).

Max-Constrained Adaptive α The user has a fixed pre-set maximum tolerance level $\alpha_{\max} \in [0, 1]$. Since the scores in the filtered set at $\alpha_{\max}$ could be $\leq \alpha_{\max}$, the user updates their tolerance level α to the maximum score in the corresponding filtered set $\mathbb{S}_{\alpha_{\max}}(x, g_\pi(x))$,

$$\alpha\left(x, \mathbb{S}\left(x, g_\pi\left(x\right)\right)\right) = \max_{(\cdot, v) \in \mathbb{S}_{\alpha_{\max}}(x, g_\pi(x))} v.$$

Fractional Inclusion Alternatively, the user might want to include $\lambda \in [0, 1]$ fraction of all responses in the set $\mathbb{S}(x, g_\pi(x))$. Then, the user's tolerance level α is the maximum score in the corresponding filtered set,

$$\alpha\left(x, \mathbb{S}\left(x, g_\pi\left(x\right)\right)\right) = \max_{(\cdot, v) \in \mathbb{S}_\alpha(x, g_\pi(x))} v$$
$$\text{s.t. } \left|\mathbb{S}_\alpha\left(x, g_\pi\left(x\right)\right)\right| = \lceil \lambda \cdot \left|\mathbb{S}\left(x, g_\pi\left(x\right)\right)\right| \rceil.$$

Note that by setting $\lambda = 1$, the user could get post-hoc error guarantees for the full generated response itself.

After choosing a post-hoc α strategy, the user gets the filtered $\mathbb{S}_{\alpha(x, \mathbb{S}(x, g_\pi(x)))}(x, g_\pi(x))$ for downstream use. The user could, for example, treat the longest response in the filtered set as the default response, as done by Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025), now with added guarantees.

3 RELATED WORK

We begin with two key definitions. Consider a non-negative random variable $R \in \mathbb{R}_{\geq 0}$. It is a *p-variable* if $\mathbf{P}\{R \leq \alpha\} \leq \alpha$, for all $\alpha \in \mathbb{R}_{\geq 0}$. And, it is an *e-variable* if $\mathbf{E}[R] \leq 1$ (which, with Markov's inequality, gives $\mathbf{P}\{1/R \leq \alpha\} \leq \alpha$, for all $\alpha \in \mathbb{R}_{\geq 0}$). Furthermore, its realized value is called a p- and e-value, respectively.

While p-values have been used for hypothesis testing (Neyman and Pearson, 1933; Wald, 1939), recent developments highlight the benefits of e-values (Shafer and Vovk, 2019; Wasserman et al., 2020; Shafer, 2021; Vovk and Wang, 2021; Wang and Ramdas, 2022; Grünwald et al., 2024; Ramdas and Wang, 2025). Notably, Grünwald (2024) emphasizes their use in post-hoc α settings. Since we are also interested in post-hoc α's, we base our scores on e-values to attain statistical guarantees.

Closest to our work is that of Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025). Mohri and Hashimoto (2024); Rubin-Toles et al. (2025) adapt ideas from conformal prediction (Shafer and Vovk, 2008; Vovk et al., 2022), typically used to construct prediction sets for supervised learning problems, to filter LLM outputs to construct response sets $\mathbb{S}_\alpha(x, g_\pi(x))$ for a fixed pre-set α. In both these works, the dependence on p-values is implicit through their use of the nested conformal framework (Gupta et al., 2022). Additionally, rather than a single fixed α, Cherian et al. (2024) consider a functional α that can vary to improve fractional inclusion, but is required to be independent of the scores. Consequently, these works satisfy Eq. (2), but not its post-hoc generalization in Eq. (3). We therefore design our scores to achieve the latter for any generative model and its outputs. Furthermore, our theoretical results extend to the assessment of response sets that are larger than those of Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025).

4 E-SCORES

We now describe the scores we propose to achieve Eq. (3). We defer the theoretical results that justify our design choices to Section 6.2; but intuitively, our scores must be reciprocals of the corresponding e-values. Consequently, we call our proposed scores the *e-scores*.

The functional form of our e-scores is influenced by Gammerman et al. (1998), who construct e-values for supervised learning under exchangeable data (used by Balinsky and Balinsky (2024); Vovk (2025); Gauthier et al. (2025) for the same). Specifically, we define our e-score for each test response $\mathbf{y}^{n+1} \in \mathbb{Y}\left(g_\pi\left(x^{n+1}\right)\right)$,

$$s_{\text{e-score}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) = \left(\frac{(n+1) \cdot f\left(x^{n+1}, \mathbf{y}^{n+1}\right)}{f\left(x^{n+1}, \mathbf{y}^{n+1}\right) + \sum_{i=1}^{n} f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)}\right)^{-1}, \tag{4}$$

where f is any function mapping a prompt x and response $\mathbf{y}$ to a non-negative value $f(x, \mathbf{y}) \in \mathbb{R}_{\geq 0}$, and,

$$f^*\left(x, \mathbb{O}\left(x, g_\pi\left(x\right)\right)\right) = \max_{(\mathbf{y}, c) \in \mathbb{O}(x, g_\pi(x)) : c = 0} f(x, \mathbf{y}),$$

is the maximum incorrect response value (set to 0 in the absence of an incorrect response).[1] As a result, our e-scores compare a test response value with the incorrect calibration response values. The specific definition of f^* provides guarantees pertaining to the inclusion of any incorrect response (cf. Section 6.2), similar to the non-conformity functions in Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025).

For our e-scores to be measures of incorrectness, $f(x, \mathbf{y})$ should intuitively be a proxy for the oracle $o(x, \mathbf{y})$: high for correct responses and low for incorrect responses. If the oracle were known, $f_o(x, \mathbf{y})$ could be any monotonically increasing transformation of the oracle. This includes, but is not limited to: (i) $o(x, \mathbf{y})$, (ii) $(1 - o(x, \mathbf{y}))^{-1}$, and (iii) $o(x, \mathbf{y}) \cdot (1 - o(x, \mathbf{y}))^{-1}$.

However, since the oracle is unknown, we approximate it with $\hat{o}$. Obtaining such an estimator is a binary classification problem (since the oracle is binary), where $\hat{o}$ predicts the probability of correctness, now in the range $[0, 1]$. Note that the data used for training $\hat{o}$ should be independent of the test and calibration data. With an estimator $\hat{o}$, the above transformation options for $f_{\hat{o}}(x, \mathbf{y})$ translate to e-scores with different ranges,

$$f_{\hat{o}}(x, \mathbf{y}) = \begin{cases} \hat{o}(x, \mathbf{y}) \in [0, 1] & \text{(for e-score 1)} \\ (1 - \hat{o}(x, \mathbf{y}))^{-1} \in [1, \infty] & \text{(for e-score 2)} \\ \hat{o}(x, \mathbf{y}) \cdot (1 - \hat{o}(x, \mathbf{y}))^{-1} \in [0, \infty] & \text{(for e-score 3)} \end{cases}. \tag{5}$$

[1] We follow the convention $a/0 = 0$ if $a = 0$, otherwise $\pm\infty$.

