Knowing When You Don't Know: Separating Access Failures from Knowledge Gaps for Closed-Book Hallucination Mitigation

Anonymous ACL submission

Abstract

Small language models in closed-book QA hallucinate confidently because they cannot distinguish facts they know from forgotten facts. We frame this failure as two distinct causes: access failures, where a fact is stored in parameters but not surfaced under standard decoding, and knowledge gaps, where the fact is simply absent. We propose a pipeline targeting each cause directly. For access failures, Chain-of-Thought distillation from a 235B teacher under an "internalized knowledge" constraint recovers a 5.4-point EM gain on HotPotQA. For knowledge gaps, we label hallucinated training examples as "I don't know" and fine-tune under a class-balanced abstention objective, then refine with Reinforcement Learning from Verifier Feedback (RLVF). RLVF answers +5 percentage points more questions than the supervised abstention model while raising abstention precision to 97.84 percent; the full pipeline cuts confident hallucinations by up to 59 percent relative to the base model. As a byproduct, fine-tuning rather than estimator choice is what makes unsupervised uncertainty estimation work: RAUQ and Semantic Entropy reach AUROC up to 0.86 on fine-tuned checkpoints versus chance-level 0.43-0.51 on the base model, a +0.35 absolute gain over the P(true) self-evaluation baseline of Kadavath et al. (2022).

1 Introduction

Large Language Models (LLMs) frequently suffer from *hallucination* in closed-book settings, generating plausible but factually incorrect assertions (Ji et al., 2023). This issue is particularly acute in Small Language Models (SLMs), which often lack the capacity to distinguish between what they know and what they have forgotten. This unreliability creates a critical barrier for deployment in safety-critical or privacy-constrained environments where Retrieval-Augmented Generation (RAG) is not feasible.

We frame closed-book hallucination as arising from two distinct failure modes that require separate interventions. The first is an *access failure*: the relevant fact is stored in the model's parameters, but the model cannot reliably surface it under standard decoding (Kadavath et al., 2022). The second is a *knowledge gap*: the fact is simply not in the model's parameters, and any direct answer will be a guess. Treating these symmetrically (e.g., forcing the model to always answer) encourages confident hallucination on the second case; teaching it to always abstain wastes capability on the first.

Our pipeline targets each failure mode separately. To recover from access failures, we apply **Chain-of-Thought (CoT) Distillation** from a much larger teacher to unlock latent parametric knowledge through step-by-step recall. To handle genuine knowledge gaps, we apply **Abstention-Aware Fine-Tuning** followed by **Reinforcement Learning from Verifier Feedback (RLVF)**, which teaches the model to refuse rather than fabricate. We instantiate the pipeline on Qwen2.5-7B-Instruct.

Our experiments on HotPotQA validate this two-pronged approach. CoT distillation improves Exact Match accuracy from 18.86% to 24.31%, significantly outperforming standard SFT. Abstention-Aware Fine-Tuning achieves 97.5% precision in detecting knowledge boundaries and cuts the rate of confident errors by more than half relative to the baseline. RLVF then provides a verifier-grounded mechanism for tuning the risk–coverage trade-off: it raises the answer rate by 5 percentage points while *improving* abstention precision to 97.84%, the highest of any model in our evaluation. Together the supervised and RL stages cover two complementary deployment regimes: a conservative supervised model for safety-critical use and a coverage-oriented RL-tuned model for general assistance, both substantially safer than the baseline.

As a byproduct, the same fine-tuning that produces native abstention behavior also turns the existing unsupervised uncertainty estimators of Vazhentsev et al. (2025) and Farquhar et al. (2024) from random (AUROC $= 0.43$) to predictive (up to 0.86); on the same inputs, the P(true) self-evaluation baseline of Kadavath et al. (2022) stays at chance.

Contributions. We make four contributions:

1. A **two-cause framing** of closed-book hallucination (access failure vs. knowledge gap) and a pipeline whose components map onto these causes one-to-one.

2. An **internalized-knowledge CoT distillation** recipe in which a 235B teacher generates rationales constrained to read as direct recall, with a validator that rejects citation-style traces. This recovers a 5.4-point absolute EM gain on HotPotQA under closed-book inference.

3. An **abstention-aware training pipeline** comprising (i) a knowledge-boundary mapping procedure that converts hallucinated training examples into "I don't know" supervision under an 80/20 class balance, and (ii) an RLVF stage whose reward combines a DeBERTa NLI verifier (open-book at training, closed-book at inference) with a token-level entropy confidence term. Together these cut confident hallucinations by up to 59% relative to the base model.

4. An **evaluation in standard selective-prediction terms** (AUROC, AURC, Expected Calibration Error, Risk@Coverage). We show that the standard P(true) self-evaluation baseline of Kadavath et al. (2022) is at chance (AUROC ≈ 0.51) on closed-book HotPotQA for both the base and fine-tuned model, while RAUQ (Vazhentsev et al., 2025) and Semantic Entropy (Farquhar et al., 2024), when paired with our fine-tuned checkpoints, reach AUROC up to 0.86, a $+0.35$ absolute gain over P(true). The same uncertainty estimators are anti-informative (AUROC $= 0.43$) on the un-fine-tuned base model, so the gain is attributable to the fine-tuning, not to the choice of estimator. Unsupervised Semantic Entropy is additionally an effective post-hoc filter, raising selective F1 from 0.50 to 0.70 at 50% rejection.

