Knowing When to Speak: Aligning LLM Honesty via Solvability-Aware Preference Optimization

Anonymous ACL submission

Abstract

Large Language Models often exhibit dishonest behaviors–confidently hallucinating answers to questions beyond their knowledge boundary or hedging on queries they have mastered. Although existing mitigation methods have progressed to establishing the model knowledge boundary, this boundary is still typically treated as a coarse "known vs. unknown" distinction, leaving a finer-grained question underexplored: *Does a failure reflect missing knowledge, or merely under-expressed capability?* In this paper, we propose KCAPO, a framework that aligns the model's output behavior with its latent capabilities. It estimates query-level solvability with a consistency-supervised hidden-state probe, then uses its mismatch with response confidence and factual behavior to refine the knowledge boundary into four capability–behavior cases. Based on this fine-grained characterization, we construct preference data to dynamically "Wake Up" dormant knowledge or "Suppress" blind guessing. Furthermore, we introduce an adaptive margin into the DPO objective, which prioritizes correcting responses with severe cognitive dissonance. Extensive experiments on short-form QA and refusal benchmarks show that KCAPO reduces hallucinations without compromising utility and improves selective abstention.

1 Introduction

Large Language Models have shown strong capabilities across NLP tasks (Liu et al., 2023b), but their deployment in knowledge-sensitive applications is limited by *hallucination* (Liu et al., 2023b; Tonmoy et al., 2024; Huang et al., 2025). Within this realm, a cornerstone of hallucination is *dishonesty*: LLMs may confidently fabricate answers to queries outside their knowledge boundaries, rather than expressing uncertainty.

Code and dataset are publicly available at: https://anonymous.4open.science/r/KCAPO-0422.

Specifically, the honesty of LLMs refers to the behavior of humbly acknowledging the limitations in its internal capabilities ("*knowing*") and faithfully expressing their awareness ("*saying*"), which is literally framed as two coupled criteria: (i) *self-knowledge*: recognizing the boundary between known and unknown information, and (ii) *self-expression*: responding faithfully to reflect their internal epistemic state rather than hedging unnecessarily (Yang et al., 2024; Chern et al., 2024; Askell et al., 2021). As pointed out by (Li et al., 2024), current LLMs often commit confidently to queries beyond their latent capability, while under-expressing competence on queries they are able to answer. This reveals a subtle but critical phenomenon: *not every failure indicates missing knowledge, some reflect merely under-expressed capability.*

To enable the accurate expression conditioned on the internal knowledge, prior methods have gradually shifted the focus from knowledge elicitation to knowledge-boundary perception. Inference-time interventions–such as contrastive decoding (Chuang et al., 2023; Zhang et al., 2025), self-refinement (Madaan et al., 2023), and integrative decoding (Cheng et al., 2025)–ask how to better surface knowledge already encoded in the model, but they do not explicitly determine *whether a query is internally solvable in the first place*. They improve expression without calibrating cognition. Training-based alignment methods (Ouyang et al., 2022; Cheng et al., 2024; Yang et al., 2024) move one step further by introducing the notion of a knowledge boundary, which focuses on that *models should refuse when they lack knowledge*. However, these approaches overwhelmingly treat the boundary as a **coarse binary distinction**–"known" versus "unknown" –leaving a finer-grained question unexplored: *does a failure reflect agenuine knowledge deficit, or merely a failure to express latent competence?*

Recent internal-state approaches have shown

that hidden representations can provide useful signals for estimating or calibrating knowledge boundaries (Azaria and Mitchell, 2023; Cheng et al., 2024; Shi et al., 2025; Ni et al., 2025). Moreover, Gekhman et al. (2025) observe that models may encode more factual knowledge in their parameters than they express in their outputs. Inspired by these, we argue that the LLM honesty is *not* merely to "refuse more" or "sound more uncertain", but to *synchronize* what the model internally knows with how it externally commits to an answer. Building on this view, we formulate honesty alignment as an explicit **cognition–behavior synchronization** problem, where the central question is not only whether a query is "known" or "unknown", but whether the model's response behavior faithfully reflects its internal solvability. Based on this finer-grained characterization, we propose Knowledge–Confidence Adaptive Direct Preference Optimization (KCAPO), which estimates query-level solvability, constructs preference data according to solvability–confidence mismatch, and dynamically adjusts alignment strength through a confidence-aware adaptive margin.

Specifically, we first estimate query-level *internal solvability* from hidden representations, providing a generation-independent signal of whether the model can solve a given query in principle. This estimate enables a principled distinction between epistemic regimes in which the model should answer confidently and those in which abstention is more appropriate. It then combines solvability, response confidence, and factual correctness to construct fine-grained preference pairs, refining knowledge-boundary alignment into four cognition–behavior cases: know and answer, know but abstain, do not know but answer, and do not know and abstain. These preference pairs selectively elicit under-expressed knowledge for solvable queries and suppress blind guessing for unsolvable ones. Furthermore, we introduce a mismatch-aware adaptive margin into the DPO objective to dynamically adjust alignment strength according to solvability–confidence mismatch. Extensive experiments across multiple benchmarks demonstrate that KCAPO substantially reduces dishonesty while preserving, and in many cases improving, performance on answerable queries. Additionally, compared to strong baselines, our method achieves a more balanced trade-off between factual reliability and answerability, highlighting the benefits of our framework.

2 Related Works

2.1 Honesty of Large Language Models.

Honesty is a cornerstone of aligning LLMs with human values (Askell et al., 2021), particularly in high-stakes domains like medicine, law, and finance (Thirunavukarasu et al., 2023; Dahl et al., 2024; Li et al., 2023b). Formally, honesty comprises two coupled criteria: *self-knowledge*—recognizing internal capability boundaries ("knowing what it knows")—and *self-expression*—faithfully articulating this state without deception ("saying what it thinks") (Yang et al., 2024; Kadavath et al., 2022). However, current training paradigms often induce a misalignment between these objectives and epistemic reality. First, **superficial imitation** encourages a confident persona that outstrips actual competence, creating a "style-content mismatch" (Zhou et al., 2023; Gudibande et al., 2023). This is exacerbated by **forced supervision** during SFT, where models are pressured to answer beyond their knowledge boundaries, incentivizing fabrication (Gekhman et al., 2024; Lin et al., 2024). Furthermore, preference alignment (e.g., RLHF) often inadvertently rewards **sycophancy**, prioritizing user appeal over objective truth (Sharma et al., 2023). Collectively, these factors foster a **cognitive-expressive dissonance**, where models fail to faithfully articulate their internal knowledge (Li et al., 2023a).