Algorithm 1: E-Scores

Input: Generative model π
Input: Test prompt x^{n+1}
Input: Calibration prompt x^i and labeled calibration responses $\mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)$, for $i = 1, \ldots, n$
Input: Transformed oracle estimator $f_{\hat{o}}$
Output: Test scored responses $\mathbb{S}\left(x^{n+1}, g_\pi\left(x^{n+1}\right)\right)$

$g_\pi\left(x^{n+1}\right) \leftarrow \pi\left(\cdot | x^{n+1}\right)$; `/* generation */`
$\mathbb{S} \leftarrow \emptyset$; `/* initialization */`
for $\mathbf{y}^{n+1} \in \mathbb{Y}\left(g_\pi\left(x^{n+1}\right)\right)$; `/* response set */`
do
  $v^{n+1} \leftarrow \left(\frac{(n+1) \cdot f_{\hat{o}}\left(x^{n+1}, \mathbf{y}^{n+1}\right)}{f_{\hat{o}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) + \sum_{i=1}^{n} f^*_{\hat{o}}\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)}\right)^{-1}$; `/* e-score */`
  $\mathbb{S} \leftarrow \mathbb{S} \cup \left\{\left(\mathbf{y}^{n+1}, v^{n+1}\right)\right\}$; `/* scored response */`
return $\mathbb{S}$

Therefore, with an oracle estimator $\hat{o}$ and its transformation $f_{\hat{o}}$, we can compute our proposed e-scores in Eq. (4). We summarize this e-scoring mechanism in Algorithm 1. In Section 6.2, we will show that our e-scores achieve Eq. (3) for any choice of $\hat{o}$ and $f_{\hat{o}}$, regardless of the possible errors in approximating o.

4.1 Combining E-Scores

With different options for the transformed oracle estimator in Eq. (5), the use-case would determine the choice in practice. Alternatively, without making assumptions about or restricting the use-cases, one can opt to combine multiple e-scores by first combining the underlying e-values. We use the fact that simple averaging of e-values yields an admissible e-value (Vovk and Wang, 2021). Let $s_{\text{e-score }(i)}\left(x^{n+1}, \mathbf{y}^{n+1}\right)$ for $i = 1, 2, 3$ be the three e-scores corresponding to the options in Eq. (5). Then, we can combine them into one e-score,

$$s_{\text{e-score (combined)}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) = \left(\frac{\sum_{i=1}^{3}\left(s_{\text{e-score }(i)}\left(x^{n+1}, \mathbf{y}^{n+1}\right)\right)^{-1}}{3}\right)^{-1}. \tag{6}$$

We will use this e-score by default, unless mentioned otherwise. This is also used for the example in Fig. 1.

5 EXPERIMENTAL RESULTS

We now experimentally demonstrate the efficacy of our proposed e-scores with two practical settings. We discuss the use-cases individually, then summarize the observed trends collectively. We begin by stating the baselines, the metrics for comparisons, the time and memory complexities, and the experimental setup. We also perform a worst-case analysis in Appendix B, where the post-hoc α's maximize the size distortion from Eq. (3).

Baselines We compare against p-value based scores, or *p-scores*. Analogous to our e-scores in Eq. (4), the p-scores can be defined as the corresponding p-values,

$$s_{\text{p-score}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) = \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{ \begin{array}{l} f\left(x^{n+1}, \mathbf{y}^{n+1}\right) \\ \quad \leq f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right) \end{array} \right\}}{n+1}, \tag{7}$$

comparing a test response value with the incorrect calibration response values as relative ranks (Shafer and Vovk, 2008; Vovk et al., 2022). Mohri and Hashimoto (2024); Cherian et al. (2024); Rubin-Toles et al. (2025) use such p-scores implicitly to achieve Eq. (2), which we make explicit in Appendix C. Due to the reliance on relative ranks, the choice of the transformed oracle estimator in Eq. (5) does not matter, as they are monotonically increasing transformations of each other.

We also compare with the transformed oracle estimators in Eq. (5) directly, without any conversion to e- or p-scores. These naive scores generally do not come with any statistical guarantees by themselves. We will use these scores for our worst-case analysis in Appendix B.

Metrics Our comparisons are based on the following.

- *Size distortion.* This is the most important metric, from our desideratum in Eq. (3). We report its empirical mean, which is desired to be at most 1.
- *Error vs. α.* While we aim to control size distortion, the expected error to α ratio, one might also be interested in the expected error and expected α individually. We report empirical means for both. We ideally want the observed error to be lower than the tolerance level (mean error $\leq$ mean α).
- *Precision vs. recall.* We provide guarantees on the inclusion of any incorrect response. Simultaneously, we do not wish to exclude many correct responses. As a result, we report the empirical means for precision (fraction of correct responses among those included) and recall (fraction of correct responses included). We compare using the precision-recall curves (higher is better for both).

Memory and Time Complexities Our e-scores are cheaper to compute than the p-scores, in memory and in time. For a given test prompt-response pair, p-scores compute relative ranks with the calibration data (cf. Eq. (7)). This requires memory and time that grows linearly in the amount of calibration data n for every individual test prompt-response pair. Conversely, our e-scores compute a sum over the calibration data (cf. Eq. (4)). This requires constant memory and time that grows linearly in n. Furthermore, this is a one time cost, amortized over all test prompt-response pairs.

Setup We randomly split the data 50-50% into the test and calibration data (no training data is required as we use pre-trained oracle estimators). The metrics are averaged over 100 such random splits. We use an NVIDIA A100 GPU for the pre-trained oracle estimators; the remaining computations run on a CPU.

5.1 Mathematical Factuality

ProcessBench (Zheng et al., 2025) is a mathematical reasoning benchmark. It contains prompts from GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), OlympiadBench (He et al., 2024), and Omni-MATH (Gao et al., 2025). The responses are first generated by 12 open-source LLMs (LLaMA Team, 2024; Yang et al., 2024a,b; Qwen Team, 2025), then separated into multiple steps/sub-responses using Qwen2.5-72B-Instruct (Qwen Team, 2025). Lastly, human experts annotate the earliest-occurring incorrect sub-response.

Let $\mathbf{y} = (y_1, y_2, \ldots)$ be a generated response and i be its annotation. Then, the responses $\mathbf{y}_{\leq j}$ for $j = i, \ldots, |\mathbf{y}|$ are deemed incorrect as they contain the incorrect i-th sub-response, whereas the responses for $j = 1, \ldots, i-1$ are correct. Fig. 1 illustrates one such example. This determines the correctness or factuality of the responses.

Zheng et al. (2025) also benchmark math-based process reward models that predict the correctness probabilities of sub-responses individually. However, we want an oracle estimator to predict the correctness probability of (partial) responses. We follow Lightman et al. (2024) to obtain this by multiplying the individual (conditional) sub-response predictions like conditional probabilities,

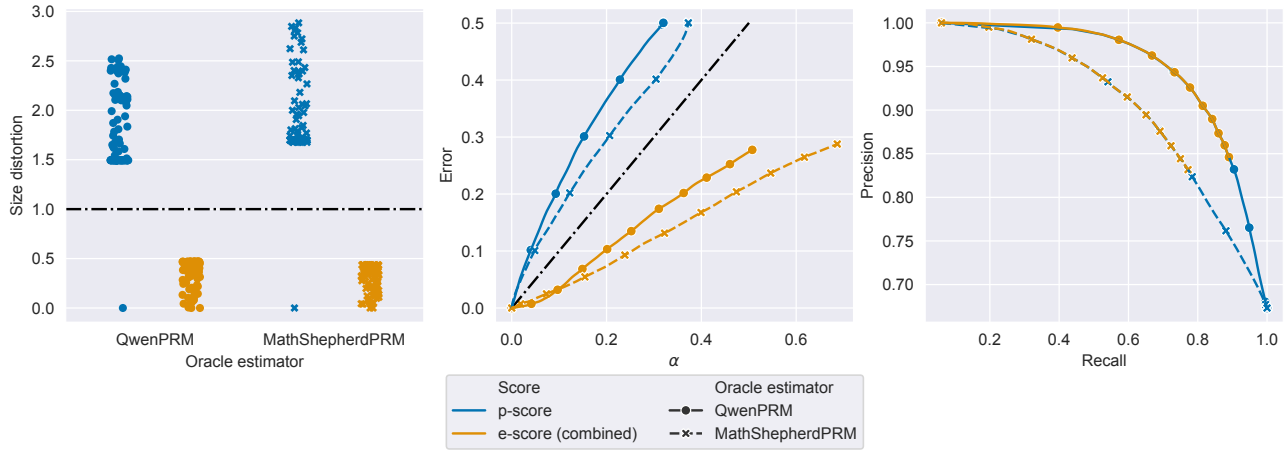
$$\hat{o}\left(x, \mathbf{y}\right) = \prod_{i=1}^{|\mathbf{y}|} \hat{o}\left(x, y_i \mid \mathbf{y}_{<i}\right).$$

We consider two process reward models for such oracle estimators: (i) Qwen2.5-Math-7B-PRM800K (or Qwen-PRM) (Zheng et al., 2025) and (ii) Math-Shepherd-PRM-7B (or MathShepherdPRM) (Wang et al., 2024).