2 Related Work

Hallucination in LLMs: Hallucination refers to the generation of text that is nonsensical or unfaithful to the provided source or established world knowledge (Ji et al., 2023). While larger models exhibit emergent self-calibration (Kadavath et al., 2022; Srivastava et al., 2023), Small Language Models (SLMs) are particularly prone to confident errors in closed-book settings, necessitating explicit intervention strategies. When relying on parametric memory, recent literature increasingly divides these errors into two distinct failure modes: *knowledge gaps*, where the model lacks the parametric facts, and *access failures*, where the model possesses the knowledge but fails to surface it under standard decoding (Meng et al., 2022; Kadavath et al., 2022). While prior work has largely isolated these failure modes analytically or focused on post-hoc detection (Azaria and Mitchell, 2023; Vazhentsev et al., 2025; Farquhar et al., 2024), our work operationalizes this dichotomy into an end-to-end training pipeline, applying targeted interventions for both causes simultaneously.

Reasoning Distillation: Chain-of-Thought (CoT) prompting significantly enhances LLM reasoning but typically requires large model scales (>100B parameters) to be effective. To transfer this capability to smaller models, Li et al. (2023) and Ho et al. (2023) proposed CoT distillation, where a student model is fine-tuned on reasoning traces generated by a large teacher. Our work extends this by imposing an *internalized-knowledge* constraint on the teacher, ensuring the distilled reasoning relies on facts the student can plausibly retrieve from its own weights rather than external context.

Abstention and Uncertainty: Teaching models to say "I don't know" is a classic problem in reliable QA (Rajpurkar et al., 2018). Recent work splits into two families. *Refusal-aware training* fine-tunes models to abstain on questions outside their knowledge boundary; R-Tuning (Zhang et al., 2024) and the self-evaluation framework of Kadavath et al. (2022) are the closest prior abstention baselines, the latter giving the P(true) signal we compare against. *Unsupervised uncertainty estimation* instead derives a confidence score post-hoc from generation statistics; recent approaches include Semantic Entropy over generation clusters (Farquhar et al., 2024), attention-pattern-based methods such as RAUQ (Vazhentsev et al., 2025), and verbalized uncertainty (Lin et al., 2022). We

combine both: knowledge-boundary-aware supervised fine-tuning followed by NLI-grounded RL (RLVF), and we then evaluate the resulting models with RAUQ, Semantic Entropy, and P(true) head-to-head.

3 Approach

Our pipeline is a sequence of five models. We use Qwen2.5-7B-Instruct (Qwen Team, 2025) throughout as the policy backbone; each fine-tuned variant is a LoRA / QLoRA (Hu et al., 2022) adapter trained on top of it. Two of the five are baselines (out-of-the-box and vanilla SFT), and three are our methods (CoT distillation, abstention SFT, and verifier-based RL).

- **Qwen2.5-Instruct** (baseline): the base model used out of the box.
- **Qwen2.5-SFT** (baseline): LoRA fine-tuned on 10,000 question–answer pairs from HotPotQA.
- **Qwen2.5-FCoT** (*ours*): fine-tuned on Dataset B (Section 3.1) to produce an internalized-knowledge rationale before answering.
- **Qwen2.5-Abstain** (*ours*): QLoRA fine-tuned on the answer / abstention pairs in Dataset C (Section 3.2).
- **Qwen2.5-RLVF** (*ours*): Qwen2.5-Abstain further trained with policy-gradient RL against a verifier-based reward (Section 3.3).

3.1 Finetune-CoT

To enable multi-step reasoning in the student model, we adopt the Fine-tune-CoT framework (Ho et al., 2023), which transfers reasoning ability from a large teacher to a smaller student by fine-tuning the student on rationales produced by the teacher. The standard pipeline has three stages: (1) prompting the teacher to generate step-by-step rationales, (2) curating the data by filtering samples whose final answer disagrees with the ground truth, and (3) fine-tuning the student on the validated traces.

We adapt the generation and curation stages to produce what we call an *internalized-knowledge* dataset (Dataset B). By "internalized knowledge" we mean reasoning that is phrased as direct recall from the model's own parameters rather than as citation of an external passage, e.g., "Arthur's Magazine was first published in 1844" rather than "the passage states that Arthur's Magazine was first published in 1844." This distinction matters because the student is deployed closed-book, so any habit of referring to "the passage" at training time produces incoherent outputs at inference time.

To produce this dataset, we use Qwen3-235B (Yang et al., 2025) as the teacher and provide it with the HotPotQA supporting paragraphs as context to suppress its own hallucinations. We then prompt it to frame its rationale as a step-by-step recall of facts and explicitly forbid citation markers such as "according to the passage." A validation loop rejects samples whose final answer is wrong (as in standard Fine-tune-CoT) and additionally rejects samples containing forbidden citation phrases, enforcing the internalized-knowledge style.

3.2 Abstention-Aware Fine-Tuning

To handle questions for which the model genuinely lacks the relevant knowledge, we fine-tune the model to output an explicit refusal ("I don't know") rather than guess. Forcing the model to answer every query encourages hallucination on out-of-distribution facts; abstention-aware fine-tuning instead calibrates the model's behavior to its actual competence.

We first aimed to map the *knowledge boundary* of Qwen2.5: the partition of HotPotQA training examples into those the model can answer from its own parameters and those on which it hallucinates. We hypothesize that if Qwen2.5-FCoT generates an incorrect answer for a training example x under closed-book inference, that example lies outside the model's knowledge boundary and is hallucination-prone. We performed closed-book inference on the first $N = 40{,}000$ examples of the HotPotQA training set. For every example, we compared the model's predicted answer $\hat{y}$ against the ground truth y using Exact Match (EM) scoring. This effectively partitioned our training data into two sets:

- **Correct set** (S_{correct}): samples where the model's parametric knowledge sufficed ($\text{EM}(\hat{y}, y) = 1$).
- **Hallucination set** ($S_{\text{hallucination}}$): samples where the model produced a confident but incorrect answer ($\text{EM}(\hat{y}, y) = 0$).