2.2 Strategies for Improving Honesty

Existing approaches generally fall into two paradigms: training-free inference interventions and training-based alignment (Li et al., 2024). The former exploits **prompting** for uncertainty elicitation (Tian et al., 2023), **sampling consistency** for confidence estimation (Wang et al., 2022), or **decoding-time interventions** (e.g., ITI (Li et al., 2023a), DoLa (Chuang et al., 2023)) to steer activation patterns towards truthfulness without modifying parameters. Conversely, training-based methods fundamentally instill honest behaviors via **Self-aware Fine-tuning (SFT)** (Yang et al., 2024) or **Reinforcement Learning (RL)** (Xu et al., 2024), utilizing preference optimization (e.g., DPO (Rafailov et al., 2023; Yang et al., 2025)) to teach models to refuse unknown queries. However, these approaches often rely on static external supervision or unilateral optimization, failing to dynamically synchronize the model's evolving internal solvability with its output policy, which limits their

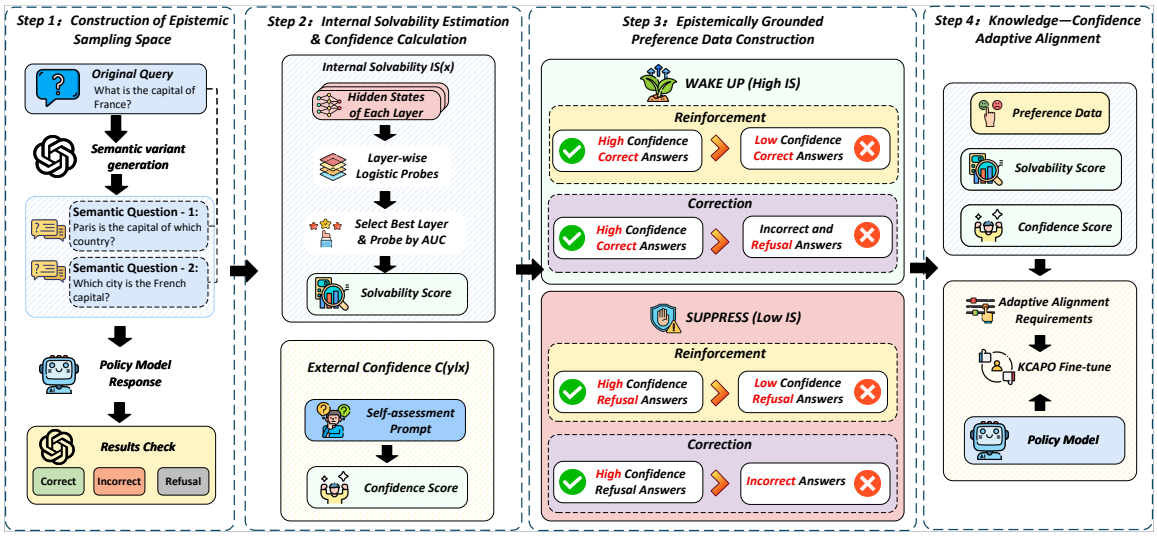

Figure 1: A diagram illustrating the four steps of KCAPO: (i): Generate initial responses and check the facts. (ii): Collect metrics on problem solvability and response confidence. (iii): Create pairwise preference data. (iv): Fine-tune LLM with KCAPO.

effectiveness in handling the complex "knowing-but-not-saying" dissonance.

3 Methodology

To *synchronize* the cognitive gap between internal cognition and external behavior, we propose **KCAPO**. As illustrated in Figure 1, KCAPO operates in four stages: (i): We generate semantic variants of each query, collect the model's responses, and verify their factual correctness to build a consistency-based sampling space. (ii): Estimating query-level *internal solvability* via hidden-state probing and measure the *external confidence* of each response; (iii): Constructing preference pairs using solvability, confidence, and initial response behavior to encourage answering solvable queries and abstaining from unsolvable ones; (iv): Fine-tuning the model with a mismatch-aware adaptive DPO objective, assigning stronger alignment pressure to responses with larger solvability–confidence mismatches.

3.1 Construction of Epistemic Sampling Space

A single response is an unreliable indicator of internal cognition, since correctness may arise from guessing and failure may arise from under-expression. We therefore construct a consistency-based sampling space: a reliably solvable factual query should remain correctly answered under semantically equivalent reformulations, while guesses or surface-pattern matches are more likely to fail under perturbations. For each original query x_i, we leverage a separate and superior LLM (e.g., GPT-5.5) to generate K variants, denoted as $X_{var} = \{x_1^i, \ldots, x_K^i\}$. These variants include *syntactic restructuring*, *lexical substitution*, and *answer-preserving distractor perturbations*, which change the surface form while preserving the gold answer. Details are in the Appendix A.1.

We prompt the policy model π_θ to answer each query in $x_i \cup X_{\text{var}}^i$, and use a stronger LLM judge to evaluate each response against the gold answer as *Correct*, *Incorrect*, or *Refusal*. This yields a sampling sub-space $\mathcal{S}_i = \{(x_k^i, y_k^i, l_k^i)\}_{k=0}^K$ for each knowledge point i, where x_k^i, y_k^i, and l_k^i denote the query variant, model initial response, and judge label, respectively. Aggregating all sub-spaces gives $\mathcal{S}_{\text{total}}$, which provides consistency-based supervision for estimating *Internal Solvability*.

3.2 Quantification of Solvability and Confidence

To align the model's response behavior with its internal cognition, we quantify two complementary signals: **Internal Solvability** ($IS(x)$), which estimates whether a query can be reliably solved by the model, and **External Confidence** ($C(y|x)$), which measures how strongly the model commits to its generated response.

Internal Solvability Estimation. Motivated by findings that hidden states encode factuality and uncertainty signals (Li et al., 2023a; Azaria and Mitchell, 2023; Kadavath et al., 2022), we estimate

solvability from internal representations rather than single-generation outcomes which may be affected by decoding noise or under-expression. Specifically, we use the consistency labels derived in Section 3.1 as behavioral supervision to train a diagnostic probe. We assign $y_{\text{solv}} \in \{0, 1\}$ from cross-variant consistency, with consistent correctness labeled solvable and consistent failure labeled unsolvable. For each original query x, we extract the last-token hidden state $h_l(x_i) \in \mathbb{R}^d$ from layer l and train a logistic regression probe:

$$IS_l(x) = \sigma\big(w_l^\top h_l(x) + b_l\big) \tag{1}$$

We select the best layer l^* by AUC and compute the final generation-independent solvability score $IS(x)$ as the average IS_{l^*} over the original query and its variants, only for knowledge points whose sampling sub-spaces exhibit inconsistent response behaviors.

External Confidence Estimation. To characterize the model's capability–behavior state, we follow prior work (Tian et al., 2023; Xiong et al., 2023; Kadavath et al., 2022) on verbalized confidence and prompt the model to rate its confidence in response y from 0 to 1, yielding $C(y|x)$. Importantly, We do not treat $C(y|x)$ as evidence of the model's knowledge state; instead, it serves as an external-expression signal for mismatch severity. High-confidence errors on low-solvability queries indicate blind guessing, whereas low confidence or refusal on high-solvability queries indicates under-expressed capability. This signal guides adaptive optimization rather than deciding what the model knows.

3.3 Epistemically Grounded Preference Data Construction

Guided by internal solvability $IS(x)$, external confidence $C(y|x)$, and initial response behavior, we construct preference pairs under two regimes, **WAKE UP** and **SUPPRESS**, with two strategies, *Correction* and *Reinforcement*. This design refines the knowledge boundary from "known vs. unknown" into four capability–behavior states: *know and answer*, *know but abstain*, *do not know but answer*, and *do not know and abstain*. Preference pairs are drawn from sampling sub-spaces with inconsistent responses across semantic variants in Section 3.1, forming the final preference dataset $\mathcal{D} = \{(x, y_w, y_l)\}$ for KCAPO, where y_w represents the epistemically appropriate behavior and y_l represents the corresponding cognitive misalignment.

WAKE UP Regime. For high-solvability queries ($IS(x) > \tau_{\text{high}}$), the model has a high potential to solve them. This regime covers two capability–behavior states: "*know and answer*" and "*know but abstain*": for correct responses, we construct **Reinforcement Pairs**, where y_w is selected from high-confidence correct responses and y_l is another correct response with lower confidence. For incorrect responses and refusals, we construct **Correction Pairs**, where y_w is a correct response and y_l is selected from incorrect responses or refusals.

