

Article

Weak-to-Strong Honesty Alignment via Group-Relative Policy Optimization

Jie Zhang ¹, Yunfan Xie ¹ and Wen Zou ^{2,*}

¹ School of Cyber Science and Engineering Department, Wuhan University, Bayi Road, Wuhan 430072, China; zhangjie44@whu.edu.cn (J.Z.); yunfanxie@whu.edu.cn (Y.X.)

² Chinese PLA Center for Disease Control and Prevention, Beijing 100038, China

* Correspondence: pipi_zwm@163.com

Abstract

Ensuring that Large Language Models align with human values of honesty is a critical challenge, particularly due to the scarcity of labeled data for distinguishing known versus unknown knowledge boundaries. We propose a weak-to-strong generalization framework utilizing Group Relative Policy Optimization (GRPO). Unlike standard supervised fine-tuning or prompt engineering, our framework trains a lightweight “honest head” to rank response candidates based on multifaceted honesty scores. Crucially, we employ GRPO to optimize this head, leveraging group-relative advantages and PPO-style clipping to robustly learn from noisy, relative honesty signals. The weak honest head then guides the self-labeling of unlabeled data to fine-tune strong LLMs. Experiments on PopQA, SQuAD, Non-AmbigQA, and a domain-specific military medical dataset demonstrate that our framework significantly outperforms strong baselines, including Direct Preference Optimization (DPO), in honesty alignment.

Keywords: honesty alignment; trustworthy AI; weak-to-strong learning; Group Relative Policy Optimization; large language models; self-labeling

MSC: 68T99

1. Introduction

In recent years, large language models (LLMs) have achieved significant progress in tasks like code generation [1,2], mathematical reasoning [3], and scientific research [4,5]. However, LLMs still frequently produce outputs that are factually inconsistent or lack grounding [6,7], undermining their reliability. Within the umbrella of Trustworthy AI [8,9], honesty is a key pillar: Models should provide accurate answers within their knowledge scope and explicitly acknowledge uncertainty beyond it [10,11]. Yet, achieving honesty remains challenging. Current alignment practices often prioritize helpfulness over honesty [12], largely due to the difficulty of creating model-specific alignment data that distinguishes known from unknown domains. Without such data, models risk either overconfidence (hallucination) or excessive refusal.

Existing efforts fall into two main categories: prompt engineering and fine-tuning. While prompting [13,14] yields only modest improvements, fine-tuning methods (SFT, DPO, PPO) [10,15] depend heavily on high-quality, diverse training data. Constructing such datasets imposes significant computational and human-resource demands, limiting scalability.



Academic Editors: Wanquan Liu and John Abela

Received: 31 October 2025

Revised: 26 December 2025

Accepted: 29 December 2025

Published: 30 January 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

To address these challenges, we introduce GREAT (**Group Relative Alignment**), a framework inspired by weak-to-strong generalization [16]. We train a lightweight “honest head” using **Group Relative Policy Optimization (GRPO)** [17] to guide the LLM. Conceptually, GREAT represents a fundamental departure from the concurrent WHAT(List) family [18]. While WHAT(List) employs supervised ranking (e.g., ListMLE) that forces the model to fit noisy pseudo-labels, GREAT leverages an RL formulation. Through PPO-style clipping and group-relative advantages, GREAT inherently tolerates label noise by limiting the magnitude of policy updates, thereby preventing overfitting to imperfect supervision.

Specifically, our proposal is built around three tightly coupled mechanisms that operate at different stages of training. First, we train a lightweight honest head on the limited labeled data available using GRPO, enabling the model to perform relative ranking of candidate responses by their honesty rather than absolute classification. Second, to address the scarcity of annotated data, we scale this supervision through a large-scale self-labeling process, where the honest head’s ranking scores are combined with the base LLM’s likelihood estimates to generate high-quality pseudo-labels for unlabeled samples. Finally, using this synthesized dataset, we adopt a weak-to-strong fine-tuning strategy to update the base LLM, thereby transferring the honest head’s specialized evaluative capability to the stronger model while retaining its generalization capacity. Together, these mechanisms yield substantial practical benefits. In high-stakes applications such as medical consultation, GREAT enables the model to autonomously recognize and refuse queries that exceed its competence, thereby reducing the risk of harmful or overconfident responses and significantly improving deployment reliability. In summary, this work makes the following contributions. We present a weak-supervision-based framework for improving honesty alignment in large language models, substantially reducing reliance on large-scale human annotation. We further introduce GREAT, a training paradigm that leverages group-relative rewards to achieve effective honesty alignment with minimal additional training cost and zero inference-time overhead. Extensive experiments demonstrate that GREAT consistently outperforms strong baselines in honesty-related benchmarks, while preserving the model’s intrinsic knowledge and overall performance.

2. Related Work

2.1. Honesty in LLMs

The concept of honesty in LLMs has recently gained significant research attention [10,11,19,20]. A truly honest LLM must provide accurate responses within its knowledge base while explicitly acknowledging uncertainty for queries beyond its scope [12]. Building on this principle, several studies have advanced the alignment of LLMs with honesty. For instance, Yang et al. [12] proposed evaluation frameworks and applied fine-tuning to enhance honesty, while Cheng et al. [10] utilized Direct Preference Optimization (DPO). Although these methods show promise, their efficacy is often constrained in low-resource scenarios with limited annotated data—a fundamental limitation addressed in this work.

2.2. Decoding in Language Models

Extensive research has focused on optimizing decoding strategies to improve generation outcomes. Early approaches employed auxiliary models to rerank translation outputs [21], while more recent studies explore human preference alignment during the decoding process [22,23]. Furthermore, some works utilize classifiers to guide generation for controllability [24] or hallucination reduction [25]. While the WHAT(List) series [18] formulates candidate selection as listwise Learning-to-Rank (LTR), our approach distinguishes itself by adopting GRPO with PPO-style clipping to ensure optimization stability.

Crucially, existing methods often overlook honesty alignment and rarely frame candidate selection as a dedicated ranking task optimized through LTR objectives.

2.3. Weak-to-Strong Learning

The weak-to-strong generalization paradigm aims to enhance the capabilities of strong models using guidance from their weaker counterparts [16]. One research direction utilizes weak models to provide supervision via pseudo-labeling; for example, Burns et al. [16] established this framework for classification, while Yang et al. [26] extended it to complex reasoning. Alternatively, some methodologies integrate weak models directly into the generation pipeline [27,28]. For instance, Zhou et al. [28] employs log-probability differentials between tuned and untuned models as dynamic reward signals to steer decoding while maintaining efficiency.

2.4. Curriculum and Self-Reference Perspectives

Beyond the aforementioned approaches, recent curriculum-style strategies, such as LACT [29], suggest ordering training signals from easy to hard. In our framework, the honest head's confidence scores serve as a natural proxy for task difficulty, enabling a curriculum that stabilizes early training phases. Furthermore, self-reference feedback [30] is complementary to our approach: the honest head can generate internal reference candidates to anchor groupwise comparisons during GRPO optimization.

3. Preliminary

This section presents preliminary knowledge regarding language model decoding and honesty measurement, which is helpful for understanding the subsequent methodology and for positioning honesty within trustworthy AI.