Using this partition, we constructed a supervised fine-tuning dataset (Dataset C) that reinforces known knowledge while penalizing hallucinations. For examples in $S_{\text{hallucination}}$ we replaced

the ground-truth answer with "I don't know," and for examples in S_{correct} we retained the original answer. A naive version of this objective is degenerate: if every hard question is relabeled to "I don't know," the model can drive its loss to zero by always refusing, a failure mode we refer to as the *laziness collapse*. To prevent this, we enforced an 80/20 class balance by randomly sampling 8,000 examples from S_{correct} and 2,000 from $S_{\text{hallucination}}$, yielding a final 10,000-example dataset. We then fine-tuned Qwen2.5-Instruct with QLoRA on this mixture.

3.3 Reinforcement Learning from Verifier Feedback

To refine the abstention boundary learned during supervised fine-tuning, we further train Qwen2.5-Abstain with policy-gradient RL against a reward derived from a verifier and from the model's own confidence. Although the policy is queried closed-book at training time, the verifier itself is *not*: it is given the HotPotQA supporting paragraphs, so the reward measures semantic alignment with grounded evidence rather than exact-match against a single gold string.

FactualityVerifier. We use a DeBERTa-based NLI model (He et al., 2021, 2023) that takes a premise and a hypothesis and returns class probabilities over $\{\text{entailment}, \text{neutral}, \text{contradiction}\}$. For each training example we concatenate the supporting paragraph titles and sentences from HotPotQA (Yang et al., 2018) into a single premise, treat the model's generated answer as the hypothesis, and read off the entailment and contradiction probabilities p_{ent} and p_{cont}. We define a grounded factuality score

$$f = p_{\text{ent}} - p_{\text{cont}} \in [-1, 1],$$

so that $f > 0$ corresponds to answers that the supporting paragraphs entail and $f < 0$ to answers they contradict. Thus RLVF optimizes semantic factuality against the verifier rather than string-level agreement with a gold label.

Confidence Term. We reward the policy for being correct *confidently*, with confidence defined as the complement of mean per-token entropy normalized by the maximum possible entropy:

$$\text{conf} = 1 - \frac{H_{\text{avg}}}{H_{\max}} \in [0, 1],$$

where $H_{\text{avg}} = \frac{1}{T}\sum_t H(p_t)$ is the mean Shannon entropy of the policy's next-token distribution p_t over the T generated tokens, and $H_{\max} = \log|V|$. We use this token-level form (rather than the semantic-entropy form of Farquhar et al. (2024), which inspired it) because it can be computed cheaply on-policy during RL rollouts. As the reward function below specifies, confidence amplifies reward only when factuality is positive, preventing the policy from being rewarded for confidently wrong answers.

Abstention Detection. We do not detect abstentions by string matching, which is brittle to paraphrases ("I'm not sure," "I don't have that information," etc.). Instead, our *AbstentionClassifier* reuses the NLI verifier: it treats the generated answer as a premise and a canonical abstention statement ("The respondent does not know the answer") as the hypothesis, and flags the answer as an abstention if the entailment probability exceeds a fixed threshold. This is analogous in spirit to prior work on verbalized uncertainty (Lin et al., 2022). Detected abstentions receive a fixed reward of -1.0, making refusal mildly costly but strictly cheaper than a contradiction.

Reward Function. For non-abstaining outputs, the reward depends only on verifier factuality and confidence:

$$R = \begin{cases} 10\, f(\lambda_{\text{base}} + \lambda_{\text{conf}} \cdot \text{conf}) & f \geq 0, \\ 5\, f(\lambda_{\text{base}} + \lambda_{\text{conf}} \cdot \text{conf}) & f < 0. \end{cases}$$

λ_{base} and λ_{conf} are fixed scalars controlling the reward floor and the maximum confidence bonus. The asymmetric $10\times/5\times$ scaling rewards grounded answers more strongly than it punishes contradictions of equal magnitude, which biases the policy toward attempting answers and counterbalances the strong incentive (from the conservative SFT initialization) to default to abstention. The choice to ground the reward in the NLI verifier rather than Exact Match avoids penalizing valid paraphrases of the gold answer (Ji et al., 2023), and the warm-start from Qwen2.5-Abstain (rather than from the base) gives the RL stage a pre-existing abstention boundary to refine, which we found necessary for stable early optimization (Ho et al., 2023).

3.4 Discrete Semantic Entropy

We use Discrete Semantic Entropy (SE) (Farquhar et al., 2024) as a post-hoc, training-free signal for residual hallucinations after the abstention pipeline. We adopt the discrete (black-box) form rather than

RAUQ (Vazhentsev et al., 2025) because the inference API for our larger model does not expose internal attention weights. For each question x we draw $M = 10$ stochastic generations and cluster them by bidirectional NLI entailment (`microsoft/deberta-large-mnli`). The probability of each semantic cluster C is approximated by Monte Carlo:

$$P(C \mid x) \approx \frac{1}{M} \sum_{i=1}^{M} \mathbb{I}(s_i \in C), \qquad (1)$$

and the total uncertainty is the entropy over clusters, $H(x) = -\sum_C P(C \mid x) \log P(C \mid x)$. We use $H(x)$ as a per-question uncertainty score, sort the validation set in descending order, and rank-reject the top-k most uncertain samples to produce rejection curves (Section 4.6).

4 Experiments

4.1 Data

We use the training and validation splits of the HotPotQA (Yang et al., 2018) distractor dataset (English only). The distractor setting has question-answer pairs with ten paragraphs as context, two of which are relevant to answering the question. Since we ran our tests in a closed-book setting, we discarded all ten context paragraphs for our finetuning, RL, and evaluation methods. We define 3 training datasets:

Dataset A1/A2: The first and second 10,000 question-answer pairs of HotPotQA (distractor setting, train split), respectively.