SUPPRESS Regime. Conversely, for queries with low internal solvability ($IS(x) < \tau_{\text{low}}$), the model is considered to have a low probability of resolving these queries. This regime covers "*do not know and answer*" and "*do not know and abstain*": For incorrect responses, we construct **Correction Pairs** where y_w is an abstention and y_l is selected from initial incorrect answers; and for existing refusals, we construct **Reinforcement Pairs** where y_w is selected from high-confidence abstentions and y_l is another abstention with lower confidence.

To mitigate data imbalance, we use group-balanced sampling, drawing mini-batches uniformly across preference types and reusing smaller groups when necessary. Details of the preference data construction are in the Appendix B.

3.4 Knowledge–Confidence Adaptive Alignment

The preference pairs constructed in Section 3.3 specify which response should be preferred, but standard DPO treats all preference pairs with the same optimization strength. This is insufficient for capability–behavior mismatch: a low-confidence mistake and a high-confidence hallucination should not receive the same correction pressure. Inspired by margin-based preference optimization (Meng et al., 2024; Wu et al., 2024; Rho, 2025), we introduce a mismatch-aware adaptive margin that scales the DPO objective according to the severity of solvability–confidence conflict.

Mismatch-Aware Adaptive Margin. For each preference pair (x, y_w, y_l), where y_w and y_l denote the preferred and dispreferred responses, we define an adaptive margin:

$$\mathcal{M}(x, y_l) = C(y_l|x) \cdot \alpha(x), \tag{2}$$

where $C(y_l|x)$ is the external confidence of the dispreferred response, and $\alpha(x)$ measures how strongly the query belongs to a mismatch-sensitive correction case:

$$\alpha(x) = \begin{cases} IS(x), & x \in \mathcal{D}_{\text{wake}} \setminus \mathcal{D}_{\text{reinf}}, \\ \gamma(1 - IS(x)), & x \in \mathcal{D}_{\text{supp}} \setminus \mathcal{D}_{\text{reinf}}, \\ 0, & x \in \mathcal{D}_{\text{reinf}}. \end{cases} \tag{3}$$

Here, γ controls the relative strength of SUPPRESS, and $\mathcal{D}_{\text{reinf}}$ denotes reinforcement pairs. The margin is applied only to correction pairs, where the dispreferred response reflects a clear mismatch: either a solvable query is not properly expressed, or an unsolvable query is answered with unwarranted confidence. For reinforcement pairs, where the goal is mainly to consolidate already appropriate behavior, the margin is set to zero.

Adaptive Optimization Objective. Let

$$r_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \tag{4}$$

be the implicit reward log-ratio, and define

$$\Delta r_\theta = r_\theta(x, y_w) - r_\theta(x, y_l). \tag{5}$$

The KCAPO loss is:

$$\begin{aligned} \mathcal{L}_{\text{KCA}} = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \big[\log \sigma \big(\beta \big(\Delta r_\theta \\ -\lambda_t \mathcal{M}(x, y_l)\big)\big)\big]. \end{aligned} \tag{6}$$

Compared with standard DPO, the adaptive margin requires a larger reward gap when the dispreferred response is highly confident and strongly conflicts with the estimated solvability. Thus, KCAPO assigns stronger alignment pressure to severe mismatch cases, while leaving ordinary reinforcement pairs close to standard DPO.

4 Experiments

To evaluate the efficacy of KCAPO in aligning the model's behavior with its internal knowledge boundaries, we specifically focus on short-form open-ended generation tasks which allow us to precisely attribute errors to *knowledge deficits* rather than generation artifacts, serving as a direct test for the model's factual grounding.

4.1 Setup

Datasets. We evaluate our method on two categories of benchmarks. First, for the **Trade-off Between Factuality and Honesty**, we employ **HalluLensQA** (Bang et al., 2025), which features dynamically generated questions to minimize data leakage; **TriviaQA** (Joshi et al., 2017), a standard reading comprehension benchmark; and **SimpleQA** (Wei et al., 2024), consisting of high-quality questions with indisputable answers. Second, to assess the **Rationality of Refusals**, we utilize **FalseQA** (Hu et al., 2023), **NEC** (Liu et al., 2023a), and **RefuNQ** (Kwiatkowski et al., 2019). These datasets serve as stress tests for discriminating solvable boundaries from unanswerable premises. Regarding data splits, our training data consists of **7,538** samples constructed via the HalluLens pipeline. For evaluation, we use a held-out set of **1,889** HalluLens samples and randomly sampled 1000 instances each from TriviaQA and SimpleQA to assess generalization ability. For refusal benchmarks, to ensure a balanced evaluation, we randomly sample 500 answerable and 500 unanswerable queries separately from each dataset. More detailed information is in the Appendix C.1.

Evaluation Metrics. Responses are classified into *Correct*, *Incorrect*, and *Refusal*. We report: (1) **Correct Answer Rate (Acc):** The overall utility, measured by the proportion of correct responses across all samples. (2) **Hallucination Rate (HR):** The reliability of generated content, defined as the proportion of incorrect answers among *attempted* responses (excluding refusals) (Bang et al., 2025). For refusal quality, we employ: (3) **RF_UN:** The refusal rate on unanswerable questions (Recall). (4) **F1 Score:** The harmonic mean of precision and recall for identifying unanswerable queries. (5) **Refusal Δ (RFΔ):** The difference between refusal rates on unanswerable versus answerable questions, quantifying the model's discriminative capability. Detailed explanation is in the Appendix D

Baselines. We compare our methods with the following approaches: (1) **Base:** Set the model temperature to 0 to obtain the model's most confident output; (2) **DoLa** (Chuang et al., 2023): A decoding strategy that amplifies factual knowledge by contrasting output distributions between mature intermediate and final layers; (3) **Self-Reflection (SR)** (Madaan et al., 2023): A prompting-based refinement approach which concatenates multiple sampled responses as context; (4) **Integrative Decoding (ID)** (Cheng et al., 2025): A method designed to unlock the potential of self-consistency by constructing a set of inputs prepended with previ-

Backbone	Method	HalluLensQA		TriviaQA		SimpleQA	
		HR ↓	Acc ↑	HR ↓	Acc ↑	HR ↓	Acc ↑
Llama-3.1-8B-Instruct	Base	0.472	0.128	0.2	0.709	0.873	0.106
	DoLa	0.664	0.154	0.227	**0.750**	0.889	0.093
	SR	0.813	0.139	0.258	0.719	0.912	0.047
	ID	0.668	0.126	0.23	0.72	0.89	0.1
	AH	0.575	0.146	0.247	0.706	0.873	0.09
	DPO	0.456	0.138	0.194	0.712	0.84	0.1
	KCAPO	**0.423**	**0.156**	**0.183**	0.724	**0.806**	**0.12**
Qwen2.5-7B-Instruct	Base	0.847	0.137	0.459	0.535	0.95	0.042
	DoLa	0.852	0.147	0.438	0.544	0.959	0.041
	SR	0.851	0.139	0.440	0.547	0.951	**0.047**
	ID	0.839	0.14	0.442	0.505	0.953	0.043
	AH	0.842	0.114	0.456	0.531	0.952	0.029
	DPO	0.831	0.139	0.45	0.54	0.96	0.039
	KCAPO	**0.822**	**0.147**	**0.435**	**0.554**	**0.947**	0.041
Mistral-7B-Instruct	Base	0.798	0.19	0.385	0.609	0.945	0.049
	DoLa	0.784	0.206	0.381	**0.621**	0.94	**0.058**
	SR	0.795	0.198	0.380	0.614	0.945	0.054
	ID	0.786	0.206	0.382	0.61	0.948	0.052
	AH	0.779	0.186	0.379	0.605	0.95	0.047
	DPO	0.79	0.193	0.382	0.61	0.938	0.048
	KCAPO	**0.773**	**0.21**	**0.375**	0.618	**0.935**	0.047

Table 1: Main results of experiment on the Trade-off Between Factuality and Honesty. Best results for each backbone and dataset are highlighted in **bold**.

ously sampled responses; (5) **Alignment for Honesty (AH)** (Yang et al., 2024): A fine-tuning framework that trains the model to proactively refuse questions when it lacks internal knowledge. (6) **DPO** (Rafailov et al., 2023): Using the same preference data as KCAPO, the model was directly fine-tuned using standard DPO. Detailed instructions on baseline selection are in the Appendix E.