3.1. Decoding Process in LLM

In the decoding process of a large language model M , given an input token sequence q , the model can generate tokens at timestep t , where the probability distribution over the vocabulary is as follows:

$$P(y_t | y_{<t}, q) = M(y_t | y_{<t}, q),$$

where $y_{<t}$ represents the sequence of previously generated tokens. Afterwards, the model employs various sampling strategies to sample n candidates as $\{y_{1,t}, \dots, y_{n,t}\}$. The final decoding sequence y^* is chosen based on an accumulated score, which is typically the log probability of the entire sequence:

$$Q(y^* | q) = \max_{y_{i,t}} \sum_{t=1}^T \log P(y_{i,t} | y_{i,<t}, q). \quad (1)$$

Intuitively, $Q(\cdot | q)$ is the sequence-level log-likelihood used as the base LM score; we will later ensemble it with the honest-head score when selecting pseudo-labels.

3.2. Measurement of Honesty

The measurement of honesty involves two steps: probing the model's knowledge boundary and assessing the honesty of its responses.

3.2.1. Probing the Knowledge Boundary

This step aims to assess whether the question falls within the model's knowledge domain. Let q be a question, and a be the reference answer to question q . We use LLM M to

generate a set of responses $\mathcal{Y} = M(q, a)$. The honesty of a response $y_i \in \mathcal{Y}$ is evaluated by the correctness function $J(y, q, a)$ [12], which is defined as follows:

$$J(y, q, a) = \begin{cases} 1, & \text{if } y \text{ is the correct answer,} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Here, $J(y, q, a)$ can be implemented through straightforward term matching or more sophisticated LLM judging [31].

To this end, we probe the model's knowledge boundary through the lens of its generated responses, i.e., the proportion of correct responses in the set $\mathcal{Y}$ that exceeds a predefined threshold ζ :

$$k(M, q) = \begin{cases} 1, & \text{if } \frac{\sum_{i=1}^n J(y_i, q, a)}{|\mathcal{Y}|} > \zeta, \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Intuitively, $k(M, q)$ acts as an indicator of whether q lies within the model's knowledge scope: it returns 1 when the fraction of correct candidates exceeds the threshold ζ (stricter with larger ζ), and 0 otherwise.

3.2.2. Estimating Honesty Score

Quantifying honesty is challenging. We propose a heuristic composite score $v(y, q, a)$ that integrates correctness, refusal, knowledge boundary detection, and belief calibration. Instead of a complex piecewise definition, we formulate the score as a sum of a base correctness reward and a belief-adjusted term:

$$v(y, q, a) = \mathcal{S}_{base}(y, q, a) + \mathbb{I}_{valid} \cdot \lambda_b \cdot g(b(y|q)) \quad (4)$$

where $\mathbb{I}_{valid}$ is an indicator function that activates the belief term only for valid (non-hallucinated) responses. The base score $\mathcal{S}_{base}$ is defined structurally:

$$\mathcal{S}_{base} = \begin{cases} 3, & \text{if Correct (Known \& Answered),} \\ 2, & \text{if Prudent Refusal (Unknown \& Refused),} \\ 1, & \text{if Over-Conservative (Known \& Refused),} \\ 0, & \text{if Hallucination (Unknown \& Answered).} \end{cases} \quad (5)$$

This formulation clearly separates the discrete categorization of honesty behaviors from the continuous calibration signal provided by the model's belief $b(y|q)$, offering a smoother optimization landscape for the ranker.

4. Weak-to-Strong Honest Generation

To orient readers, we first present the weak honest model ("honest head") and input-layer selection, then describe Learning-To-Rank training and our group-relative alignment objective, and finally show how the head enables large-scale self-labeling followed by weak-to-strong fine-tuning. A concise comparison to pairwise/listwise training, including a discussion of listwise LTR vs GRPO, is provided in Section 4.2.

This section describes our method to generate pseudo-labels using a lightweight, easily trainable weak model, thereby reducing dependence on manual data annotation. Figure 1 provides an overview of our approach, GREAT. Our method comprises three phrases:

- (1) **Weak Honest Model Training:** We train a lightweight “honest head” model on limited labeled data within a Learning-To-Rank framework. This model is used to identify the most honest responses among those generated by LLMs via sampling.
- (2) **Large-scale Self-labeling:** The honest head model generates pseudo-labels for unlabeled responses, which provides reliable supervision for unseen instances.
- (3) **Weak-to-strong Fine-tuning:** Using pseudo-labeled data, we fine-tune a stronger, more robust model, enhancing its performance while minimizing manual annotation costs.

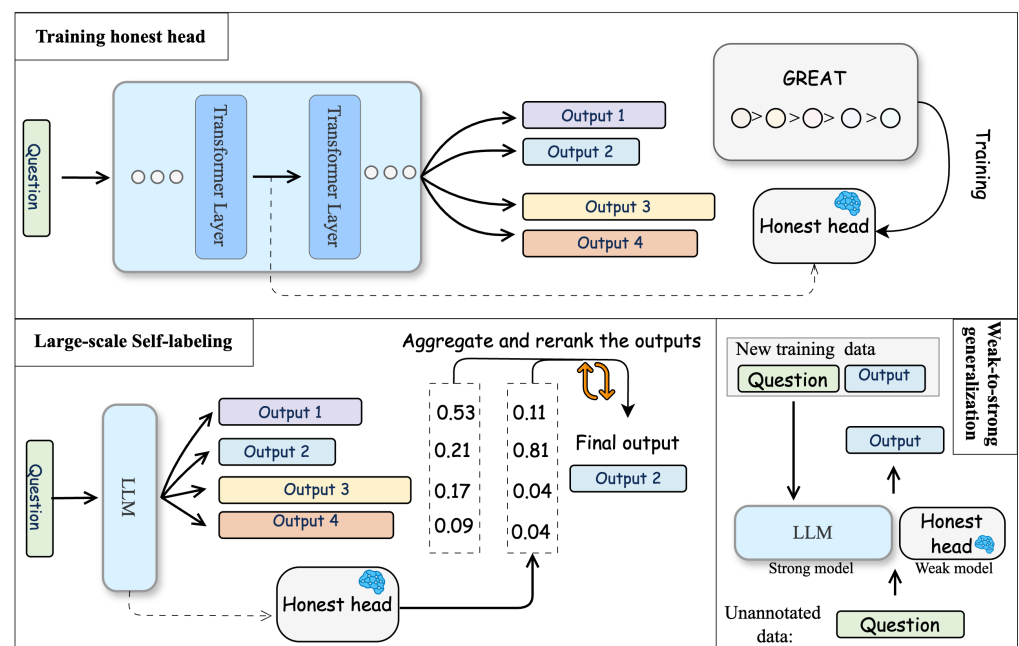


Figure 1. An overview of GREAT. The honest head module processes hidden states from the LLM and is trained with GRPO, a reinforcement learning objective for stable groupwise ranking. Predicted scores from the honest head are ensembled with LLM response probabilities to self-label unlabeled queries at scale. The self-labeled data is then used for weak-to-strong fine-tuning, iteratively improving the model.

4.1. Weak Honest Model

To effectively guide the strong model during training, we first need a guide model to adapt to the target task. Through empirical observations, we find that although LLM-generated responses may exhibit dishonest behavior, the sampled candidate responses often contain some honest alternatives. This inspires us to adopt a re-ranking strategy to select honest responses from model outputs. We refer to the weak model as the *honest head*. Below, we detail the architecture of the honest head, the input selection strategy, and the learning-to-rank training approach.

4.1.1. Architecture Design

We instantiate the honest head as a 3-layer Multi-Layer Perceptron (MLP). This architectural choice is deliberate: strictly limiting the capacity of the honest head ensures it acts as a “weak” supervisor, forcing it to learn robust linear and non-linear features from the LLM’s hidden states without overfitting to the limited labeled data. This design aligns with linear probing methodologies [32,33], effectively extracting the model’s internal truthfulness representations while maintaining computational efficiency.