Dataset B: The first 10,000 HotPotQA distractor/train examples augmented with teacher-generated rationales under our internalized-knowledge constraint (Section 3.1). A representative entry:

- **Q:** Which magazine was started first, Arthur's Magazine or First for Women?
- **R:** *Step 1*: The question asks ... *Step 2*: Arthur's Magazine was published starting in 1844 . . . *Step 3*: Since 1844 is earlier than the 1980s, the answer is Arthur's Magazine.[1]
- **A:** Arthur's Magazine.

Dataset C: 10,000 question–answer or question–abstention pairs sampled from the first 40,000 HotPotQA distractor/train examples (Section 3.2). A representative abstention entry:

- **Q:** Who wrote the 1971 drama which stars the winner of the 1997 BAFTA Fellowship?
- **A:** I don't know.

[1] Reasoning trace abbreviated for space.

4.2 Evaluation method

We measure answering quality with Exact Match (EM) and token-level F1 between the gold answer and the model's prediction. For abstention models, we report two additional quantities: the *abstention rate*, defined as the fraction of validation questions for which the model output an "I don't know"-style response (as judged by the AbstentionClassifier), and the *selective EM/F1*, defined as EM/F1 computed only on the answered subset.

To assess whether high-uncertainty samples are also high-error samples, we additionally compute two unsupervised uncertainty signals on each generation: RAUQ (Vazhentsev et al., 2025), which aggregates attention-weight patterns across heads to score how anomalous a generation is, and Semantic Entropy (SE) (Farquhar et al., 2024), defined below in Section 3.4. We sort model responses by each uncertainty score and report *rejection curves*: EM/F1 on the retained subset as a function of the fraction of most-uncertain samples discarded. These curves serve as a non-parametric upper bound on what a perfectly calibrated abstention policy could achieve.

4.3 Experimental details

Models and infrastructure. All rows in the main results table use **Qwen2.5-7B-Instruct** (bfloat16) as the policy backbone; SFT, FCoT, Abstain, and RLVF are LoRA / QLoRA adapters (4-bit NF4 base, FP16 adapters) on top of it. A separate scaled re-implementation of CoT distillation and abstention SFT (with the full 80K HotPotQA training set rather than the 10K subset) was performed on **Qwen3-8B** (Yang et al., 2025) on a higher-throughput platform that did not host a Qwen2.5 checkpoint. We verified on the closed-book validation split that Qwen3-8B and Qwen2.5-7B agree on EM/F1 to within ± 1 point on the same training data, but the Qwen3-8B re-runs are not reported in the main table to keep the comparison consistent across rows.

Training configuration. We applied LoRA ($r = 16$, $\alpha = 32$, dropout 0.05) to all linear projection layers. Optimization used `adamw_torch` with a cosine scheduler (3% warmup) and an effective batch size of 16. All fine-tuning runs were

one epoch. **Supervised baselines (SFT, FCoT, Abstain)** used learning rate 2×10^{-4}. SFT used standard chat formatting (max length 512) with user prompts masked from the loss. FCoT and Abstain used a `Question ### Rationale -> Answer` delimiter format (max length 2048) so the rationale and answer can be parsed deterministically at inference. Qwen2.5-SFT was trained on Dataset A1, Qwen2.5-FCoT on Dataset B, and Qwen2.5-Abstain on Dataset C. **Qwen2.5-RLVF** was initialized from the Qwen2.5-Abstain checkpoint and trained with on-policy rollouts on Dataset A2 at learning rate 1×10^{-5} and weight decay 0.01.

Compute budget. Total compute across all fine-tuning runs and evaluation passes is estimated at approximately 30–50 A100 GPU-hours; each supervised fine-tuning run on 10K examples took ∼1–3 hours, and evaluation scoring runs took ∼15–30 minutes per checkpoint.

Inference. All evaluation was closed-book with greedy decoding. Standard (non-reasoning) models received a concise system prompt restricting outputs to short entities (max 50 tokens). Reasoning models were prompted with the ### delimiter to trigger a rationale (max 512 tokens) and the final answer was parsed from after ->.

4.4 Results

Generation performance and error correction. Our fine-tuning experiments show clear improvements over the Qwen2.5-Instruct baseline across all metrics. As shown in Table 8, FCoT achieves the highest Exact Match of 24.31%, a 5.4-point absolute gain over the base model's 18.86%. FCoT also corrects more baseline errors than plain SFT (697 vs. 477) at a comparable regression count (294 vs. 249), so the net gain is real. Strikingly, our 7B FCoT essentially matches closed-book PaLM-62B (a model with $\sim 9\times$ more parameters) (Sun et al., 2023), reaching 92% of its EM (24.31 vs. 26.46; Table 1). The comparison is not strictly like-for-like (different architecture, different data), but it indicates that under closed-book reasoning, capacity is not the only path to performance: an internalized-knowledge CoT recipe can close most of the parameter-count gap.

Abstention precision and recall. Table 7 reports precision and recall of each abstention model against four reference comparators (base, SFT, FCoT, and the "Hardest" subset of questions all three reference models answered incorrectly). Precision is the fraction of abstentions that correspond to questions the reference model got wrong (i.e., justified refusals); recall is the fraction of reference-model errors the abstention model successfully refused. Both abstention-trained variants achieve very high precision: **Qwen2.5-RLVF reaches 97.84% precision against the base model, the highest of any model in our evaluation**, narrowly above the 97.49% of Qwen2.5-Abstain. In other words, when our final model refuses, it is almost certainly right to do so.