Implementation Details. We validate KCAPO across three diverse instruction-tuned backbones: **Llama-3.1-8B-Instruct** (Dubey et al., 2024), **Qwen2.5-7B-Instruct** (Team et al., 2024), and **Mistral-7B-Instruct** (Chaplot, 2023). To derive solvability labels, we generate $K = 3$ semantic variants for each query using **GPT-5.5**, and use the same model as the correctness judge. A subsequent manual audit showed an estimated labeling error rate of only 3%. All experiments are repeated with three random seeds, and we report the mean and standard deviation. We set τ_{high}=0.6 and τ_{low}=0.4. Detailed information is in the Appendix F.

4.2 Main Results

Table 1 presents the main evaluation results. As observed, **KCAPO** demonstrates superior capability across all three backbones. Most notably, our method achieves the **lowest Hallucination Rate (HR)** and the **highest Accuracy (Acc)** in the majority of experimental settings. Especially its performance on Llama: KCAPO significantly reduces the Hallucination Rate from 0.472 (Base) to **0.423**, while simultaneously boosting accuracy from 0.128 to **0.156**. This dual improvement empirically validates the effectiveness of our alignment strategy. Although KCAPO does not achieve the highest accuracy in every setting, it exhibits a more favorable reliability–utility trade-off. Compared with the **Base model**, KCAPO consistently reduces the hallucination rate across all evaluated datasets and backbones, while improving or largely preserving accuracy in most cases. This suggests that KCAPO does not merely induce conservative refusals, but helps align the model's responses with its internal solvability. In contrast, inference-time strategies such as DoLa may improve accuracy in some settings, but they can also increase hallucination rates relative to the Base model, indicating a less stable factuality–honesty balance. Overall, KCAPO provides a more reliable alignment behavior by suppressing hallucinations while maintaining competitive utility. Notably, the performance gain of standard DPO over the base model verifies the efficacy of our epistemically grounded data construction. Furthermore, KCAPO's further superiority over DPO demonstrates the effectiveness of our alignment strategy. In conclusion, KCAPO effectively balances honesty and factuality.

4.3 Rationality of Refusals

As shown in Table 2, KCAPO demonstrates superior performance across all three benchmarks,

BackBone	Method	FalseQA			NEC			RefuNQ		
		RF_UN ↑	F1	RFΔ ↑	RF_UN ↑	F1 ↑	RFΔ ↑	RF_UN ↑	F1 ↑	RFΔ ↑
llama3.1-8b-instruct	Base	0.233	0.353	0.149	0.788	0.824	0.664	0.614	0.693	0.456
	DoLa	0.246	0.374	0.178	0.63	0.735	0.546	0.582	0.697	0.496
	SR	0.206	0.316	0.11	0.61	0.719	0.524	0.542	0.658	0.438
	ID	0.224	0.343	0.142	0.65	0.746	0.558	0.573	0.688	0.481
	AH	0.234	0.361	0.172	**0.808**	0.816	0.636	0.61	0.718	0.522
	KCAPO	**0.306**	**0.442**	**0.226**	0.8	**0.839**	**0.694**	**0.632**	**0.7336**	**0.541**
Qwen2.5-7B-Instruct	Base	0.34	0.487	0.286	0.746	0.818	0.67	0.592	0.708	0.514
	DoLa	0.234	0.373	0.214	0.402	0.561	0.372	0.22	0.352	0.192
	SR	0.172	0.285	0.138	0.446	0.601	0.408	0.332	0.488	0.304
	ID	0.323	0.478	0.297	0.432	0.583	0.384	0.32	0.473	0.288
	AH	0.264	0.399	0.206	0.628	0.733	0.544	0.372	0.523	0.324
	KCAPO	**0.358**	**0.516**	**0.33**	**0.774**	**0.847**	**0.722**	**0.598**	**0.730**	**0.559**
Mistral-7B-Instruct	Base	0.324	0.474	0.282	0.432	0.592	0.406	0.45	0.6	0.4
	DoLa	0.276	0.420	0.24	0.262	0.410	0.246	0.248	0.391	0.228
	SR	0.09	0.163	0.078	0.1	0.180	0.094	0.084	0.154	0.078
	ID	0.324	0.474	0.282	0.324	0.475	0.286	0.432	0.586	0.39
	AH	0.354	0.506	0.309	0.65	0.767	0.606	0.45	0.604	0.41
	KCAPO	**0.398**	**0.549**	**0.348**	**0.704**	**0.808**	**0.666**	**0.68**	**0.786**	**0.63**

Table 2: Evaluation of refusal rationality on FalseQA, NEC, and RefuNQ. Bold indicates the best performance. Our method consistently achieves the highest F1 and RFΔ, verifying its effectiveness in aligning internal knowledge boundaries.

verifying that our refusal behavior is both precise and context-aware. On the one hand, KCAPO almost achieves the highest refusal rates on unanswerable queries (RF_UN) across three benchmarks. This proves effective SUPPRESS of responses to nonsensical premises. On the other hand, high refusal rates do not come at the cost of utility. The consistent top-tier performance in **F1 Score** and **Refusal** Δ confirms that KCAPO successfully delineates the knowledge boundary—learning to actively refuse unknowns while retaining the capacity to answer solvable queries.

4.4 Reliability Analysis of the Solvability Probe

A model is more likely to respond correctly when it has the potential to solve the problem, and more likely to respond incorrectly when it lacks the potential. To rigorously validate that the trained probe $IS(x)$ serves as a reliable proxy for the model's internal knowledge state, we conducted an analysis on a held-out dataset. We utilized the HalluLensQA pipeline to synthesize **8,000** additional QA pairs (Wikipedia documents used to generate the question are strictly **excluded** during the probe training phase.). As visualized in Figure 2, in the low-solvability regime ($IS(x) < 0.5$), incorrect responses significantly predominate, reflecting that most questions in this range are difficult for the model to answer correctly. A critical **crossover point** is observed at $IS(x) \approx 0.5$, where the probability densities of correct and incorrect responses intersect. Beyond this threshold, the trend reverses: the density of correct responses consistently surpasses that of incorrect ones. This result empirically verifies that the learned internal solvability score $IS(x)$ is a robust indicator of the model's internal cognitive state, justifying its use in categorizing WAKE UP and SUPPRESS regimes for alignment.