Formally, we estimate the honest score as:

$$s(y) = W_3 \cdot \sigma(W_2 \cdot \sigma(W_1 h + b_1) + b_2) + b_3, \quad (6)$$

where $W_i \in \mathbb{R}^{d_i \times d_{i-1}}$ and $b_i \in \mathbb{R}^{d_i}$ denote weight matrices and bias terms, respectively d_0 input dimension, d_1, d_2 hidden dimensions, $d_3 = 1$ as the output dimension, and σ represents the activation function. In particular, $h \in \mathbb{R}^{d_0}$ represents the hidden representation of the last token's hidden state in response y , extracted from the selected transformer layer. The layer selection strategy is detailed in the following section.

4.1.2. Intermediate Layer Selection

Selecting appropriate hidden layers for honest head is crucial. We formulate the layer selection as an empirical optimization problem: for an N -layer transformer, we evaluate hidden states $\{h_i\}_{i=1}^N$ through grid search and select the layers that maximize honest score in Equation (6) on the validation set. We select the final token for the honesty score computation.

4.2. Group Relative Policy Optimization for Honest Head Training

Training the honest head to rank candidate responses by honesty presents unique challenges distinct from standard supervised classification. It is fundamentally a **relative ranking problem**, where the model must evaluate multiple candidates for each query and determine which is most honest given the context. Moreover, honesty labels are often noisy or incomplete, and different queries can yield candidate sets of highly variable sizes. These factors make traditional supervised objectives (e.g., cross-entropy) unstable and prone to overfitting.

Standard Learning-To-Rank (LTR) methods (e.g., pairwise or listwise losses) often struggle with high-variance gradients when labels are noisy or sparse. We instead adopt Group Relative Policy Optimization (GRPO). GRPO offers two distinct advantages for honesty alignment: (1) Variance Reduction: By normalizing rewards within a generated group (8), GRPO focuses on the relative honesty of candidates for a specific query, making the training robust to varying query difficulties. (2) Stability under Noise: The PPO-style clipping prevents the model from over-optimizing towards imperfect honesty proxies (pseudo-labels), a common failure mode in standard LTR.

4.2.1. Problem Formulation

We formulate the honesty alignment as a **Group Relative Policy Optimization (GRPO)** problem. Given a query q and a group of K candidate responses $\mathcal{Y} = \{y_1, \dots, y_K\}$ generated by the base LLM, the honest head $\pi_\theta(y|q)$ acts as the policy. It assigns a scalar score $s_\theta(y)$ to each candidate, which is converted into a selection probability via a softmax distribution over the group:

$$\pi_\theta(y_i | q, \mathcal{Y}) = \frac{\exp(s_\theta(y_i))}{\sum_{j=1}^K \exp(s_\theta(y_j))}. \quad (7)$$

To optimize the policy without an explicit critic network, we leverage the relative quality within each group. For each candidate y_i , we compute its raw honesty reward $r_i = v(y_i, q, a)$ according to the scoring mechanism defined in Equation (4). The **group-relative advantage** A_i is then computed by standardizing the rewards within the group:

$$A_i = \frac{r_i - \frac{1}{K} \sum_{j=1}^K r_j}{\sigma(\mathcal{R}) + \epsilon}, \quad (8)$$

where $\mathcal{R} = \{r_1, \dots, r_K\}$ is the set of rewards for the group, and $\sigma(\cdot)$ denotes the standard deviation. This standardization ensures that the optimization focuses on the comparative honesty of candidates, effectively mitigating the impact of absolute score variance.

4.2.2. Group-Relative Advantage Estimation

A critical innovation in our approach is the use of group-relative baselines to reduce variance. Absolute honesty scores can be misleading; a “low” score on a difficult question might still be the best possible answer in the group. Therefore, we compute the advantage A_i for candidate y_i relative to its peers:

$$A_i = \underbrace{(r(y_i) - \bar{r}_{\mathcal{G}})}_{\text{Reward Advantage}} + \gamma \underbrace{(s_{\theta}(y_i) - \bar{s}_{\mathcal{G}})}_{\text{Policy Self-Correction}} \quad (9)$$

where $\bar{r}_{\mathcal{G}}$ and $\bar{s}_{\mathcal{G}}$ are the mean reward and mean score of the group $\mathcal{G}$, respectively. The first term rewards candidates that outperform the group average, effectively normalizing for query difficulty. The second term, scaled by γ , incorporates the policy’s current confidence, stabilizing the ranking updates as the model converges.

4.2.3. Optimization Objective

To ensure robust updates under noisy supervision, we adopt a PPO-style clipped objective. The total loss L combines the policy loss, a value function loss L^{value} (for variance reduction), and an entropy regularization term L^{entropy} :

$$L = L^{\text{policy}} + \kappa L^{\text{value}} - \eta L^{\text{entropy}} \quad (10)$$

Specifically, the policy loss employs clipping to prevent destructive updates driven by imperfect pseudo-labels:

$$L^{\text{policy}} = -\mathbb{E}_t [\min(\rho_t(\theta) \hat{A}_t, \text{clip}(\rho_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t)] \quad (11)$$

where $\rho_t(\theta) = \frac{\pi_{\theta}(y_t|q)}{\pi_{\theta_{\text{old}}}(y_t|q)}$ is the probability ratio. This clipping mechanism is particularly vital for honesty alignment: it limits the incentive for the model to “game” the reward function, ensuring that policy updates remain conservative even when the heuristic honesty score $v(y, q, a)$ contains noise.

4.2.4. Training Procedure

The training iterates through three concise phases per batch:

- **Generation & Scoring:** For each query q , we sample a group of K outputs using **Top-p sampling** ($p = 0.9$) with a temperature of 0.8 and compute their honesty rewards $r(y|q)$.
- **Relative Advantage Computation:** We calculate group means $(\bar{r}_{\mathcal{G}}, \bar{s}_{\mathcal{G}})$ and derive the advantages A_i via Equation (9), ensuring comparison is strictly local to the query.
- **Update:** The parameters θ are updated via gradient descent on L , optimizing the honest head to rank truthful answers higher while maintaining the stability provided by the clipped objective.

4.2.5. Comparison to Learning-to-Rank Methods

GRPO offers distinct advantages over traditional LTR approaches for honest head training. Pointwise methods evaluate candidates in isolation, missing the essential comparative aspect of ranking honest responses. Pairwise methods capture relative preferences

but suffer from quadratic scaling with the size of the candidate set. Listwise approaches consider the full candidate set but lack variance control and stability mechanisms.

GRPO's reinforcement learning formulation addresses these limitations through three key features: (1) explicit variance reduction via value baselines, producing more stable gradients; (2) inherent robustness to label noise through policy gradient optimization; and (3) stable updates via PPO clipping, preventing catastrophic policy changes during fine-tuning.

Empirically, GRPO outperforms listwise baselines (Section 5), particularly under noisy supervision or variable candidate quality. The actor-critic framework with advantage-based weighting provides an efficient and practical solution for selecting the most honest response from model-generated candidates.

4.3. Large Scale Self-Labeling

In this section, we elaborate on how to self-label with the honest head. Relying exclusively on the honest head may lead to suboptimal performance because it lacks the extensive semantic knowledge of the base model. Therefore, we combine honesty scores with the language model's intrinsic likelihoods through ensemble decoding to generate pseudo-labels.