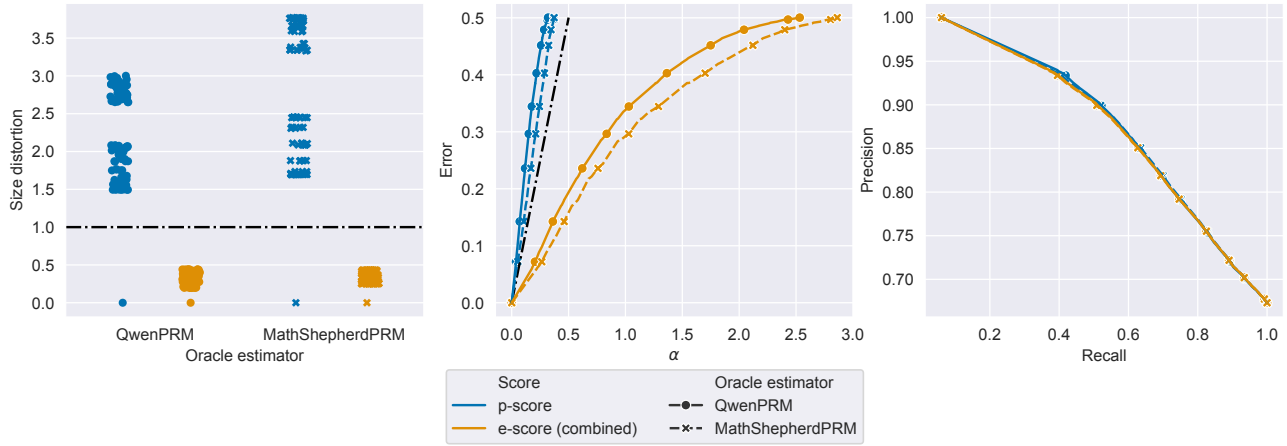
We examine both the post-hoc α strategies described in Section 2.4, and illustrate the results in Fig. 2. Furthermore, the example in Fig. 1 is from this use-case.

5.2 Property Constraints Satisfaction

UltraFeedback (Cui et al., 2024) is a diverse and fine-grained preference dataset. It contains prompts from 6 benchmarks (we will only use Evol-Instruct (Xu et al., 2024) and TruthfulQA (Lin et al., 2022)) and responses from 17 commercial and open-source LLMs (Chiang et al., 2023; Tunstall et al., 2023; Taori et al., 2023; Touvron et al., 2023; Biderman et al., 2023; Almazrouei et al., 2023; Ding et al., 2023; OpenAI, 2024a; Xu et al.,



(a) **Max-constrained adaptive α strategies.** We set $\alpha_{\max} = 0, .01, .02, \ldots, .99, 1$ (cf. Section 2.4).



(b) **Fractional inclusion strategies.** We set $\lambda = 0, .01, .02, \ldots, .99, 1$ (cf. Section 2.4).

Figure 2: **Scores for mathematical factuality.** We use the setting in Section 5.1 to compare our proposed e-scores (in orange) against p-scores (in blue). The left graphs plot size distortion (cf. Eq. (3)). The center graphs plot mean error vs. mean α (where the dashed black line is the identity line). The right graphs plot mean precision vs. mean recall (with the e-scores curves overlapping and hiding part of or the entire p-scores curves).

2024). It also employs GPT-4 (OpenAI, 2024a) to provide a rating for every response on four different criteria: helpfulness, honesty, instruction-following, and truthfulness; these are numerical ratings from 1 to 5.

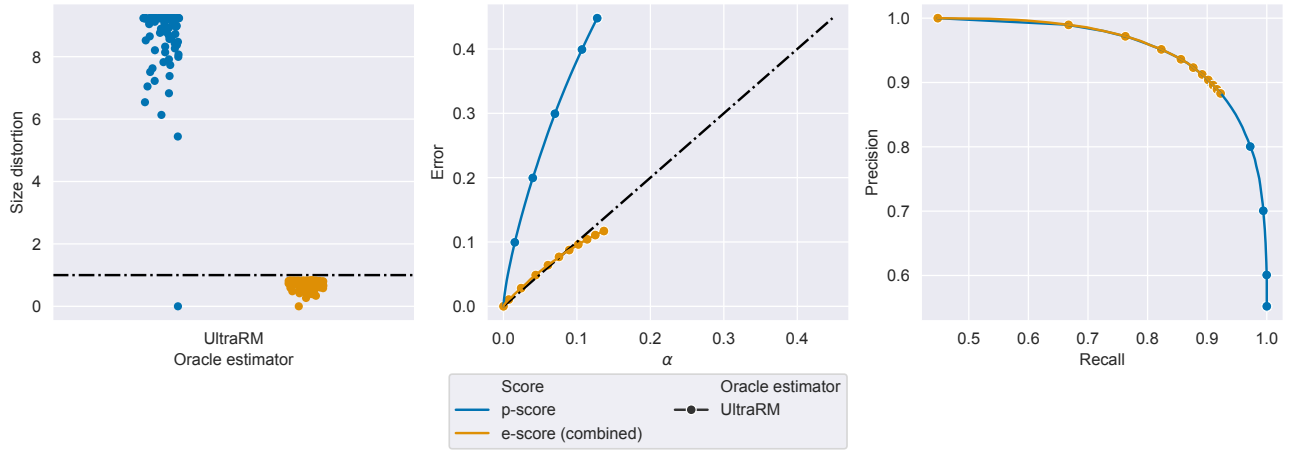
In practice, users are often interested in responses that satisfy certain desirable properties; this is equivalent to thresholding or constraining the property ratings to their corresponding desirable values (Dhillon et al., 2025). We use such a constraining mechanism to define the correctness of responses for two different use-cases.

Helpfulness and Instruction-Following We define a response to be correct if both its helpfulness and instruction-following ratings are either 4 or higher. We use prompts from the Evol-Instruct benchmark (Xu et al., 2024). We illustrate these results in Fig. 3a.

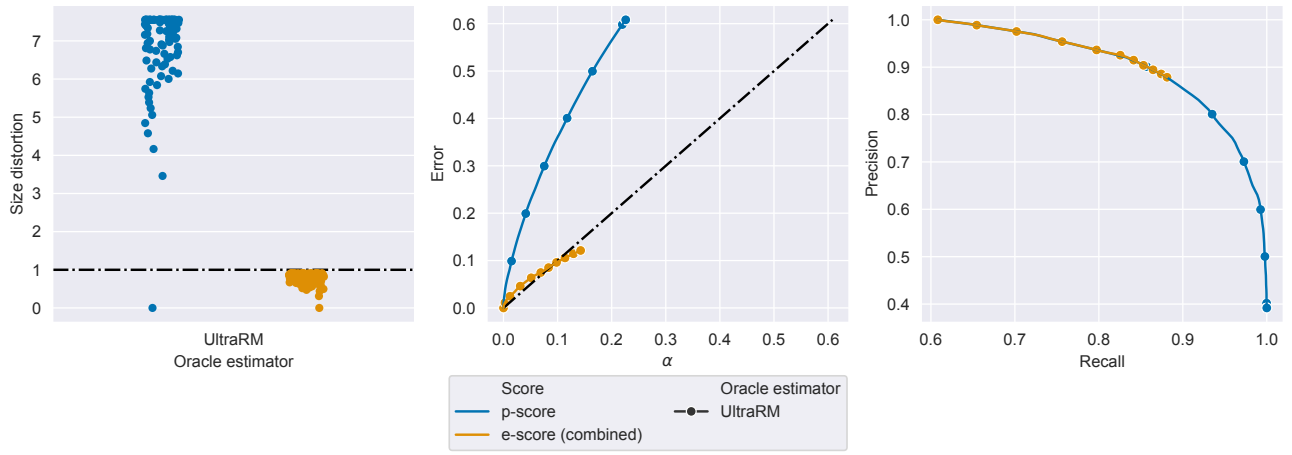
Honesty and Truthfulness We define a response to be correct if both its honesty and truthfulness ratings are 5. We use prompts from the TruthfulQA benchmark (Lin et al., 2022). We illustrate these results in Fig. 3b.