4.5 Alignment Analysis of Internal State and External Confidence

To verify whether KCAPO achieves genuine cognitive calibration rather than superficial behavioral mimicking, we examine the alignment between the model's internal solvability and its expressed confidence. We introduce the **Self-Verification Score (SVS)**, calculated as the log-probability difference between "True" and "False" tokens in a self-evaluation prompt, to quantify subjective certainty. In order to effectively analyze the internal effects of alignment, we restrict our analysis to **Persistent Samples** where response correctness remains unchanged post-alignment. As visualized in Figure 3, KCAPO induces a profound realignment. For *Persistent Correct* samples, the solvability distribution shifts noticeably rightward, indicating that the model sharpens its internal representations for known facts. In stark contrast, for *Persistent Incorrect* samples, the SVS distribution collapses toward zero while solvability scores converge to the uncertainty region. This transition from "confidently wrong" to "ambivalently wrong" demonstrates that even when the model fails to explicitly refuse, KCAPO successfully erodes the internal confidence underlying hallucinations. Notably, the SVS for correct responses remains calibrated rather

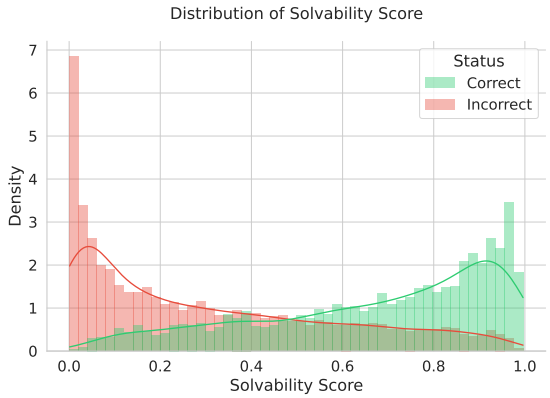

Figure 2: The statistical chart shows the number of correct and incorrect model responses within the solvability score interval.

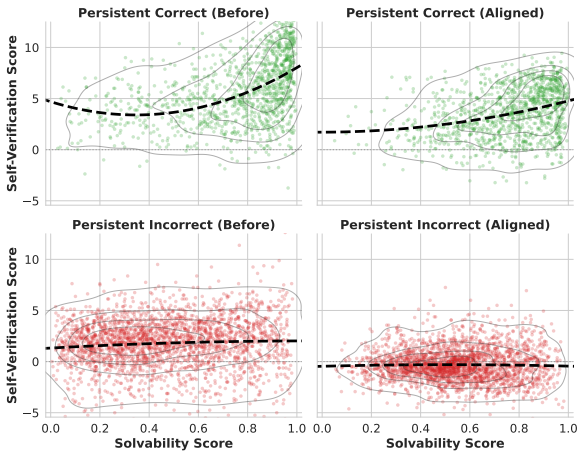

Figure 3: Distribution of Internal Solvability vs. Self-Verification Scores (SVS) for persistent samples.

than saturating at extreme values; we argue this reflects a positive mitigation of the model's general tendency toward overconfidence, ensuring a more honest assessment of its own capabilities.

4.6 Ablation Study

To verify the contribution of each component, we conducted ablation studies on Llama-3.1-8B-Instruct (Table 3). The **WAKE UP** and Reinforcement training data pairs are critical for maintaining model utility. Removing the WAKE UP training data pairs (**w/o WAKE UP**) leads to an *overly conservative* model: while hallucination drops, the **Refusal Rate (RR)** spikes to **0.776**, and Accuracy collapses to **0.124**—even lower than the Base model. This aligns with the results of excluding Reinforcement pairs (**w/o Reinf.**), which causes a distinct drop in Accuracy (0.156 → 0.143). These results confirm that without positive reinforcement of known facts, the model succumbs to indiscriminate refusal and fails to consolidate its internal confidence. Conversely, the **SUPPRESS** and Correction training data pairs are the primary drivers for safety. Removing the SUPPRESS pairs (**w/o SUPPRESS**) results in a severe regression in reliability, with the Hallucination Rate (HR) surging to **0.506** (worse than Base) and a diminished Refusal Δ (0.141). Similarly, removing Correction pairs (**w/o Corr.**) causes the HR to spike to 0.49. This demonstrates that simply reinforcing correct knowledge is insufficient; the model explicitly requires negative constraints to delineate the boundary of ignorance. Finally, we validate our algorithmic design and data strategy. Compared to standard DPO (**w/ DPO**), our adaptive margin strategy achieves superior performance across all metrics (HR: 0.423 vs. 0.456; RFΔ: 0.226 vs. 0.179), proving that weighting violations by epistemic severity yields a sharper knowledge boundary. Regarding data sources, we trained a variant on the TriviaQA dataset (**w/ TriviaQA**) and its performance did not show a significant change compared to using HalluLensQA. This demonstrates that our method has a certain degree of universality across different datasets.

Method	HalluLens QA			FalseQA			
	HR ↓	RR	Acc ↑	RF_AN ↓	RF_UN ↑	F1 ↑	RFΔ ↑
Base	0.472	0.755	0.128	0.084	0.233	0.354	0.149
w/o WAKE UP	0.434	0.776	0.124	0.096	0.268	0.393	0.172
w/o SUPPRESS	0.506	0.730	0.134	0.082	0.223	0.342	0.141
w/o Reinf.	0.452	0.740	0.143	0.083	0.282	0.413	0.199
w/o Corr.	0.490	0.762	0.116	0.092	0.243	0.364	0.151
w/ DPO	0.456	0.742	0.138	0.085	0.264	0.391	0.179
w/ TriviaQA	**0.420**	0.740	0.152	0.082	0.300	0.4	0.218
KCAPO	0.423	0.732	**0.156**	**0.080**	**0.306**	**0.442**	**0.226**

Table 3: Ablation study on Llama-3.1-8B-Instruct.

5 Conclusion

We reframed LLM honesty alignment as synchronizing internal solvability with external expression, moving beyond the coarse "known vs. unknown" dichotomy. Specifically, we propose a novel framework KCAPO, which estimates query-level solvability via hidden-state probing and constructs fine-grained preference pairs across four capability–behavior regimes, coupled with a mismatch-aware adaptive DPO margin. Experiments across three backbones and six benchmarks show that KCAPO reduces hallucination while preserving answer utility, outperforming both inference-time elicitation and standard alignment methods.

Limitations

While KCAPO demonstrates promising results, we acknowledge several limitations that point to directions for future research. Our current solvability estimation relies on a linear logistic regression probe using last-token hidden states. While effective for short-form QA tasks, this approach may be insufficient for capturing complex, multi-hop reasoning dependencies or long-context generation where the knowing" signal is distributed across multiple tokens. Future work could explore more sophisticated probing architectures, such as non-linear probes or attention-head analysis, to enhance robustness across complex domains. Our evaluation primarily relies on GPT-5.5 as a proxy judge for both training data construction and final assessment. While this is a standard practice in recent literature, it inevitably introduces potential biases and label noise inherent to the judge model. Furthermore, our experiments are currently focused on short-form knowledge-sensitive tasks. Extending KCAPO to open-ended creative generation or dialogue agents where "truth" is less well-defined remains an open challenge.

References

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Benjamin Mann, Nova DasSarma, and 1 others. 2021. A general language assistant as a laboratory for alignment. *CoRR*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: LLM hallucination benchmark. *arXiv preprint arXiv:2504.17550*.

Devendra Singh Chaplot. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can AI assistants know what they don't know? In *Proceedings of the International Conference on Machine Learning*, pages 8184–8202.

Yi Cheng, Xiao Liang, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, and 1 others. 2025. Integrative decoding: Improving factuality via implicit self-consistency. In *Proceedings of the 13th International Conference on Learning Representations*.

Steffi Chern, Zhulin Hu, Yuqing Yang, Ethan Chern, Yuan Guo, Jiahe Jin, Binjie Wang, and Pengfei Liu. 2024. Behonest: Benchmarking honesty in large language models. *arXiv preprint arXiv:2406.13261*.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. 2023. Dola: Decoding by contrasting layers improves factuality in large language models. In *The 12th International Conference on Learning Representations*.

Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16(1):64–93.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Zorik Gekhman, Eyal Ben David, Hadas Orgad, Eran Ofek, Yonatan Belinkov, Idan Szpektor, Jonathan Herzig, and Roi Reichart. 2025. Inside-out: Hidden factual knowledge in llms. *arXiv preprint arXiv:2503.15299*.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. 2024. Does fine-tuning llms on new knowledge encourage hallucinations? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7765–7784.

Arnav Gudibande, Eric Wallace, Charlie Snell, Xinyang Geng, Hao Liu, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. The false promise of imitating proprietary llms. *arXiv preprint arXiv:2305.15717*.

Shengding Hu, Yifan Luo, Huadong Wang, Xingyi Cheng, Zhiyuan Liu, and Maosong Sun. 2023. Won't get fooled again: Answering questions with false premises. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 5626–5643.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023a. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530.

Siheng Li, Cheng Yang, Taiqiang Wu, Chufan Shi, Yuji Zhang, Xinyu Zhu, Zesen Cheng, Deng Cai, Mo Yu, Lemao Liu, and 1 others. 2024. A survey on the honesty of large language models. *Transactions on Machine Learning Research*.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. In *Proceedings of the 4th ACM international conference on AI in finance*, pages 374–382.

Sheng-Chieh Lin, Luyu Gao, Barlas Oguz, Wenhan Xiong, Jimmy Lin, Wen-tau Yih, and Xilun Chen. 2024. Flame: Factuality-aware alignment for large language models. *Advances in Neural Information Processing Systems*, 37:115588–115614.

Genglin Liu, Xingyao Wang, Lifan Yuan, Yangyi Chen, and Hao Peng. 2023a. Examining llms' uncertainty expression towards questions outside parametric knowledge. *arXiv preprint arXiv:2311.09731*.

Yiheng Liu, Tianle Han, Siyuan Ma, Jiayue Zhang, Yuanyuan Yang, Jiaming Tian, Hao He, Antong Li, Mengshen He, Zhengliang Liu, and 1 others. 2023b. Summary of chatgpt-related research and perspective towards the future of large language models. *Meta-radiology*, 1(2):100017.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235.

Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. 2025. Towards fully exploiting llm internal states to enhance knowledge boundary perception. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 24315–24329.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

Hyung Gyu Rho. 2025. Margin adaptive dpo: Leveraging reward model for granular control in preference optimization. *arXiv preprint arXiv:2510.05342*.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, and 1 others. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Zeyu Shi, Ziming Wang, Tianyu Chen, Shiqi Gao, Haoyi Zhou, Qingyun Sun, and Jianxin Li. 2025. Fine-tuned llms know they don't know: A parameter-efficient approach to recovering honesty. *arXiv preprint arXiv:2511.12991*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

SM Towhidul Islam Tonmoy, SM Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A comprehensive survey of hallucination mitigation techniques in large language models. *CoRR*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *arXiv preprint arXiv:2411.04368*.

Junkang Wu, Xue Wang, Zhengyi Yang, Jiancan Wu, Jinyang Gao, Bolin Ding, Xiang Wang, and Xiangnan He. 2024. Alphadpo: Adaptive reward margin for direct preference optimization. In *Forty-second International Conference on Machine Learning*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training llms to refuse unknown questions using rl from knowledge feedback. *arXiv preprint arXiv:2403.18349*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2024. Alignment for honesty. *Advances in Neural Information Processing Systems*, 37:63565–63598.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. 2025. Mitigating hallucinations in large vision-language models via dpo: On-policy data hold the key. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10610–10620.

Yue Zhang, Leyang Cui, Shuming Shi, and 1 others. 2025. Alleviating hallucinations of large language models through induced hallucinations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8218–8232.

Kaitlyn Zhou, Dan Jurafsky, and Tatsunori B Hashimoto. 2023. Navigating the grey area: How expressions of uncertainty and overconfidence affect language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5506–5524.

A Sampling space construction

A.1 Variant generation details

We generate query variants in the following three ways: (1) *Syntactic Restructuring*: Modifying grammatical structures (e.g., active vs. passive voice) to reduce reliance on shallow retrieval patterns; (2) *Lexical Elucidation*: Replacing non-key entities or predicates with synonyms to ensure the model's robustness to surface lexical changes; (3) *Answer-preserving Distractor Perturbations*: Introducing subtle noise or counterfactual premises to test whether the model can maintain focus on core facts. These variants are designed to preserve the underlying answer while changing the surface form and contextual conditions of the query. The goal is to reduce the influence of accidental correctness from a single generation and to test whether the model's response remains stable around the same knowledge point. If the model consistently answers these variants correctly, we treat this as stronger behavioral evidence that the query lies within its internal solvability; if its responses fluctuate across variants, the knowledge point is considered unstable and is used to expose potential capability–behavior mismatch.

B Preference Data Construction

A detailed description of constructing various preference pairs in the sampling subspace of each knowledge point is provided in Algorithm 1. Notably, in the **SUPPRESSION** regime, we do not use any correct response to construct preference pairs. This design is consistent with the observation in Figure 2: responses in the low-solvability region are dominated by incorrect answers, indicating that occasional correct responses in this region are more likely to be unstable or noise-induced rather than reliable evidence of internal solvability.

C Datasets

C.1 Details of Evaluation Datasets

We evaluate KCAPO on two categories of benchmarks, corresponding to two complementary aspects of knowledge-boundary alignment: (1) the trade-off between factual answering and hallucination reduction, and (2) the rationality of refusal behavior.

The first category examines whether the model can answer factual questions accurately.

Algorithm 1 Preference Pair Construction for KCAPO

Require: Sampling sub-spaces $\{\mathcal{S}_i\}_{i=1}^{N}$; thresholds $\tau_{\text{low}}, \tau_{\text{high}}$; top number n
Ensure: Preference dataset $\mathcal{D}$
  $\mathcal{D} \leftarrow \emptyset$
  Define $\text{AllPairs}(\mathcal{Y}^+, \mathcal{Y}^-) = \{(x_l, y_w, y_l) \mid x_l \in x_i \cup X^i_{\text{var}},\ y_w \in \mathcal{Y}^+,\ y_l \in \mathcal{Y}^-,\ y_w \neq y_l\}$
  for each sampling sub-space $\mathcal{S}_i$ **do**
    $IS_i \leftarrow$ mean probe score over variants in $\mathcal{S}_i$
    Split responses into correct $\mathcal{Y}^c_i$, incorrect $\mathcal{Y}^w_i$, and refusal $\mathcal{Y}^r_i$
    if $IS_i > \tau_{\text{high}}$ **and** $\mathcal{Y}^c_i \neq \emptyset$ **then**
      $\mathcal{Y}^+_i \leftarrow \text{TopConf}_n(\mathcal{Y}^c_i)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \text{AllPairs}(\mathcal{Y}^+_i, \mathcal{Y}^c_i \setminus \mathcal{Y}^+_i)$ ▷ Wake-up reinforcement
      $\mathcal{D} \leftarrow \mathcal{D} \cup \text{AllPairs}(\mathcal{Y}^+_i, \mathcal{Y}^w_i \cup \mathcal{Y}^r_i)$ ▷ Wake-up correction
    else if $IS_i < \tau_{\text{low}}$ **and** $\mathcal{Y}^r_i \neq \emptyset$ **and** $\mathcal{Y}^c_i = \emptyset$ **then**
      $\mathcal{Y}^+_i \leftarrow \text{TopConf}_n(\mathcal{Y}^r_i)$
      $\mathcal{D} \leftarrow \mathcal{D} \cup \text{AllPairs}(\mathcal{Y}^+_i, \mathcal{Y}^r_i \setminus \mathcal{Y}^+_i)$ ▷ Suppression reinforcement
      $\mathcal{D} \leftarrow \mathcal{D} \cup \text{AllPairs}(\mathcal{Y}^+_i, \mathcal{Y}^w_i)$ ▷ Suppression correction
    end if
  end for
  return $\text{Shuffle}(\mathcal{D})$

HalluLensQA. HalluLensQA (Bang et al., 2025) is used as the primary benchmark for evaluating hallucination behavior under knowledge-sensitive question answering. Its dynamically generated questions help reduce potential data leakage and provide a challenging setting for testing whether a model can avoid unsupported answers. In our experiments, HalluLensQA is used both for constructing training data through the HalluLens pipeline and for held-out evaluation. This makes it suitable for assessing whether KCAPO can reduce hallucination while preserving answer utility.