Confident-error reduction across two operating points. EM penalizes wrong answers and abstentions identically and so hides the central effect we care about: how often the model emits a *confident hallucination*, which we measure directly as attempted $\times$ (1 − selective EM). The non-abstaining models all sit near 80% (Instruct 81.14%, SFT 78.06%, FCoT 75.69%) because they cannot refuse. Our abstention-trained models change this sharply at two operating points on the risk–coverage frontier (Figure 5): **Qwen2.5-Abstain** (*conservative*) answers 53.66% of questions at 97.49% abstention precision, leaving a confident-error rate of $\mathbf{33.28}\%$: a $\mathbf{59}\%$ relative reduction over the base model with EM almost unchanged (18.86 → 19.47). **Qwen2.5-RLVF** (*coverage-oriented*) answers 5 more points of questions (58.83%) while *raising* abstention precision to $\mathbf{97.84}\%$ (the highest of any model in our evaluation) and keeping confident errors at $\mathbf{38.46}\%$, a $\mathbf{53}\%$ relative reduction.

4.5 Selective Prediction Metrics

To put our results in standard selective-prediction terms, Table 2 reports AUROC of the uncertainty score against correctness (higher is better, 0.5 = random), Area Under the Risk–Coverage curve (AURC; lower is better), risk at 50% coverage (lower better), and Expected Calibration Error (ECE; lower better, with 15 equal-frequency bins). RAUQ is used as the uncertainty score for greedy-decoded models and Semantic Entropy for the RLVF model under stochastic decoding. As a prior-method comparison, we also report the P(true) self-evaluation baseline of Kadavath et al. (2022), where the same model is asked "Is the proposed answer correct?" and the probability of `Yes` (renormalized over {Yes, No}) is taken as confidence.

Model	EM	F1	Abs. Rate	Sel. EM	Sel. F1	Avg. RAUQ
Our experiments ($N = 7405$; Qwen2.5-7B-Instruct backbone)						
Qwen2.5-Instruct	18.86	27.02	–	–	–	1.25
Qwen2.5-SFT	21.94	31.41	–	–	–	1.62
Qwen2.5-FCoT	**24.31**	33.87	–	–	–	–
Qwen2.5-Abstain	19.47	25.78	46.34	37.98	48.46	1.32
Qwen2.5-RLVF	20.37	26.24	41.17	34.62	44.60	1.40
Larger-model closed-book reference						
PaLM-62B	26.46	35.67	N/A	N/A	N/A	N/A

Table 1: Main results on HotPotQA validation (distractor setting, closed-book; $N = 7405$). EM and F1 are token-level percentages. Abs. Rate is the fraction of questions answered with an "I don't know"-style refusal (judged by the AbstentionClassifier of Section 3.3); Sel. EM/F1 are computed over the non-abstained subset; Avg. RAUQ is the mean per-question RAUQ uncertainty score. PaLM-62B is included as a larger-model closed-book reference point (Sun et al., 2023); it is $\sim 9\times$ larger than our policy backbone.

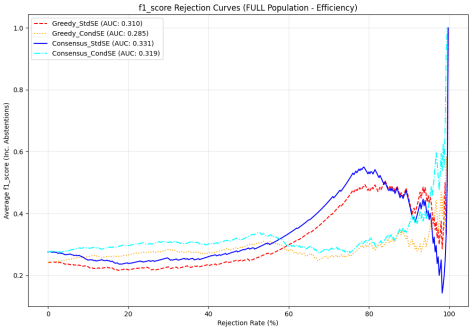
Model	Uncertainty	N	EM	F1	AUROC ↑	AURC ↓	Risk@50 ↓	ECE ↓
Prior-method baseline: P(true), Kadavath et al., 2022								
Qwen2.5-Instruct	P(true)	7405	0.182	0.262	0.508	0.819	0.810	–
Qwen2.5-SFT	P(true)	7405	0.209	0.302	0.502	0.794	0.785	–
Our methods								
Qwen2.5-Instruct	RAUQ	7405	0.182	0.262	0.431	0.807	0.865	0.523
Qwen2.5-SFT	RAUQ	7405	0.209	0.302	**0.856**	**0.606**	**0.607**	0.400
Qwen2.5-RLVF	SE	1000	0.232	0.320	0.826	0.580	0.588	0.480

Table 2: Selective prediction on HotPotQA validation. Higher AUROC is better; lower AURC, Risk@k, and ECE are better.

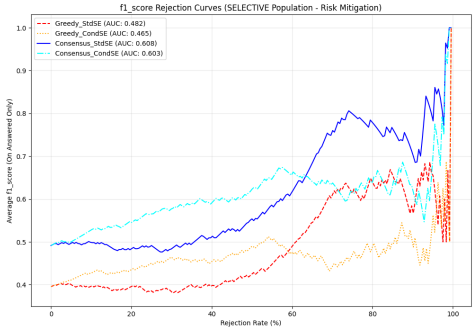
Fine-tuning is the lever, not the estimator. P(true) is at chance on both the base model (0.508) and on Qwen2.5-SFT (0.502); the self-evaluation prompt does not depend on output structure and is unaffected by fine-tuning. RAUQ on the un-fine-tuned base is also uninformative, in fact anti-informative (0.431, below chance) because the base model frequently emits 1–3-token outputs on which attention-based confidence is degenerate. The same RAUQ estimator on the supervised fine-tuned model jumps to AUROC 0.856; Semantic Entropy on the RLVF model reaches 0.826. AURC drops from 0.79–0.82 under P(true) to 0.58–0.61 under our methods, and ECE drops by 10–12 points. The $+0.35$ absolute AUROC gain is therefore attributable to fine-tuning (which produces longer, more structured outputs) rather than to the choice of estimator; we view this as evidence for a *fine-tuning-then-estimator* pipeline.