Cui et al. (2024) also provide a reward model, UltraRM, that predicts a real-valued preference for a response. We follow the Bradley-Terry model (Bradley and Terry, 1952) (although without a reference preference value, or equivalently, a reference preference of 0) and append the pre-trained UltraRM with a sigmoid operator, updating its range to $[0, 1]$. We use this as our oracle estimator.

Note that these use-cases consider the full responses with $|\mathbf{y}| = 1$. As a result, we only examine the max-constrained adaptive α strategy (cf. Section 2.4); fractional inclusion would include or exclude all responses.



(a) **Helpfulness and instruction-following.** A response is correct if its helpfulness and instruction-following ratings ≥ 4.



(b) **Honesty and truthfulness.** A response is correct if its honesty and truthfulness ratings $= 5$.

Figure 3: **Scores for property constraints satisfaction.** We use the setting in Section 5.2 to compare our proposed e-scores (in orange) against p-scores (in blue). We consider different max-constrained adaptive α strategies, setting $\alpha_{\max} = 0, .01, .02, \ldots, .99, 1$ (cf. Section 2.4). The left graphs plot size distortion (cf. Eq. (3)). The center graphs plot mean error vs. mean α (where the dashed black line is the identity line). The right graphs plot mean precision vs. mean recall (with the e-scores curves overlapping and hiding part of the p-scores curves).

5.3 Observed Trends

The trends we observe for the different experimental use-cases are consistent; we summarize them together.

Size Distortion Our proposed e-scores reliably upper bound size distortion to 1 and satisfy Eq. (3), corroborating our theory in Section 6.2. Conversely, p-scores are unable to achieve this; the only time they experimentally do is when all responses (correct and incorrect) are excluded, achieving 0 error by default.

Error vs. α Our proposed e-scores consistently obtain a mean error lower than or approximately equal to the mean tolerance α. Conversely, p-scores consistently obtain a mean error higher than the mean tolerance α.

Precision vs. Recall The precision-recall curves of our proposed e-scores overlap with those of the p-scores. In satisfying Eq. (3), the e-scores are more conservative and prefer maintaining high precision over high recall. Consequently, restricting α's to be ≤ 1 (under the max-constrained adaptive α strategies) restricts the e-score recalls compared to the p-score recalls, resulting in partial overlap. However, removing this restriction (under the fractional inclusion strategies) retains complete overlap of the e- and p-score precision-recall curves.

Oracle Estimator The choice of the oracle estimator impacts the metrics. This is best illustrated by

the precision-recall curves in Fig. 2a, where Qwen-PRM achieves higher precisions and recalls compared to MathShepherdPRM. This is expected as the former is comparatively more accurate (Zheng et al., 2025).

6 THEORETICAL RESULTS

We present our theoretical results here. We generalize the response set from Eq. (1) to a larger set, and show that our e-scores satisfy our desideratum in Eq. (3).

6.1 Super-Set of Responses

We intend to make the response set from Eq. (1) as large as possible, while maintaining statistical guarantees. This would enable the assessment of a larger set of responses, opening avenues for more diverse applications and use-cases. So, we consider the responses,

$$\mathbb{Y}\left(g_\pi\left(x\right)\right) = \cup_\sigma \left\{\mathbf{y}_{\leq i} : \mathbf{y} = \sigma\left(g_\pi\left(x\right)\right), i = 1, \ldots, |\mathbf{y}|\right\}, \tag{8}$$

where $\sigma\left(g_\pi\left(x\right)\right)$ is a permuted version of the generated response $g_\pi\left(x\right)$, and the union is over all permutations.

If we fix σ to the identity ordering only, we recover Eq. (1). Similarly, Rubin-Toles et al. (2025) restrict σ to orderings (what they call topological orderings of an approximate deducibility graph) obtained from GPT-4o (OpenAI, 2024b). Lastly, Mohri and Hashimoto (2024); Cherian et al. (2024) do not account for the inherent ordering of the sub-responses to make up a response. Therefore, Eq. (8) is a super-set of responses, containing all the response sets discussed above. In fact, we believe that it is the largest set of responses to consider when given a *single* generated response $g_\pi\left(x\right)$.

Instead of the full response set in Eq. (8), one might choose to use a sub-set depending on the use-case, while maintaining guarantees. For example, we use Eq. (1) in Section 5.1 as it is tailor-made for that benchmark.

6.2 Worst-Case Analysis and E-Values

We are interested in achieving the desideratum in Eq. (3) for any post-hoc α that a user might choose. Without restricting the user's choice, we will analyze the setting where $\alpha\left(x, \mathbb{S}\left(x, g_\pi\left(x\right)\right)\right)$ maximizes size distortion. If Eq. (3) is satisfied under this worst-case setting, it will also be satisfied under any post-hoc α.

Note that a response is included in the filtered set $\mathbb{S}_\alpha\left(x, g_\pi\left(x\right)\right)$ if and only if its score is $\leq \alpha$. As a result, we can re-write the inclusion of an incorrect response as at least one incorrect response having a score $\leq \alpha$,

$$\begin{aligned}&\exists\left(\mathbf{y}, \cdot\right) \in \mathbb{S}_\alpha\left(x, g_\pi\left(x\right)\right) \text{ s.t. } o\left(x, \mathbf{y}\right) = 0 \\ &\iff \min_{(\mathbf{y}, c) \in \mathbb{O}(x, g_\pi(x)) : c = 0} s\left(x, \mathbf{y}\right) \leq \alpha.\end{aligned}$$

Therefore, the worst-case size distortion simplifies to,

$$\begin{aligned}&\mathbf{E}\left[\max_{\alpha \in \mathbb{R}_{\geq 0}} \frac{\mathbb{1}\left\{\min_{(\mathbf{Y}, C) \in \mathbb{O}(X, g_\pi(X)) : C = 0} s\left(X, \mathbf{Y}\right) \leq \alpha\right\}}{\alpha}\right] \\ &= \mathbf{E}\left[\left(\min_{(\mathbf{Y}, C) \in \mathbb{O}(X, g_\pi(X)) : C = 0} s\left(X, \mathbf{Y}\right)\right)^{-1}\right],\end{aligned}$$

because α is set to the smallest value for which the indicator evaluates to 1, otherwise the whole term is 0.

To upper bound the above expectation by 1 is equivalent to requiring $\left(\min_{(\mathbf{y}, c) \in \mathbb{O}(x, g_\pi(x)) : c = 0} s\left(x, \mathbf{y}\right)\right)^{-1}$ to be an e-value (by definition), and hence the specific choice of our proposed e-scores in Eq. (4). Indeed, if we use our e-scores here, the above term simplifies to,

$$\begin{aligned}&\left(\min_{(\mathbf{y}^{n+1}, c^{n+1}) \in \mathbb{O}(x^{n+1}, g_\pi(x^{n+1})) : c^{n+1} = 0} s\left(x^{n+1}, \mathbf{y}^{n+1}\right)\right)^{-1} \\ &= \frac{(n+1) \cdot f^*\left(x^{n+1}, \mathbb{O}\left(x^{n+1}, g_\pi\left(x^{n+1}\right)\right)\right)}{\sum_{i=1}^{n+1} f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)},\end{aligned}$$

which is an e-value under exchangeable data for any non-negative function f (Gammerman et al., 1998).