TriviaQA. TriviaQA (Joshi et al., 2017) is a standard open-domain reading comprehension and factual question answering benchmark. Compared with HalluLensQA, TriviaQA provides a more established and widely used evaluation setting. We include it to test whether the alignment learned from our constructed training data generalizes to conventional factual QA tasks, rather than only improving performance on the data generation pipeline used for training.

SimpleQA. SimpleQA (Wei et al., 2024) consists of short, challenging factual questions with clear and verifiable answers. Because its questions are designed to have unambiguous answers, it is particularly suitable for measuring factual correctness in short-form generation. We use SimpleQA to

evaluate whether KCAPO improves factual reliability without relying on long-form reasoning or ambiguous answer criteria.

The second category examines whether the model can distinguish answerable queries from unanswerable or false-premise ones and provide reasonable rejections.

NEC. NEC (Liu et al., 2023a) evaluates whether a model can abstain from answering questions involving non-existent concepts. The intuition is that certain fabricated concepts, such as fictional names of animals, countries, foods, medicines, sports terms, or random strings, are unlikely to be present in the model's pretraining data. Therefore, questions built around such concepts should be unanswerable, and directly answering them indicates that the model is hallucinating unsupported knowledge. The original NEC benchmark contains 2,078 unanswerable questions based on non-existent concepts, together with a control group of 2,072 answerable questions constructed from the same templates but with real concepts. We use NEC to test whether KCAPO can help the model recognize queries outside its knowledge boundary and abstain instead of fabricating answers.

FalseQA. FalseQA (Hu et al., 2023) consists of questions with false premises. Unlike NEC, where the key concept itself is fabricated, FalseQA often involves known entities connected by a false relation. For example, a question such as "Why does drinking water cause cancer?" assumes a false causal relation between two known concepts. A reliable model should identify the problematic premise rather than directly answer the question. FalseQA contains 4,730 questions in total, half of which contain false premises, while the other half are answerable control questions sampled from the same templates without false assumptions. We use FalseQA to evaluate whether KCAPO improves the model's ability to challenge misleading premises and avoid answering questions that should be rejected.

RefuNQ. RefuNQ is built from Natural Questions (Kwiatkowski et al., 2019) to provide a more naturalistic testbed for refusal behavior. Its answerable partition is directly derived from NQ questions with short human-annotated answers, enabling accuracy evaluation through lexical matching. Its unanswerable partition is constructed by replacing a randomly selected noun in each original question with a generic non-existent concept from NEC, producing minimally perturbed questions that are no longer answerable. After preprocessing, RefuNQ contains 2,266 answerable questions and 2,173 unanswerable perturbed questions. We use RefuNQ to evaluate the trade-off between helpfulness and abstention: a model should answer the original NQ-style questions while refusing their unanswerable perturbations.

C.2 Why We Focus on Short-Form Factual QA

Our evaluation focuses on short-form factual QA, where each query typically targets an atomic fact and expects a concise answer, such as an entity name, date, location, or short phrase. This setting is widely used in prior work on factuality, calibration, and honesty in LLMs (Kadavath et al., 2022; Yang et al., 2024; Wei et al., 2024; Bang et al., 2025). More importantly, it directly matches the goal of KCAPO: aligning a model's response behavior with its internal solvability on fact-level knowledge.

Compared with long-form generation, short-form QA provides a cleaner testbed for knowledge-boundary alignment. In long-form responses, correctness is often entangled with partial support, missing details, discourse planning, reasoning chains, or stylistic preferences. As a result, a failure may not clearly indicate whether the model lacks the relevant knowledge, fails to express it, or simply generates an incomplete explanation. By contrast, short-form factual QA reduces these confounding factors: if the target answer is unambiguous, the model's behavior can be more directly categorized as correct, incorrect, or abstained. This makes it better suited for estimating query-level solvability and constructing capability–behavior preference pairs.

This setting also makes our semantic-variant construction and automatic correctness judgment more reliable. Since the answers are usually concrete and precise, meaning-preserving rewrites of the question are less likely to change the expected answer. Similarly, judging whether a response contains the correct short answer is substantially simpler than evaluating open-ended long-form factuality. Therefore, the short-form setting allows us to use semantic variants as a consistency test of the same knowledge point, while keeping the labeling process relatively stable.

Our empirical results further support this design

choice. As shown in Figure 2, the estimated solvability score correlates clearly with response correctness: low-solvability regions are dominated by incorrect responses, while high-solvability regions contain more correct responses. This indicates that short-form QA provides a suitable setting for learning and validating the internal solvability signal. Moreover, the main results on HalluLensQA, TriviaQA, and SimpleQA show that KCAPO reduces hallucination while preserving answer utility, and the refusal benchmarks further show improved discrimination between answerable and unanswerable queries. Together, these results suggest that short-form factual QA is an appropriate and controlled evaluation setting for studying whether a model should answer or abstain based on its internal solvability.

We acknowledge that extending KCAPO to long-form generation is an important future direction. However, long-form factuality introduces additional challenges in response decomposition, attribution, and partial correctness evaluation. We therefore treat short-form factual QA as a necessary first step for isolating knowledge-boundary alignment at the atomic-fact level before extending the framework to more complex generation scenarios.

D Metric Definitions

We classify each model response into one of three labels: *Correct*, *Incorrect*, and *Refusal*. For a dataset $\mathcal{T}$ with N samples, let $l_i \in \{\mathrm{C}, \mathrm{I}, \mathrm{R}\}$ denote the label of the response to the i-th query, where C, I, and R correspond to *Correct*, *Incorrect*, and *Refusal*, respectively.

Correct Answer Rate. Correct Answer Rate measures the overall utility of the model:

$$\mathrm{Acc} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(l_i = \mathrm{C}). \tag{7}$$

This metric counts correct answers over all samples, including those where the model refuses.

Hallucination Rate. Hallucination Rate measures the error rate among attempted answers, excluding refusals:

$$\mathrm{HR} = \frac{\sum_{i=1}^{N} \mathbb{I}(l_i = \mathrm{I})}{\sum_{i=1}^{N} \mathbb{I}(l_i \in \{\mathrm{C}, \mathrm{I}\})}. \tag{8}$$

Thus, HR reflects the factual reliability of generated answers rather than the model's tendency to abstain.