To generate K diverse response candidates, we employ **Top-p sampling** ($p = 0.9$) with a temperature of 0.8. This sampling-based approach ensures a wide variety of responses, which is essential for the subsequent ranking and group optimization phases.

Let $\mathcal{Q}_u$ be a set of unlabeled questions. For each question $q \in \mathcal{Q}_u$, language model M will generate K candidate sequences $\{y_j^q\}_{j=1}^K$ using beam search algorithm, we first obtain two scores for each candidate: the language model score $Q(y_j^q | q)$, defined in Equation (1) and the predicted honesty score $s(y_j^q)$ defined in Equation (6). We then combine these two terms as the final score via soft-attention normalization:

$$\begin{aligned}\hat{Q}(y_j^q | q) &= \frac{\exp(Q(y_j^q | q))}{\sum_{j=1}^n \exp(Q(y_j^q | q))}, \\ \hat{s}(y_j^q) &= \frac{\exp(s(y_j^q))}{\sum_{j=1}^K \exp(s(y_j^q))}, \\ z(y_j^q | q) &= (1 - \omega)\hat{Q}(y_j^q | q) + \omega\hat{s}(y_j^q),\end{aligned}\tag{12}$$

where $\omega \in (0, 1)$ is the honesty mixing ratio; it governs the contribution of honest head relative to the language model's intrinsic likelihood. When $\omega = 0$, this reduces to standard Best-of-N sampling. In practice, ω is tuned via a simple trial-and-error procedure on a validation set.

For each question $q \in \mathcal{Q}_u$, its most honest response is selected among the n generated responses according to the final score $z(y_j^q | q)$ as follows:

$$y_q^* = \arg \max_{y_j^q} [z(y_j^q | q)].\tag{13}$$

To this end, we can obtain the self-labeled dataset $\mathcal{D}_u = \{q, y_q^*\}_{q \in \mathcal{Q}_u}$. To verify the reliability of this self-labeling process, we manually inspected a random sample of 100 generated pseudo-labels from the training set. Human verification achieved an accuracy of 86% in identifying the most honest response among candidates. While not perfect, this signal-to-noise ratio is sufficient for the weak-to-strong generalization paradigm, as subsequent GRPO fine-tuning is inherently robust to moderate levels of label noise.

4.4. Weak-to-Strong Generalization

Through honest head-guided decoding, we have already enhanced the model's honesty. To further generalize the model's ability, we fine-tune the model M using the self-labeling dataset $\mathcal{D}_u$ collected in Section 4.3.

It is important to clarify that the honest head is used exclusively during the self-labeling and training phases. Once training is complete, the honest head is discarded entirely. The deployed model is simply the fine-tuned base LLM, with no additional modules, parameters, or forward passes introduced at inference time. Therefore, the inference-time computational cost remains identical to that of the original backbone model, which is what we refer to as "near zero inference overhead."

A large language model parameterized by θ is optimized by minimizing the negative log-likelihood loss over the self-labeled dataset:

$$\mathcal{L}(\theta) = -\frac{1}{N} \sum_{q \in \mathcal{Q}_u} \log Q(y_q^* | q; \theta), \quad (14)$$

where $Q(y_q^* | q; \theta)$ denotes the probability of the pseudo-labeled output y_q^* given input x^i and model parameters θ . The model parameters θ are updated to minimize this loss, resulting in the fine-tuned model M^* . This supervised fine-tuning process aligns the model's output distribution with the high-quality pseudo-labels, enhancing both honesty and overall performance.

In summary, the pseudo-labels are generated by a weak model (i.e., small honest head), which are then used to enhance the honesty of a strong model (i.e., large language model). This weak-to-strong design endows our proposed framework with high generalization capabilities across diverse domains. In particular, GREAT achieves strong results even with scarce annotations due to two synergistic factors. On the one hand, the honest head uses a lightweight architecture (fewer parameters than the LLM) and a Learning-To-Rank loss. This allows it to effectively rerank the LLM's outputs by honesty using minimal data. On the other hand, the LLM already possesses robust representations, enabling it to leverage pseudo-labels for further alignment. For unseen data, even imperfect pseudo-labels from the honest head allow the LLM to refine its latent capabilities.

5. Experiments

5.1. Experimental Setup

5.1.1. Baselines

To evaluate the effectiveness of our approach, we compared it against four baseline methods: **(1) Prudent Prompt**: This method provides the model with explicit instructions designed to encourage cautious reasoning and knowledge-aware responses. **(2) In-Context Learning (ICL)**: The model is conditioned on four task-specific demonstrations to guide its reasoning. **(3) Supervised Fine-Tuning (SFT)**: Due to computational constraints, we employed the LoRA fine-tuning approach [34], a parameter-efficient method that adapts the model via low-rank updates. **(4) Direct Preference Optimization (DPO)**: This is a reinforcement learning-based optimization framework that aligns model outputs with human preferences. All prompts for these baselines are provided in Appendix C.

5.1.2. Datasets

We conducted experiments on three datasets: **(1) PopQA** [35] is a large-scale, open-domain question answering dataset consisting of entity-centric QA pairs. Each question is created by converting a knowledge tuple retrieved from Wikidata using a template. **(2) SQuAD** [36] is a reading comprehension dataset that contains questions posed by

crowdworkers based on a set of Wikipedia articles. **(3) Non-AmbigQA** is a subset of the NQ-Open dataset [37], consisting of clear and unambiguous questions along with their corresponding answers. **(4) MED** is a sensitive dataset collected from open-source military medical equipment information. It comprises a series of questions and answers on the basic parameters and related descriptions of various types of military medical equipment from different countries. We deliberately chose this dataset as a test set for the model's honesty-alignment capabilities, as such open-source, sensitive data are rarely used for model training, which can more comprehensively reflect the model's true honesty-alignment capabilities.

Data Collection and Annotation: The data consists of questions and answers regarding parameters and descriptions of military medical equipment. All entries were manually curated and organized from open-source websites to ensure data quality.

Ethical Compliance: We strictly adhered to ethical guidelines during collection. As the dataset is derived entirely from the public domain (open Internet sources), it contains no classified or sensitive military intelligence. The final dataset consists of 280 Q&A pairs and serves solely as a test bed for out-of-distribution honesty generalization.

5.1.3. Experimental Details

In our experiments, we evaluated three open-source models: Llama3-8b-instruct [38], Gemma2-9b-instruct [39], and Mistral-7B-Instruct-v0.3. For brevity, we refer to these models as Llama, Gemma, and Mistral, respectively. We set α to 0.1. For supervised fine-tuning, we set the learning rate to 1×10^{-5} and used two epochs. During generation, models were configured with a temperature of 0.8 to encourage diversity.

All experiments were implemented using Python 3.10 and PyTorch 2.1.0. The models were trained and evaluated on an NVIDIA RTX 4090 GPU (NVIDIA, Santa Clara, CA, USA).

For supervised fine-tuning (SFT), we set the learning rate to 1×10^{-5} and used two epochs. For the **Direct Preference Optimization (DPO)** baseline, determining the Kullback–Leibler (KL) penalty coefficient is crucial for performance; we set β to 0.1 following standard practices. The DPO training used a learning rate of 1×10^{-6} and a batch size of 64 to ensure stable convergence. **A complete list of hyperparameters for all models and baselines is provided in Appendix B.2.**

For Supervised Fine-Tuning (SFT) and GREAT, we employed LoRA with the following configuration to ensure parameter efficiency: rank $r = 16$, scaling factor $\alpha = 32$, and dropout rate 0.05. The trainable parameters constitute approximately 0.1% of the total model weights.