Lastly, since our e-scores satisfy Eq. (3) under this worst-case setting, they will satisfy Eq. (3) under any post-hoc α. We summarize this theoretical result below, and provide the detailed derivation in Appendix A.

Theorem 1. *If the test and the calibration prompts are exchangeable, then, our proposed e-scores in Eqs. (4) and (6) upper bound the size distortion (marginal over the test and the calibration prompts) by 1, as in Eq. (3).*

7 CONCLUSIONS

In this paper, we studied the problem of achieving statistical guarantees for a post-hoc notion of error, namely size distortion, for generative model outputs. We proposed e-scores, based on e-values, as measures of incorrectness. We proved theoretically that our proposed e-scores achieve the desired post-hoc guarantees, which we corroborated with experimental results. In doing so, e-scores provide users the flexibility of choosing data-dependent tolerance levels α. We also showed that our guarantees extend to a large set of responses, opening up possibilities for more diverse applications.

Future Work Our experiments demonstrated that the choice of the oracle estimator impacts metrics such as the precision-recall curves. While we used pre-trained estimators, they could be trained for specific applications to strengthen the e-scores. Furthermore, while we considered size distortion as our post-hoc notion of error, other candidates exist, although they are not well understood. Koning (2024) discusses some alternatives, which could be investigated in the future.

References

E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, Étienne Goffinet, D. Hesslow, J. Launay, Q. Malartic, D. Mazzotta, B. Noune, B. Pannier, and G. Penedo. The Falcon series of open language models, 2023. URL https://arxiv.org/abs/2311.16867. 6

A. A. Balinsky and A. D. Balinsky. Enhancing conformal prediction using e-test statistics. In S. Vantini, M. Fontana, A. Solari, H. Boström, and L. Carlsson, editors, *Proceedings of the Thirteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 230 of *Proceedings of Machine Learning Research*, pages 65–72. PMLR, 09–11 Sep 2024. URL https://proceedings.mlr.press/v230/balinsky24a.html. 5, 14

S. Biderman, H. Schoelkopf, Q. G. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, and O. Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/biderman23a.html. 6

R. A. Bradley and M. E. Terry. Rank analysis of incomplete block designs: The method of paired comparisons. *Biometrika*, 39(3–4):324–345, 12 1952. ISSN 0006-3444. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324. 7

D. R. Carney. My position on "Power Poses", 2016. URL https://faculty.haas.berkeley.edu/dana_carney/pdf_my%20position%20on%20power%20poses.pdf. 1

J. J. Cherian, I. Gibbs, and E. J. Candès. Large language model validity via enhanced conformal prediction methods. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 114812–114842. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/d02ff1aeaa5c268dc34790dd1ad21526-Paper-Conference.pdf. 1, 2, 3, 4, 5, 6, 9, 16

W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing GPT-4 with 90%* ChatGPT quality, March 2023. URL https://lmsys.org/blog/2023-03-30-vicuna/. 6

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168. 6

G. Cui, L. Yuan, N. Ding, G. Yao, B. He, W. Zhu, Y. Ni, G. Xie, R. Xie, Y. Lin, Z. Liu, and M. Sun. UltraFeedback: Boosting language models with scaled AI feedback. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html. 6, 7

G. S. Dhillon, X. Shi, Y. W. Teh, and A. Smola. L3Ms — Lagrange large language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu, editors, *International Conference on Representation Learning*, volume 2025, pages 58300–58314, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/92d3d2a9801211ca3693ccb2faa1316f-Paper-Conference.pdf. 3, 7

N. Ding, Y. Chen, B. Xu, Y. Qin, S. Hu, Z. Liu, M. Sun, and B. Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3029–3051, Singapore, Dec. 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.183. URL https://aclanthology.org/2023.emnlp-main.183/. 6

A. Gammerman, V. Vovk, and V. Vapnik. Learning by transduction. In G. F. Cooper and S. Moral, editors, *Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 148–155. Morgan Kaufmann Publishers Inc., 24–26 July 1998. 5, 9, 14

B. Gao, F. Song, Z. Yang, Z. Cai, Y. Miao, Q. Dong, L. Li, C. Ma, L. Chen, R. Xu, Z. Tang, B. Wang, D. Zan, S. Quan, G. Zhang, L. Sha, Y. Zhang, X. Ren, T. Liu, and B. Chang. Omni-MATH: A universal olympiad level mathematic benchmark for large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=yaqPf0KAlN. 6

E. Gauthier, F. Bach, and M. I. Jordan. Backward conformal prediction. In *The Thirty-ninth Annual Conference on Neural Information Processing Sys-*

tems, 2025. URL https://openreview.net/forum?id=23ichdd74N. 5

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities, 2025. URL https://arxiv.org/abs/2507.06261. 1

P. Grünwald, R. de Heide, and W. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 03 2024. ISSN 1369-7412. doi: 10.1093/jrsssb/qkae011. URL https://doi.org/10.1093/jrsssb/qkae011. 4

P. D. Grünwald. Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121 (39):e2302098121, 2024. doi: 10.1073/pnas.2302098121. URL https://www.pnas.org/doi/abs/10.1073/pnas.2302098121. 4

C. Gupta, A. K. Kuchibhotla, and A. Ramdas. Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition*, 127:108496, 2022. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2021.108496. URL https://www.sciencedirect.com/science/article/pii/S0031320321006725. 4

C. He, R. Luo, Y. Bai, S. Hu, Z. Thai, J. Shen, J. Hu, X. Han, Y. Huang, Y. Zhang, J. Liu, L. Qi, Z. Liu, and M. Sun. OlympiadBench: A challenging benchmark for promoting AGI with olympiad-level bilingual multimodal scientific problems. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3828–3850, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.211. URL https://aclanthology.org/2024.acl-long.211/. 6

D. Hendrycks, C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. Measuring mathematical problem solving with the MATH dataset. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/be83ab3ecd0db773eb2dc1b0a17836a1-Paper-round2.pdf. 6

L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), Jan. 2025. ISSN 1046-8188. doi: 10.1145/3703155. URL https://doi.org/10.1145/3703155. 1

N. W. Koning. Post-hoc α hypothesis testing and the post-hoc p-value, 2024. URL https://arxiv.org/abs/2312.08040. 2, 4, 9

H. Lightman, V. Kosaraju, Y. Burda, H. Edwards, B. Baker, T. Lee, J. Leike, J. Schulman, I. Sutskever, and K. Cobbe. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi. 6

S. Lin, J. Hilton, and O. Evans. TruthfulQA: Measuring how models mimic human falsehoods. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.229. URL https://aclanthology.org/2022.acl-long.229/. 6, 7

LLaMA Team. The LLaMA 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783. 6

C. Mohri and T. Hashimoto. Language models with conformal factuality guarantees. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36029–36047. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/mohri24a.html. 1, 2, 3, 4, 5, 6, 9, 16

J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A*, 231(694–706):289–337, 1933. doi: 10.1098/rsta.1933.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1933.0009. 4

OpenAI. GPT-4 technical report, 2024a. URL https://arxiv.org/abs/2303.08774. 1, 6, 7

OpenAI. GPT-4o system card, 2024b. URL https://arxiv.org/abs/2410.21276. 1, 9

OpenAI. OpenAI o1 system card, 2024c. URL https://arxiv.org/abs/2412.16720. 1

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115. 6

A. Ramdas and R. Wang. Hypothesis testing with e-values. *Foundations and Trends® in Statistics*, 1 (1–2):1–390, 2025. ISSN 2978-4212. doi: 10.1561/3600000002. URL http://dx.doi.org/10.1561/3600000002. 4