4.6 Semantic-Entropy Rejection Curves

Semantic Entropy is a strong proxy for correctness: correct answers cluster tightly around SE ≈ 0 while incorrect answers have median SE ≈ 0.9 (Appendix A.2). It is also informative about the model's own refusals: both standard SE and conditional SE (which strips the "I don't know" cluster) are shifted higher for the model's abstained subset than for its answered subset, indicating the native abstention mechanism aligns with semantic uncertainty even when it does not perfectly separate the two distributions. Figure 1 and Table 6 show the efficiency of using SE for post-hoc rejection. We compare a *Greedy* strategy (confidence from the single most likely sequence) against a *Consensus* strategy (marginalising probability mass over all sampled generations in the same semantic cluster). Consensus dominates Greedy at every rejection rate: at 0% rejection it reaches selective F1 0.50 vs. Greedy's 0.40, and at 50% rejection it reaches selective F1 **0.70**, confirming that aggregating semantic consistency across multiple samples is a more robust hallucination signal than a single-path decode.



Full Population (Efficiency)



Selective Population (Risk Mitigation)

Figure 1: F1 rejection curves: efficiency on the full dataset (left) vs. risk mitigation on answered questions (right). Both panels rank questions by Semantic Entropy and remove the top-$k\%$; Consensus dominates Greedy at every rejection rate, reaching selective F1 0.70 at 50% rejection.

5 Analysis

5.1 Reasoning Trace: Qualitative Check

Qualitative inspection of FCoT outputs (Appendix A.1) confirms that the model learned the internalized-knowledge style; e.g., on "Were Scott Derrickson and Ed Wood of the same nationality?," FCoT retrieves both nationalities from parameters and corrects the baseline's *No* to *yes* without ever referring to "the passage." Regressions cluster into two modes. **Hallucination via reasoning**: step-by-step generation can induce false details (the model claims Rose McGowan was best known for *Buffy the Vampire Slayer*, conflating two different actresses and shows, overriding the base model's correct *Charmed* answer). **Granularity mismatch**: the trace can over-specify (deducing "film director" against the rigid gold "director"); this is a limitation of Exact Match on reasoning outputs rather than of the model, and supports our decision to also report token-level F1.

5.2 SFT vs. RLVF: Complementary Operating Points

The two stages produce models with qualitatively different behaviour. On *"What army did the namesake of the ship launched as the München in 1930 fight in during the American Revolutionary War?"*, the supervised model abstains; RLVF answers *Continental Army* (correct), recovering parametric knowledge the conservative model declines to commit to. On *"What government position was held by the woman who portrayed Corliss Archer in Kiss and Tell?"*, the supervised model again abstains; RLVF answers *First Lady* (the correct answer is *Chief of Protocol*), a confident error from the same reward shape that recovered the previous case. The two stages thus expose complementary operating points: the supervised model fits safety-critical deployments (97.5% abstention precision, lowest confident-error rate), and the RL-tuned model fits general-assistant deployments (97.84% abstention precision, +5-point coverage). Both come from the same pipeline and differ only in whether the abstention boundary is fixed by supervised labels or refined against a verifier reward; the RL stage's methodological contribution is the reward-shaped lever, which a sweep of the abstention penalty would let one move along.

6 Conclusion

We decomposed closed-book hallucination into access failures and knowledge gaps and built a pipeline that addresses each: CoT distillation (5.4-point EM gain), abstention-aware SFT (97.5% abstention precision), and verifier-grounded RL (97.84% abstention precision at +5 points of coverage). The supervised and RL stages expose two complementary operating points on the risk–coverage frontier and both cut confident hallucinations by up to 59% over the base model. In standard selective-prediction terms, fine-tuning, not the choice of uncertainty estimator, lifts RAUQ and Semantic Entropy from chance to AUROC up to 0.86, a +0.35 absolute gain over the P(true) self-evaluation baseline. Natural follow-ups include a hyperparameter sweep over the abstention penalty to densify the risk–coverage curve, and extension to open-book settings where the model must discriminate parametric knowledge from retrieved context.

Limitations

Benchmark scope. Our experiments are on HotPotQA distractor in a closed-book setting, a short-form English factoid benchmark over Wikipedia-derived text. Generalization to long-form, non-English, or domain-specialised QA is a natural extension and remains open.

Inference cost of Semantic Entropy. Post-hoc SE filtering requires $M{=}10$ stochastic samples plus pairwise NLI clustering per question. This is an effective filter but multiplies inference compute over a single greedy decode; cheaper proxies (e.g., RAUQ as a pre-filter) are a useful direction.

Ethical Considerations

Residual confident errors. Even after RLVF training, the model still emits confident wrong answers on a substantial fraction of attempted questions (approximately 38% of the questions it chooses to answer). Users or system designers who focus on the high abstention precision (97.84%) risk underestimating this residual error rate. Deployment in safety-critical applications (medical triage, legal advice, etc.) should include independent verification and should not rely solely on the model's own refusal signal.

Domain shift. The knowledge boundary is calibrated on Wikipedia-derived multi-hop factoid QA. Applying these checkpoints to out-of-domain corpora without re-calibrating the abstention objective could cause systematic miscoverage: the model may answer confidently in domains where it has no reliable parametric knowledge, or over-refuse in domains that overlap with its training distribution in unexpected ways.

References

Amos Azaria and Tom Mitchell. 2023. The internal state of an LLM knows when it's lying. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*.

Namgyu Ho, Laura Schmid, and Se-Young Yun. 2023. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 14852–14882. Association for Computational Linguistics.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Liunian Harold Li, Jack Hessel, Youngjae Yu, Xiang Ren, Kai-Wei Chang, and Yejin Choi. 2023. Symbolic chain-of-thought distillation: Small models can also "think" step-by-step. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL), Volume 1: Long Papers*, pages 2665–2679. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), Volume 2: Short Papers*, pages 784–789. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex

Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Zhiqing Sun, Xuezhi Wang, Yi Tay, Yiming Yang, and Denny Zhou. 2023. Recitation-augmented language models. In *The Eleventh International Conference on Learning Representations*.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and Artem Shelmanov. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for LLMs. *Preprint*, arXiv:2505.20045.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380. Association for Computational Linguistics.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-Tuning: Instructing large language models to say "I don't know". In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 7106–7132. Association for Computational Linguistics.