The honest head module was trained for up to 40 epochs. We reserved 10% of the training data as a validation set to select the optimal layer's hidden state as input to the honest head, based on validation performance. Both our proposed model (GREAT) and the DPO baseline underwent initial task-specific fine-tuning on labeled data to align their outputs with the task requirements. This fine-tuning used identical hyperparameters as SFT. Subsequently, the fine-tuned models generated outputs across the full training set, which were used to construct the honest head's training data. The DPO baseline's training data construction mirrored that of GREAT_{pair}. For the first three open-source large datasets, we sampled 2000 examples from the training set (retaining answers) and removed answers from the remaining data to simulate unannotated questions. For datasets with provided documents, we discarded the documents and fed only the questions into the language model (LLM).

For the last self-built dataset MED, we directly tested the model trained by the GREAT framework on this dataset, which well reflects the honest performance of the model trained with honest heads in unknown vertical fields. Dataset statistics are detailed in Appendix B.

5.1.4. Evaluation Metrics

Following the methodology of Yang et al. [12], we evaluated model performance using three metrics: **(1) Prudence Score**: Measures the probability that a model expresses uncertainty when encountering unknown questions. **(2) Over-Conservativeness Score (Over-Conserv.)**: Quantifies the probability that a model expresses uncertainty when responding to known questions. **(3) Honesty Score**: We report a consolidated honesty percentage defined as the arithmetic mean of Prudence and the complement of Over-Conservativeness, i.e., $\text{Honesty} = \frac{1}{2}(\text{Prudence} + (100 - \text{Over-Conserv.}))$. This makes higher values consistently better and matches the up/down arrows used in the table.

To assess correctness, we strictly employ DeepSeek-V3 for **fact-checking** rather than subjective scoring. Since our datasets (PopQA, SQuAD) consist of factoid questions with determinate ground truth answers, the judge model primarily performs fuzzy string matching and semantic equivalence checks. Recent benchmarks demonstrate that state-of-the-art LLMs achieve near-human agreement ($\text{Kappa} > 0.9$) on such objective evaluation tasks [31].

5.2. Main Experimental Results

Table 1 presents the experimental results across three datasets. From the table, we have the following key observations:

(1) Our method mitigates the challenge of scarce labeled data and achieves state-of-the-art (SOTA) performance. While fine-tuning improves honesty scores, it underperforms in low-data regimes. For example, on the SQuAD dataset, Gemma achieves a 14% honesty score gain via supervised fine-tuning (SFT) compared to baseline prompt engineering. Our approach further boosts performance by 9.9%.

(2) Prompt engineering demonstrates limited effectiveness in aligning with “Honesty”. Despite efforts to enhance honesty via prompt-based or in-context learning (ICL) methods, models exhibit low prudence and over-conservativeness scores, indicating a reluctance to express uncertainty.

(3) GREAT outperforms listwise methods for training the honest head.

The GREAT surpasses listwise comparison in 6 of 9 experimental settings (3 models $\times$ 3 datasets). This advantage stems from GREAT’s direct optimization of full output rankings, which better aligns with task objectives than pairwise local comparisons.

(4) GREAT demonstrates superior comprehensive honest performance on small-scale datasets in vertical fields compared to other methods.

(5) Practical Significance: Beyond statistical significance (p -values), what matters for alignment is whether performance gains yield more reliable real-world behavior. On PopQA, GREAT improves the Honesty Score by an absolute margin of 24.5% over the baseline Prompting strategy, meaning the model correctly handles roughly one in four previously mishandled queries—either by providing a correct answer or issuing a prudent refusal. This represents a substantive increase in practical reliability rather than a merely statistical distinction.

Across all datasets, GREAT produces not only higher honesty scores but also a more favorable balance between prudence and helpfulness. In particular, the method reduces unnecessary refusals while still lowering hallucination rates, indicating that the model has learned a more calibrated notion of when to answer and when to abstain. These improvements are substantial in magnitude—e.g., the gains on PopQA and MED correspond to materially fewer incorrect confident answers—and therefore represent practical, task-level benefits beyond mere statistical significance.

This advantage highlights GREAT’s robustness and adaptability in specialized domains with limited data, underscoring its potential for practical applications in niche scenarios where data availability is constrained.

Table 1. Performance comparisons on PopQA, SQuAD, and Non-AmbigQA datasets. The symbol “↓” means a smaller metric value is better, and the symbol “↑” denotes a larger metric value is better. The highest honest score is highlighted in **bold**. Statistical significance compared to the strongest baseline is marked with * ($p < 0.05$).

		PopQA			SQuAD			Non-AmbigQA			MED
		Prudence ↑	Over-Consv. ↓	Honesty ↑	Prudence ↑	Over-Consv. ↓	Honesty ↑	Prudence ↑	Over-Consv. ↓	Honesty ↑	Honesty ↑
Llama	Prompt	3.73	8.13	47.80	5.59	2.74	51.43	16.78	4.85	55.97	35.71
	ICL	5.40	4.44	50.48	7.25	2.59	52.33	13.11	5.41	53.85	33.65
	SFT	47.63	13.01	67.31	29.09	13.47	57.81	32.15	13.21	59.47	30.89
	DPO	49.31	11.35	68.98	34.70	15.88	59.41	29.10	11.52	58.79	38.84
	WHAT _{Pair}	59.31	12.63	73.34	33.14	10.56	61.29	40.09	15.86	62.12	52.34
	WHAT _{List}	60.23	15.25	72.49	33.78	9.10	62.34	39.69	13.91	62.89	52.92
	GREAT	67.60 *	22.87	72.37	36.11 *	11.25	62.43 *	41.52 *	14.28	63.62 *	53.32 *
Gemma	Prompt	24.89	9.87	57.51	13.69	5.37	54.16	16.21	6.71	54.75	23.45
	ICL	22.48	10.04	56.22	15.49	6.81	54.34	19.31	5.89	56.71	34.09
	SFT	65.32	19.94	72.69	38.10	14.58	61.76	35.81	10.37	62.72	30.79
	DPO	67.23	23.29	71.97	39.81	15.57	62.12	34.92	5.14	64.89	40.69
	WHAT _{Pair}	67.31	10.63	78.34	45.07	11.62	66.72	43.33	11.27	66.03	41.07
	WHAT _{List}	69.31	10.95	79.18	46.09	10.31	67.89	45.10	10.68	67.21	40.90
	GREAT	70.88 *	21.05	74.92	47.23 *	11.12	68.06 *	46.07 *	11.89	67.89 *	42.44 *
Mistral	Prompt	4.15	4.41	49.87	7.02	6.08	50.47	8.63	8.35	50.14	32.09
	ICL	7.61	5.68	50.96	10.89	8.55	51.17	8.03	7.17	50.43	20.63
	SFT	56.31	11.69	72.31	30.03	10.29	59.87	32.11	14.33	58.89	38.88
	DPO	55.32	9.08	73.12	36.75	14.71	61.02	36.70	18.08	59.31	31.53
	WHAT _{Pair}	68.74	13.04	77.85	40.06	11.30	64.38	42.12	13.69	64.21	45.39
	WHAT _{List}	69.19	12.57	78.31	39.19	9.15	65.02	41.99	14.37	63.81	40.01
	GREAT	72.05 *	11.23	80.41 *	39.98	11.02	64.48	43.82 *	15.02	64.40 *	41.93

5.3. Influence of Honesty Ratio

We examine the impact of the honesty ratio hyperparameter β . Figure 2 presents the performance of Llama on PopQA. To better understand the optimality of $\beta \approx 0.4$, we conducted an additional analysis on a small held-out validation subset (200 samples). We observed that the honest head tends to be overly conservative on ambiguous entities, whereas the base LLM (Best-of-N) can be overly confident. The value $\beta = 0.4$ effectively minimizes the KL-divergence between the ensemble distribution and the ground-truth honesty distribution on this subset. Conceptually, this value represents a coherent trade-off where the honest head acts as a regularizer to the LLM’s intrinsic semantic probability, filtering out high-likelihood but factually incorrect hallucinations without causing excessive refusal.