M. Rubin-Toles, M. Gambhir, K. Ramji, A. Roth, and S. Goel. Conformal language model reasoning with coherent factuality. In *The Thirteenth International*

Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=AJpUZd8Clb. 1, 2, 3, 4, 5, 6, 9, 16

G. Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184 (2):407–431, 05 2021. ISSN 0964-1998. doi: 10.1111/rssa.12647. URL https://doi.org/10.1111/rssa.12647. 4

G. Shafer and V. Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(12): 371–421, 2008. URL http://jmlr.org/papers/v9/shafer08a.html. 1, 3, 4, 6, 15

G. Shafer and V. Vovk. *Game-Theoretic Foundations for Probability and Finance*. John Wiley & Sons, Ltd, 2019. ISBN 9781118548035. doi: 10.1002/9781118548035. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118548035. 4

R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Stanford Alpaca: An instruction-following LLaMA model, 2023. URL https://github.com/tatsu-lab/stanford_alpaca. 6

H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. LLaMA 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288. 6

L. Tunstall, N. Lambert, N. Rajani, E. Beeching, T. Le Scao, L. von Werra, S. Han, P. Schmid, and A. Rush. Creating a coding assistant with StarCoder. *Hugging Face Blog*, 2023. URL https://huggingface.co/blog/starchat-alpha. 6

V. Vovk. Conformal e-prediction. *Pattern Recognition*, 166:111674, 2025. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.111674. URL https://www.sciencedirect.com/science/article/pii/S0031320325003346. 5, 14

V. Vovk and R. Wang. E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754, 2021. doi: 10.1214/20-AOS2020. URL https://doi.org/10.1214/20-AOS2020. 4, 5

V. Vovk, A. Gammerman, and G. Shafer. *Algorithmic Learning in a Random World*. Springer International Publishing, Cham, 2022. ISBN 978-3-031-06649-8. doi: 10.1007/978-3-031-06649-8. URL https://doi.org/10.1007/978-3-031-06649-8. 1, 3, 4, 6, 15

A. Wald. Contributions to the theory of statistical estimation and testing hypotheses. *The Annals of Mathematical Statistics*, 10(4):299–326, 1939. doi: 10.1214/aoms/1177732144. URL https://doi.org/10.1214/aoms/1177732144. 4

P. Wang, L. Li, Z. Shao, R. Xu, D. Dai, Y. Li, D. Chen, Y. Wu, and Z. Sui. Math-Shepherd: Verify and reinforce LLMs step-by-step without human annotations. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9426–9439, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.510. URL https://aclanthology.org/2024.acl-long.510/. 6

R. Wang and A. Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):822–852, 01 2022. ISSN 1369-7412. doi: 10.1111/rssb.12489. URL https://doi.org/10.1111/rssb.12489. 4

L. Wasserman, A. Ramdas, and S. Balakrishnan. Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890, 2020. doi: 10.1073/pnas.1922664117. URL https://www.pnas.org/doi/abs/10.1073/pnas.1922664117. 4

J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf. 3

C. Xu, Q. Sun, K. Zheng, X. Geng, P. Zhao, J. Feng, C. Tao, Q. Lin, and D. Jiang. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CfXh93NDgH. 6, 7

A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang,

J. Ma, J. Yang, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, X. Liu, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Guo, and Z. Fan. Qwen2 technical report, 2024a. URL https://arxiv.org/abs/2407.10671. 6

A. Yang, B. Zhang, B. Hui, B. Gao, B. Yu, C. Li, D. Liu, J. Tu, J. Zhou, J. Lin, K. Lu, M. Xue, R. Lin, T. Liu, X. Ren, and Z. Zhang. Qwen2.5-math technical report: Toward mathematical expert model via self-improvement, 2024b. URL https://arxiv.org/abs/2409.12122. 6

C. Zheng, Z. Zhang, B. Zhang, R. Lin, K. Lu, B. Yu, D. Liu, J. Zhou, and J. Lin. ProcessBench: Identifying process errors in mathematical reasoning. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1009–1024, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/. 2, 6, 9

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. **Yes, in Sections 2, 4 and 6 and Appendix A.**
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. **Yes, in Sections 5 and 6 and Appendices A and B.**
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. **No, we are unable to release the code at this time. However, we included all details required to replicate our results.**

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. **Yes, in Sections 2, 4 and 6 and Appendix A.**
   (b) Complete proofs of all theoretical results. **Yes, in Appendix A.**
   (c) Clear explanations of any assumptions. **Yes, in Sections 2, 4 and 6 and Appendix A.**

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). **No, we are unable to release the code at this time. However, we included all details required to replicate our results.**
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). **Not applicable; we used pre-trained models.**
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). **Yes, in Section 5 and Appendix B.**
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). **Yes, in Section 5.**

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. **Yes, in Section 5 and Appendix B.**
   (b) The license information of the assets, if applicable. **Not Applicable.**
   (c) New assets either in the supplemental material or as a URL, if applicable. **Not Applicable.**
   (d) Information about consent from data providers/curators. **Not Applicable.**
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. **Not Applicable.**

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. **Not Applicable.**
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. **Not Applicable.**
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. **Not Applicable.**

A PROOFS

We include the detailed derivation of Theorem 1 below. For convenience, we first re-state the theorem.

Theorem 1. *If the test and the calibration prompts are exchangeable, then, our proposed e-scores in Eqs. (4) and (6) upper bound the size distortion (marginal over the test and the calibration prompts) by 1, as in Eq. (3).*

Proof. Note that a response is included in the filtered set $\mathbb{S}_\alpha\left(x, g_\pi\left(x\right)\right)$ if and only if its score is $\leq \alpha$. As a result, we can re-write the inclusion of an incorrect response as at least one incorrect response having a score $\leq \alpha$,

$$\exists\left(\mathbf{y}, \cdot\right) \in \mathbb{S}_\alpha\left(x, g_\pi\left(x\right)\right) \text{ s.t. } o\left(x, \mathbf{y}\right) = 0 \iff \min_{(\mathbf{y}, c) \in \mathbb{O}(x, g_\pi(x)): c=0} s\left(x, \mathbf{y}\right) \leq \alpha.$$

Therefore, the size distortion for any post-hoc α strategy is upper bound by the following (simplified) worst-case,

$$\begin{aligned}
&\mathbf{E}\left[\frac{\mathbb{1}\left\{\exists\left(\mathbf{Y}, \cdot\right) \in \mathbb{S}_{\alpha(X, \mathbb{S}(X, g_\pi(X)))}\left(X, g_\pi\left(X\right)\right) \text{ s.t. } o\left(X, \mathbf{Y}\right) = 0\right\}}{\alpha\left(X, \mathbb{S}\left(X, g_\pi\left(X\right)\right)\right)}\right] \\
&\quad \leq \mathbf{E}\left[\max_{\alpha \in \mathbb{R}_{\geq 0}} \frac{\mathbb{1}\left\{\exists\left(\mathbf{Y}, \cdot\right) \in \mathbb{S}_\alpha\left(X, g_\pi\left(X\right)\right) \text{ s.t. } o\left(X, \mathbf{Y}\right) = 0\right\}}{\alpha}\right] \\
&\quad = \mathbf{E}\left[\max_{\alpha \in \mathbb{R}_{\geq 0}} \frac{\mathbb{1}\left\{\min_{(\mathbf{Y}, C) \in \mathbb{O}(X, g_\pi(X)): C=0} s\left(X, \mathbf{Y}\right) \leq \alpha\right\}}{\alpha}\right] \\
&\overset{(i)}{=} \mathbf{E}\left[\left(\min_{(\mathbf{Y}, C) \in \mathbb{O}(X, g_\pi(X)): C=0} s\left(X, \mathbf{Y}\right)\right)^{-1}\right] = \mathbf{E}\left[\max_{(\mathbf{Y}, C) \in \mathbb{O}(X, g_\pi(X)): C=0} \left(s\left(X, \mathbf{Y}\right)\right)^{-1}\right],
\end{aligned} \tag{9}$$

where the equality (i) is achieved by setting α to the smallest value for which the indicator function evaluates to 1, otherwise the whole term is 0. We are interested in upper bounding the above expectation by 1 to satisfy Eq. (3).