Refusal Rate on Unanswerable Questions. For refusal benchmarks, each dataset is divided into answerable and unanswerable subsets, denoted as $\mathcal{A}$ and $\mathcal{U}$, respectively. The refusal rate on unanswerable questions is:

$$\mathrm{RF_UN} = \frac{\sum_{i \in \mathcal{U}} \mathbb{I}(l_i = \mathrm{R})}{|\mathcal{U}|}. \tag{9}$$

This is equivalent to the recall of detecting unanswerable queries, where refusal is treated as the model's prediction that the query should not be answered.

Refusal Rate on Answerable Questions. We also compute the refusal rate on answerable questions:

$$\mathrm{RF_AN} = \frac{\sum_{i \in \mathcal{A}} \mathbb{I}(l_i = \mathrm{R})}{|\mathcal{A}|}. \tag{10}$$

A lower $\mathrm{RF_AN}$ indicates that the model is less likely to unnecessarily refuse answerable queries.

F1 Score for Unanswerable Detection. We treat unanswerable-query detection as a binary classification problem, where refusal is the positive prediction. The precision and recall are computed as:

$$\mathrm{Precision} = \frac{\sum_{i \in \mathcal{U}} \mathbb{I}(l_i = \mathrm{R})}{\sum_{i \in \mathcal{A} \cup \mathcal{U}} \mathbb{I}(l_i = \mathrm{R})}, \tag{11}$$

$$\mathrm{Recall} = \mathrm{RF_UN}. \tag{12}$$

The F1 score is then:

$$\mathrm{F1} = \frac{2 \cdot \mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{13}$$

Refusal Discrimination Gap. Finally, we report the refusal discrimination gap:

$$\mathrm{RF}\Delta = \mathrm{RF_UN} - \mathrm{RF_AN}. \tag{14}$$

A larger $\mathrm{RF}\Delta$ indicates better selective abstention: the model refuses more often on unanswerable queries while avoiding unnecessary refusals on answerable ones.

E Rationale for Baseline Selection

We select baselines to cover the main lines of prior work related to hallucination mitigation, knowledge elicitation, and answer/refusal alignment. These baselines allow us to evaluate whether KCAPO improves over methods that only refine model outputs at inference time, methods that explicitly introduce refusal behavior, and standard preference optimization without our adaptive objective.

Base Model. The base model is evaluated with greedy decoding by setting the temperature to 0. This baseline reflects the model's original answer/refusal behavior without additional inference-time intervention or fine-tuning. It serves as the reference point for measuring whether an alignment method improves factuality, utility, and refusal behavior.

Inference-time Knowledge Elicitation Methods. We include DoLa, Self-Reflection (SR), and Integrative Decoding (ID) as representative inference-time baselines because they share a common motivation: when a model already encodes useful knowledge, can we better elicit, refine, or aggregate this knowledge at response time to reduce hallucination? This line of work is closely related to our setting, since it also aims to recover factual answers that may be latent in the model. However, its primary focus is on improving the expression of existing knowledge, rather than deciding whether a query is internally solvable or whether the model should answer or abstain.

DoLa (Chuang et al., 2023) represents decoding-time knowledge elicitation. It is motivated by the observation that factual knowledge may be more clearly reflected in certain intermediate layers than in the final output distribution, and therefore contrasts layer-wise predictions to encourage more factual generations. We include DoLa to test whether better exposing encoded factual knowledge during decoding is sufficient to improve the factuality–abstention trade-off.

Self-Reflection (SR) (Madaan et al., 2023) represents response-time refinement. It is motivated by the idea that a model may be able to detect and revise its own flawed outputs when given additional opportunities to reflect on generated responses. We include SR to examine whether hallucination can be reduced by post-generation self-correction, without explicitly modeling whether the original query is within the model's internal capability.

Integrative Decoding (ID) (Cheng et al., 2025) represents consistency-based response integration. It is motivated by the view that different sampled responses may expose complementary pieces of the model's latent knowledge, and that integrating these responses can improve factual reliability. We include ID because it is particularly close to the idea of exploiting response variability, but its goal is still to improve answer generation rather than to align answer/refusal behavior with an explicit solvability signal.

Together, these baselines test a key distinction in motivation: inference-time methods ask how to better express knowledge that may already be present, whereas KCAPO asks a finer-grained question—whether a failure reflects missing knowledge or under-expressed capability, and how the model's answer/refusal behavior should be aligned accordingly.

Knowledge-boundary Alignment Baseline. We include Alignment for Honesty (AH) (Yang et al., 2024) because it represents a further step beyond inference-time knowledge elicitation. While DoLa, SR, and ID mainly ask how to better surface knowledge already encoded in the model, AH is motivated by a different question: how can a model recognize when it should refuse rather than answer? It therefore introduces the notion of a knowledge boundary and trains models to abstain when they lack internal knowledge.

This baseline is particularly relevant to KCAPO, since both methods aim to improve answer/refusal behavior rather than only refine generated answers. However, AH primarily frames the problem as distinguishing known from unknown queries. KCAPO further refines this motivation by asking whether a failure reflects missing knowledge or under-expressed capability, and aligns internal solvability with external confidence and response behavior through fine-grained preference optimization.

Standard Preference Optimization. We include standard DPO (Rafailov et al., 2023) using the same preference data as KCAPO. This comparison isolates the contribution of our mismatch-aware adaptive margin. If KCAPO outperforms DPO under the same preference pairs, the improvement cannot be attributed solely to the constructed data, but also to dynamically adjusting the alignment strength according to solvability–confidence mismatch.

F Implementation Details and Hyperparameter Settings

F.1 Backbone Models and Hardware

We evaluate KCAPO on three instruction-tuned backbones: **Llama-3.1-8B-Instruct** (Dubey et al., 2024), **Qwen2.5-7B-Instruct** (Team et al., 2024), and **Mistral-7B-Instruct** (Chaplot, 2023). All experiments are conducted on two NVIDIA RTX 4090 GPUs. To reduce memory cost, we fine-tune

the models using Low-Rank Adaptation (LoRA) and enable bfloat16 training.

For solvability estimation, we generate $K = 3$ semantic variants for each query and use a stronger LLM judge to evaluate response correctness. The probe thresholds are set to $\tau_{\text{high}} = 0.6$ and $\tau_{\text{low}} = 0.4$, which define the high- and low-solvability regions used for preference construction. All experiments are repeated with three random seeds, and we report the mean and standard deviation.

F.2 Training Hyperparameters

Table 4 summarizes the training hyperparameters used in our experiments. We set $\tau_{\text{high}} = 0.6$ and $\tau_{\text{low}} = 0.4$ to define high- and low-solvability regions. Since the solvability probe produces a probability-like score, 0.5 serves as the natural decision boundary. Rather than splitting at 0.5, we leave a buffer region around the boundary and only assign regime labels to queries with more confident solvability estimates. This reduces noisy preference construction for ambiguous samples and ensures that WAKE UP and SUPPRESS pairs are built from clearer capability states.

We set the adaptive margin weight to $\lambda = 0.3$. This value allows the solvability–confidence mismatch to directly affect the DPO objective while preserving the original preference signal. If λ is too small, KCAPO degenerates toward standard DPO; if it is too large, the optimization may overemphasize confidence penalties.

Hyperparameter	Value
Effective batch size	64
Learning rate	5×10^{-6}
Training epochs	3
DPO temperature β	0.1
KCAPO margin weight λ	0.3
λ warmup ratio	0.1
LoRA rank r	16
LoRA alpha	32
LoRA dropout	0.05
LoRA target modules	q, k, v, o projections; gate/up/down projections
Precision	bf16

Table 4: Default training hyperparameters for KCAPO. The effective batch size is computed as per-device batch size $\times$ gradient accumulation steps $\times$ number of GPUs.