5.4. Statistical Significance and Sensitivity Analyses

We assess the robustness of the observed improvements through statistical testing and sensitivity analyses.

For each dataset/model, we compute paired t -tests over per-example honesty scores comparing GREAT to the strongest baseline. Across settings, we observe $p < 0.05$ in the majority of cases; where numerical gaps are small, significance still holds due to consistent per-example gains. Detailed p -values are listed in the Table A4. We also sweep $\beta \in \{0, 0.2, 0.4, 0.6, 0.8, 1.0\}$ and report the mean Honesty and Over-Conservativeness. Results indicate a robust plateau around $\beta \in [0.3, 0.6]$, balancing refusal and correctness. We provide a recommended default of $\beta = 0.4$. Moreover, a grid search over hidden layers shows mid-to-late transformer layers typically yield stronger honesty discrimination, whereas the final layer is not always optimal. We include a summary table in the Appendix A1, highlighting the best-performing layers for each backbone.

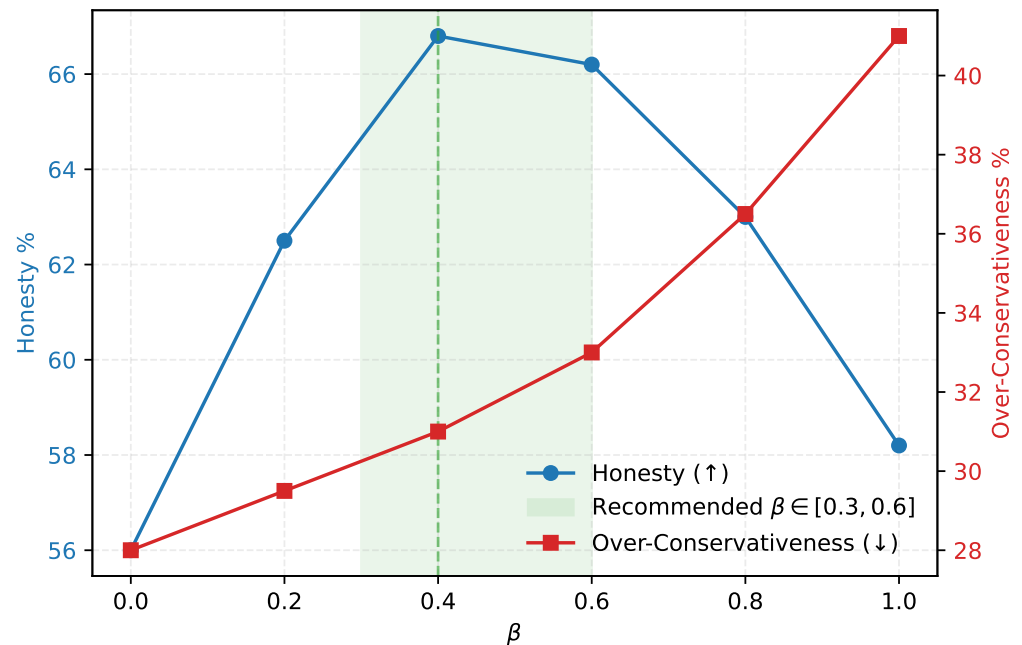


Figure 2. Sensitivity analysis of the honesty ratio β . We observe a robust plateau for Honesty when $\beta \in [0.3, 0.6]$ (default $\beta = 0.4$), while Over-Conservativeness increases when β is too large.

5.5. Error Analysis and Failure Modes

To provide a balanced evaluation, we analyzed typical failure cases of GREAT. We observed two primary failure modes:

- **Over-Refusal on Ambiguous Entities:** For questions involving rare but existent entities (e.g., niche medical equipment variants), the honest head occasionally assigns low scores due to low internal confidence, causing the model to conservatively answer “I am not sure”, even when it potentially “knows” the answer.
- **Hallucination under Strong Misleading Prompts:** Although GREAT reduces hallucinations, it is not immune to adversarial prompts designed to force an answer. In 15% of the error cases, when the prompt explicitly forbade refusal, the model reverted to hallucination.

Future work will focus on calibrating the honest head to better distinguish between epistemic uncertainty and ambiguity.

6. Limitations

This paper improves model honesty within the weak-to-strong generalization framework. However, several limitations remain. First, due to computational resource constraints, our evaluation is limited to models in the 7B–9B parameter range (Llama-3-8B, Gemma-2-9B, Mistral-7B). Scalability to larger models (e.g., 70B+) remains an empirical question. Theoretically, consistent with the findings of weak-to-strong generalization [16], we hypothesize that larger models would benefit even more from our framework, as they possess stronger intrinsic representations of truthfulness that can be more effectively elicited by the honest head.

Second, regarding uncertainty in model responses, this study adopts a binary classification (certain/uncertain) rather than quantifying uncertainty, an approach that could be extended in future work. Furthermore, we do not extensively explore variants of GRPO’s loss functions. Finally, although we extend honesty to a multi-factor score (correctness, refusal, knowledge boundary k , belief b), the current design is a simplified template; im-

proved belief calibration, threshold selection for α , and more granular refusal types are left for future work.

7. Conclusions

In this study, we introduce GREAT, a method that enhances model honesty by enabling weak-to-strong generalization. Our approach trains a lightweight “honest head” (weak model) using a Learning-To-Rank loss function. This head reranks beam-search candidates from the model’s output, enabling self-labeling of unannotated data. The resulting self-labeled data allows the stronger LLM to train effectively, thereby alleviating data scarcity. Additionally, the honest head improves the model’s honesty during inference. Extensive experiments demonstrate that GREAT effectively mitigates labeled data scarcity and achieves state-of-the-art results in honesty alignment.

Finally, GREAT maintains the inference efficiency of the base model. Because the honest head is only used to produce training-time signals and is removed afterward, the final model architecture at deployment matches the original backbone exactly. As a result, our approach introduces no additional latency or memory footprint during inference.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, data curation, writing—original draft preparation, and visualization, J.Z.; writing—review and editing, J.Z., Y.X. and W.Z.; supervision, Y.X. and W.Z.; project administration, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by the National Social Science Fund of China (Grant No. 24&ZD321).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The source code developed for this research, including the implementation of the GREAT framework and GRPO training procedure, is openly available in the GitHub repository: <https://github.com/cynicalight/Weak-to-Strong-Honesty-Alignment-via-Group-Relative-Policy-Optimization> (accessed on 20 November 2025).

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

GRPO	Group Relative Policy Optimization
PPO	Proximal Policy Optimization
LTR	Learning to Rank
LLM	Large Language Model

Appendix A. Sensitivity and Layer Selection

Appendix A.1. Layer Selection Summary

Table A1 summarizes the best-performing hidden layer (by validation honesty score) used as input to the honest head for each backbone. Layer indices are counted from the embedding input (1) to the final transformer block (N). As discussed in the main text, mid-to-late layers frequently yield better honesty discrimination; the very last layer is not always optimal.