Now, we use the definition of our proposed e-scores from Eq. (4). Note that our e-scores depend on the calibration data; we will make this dependence explicit in the following. Then, the worst-case size distortion simplifies to,

$$\begin{aligned}
&\mathbf{E}\left[\max_{(\mathbf{Y}^{n+1}, C^{n+1}) \in \mathbb{O}(X^{n+1}, g_\pi(X^{n+1})): C^{n+1}=0} \left(s_{\text{e-score}}\left(X^{n+1}, \mathbf{Y}^{n+1}; X^1, \ldots, X^n\right)\right)^{-1}\right] \\
&= \mathbf{E}\left[\max_{(\mathbf{Y}^{n+1}, C^{n+1}) \in \mathbb{O}(X^{n+1}, g_\pi(X^{n+1})): C^{n+1}=0} \frac{(n+1) \cdot f\left(X^{n+1}, \mathbf{Y}^{n+1}\right)}{f\left(X^{n+1}, \mathbf{Y}^{n+1}\right) + \sum_{i=1}^n f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}\right].
\end{aligned}$$

Note that for $a, b \in \mathbb{R}_{\geq 0}$, the ratio $a/(a+b)$ is a monotonically non-decreasing transformation of a because the derivative with respect to a (i.e., $b/(a+b)^2$) is non-negative. Consequently, the above maximum is achieved at $f^*\left(x, \mathbb{O}\left(x, g_\pi\left(x\right)\right)\right) = \max_{(\mathbf{y}, c) \in \mathbb{O}(x, g_\pi(x)): c=0} f\left(x, \mathbf{y}\right)$. Therefore, the worst-case size distortion simplifies to,

$$\begin{aligned}
&\mathbf{E}\left[\max_{(\mathbf{Y}^{n+1}, C^{n+1}) \in \mathbb{O}(X^{n+1}, g_\pi(X^{n+1})): C^{n+1}=0} \frac{(n+1) \cdot f\left(X^{n+1}, \mathbf{Y}^{n+1}\right)}{f\left(X^{n+1}, \mathbf{Y}^{n+1}\right) + \sum_{i=1}^n f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}\right] \\
&\quad = \mathbf{E}\left[\frac{(n+1) \cdot f^*\left(X^{n+1}, \mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right)\right)}{f^*\left(X^{n+1}, \mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right)\right) + \sum_{i=1}^n f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}\right] \\
&\quad = \mathbf{E}\left[\frac{(n+1) \cdot f^*\left(X^{n+1}, \mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right)\right)}{\sum_{i=1}^{n+1} f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}\right].
\end{aligned}$$

Lastly, we assume that the test and the calibration prompts are exchangeable, i.e., for any permutation σ over the indices $\{1, \ldots, n+1\}$, the ordering of the permuted prompts is equal in distribution to the un-permuted prompts,

$$\left(X^{\sigma(1)}, \ldots, X^{\sigma(n+1)}\right) \overset{d}{=} \left(X^1, \ldots, X^{n+1}\right).$$

We can follow arguments similar to those made by Gammerman et al. (1998); Balinsky and Balinsky (2024); Vovk (2025) to show that the above expectation is ≤ 1 under exchangeability. Specifically, we define random variables,

$$R^i = \frac{(n+1) \cdot f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}{\sum_{j=1}^{n+1} f^*\left(X^j, \mathbb{O}\left(X^j, g_\pi\left(X^j\right)\right)\right)},$$

for $i = 1, \ldots, n+1$. Under exchangeability of $X^1, \ldots X^{n+1}$, the distributions of $R^1, \ldots R^{n+1}$ are identical. Then,

$$\mathbf{E}\left[\frac{(n+1)\cdot f^*\left(X^{n+1}, \mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right)\right)}{\sum_{i=1}^{n+1} f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}\right] = \mathbf{E}\left[R^{n+1}\right] = \frac{\sum_{i=1}^{n+1}\mathbf{E}\left[R^i\right]}{n+1} = \frac{\mathbf{E}\left[\sum_{i=1}^{n+1} R^i\right]}{n+1}$$
$$= \frac{\mathbf{E}\left[\sum_{i=1}^{n+1} \frac{(n+1)\cdot f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}{\sum_{j=1}^{n+1} f^*\left(X^j, \mathbb{O}\left(X^j, g_\pi\left(X^j\right)\right)\right)}\right]}{n+1} = \mathbf{E}\left[\frac{\sum_{i=1}^{n+1} f^*\left(X^i, \mathbb{O}\left(X^i, g_\pi\left(X^i\right)\right)\right)}{\sum_{j=1}^{n+1} f^*\left(X^j, \mathbb{O}\left(X^j, g_\pi\left(X^j\right)\right)\right)}\right] \overset{(ii)}{\leq} \mathbf{E}\left[1\right] = 1,$$

where the inequality (ii) accounts for the sum being 0, making $0/0 = 0$ (by convention). Hence, our e-scores in Eq. (4) upper bound the size distortion (marginal over the test and the calibration prompts) by 1, as in Eq. (3).

Furthermore, we can use the definition of our proposed combined e-scores from Eq. (6). Note that instead of combining three e-scores, we can combine any $k \geq 1$. The worst-case size distortion from Eq. (9) simplifies to,

$$\mathbf{E}\left[\max_{\left(\mathbf{Y}^{n+1}, C^{n+1}\right)\in\mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right): C^{n+1}=0} \left(s_{\text{e-score (combined)}}\left(X^{n+1}, \mathbf{Y}^{n+1}; X^1, \ldots, X^n\right)\right)^{-1}\right]$$
$$= \mathbf{E}\left[\max_{\left(\mathbf{Y}^{n+1}, C^{n+1}\right)\in\mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right): C^{n+1}=0} \frac{\sum_{i=1}^{k}\left(s_{\text{e-score }(i)}\left(X^{n+1}, \mathbf{Y}^{n+1}; X^1, \ldots, X^n\right)\right)^{-1}}{k}\right]$$
$$\leq \frac{\sum_{i=1}^{k}\mathbf{E}\left[\max_{\left(\mathbf{Y}^{n+1}, C^{n+1}\right)\in\mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right): C^{n+1}=0}\left(s_{\text{e-score }(i)}\left(X^{n+1}, \mathbf{Y}^{n+1}; X^1, \ldots, X^n\right)\right)^{-1}\right]}{k}.$$

We have shown that the worst-case size distortion for individual e-scores (in the numerator) is ≤ 1. Then,

$$\frac{\sum_{i=1}^{k}\mathbf{E}\left[\max_{\left(\mathbf{Y}^{n+1}, C^{n+1}\right)\in\mathbb{O}\left(X^{n+1}, g_\pi\left(X^{n+1}\right)\right): C^{n+1}=0}\left(s_{\text{e-score }(i)}\left(X^{n+1}, \mathbf{Y}^{n+1}; X^1, \ldots, X^n\right)\right)^{-1}\right]}{k} \leq \frac{\sum_{i=1}^{k} 1}{k} = 1.$$

Hence, our combined e-scores in Eq. (6) upper bound the size distortion (marginally) by 1, as in Eq. (3).

□

B EXPERIMENTAL RESULTS FOR WORST-CASE ANALYSIS

We include additional experimental results here, expanding on Section 5. In particular, we perform a worst-case analysis for the different use-cases, where the post-hoc α strategy maximizes size distortion, which we quantified in Eq. (9). We begin by stating the common baselines considered, in addition to the p-scores from Section 5.