A Appendix

A.1 Qualitative Reasoning-Trace Examples

Tables 3 and 4 below give the full reasoning traces summarised in Section 5.1.

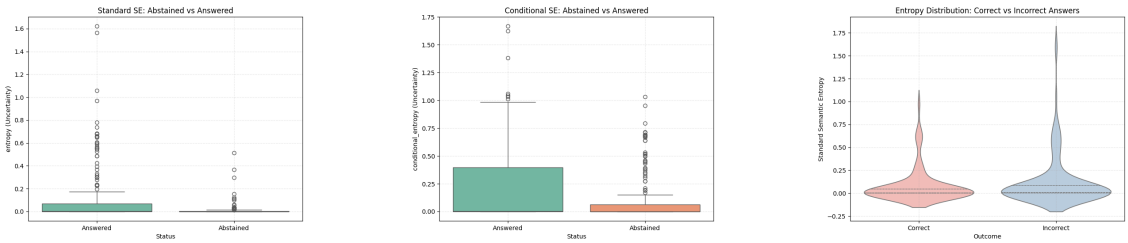
Context (Base Wrong → CoT Correct)	Generated Reasoning Trace
Q: Were Scott Derrickson and Ed Wood of the same nationality? **Gold:** yes **Base Pred:** No (Wrong) **CoT Pred:** yes (Correct)	Step 1: The question asks about the nationalities of Scott Derrickson and Ed Wood. Step 2: Scott Derrickson is an American film director. . . while Ed Wood was an American filmmaker. . . Step 3: Since both individuals were born and active in the United States. . . the answer is yes.

Table 3: Example of successful internalized-knowledge retrieval by Qwen2.5-FCoT.

Context (Base Correct → CoT Wrong)	Reasoning Trace Fault
Failure Mode 1: Hallucination via reasoning	
Q: What WB supernatural drama series was Jawbreaker star Rose McGowan best known for being in? **Gold/Base:** Charmed **CoT Pred:** Buffy the Vampire Slayer	Step 2: Rose McGowan is best known for her role as Cordelia Chase in the WB supernatural drama series "Buffy the Vampire Slayer." *(Error: hallucinated role / show.)*
Failure Mode 2: Granularity mismatch	
Q: What occupation do Chris Menges and Aram Avakian share? **Gold/Base:** Director **CoT Pred:** film director	Step 3: Since both individuals are primarily recognized for directing films, their shared occupation is film director. *(Error: trace too specific for gold label.)*

Table 4: CoT failure modes (regression cases).

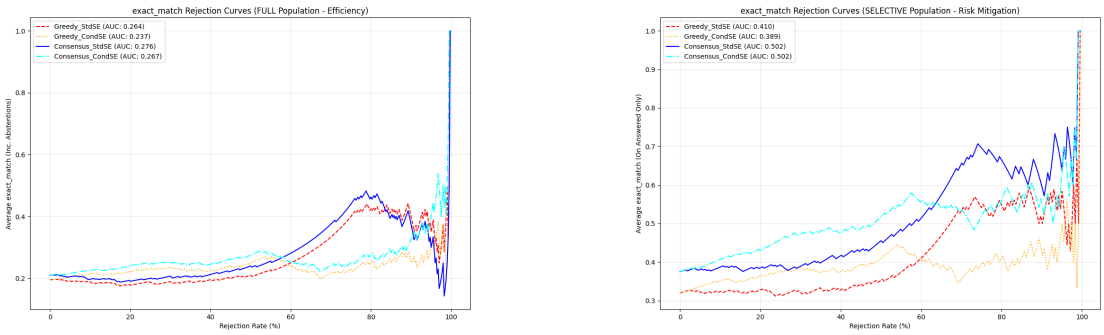
A.2 Abstention Model vs. SE Analysis



Std SE: Abstained vs. Answered Cond SE: Abstained vs. Answered Std SE: Correct vs. Incorrect

Figure 2: Distribution of Semantic Entropy across outcomes.

A.3 Additional SE Rejection Curves



Full Population (Efficiency) Selective Population (Risk Mitigation)

Figure 3: Exact Match (EM) rejection curves.

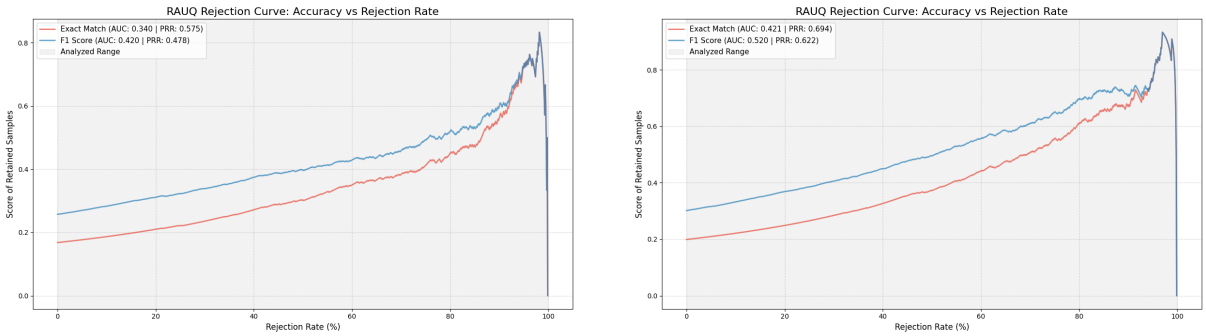
Table 5: Impact of Semantic Entropy-based Rejection on Exact Match ($N = 2500$)