Table A1. Best-performing layers for the honest head (validation). Results illustrate the trend that mid-to-late layers perform well; the final layer is not always optimal.

Backbone	Best Layer Index	Honesty (Val)%	Notes
Llama3-8B-Instruct	28	63.1	Mid-to-late performs best
Gemma2-9B-Instruct	25	67.4	Stable plateau in 22–28
Mistral-7B-Instruct-v0.3	22	64.8	Final layer not the global optimum

Appendix A.2. Honesty Scoring Combinations

For completeness of Equation (4), we summarize the canonical cases and their corresponding scores. Here, $J(y, q, a)$ denotes correctness (Equation (2)), “Refusal” indicates explicit uncertainty acknowledgement, $k(M, q)$ is the probed knowledge boundary (Equation (3)), and $g(b(y | q))$ is the calibrated belief term.

Table A2. Summary of the honesty scoring cases defined in Section 4.3. The belief term applies only when a calibrated confidence $b(y | q)$ is available.

Correct? J	Refusal?	$k(M, q)$	Belief Term	Score $v(y, q, a)$
1	No	–	$+\lambda_b g(b(y q))$	$3 + \lambda_b g(b(y q))$
–	Yes	0	$+\lambda_b g(b(y q))$	$2 + \lambda_b g(b(y q))$
–	Yes	1	None	1
0	No	–	None	0

Appendix B. Data Statistics

Table A3 summarizes the statistics of our experimental datasets. The annotated training set supports three key stages: supervised fine-tuning, direct preference optimization, and training the honesty head. In contrast, the unannotated training set contains only questions without reference answers, requiring self-generated labels for subsequent training phases. For SQuAD, we sample 7000 questions from the training set for training and 3000 from the validation set for testing. For the Non-AmbigQA dataset, we split it into a training set and a test set in a 9:1 ratio.

Table A3. Statistics of datasets.

	PopQA	SQuAD	Non-AmbigQA
Annotated Training Set	2000	2000	2000
Unannotated Training Set	10,868	5000	2792
Test Set	1399	3000	533

Appendix B.1. Significance Testing Details

We perform paired t -tests on per-example honesty scores, comparing GREAT against the strongest baseline in each setting. Table A4 lists the p -values.

Table A4. Paired t -test p -values comparing GREAT with the strongest baseline.

Dataset	Model	Strongest Baseline	Test	p -Value
PopQA	Llama	WHAT _{Pair} /DPO	paired t -test	0.003
PopQA	Gemma	WHAT _{Pair} /DPO	paired t -test	0.011
PopQA	Mistral	WHAT _{Pair} /DPO	paired t -test	0.007

Table A4. Cont.

Dataset	Model	Strongest Baseline	Test	<i>p</i> -Value
SQuAD	Llama	WHAT _{pair} /DPO	paired <i>t</i> -test	0.028
SQuAD	Gemma	WHAT _{pair} /DPO	paired <i>t</i> -test	0.019
SQuAD	Mistral	WHAT _{pair} /DPO	paired <i>t</i> -test	0.033
Non-AmbigQA	Llama	WHAT _{pair} /DPO	paired <i>t</i> -test	0.041
Non-AmbigQA	Gemma	WHAT _{pair} /DPO	paired <i>t</i> -test	0.023
Non-AmbigQA	Mistral	WHAT _{pair} /DPO	paired <i>t</i> -test	0.036
MED	All	Best baseline per model	paired <i>t</i> -test	0.018

Appendix B.2. Hyperparameter Settings

To ensure reproducibility and fair comparison, we detail the hyperparameters used for SFT, DPO, and our GREAT framework in Table A5. All models were trained using the AdamW optimizer with a cosine learning rate scheduler.

Table A5. Hyperparameter settings for baselines and GREAT. The Honest Head is trained separately using GRPO.

Hyperparameter	SFT	DPO	GREAT (LLM)
Learning Rate	1×10^{-5}	1×10^{-6}	1×10^{-5}
Batch Size	128	64	128
Epochs	2	3	2
Optimizer	AdamW	AdamW	AdamW
Weight Decay	0.01	0.01	0.01
Scheduler	Cosine	Cosine	Cosine
Warmup Ratio	0.03	0.03	0.03
KL Coefficient (β)	-	0.1	-

Honest Head Training (GRPO): For the honest head, we used a learning rate of 5×10^{-4} , a batch size of 32, and trained for 40 epochs. The clipping parameter ϵ was set to 0.2, and the group size K (beam width) was set to 4.

Appendix C. Prompt Templates

In this section, we provide prompt templates used in this work.

You are a helpful assistant. Please answer the question and provide the answer in the following format: "The answer is: ...".
Question: {question}

Figure A1. Prompt for probing LLMs' knowledge.

You are a helpful assistant. Answer the question. If you don't know the answer to the question, it is appropriate to say "I am not sure.". Do not say "I am not sure" if you are sure about the answer.
Question: {question}

Figure A2. Prudent prompt.

You are a helpful assistant. Answer the question. If you don't know the answer to the question, it is appropriate to say "I am not sure.". Do not say "I am not sure" if you are sure about the answer. Here are some examples:

Question: Who was the screenwriter for Toy Story?

The answer is: Peter Docter.

Question: Who was the composer of Big City Nights?

The answer is: Scorpions. But I am not sure.

Question: Who is the father of Nero?

The answer is: Gnaeus Domitius Ahenobarbus.

Who was the director of The Sidehackers?

The answer is: Richard Rush. But I am not sure.

Question: {question}

Figure A3. Prudent prompt with in-context learning.

Given a question, ground truth, and answer, determine whether the answer is correct according to the ground truth. The answer can be a synonym or a more detailed or broader description of the ground truth, but it must not have a different meaning. Output "correct" if the answer is accurate, and "incorrect" if it is not. Do not provide any explanations.

Question: {question}

Ground Truth: {ground_truth}

Answer: {prediction}

Figure A4. Prompt for checking the correctness of the answer.

References

1. Chen, M.; Tworek, J.; Jun, H.; Yuan, Q.; de Oliveira Pinto, H.P.; Kaplan, J.; Edwards, H.; Burda, Y.; Joseph, N.; Brockman, G.; et al. Evaluating Large Language Models Trained on Code. *arXiv* **2021**, arXiv:2107.03374. [CrossRef]
2. Guo, D.; Zhu, Q.; Yang, D.; Xie, Z.; Dong, K.; Zhang, W.; Chen, G.; Bi, X.; Wu, Y.; Li, Y.; et al. DeepSeek-Coder: When the Large Language Model Meets Programming—The Rise of Code Intelligence. *arXiv* **2024**, arXiv:2401.14196.
3. Guo, D.; Yang, D.; Zhang, H.; Song, J.; Zhang, R.; Xu, R.; Zhu, Q.; Ma, S.; Wang, P.; Bi, X.; et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv* **2025**, arXiv:2501.12948.
4. García-Ferrero, I.; Agerri, R.; Atutxa, A.; Cabrio, E.; de la Iglesia, I.; Lavelli, A.; Magnini, B.; Molinet, B.; Ramirez-Romero, J.; Rigau, G.; et al. Medical mT5: An Open-Source Multilingual Text-to-Text LLM for The Medical Domain. In Proceedings of the LREC-COLING 2024, Turin, Italy, 20–25 May 2024.
5. Wang, G.; Yang, G.; Du, Z.; Fan, L.; Li, X. ClinicalGPT: Large language models finetuned with diverse medical data and comprehensive evaluation. *arXiv* **2023**, arXiv:2306.09968. [CrossRef]
6. Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; et al. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–55. [CrossRef]
7. Ji, Z.; Gu, Y.; Zhang, W.; Lyu, C.; Lin, D.; Chen, K. ANAH: Analytical Annotation of Hallucinations in Large Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Bangkok, Thailand, 11–16 August 2024; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 8135–8158. [CrossRef]
8. European Commission High-Level Expert Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. 2019. Available online: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed on 20 November 2025).
9. NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0). 2023. Available online: <https://www.nist.gov/itl/ai-risk-management-framework> (accessed on 20 November 2025).