Baselines In addition to the p-scores defined in Eq. (7), we also compare with their randomized version,

$$s_{\text{p-score (randomized)}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) =$$
$$\frac{u\cdot\left(1+\sum_{i=1}^{n}\mathbb{1}\left\{\begin{array}{c} f\left(x^{n+1}, \mathbf{y}^{n+1}\right) \\ = f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)\end{array}\right\}\right) + \sum_{i=1}^{n}\mathbb{1}\left\{\begin{array}{c} f\left(x^{n+1}, \mathbf{y}^{n+1}\right) \\ < f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)\end{array}\right\}}{n+1},$$

where $u \sim \mathcal{U}(0,1)$ is a uniform random sample in the range $[0,1]$. We can recover the p-scores defined in Eq. (7) as a special case of this definition by deterministically setting $u = 1$. While the non-randomized p-scores correspond to p-values, these randomized p-scores correspond to exact p-values (Shafer and Vovk, 2008; Vovk et al., 2022).[2]

We also compare with the transformed oracle estimators in Eq. (5) directly, without any conversion to e- or p-scores using the calibration data. Since we want our scores to be low for correct and high for incorrect responses (as measures of incorrectness), we define the *naive* scores to be the reciprocal of the transformed oracle estimators,

$$s_{\text{naive}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) = \left(f_{\hat{o}}\left(x^{n+1}, \mathbf{y}^{n+1}\right)\right)^{-1} = \begin{cases}\left(\hat{o}\left(x^{n+1}, \mathbf{y}^{n+1}\right)\right)^{-1} \in [1,\infty] & \text{(for naive 1)} \\ 1-\hat{o}\left(x^{n+1}, \mathbf{y}^{n+1}\right) \in [0,1] & \text{(for naive 2)} \\ \left(\hat{o}\left(x^{n+1}, \mathbf{y}^{n+1}\right)\right)^{-1}\cdot\left(1-\hat{o}\left(x^{n+1}, \mathbf{y}^{n+1}\right)\right) \in [0,\infty] & \text{(for naive 3)}\end{cases}.$$

[2]Consider a non-negative random variable $R \in \mathbb{R}_{\geq 0}$. It is an *exact p-variable* if $\mathbf{P}\{R \leq \alpha\} = \alpha$, for all $\alpha \in [0,1]$.

Table 1: **Scores for the worst-case size distortion analysis.** We use the mathematical factuality (cf. Section 5.1) and property constraints satisfaction (cf. Section 5.2) settings to compare our proposed e-scores against p-scores and naive scores. We consider the worst-case that maximizes size distortion (cf. Eq. (9)). We report the mean and the inter-quartile range (which depicts the 25-th and 75-th quantiles) of the size distortion.

Score	**Worst-case size distortion**			
	Mathematical factuality		**Property constraints satisfaction**	
	QwenPRM	**MathShepherd-PRM**	**Helpfulness and instruction-following**	**Honesty and truthfulness**
naive (1)	$0.24_{(0.00\text{-}0.46)}$	$0.20_{(0.00\text{-}0.37)}$	$0.07_{(0.00\text{-}0.01)}$	$0.09_{(0.00\text{-}0.03)}$
naive (2)	$1.89_{(0.00\text{-}1.86)}$	$1.30_{(0.00\text{-}1.59)}$	$2.25_{(0.00\text{-}1.01)}$	$5.42_{(0.00\text{-}1.03)}$
naive (3)	$1.39_{(0.00\text{-}0.86)}$	$0.80_{(0.00\text{-}0.59)}$	$1.80_{(0.00\text{-}0.01)}$	$4.82_{(0.00\text{-}0.03)}$
p-score	$7.21_{(0.00\text{-}4.00)}$	$7.46_{(0.00\text{-}4.03)}$	$9.60_{(0.00\text{-}4.00)}$	$7.55_{(0.00\text{-}3.98)}$
p-score (randomized)	$15.80_{(0.00\text{-}4.00)}$	$13.96_{(0.00\text{-}4.03)}$	$15.91_{(0.00\text{-}4.00)}$	$14.92_{(0.00\text{-}3.99)}$
e-score (1)	$1.00_{(0.00\text{-}1.95)}$	$1.01_{(0.00\text{-}1.90)}$	$1.00_{(0.00\text{-}0.19)}$	$1.00_{(0.00\text{-}0.34)}$
e-score (2)	$0.79_{(0.00\text{-}0.79)}$	$0.73_{(0.00\text{-}0.89)}$	$0.80_{(0.00\text{-}0.38)}$	$0.97_{(0.00\text{-}0.23)}$
e-score (3)	$1.01_{(0.00\text{-}0.63)}$	$1.01_{(0.00\text{-}0.75)}$	$1.00_{(0.00\text{-}0.01)}$	$1.05_{(0.00\text{-}0.01)}$
e-score (combined)	$0.94_{(0.00\text{-}1.12)}$	$0.91_{(0.00\text{-}1.18)}$	$0.94_{(0.00\text{-}0.19)}$	$1.01_{(0.00\text{-}0.19)}$

These naive scores generally do not come with any statistical guarantees by themselves. However, because the reciprocal of naive (1) is always ≤ 1, it happens to correspond to an *uninformative* e-value that is always ≤ 1 (the expectation is ≤ 1 by design). Therefore, even though naive (1) achieves the size distortion bound in Eq. (3), it regularly excludes responses (correct and incorrect) and is extremely conservative compared to our e-scores.

B.1 Worst-Case Size Distortion Analysis

We analyze the worst-case size distortion in Eq. (9). Table 1 illustrates the results for both our experimental use-cases: mathematical factuality (cf. Section 5.1) and property constraints satisfaction (cf. Section 5.2). Our proposed e-scores (and naive (1)) reliably upper bound the worst-case size distortion to 1 and satisfy Eq. (3), corroborating our theory in Theorem 1. Conversely, p-scores and other naive scores are unable to achieve this.

C IMPLICIT P-SCORES IN RELATED WORK

Here we highlight the implicit role of p-scores (cf. Eq. (7)), and hence p-values, in the works most closely related to ours (Mohri and Hashimoto, 2024; Cherian et al., 2024; Rubin-Toles et al., 2025), making it explicit. To begin with, these works compute the calibration values $f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)$ for $i = 1, \ldots, n$. Given a fixed user-defined $\alpha \in [1/(n+1), 1]$, they compute a threshold τ_α set to the $\lceil (1-\alpha)\cdot(n+1)\rceil$-th smallest of the calibration values above. Then, a test response $\mathbf{y}^{n+1}$ is included in the returned set if $f\left(x^{n+1}, \mathbf{y}^{n+1}\right)$ is larger than this threshold,

$$
\begin{aligned}
& f\left(x^{n+1}, \mathbf{y}^{n+1}\right) > \tau_\alpha \\
\iff & \sum_{i=1}^{n} \mathbb{1}\left\{f\left(x^{n+1}, \mathbf{y}^{n+1}\right) > f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)\right\} \geq (1-\alpha)\cdot(n+1) \\
\iff & \sum_{i=1}^{n} \mathbb{1}\left\{f\left(x^{n+1}, \mathbf{y}^{n+1}\right) \leq f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)\right\} \leq \alpha\cdot(n+1) - 1 \\
\iff & \frac{1 + \sum_{i=1}^{n} \mathbb{1}\left\{f\left(x^{n+1}, \mathbf{y}^{n+1}\right) \leq f^*\left(x^i, \mathbb{O}\left(x^i, g_\pi\left(x^i\right)\right)\right)\right\}}{n+1} \leq \alpha \\
\iff & s_{\text{p-score}}\left(x^{n+1}, \mathbf{y}^{n+1}\right) \leq \alpha.
\end{aligned}
$$

In our setup, this is equivalent to returning the filtered set $\mathbb{S}_\alpha\left(x^{n+1}, g_\pi\left(x^{n+1}\right)\right)$ using p-scores. We again highlight that these approaches satisfy Eq. (2), but not its post-hoc generalization in Eq. (3); for that, we propose e-scores.