Strategy	Rejection Threshold	Full EM	Selective EM
Greedy	0%	0.1796	0.3065
Greedy	50%	0.2912	0.4863
Consensus	0%	0.2124	0.3884
Consensus	50%	0.3256	**0.5944**

Table 6: Impact of Semantic Entropy-based Rejection on F1 Scores ($N = 2500$)

Strategy	Rejection Threshold	Full F1	Selective F1
Greedy	0%	0.2329	0.3974
Greedy	50%	0.3482	0.5771
Consensus	0%	0.2719	0.4972
Consensus	50%	0.3828	**0.6991**

A.4 Supplementary RAUQ Rejection Curves



Vanilla Qwen2.5-7B Fine-tuned

Figure 4: RAUQ rejection curves: both models improve as high-uncertainty samples are rejected, and the fine-tuned model retains higher quality across the curve.

A.5 Confusion Matrices

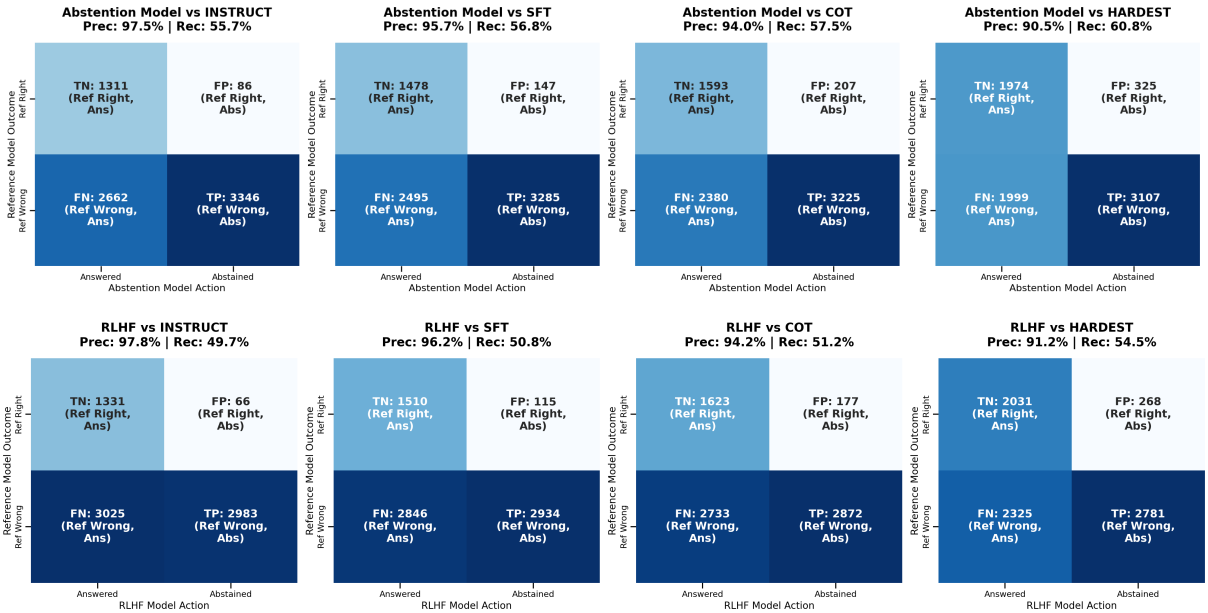


Figure 5: Abstention performance: supervised fine-tuning vs. RLVF. **Top**: confusion matrices for the supervised abstention model (Qwen2.5-Abstain), trained with explicit abstention labels on questions where baseline models failed. **Bottom**: confusion matrices for the RLVF model, trained with policy-gradient RL against the DeBERTa NLI verifier reward (Section 3.3). Each panel shows how well the model's abstention decisions align with questions that the four reference models answered incorrectly.

A.6 Additional Tables

Table 7: Performance Analysis of Abstention Mechanism Across Reference Models

Model	Ref. Model	Precision	Recall	TP (Good Abs)	FP (Bad Abs)	FN (Missed)
Qwen2.5-Abstain	Base	**97.49%**	55.69%	3,346	**86**	2,662
	SFT	95.72%	56.83%	3,285	147	2,495
	CoT	93.97%	57.54%	3,225	207	2,380
	Hardest	90.53%	**60.85%**	3,107	325	**1,999**
Qwen2.5-RLVF	Base	**97.84%**	49.65%	2,983	**66**	3,025
	SFT	96.23%	50.76%	2,934	115	2,846
	CoT	94.19%	51.24%	2,872	177	2,733
	Hardest	91.21%	54.47%	2,781	268	2,325

Table 8: Overall Model Performance and Error Analysis Relative to Base Model

Metric	Base Model	SFT Model	CoT Model
Overall Accuracy (Exact Match)	18.87%	21.94%	**24.31%**
Corrections (Base Wrong → Right)	–	477	**697**
Regressions (Base Right → Wrong)	–	**249**	294

A.7 Licenses and Intended Use

All artifacts are used under terms consistent with academic non-commercial research. HotpotQA is released under CC BY-SA 4.0; our derived datasets are released under the same license. Qwen2.5-7B-Instruct and Qwen3-8B are released under the Qwen License (non-commercial research use); our fine-tuned LoRA adapters are released on the same terms. The DeBERTa model variants used for the FactualityVerifier and semantic clustering (`microsoft/deberta-large-mnli`, `MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli`, `cross-encoder/nli-deberta-v3-xsmall`) are released under MIT or Apache 2.0 licences, permitting research use. No artifact is deployed in a commercial or production context; all use is consistent with each source's intended research purpose.