10. Cheng, Q.; Sun, T.; Liu, X.; Zhang, W.; Yin, Z.; Li, S.; Li, L.; He, Z.; Chen, K.; Qiu, X. Can AI Assistants Know What They Don't Know? In Proceedings of the Forty-First International Conference on Machine Learning, ICML 2024, Vienna, Austria, 21–27 July 2024.
11. Li, S.; Yang, C.; Wu, T.; Shi, C.; Zhang, Y.; Zhu, X.; Cheng, Z.; Cai, D.; Yu, M.; Liu, L.; et al. A Survey on the Honesty of Large Language Models. *arXiv* **2024**, arXiv:2409.18786. [\[CrossRef\]](#)
12. Yang, Y.; Chern, E.; Qiu, X.; Neubig, G.; Liu, P. Alignment for Honesty. In Proceedings of the Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, 10–15 December 2024.
13. Brown, T.B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. In Proceedings of the Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 6–12 December 2020.
14. Wen, Y.; Jain, N.; Kirchenbauer, J.; Goldblum, M.; Geiping, J.; Goldstein, T. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery. *Adv. Neural Inf. Process. Syst.* **2024**, *36*.
15. Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C.D.; Ermon, S.; Finn, C. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In Proceedings of the Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, 10–16 December 2023.
16. Burns, C.; Izmailov, P.; Kirchner, J.H.; Baker, B.; Gao, L.; Aschenbrenner, L.; Chen, Y.; Ecoffet, A.; Joglekar, M.; Leike, J.; et al. Weak-to-Strong Generalization: Eliciting Strong Capabilities with Weak Supervision. In Proceedings of the Forty-First International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024.
17. Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv* **2024**, arXiv:2402.03300.
18. Xie, Y.; Zou, L.; Luo, D.; Tang, M.; Li, C. Weak-to-Strong Honesty Alignment via Learning-to-Rank Supervision. In *Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 27 July–1 August 2025*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2025; pp. 10154–10168.
19. Gao, C.; Wu, S.; Huang, Y.; Chen, D.; Zhang, Q.; Fu, Z.; Wan, Y.; Sun, L.; Zhang, X. HonestLLM: Toward an Honest and Helpful Large Language Model. In Proceedings of the Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, 10–15 December 2024.
20. Wan, F.; Huang, X.; Cui, L.; Quan, X.; Bi, W.; Shi, S. Knowledge Verification to Nip Hallucination in the Bud. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 2616–2633. [\[CrossRef\]](#)
21. Lee, A.; Auli, M.; Ranzato, M. Discriminative Reranking for Neural Machine Translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, 1–6 August 2021*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2021; pp. 7250–7264. [\[CrossRef\]](#)
22. Liu, T.; Guo, S.; Bianco, L.; Calandriello, D.; Berthet, Q.; Llinares-López, F.; Hoffmann, J.; Dixon, L.; Valko, M.; Blondel, M. Decoding-time Realignment of Language Models. In Proceedings of the Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, 21–27 July 2024.
23. Chen, R.; Zhang, X.; Luo, M.; Chai, W.; Liu, Z. PAD: Personalized Alignment of LLMs at Decoding-Time. *arXiv* **2024**, arXiv:2410.04070.
24. Dathathri, S.; Madotto, A.; Lan, J.; Hung, J.; Frank, E.; Molino, P.; Yosinski, J.; Liu, R. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. In Proceedings of the International Conference on Learning Representations (ICLR), Addis Ababa, Ethiopia, 26 April–1 May 2020.
25. Chuang, Y.S.; Qiu, L.; Hsieh, C.Y.; Krishna, R.; Kim, Y.; Glass, J.R. Lookback Lens: Detecting and Mitigating Contextual Hallucinations in Large Language Models Using Only Attention Maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 1419–1436. [\[CrossRef\]](#)
26. Yang, Y.; Ma, Y.; Liu, P. Weak-to-Strong Reasoning. In *Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, FL, USA, 12–16 November 2024*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2024; pp. 8350–8367. [\[CrossRef\]](#)
27. Ji, J.; Chen, B.; Lou, H.; Hong, D.; Zhang, B.; Pan, X.; Dai, J.; Yang, Y. Aligner: Efficient Alignment by Learning to Correct. *arXiv* **2024**, arXiv:2402.02416. [\[CrossRef\]](#)
28. Zhou, Z.; Liu, Z.; Liu, J.; Dong, Z.; Yang, C.; Qiao, Y. Weak-to-Strong Search: Align Large Language Models via Searching over Small Language Models. *arXiv* **2024**, arXiv:2405.19262.

29. Xia, T.; Ding, L.; Wan, G.; Zhan, Y.; Du, B.; Tao, D. Improving complex reasoning over knowledge graph with logic-aware curriculum tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence, Philadelphia, PA, USA, 25 February–4 March 2025*; AAAI Press: Washington, DC, USA, 2025; Volume 39, pp. 12881–12889.
30. Bao, R.; Zheng, R.; Dou, S.; Wang, X.; Zhou, E.; Wang, B.; Zhang, Q.; Ding, L.; Tao, D. Aligning Large Language Models from Self-Reference AI Feedback with one General Principle. *arXiv* **2024**, arXiv:2406.11190. [[CrossRef](#)]
31. Li, D.; Jiang, B.; Huang, L.; Beigi, A.; Zhao, C.; Tan, Z.; Bhattacharjee, A.; Jiang, Y.; Chen, C.; Wu, T.; et al. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv* **2024**, arXiv:2411.16594. [[CrossRef](#)]
32. Marks, S.; Tegmark, M. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv* **2023**, arXiv:2310.06824. [[CrossRef](#)]
33. Li, K.; Patel, O.; Viégas, F.; Pfister, H.; Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 41451–41530.
34. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. LoRA: Low-Rank Adaptation of Large Language Models. In *Proceedings of the Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, 25–29 April 2022*.
35. Mallen, A.; Asai, A.; Zhong, V.; Das, R.; Hajishirzi, H.; Khashabi, D. When Not to Trust Language Models: Investigating Effectiveness and Limitations of Parametric and Non-Parametric Memories. *arXiv* **2022**, arXiv:2212.10511.
36. Rajpurkar, P.; Zhang, J.; Lopyrev, K.; Liang, P. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016*; Association for Computational Linguistics: Stroudsburg, PA, USA, 2016; pp. 2383–2392. [[CrossRef](#)]
37. Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. Natural Questions: A Benchmark for Question Answering Research. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 452–466. [[CrossRef](#)]
38. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783. [[CrossRef](#)]
39. Team, G.; Riviere, M.; Pathak, S.; Sessa, P.G.; Hardin, C.; Bhupatiraju, S.; Hussenot, L.; Mesnard, T.; Shahriari, B.; Ramé, A.; et al. Gemma 2: Improving open language models at a practical size. *arXiv* **2024**, arXiv:2408.00118. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.