Entrospect: Information-Theoretic Self-Reflection Elicits Better Response Refinement of Small Language Models

Anonymous ACL submission

Abstract

Self-reflection helps de-hallucinate Large Language Models (LLMs). However, the effectiveness of self-reflection remains insufficiently validated in the context of Small Language Models (SLMs), which exhibit limited semantic capacities. In particular, we demonstrate that the conventional self-reflection paradigm, such as Self-Refine, fails to deliver robust response refinement for models with parameter sizes of 10 billion or smaller, even when compared to generations elicited through Chain-of-Thought (CoT) prompting. To improve SLMs' self-reflection, we redesign Self-Refine and introduce *Entrospect*[1] (*Entro*py-aware In-tro*spect*ion), an information-theoretic framework based on prompt engineering.

We evaluated *Entrospect* using accuracy and average time consumption metrics to comprehensively assess its precision and computational efficiency. Experiments conducted across four distinct SLMs and four baseline methods demonstrate that *Entrospect* achieves state-of-the-art performance on validation tasks. Notably, under identical model and data settings, *Entrospect* delivers a remarkable improvement of up to 36.2% in reasoning accuracy while enhancing computational efficiency by as much as 10 times compared to its predecessor, Self-Refine.

1 Introduction

Large Language Models have advanced rapidly, impacting many fields with improved natural language generation (Brown et al., 2020; Chang et al., 2024). However, their tendency to produce hallucinations—especially counterfactual ones—poses a critical challenge to reliability (Zhang et al., 2023; Huang et al., 2023). Hallucinations occur when models generate factually incorrect or nonsensical outputs, undermining their trustworthiness and hindering real-world adoption. Addressing this issue is essential for improving their practical utility and acceptance (Weidinger et al., 2021, 2022).

[1]The project is intended to be open-source soon after the publication. For reviewers, we attached the examples of Entrospect's outputs to the submission.

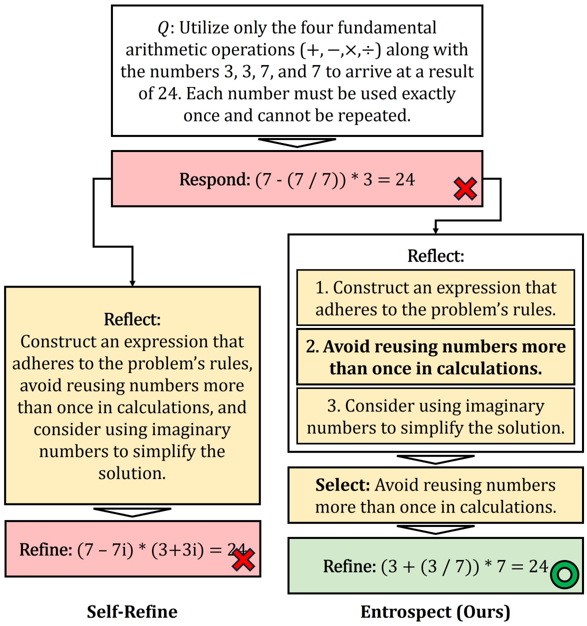

Figure 1: The single-round refinement of an initial response for the same query, comparing Self-Refine and our proposed *Entrospect*. Self-Refine fully relies on the model's self-reflected feedback, where any biases introduced during *reflect* are directly carried over into *refine*, hindering constructive improvements. On the other hand, our Entrospect identifies the optimal revision suggestion from an itemized ouput of the self-reflection, enabling Entrospect to achieve more robust and reliable response refinement.

To address these challenges, self-reflection has been proposed as a solution to counterfactual hallucinations, particularly for black-box models with inaccessible parameters (Madaan et al., 2024). However, its effectiveness is limited in Small Language Models (SLMs), which often lack sufficient semantic capabilities, inducing frequent occurences of imperfect feedback, encompassing the self-reflected revision suggestions. Given the widespread use of SLMs in resource-constrained environments (Li et al., 2024; Wang et al., 2024), this limitation is particularly significant. In such cases, self-reflection may fail to consistently assist in the corrections of outputs, highlighting the need for more robust and scalable approaches.

Given the challenges of applying self-reflection to SLMs, a key question arises: *how might we con-*

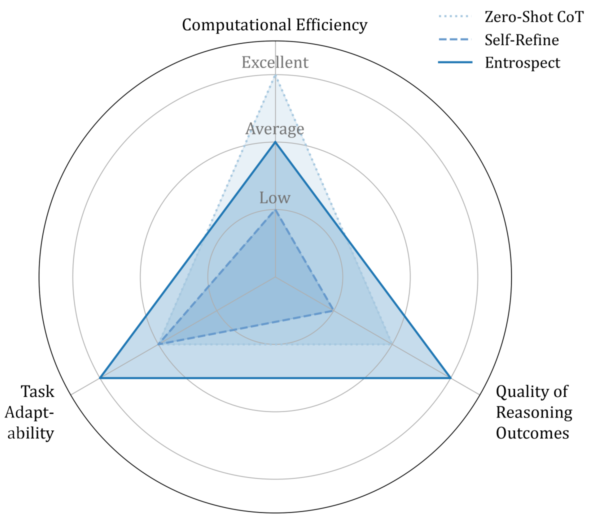

Figure 2: Entrospect contributes furtherance to the response refinement of SLMs particulary over its predecessor, Self-Refine, across three major aspects.

struct a framework that effectively integrates self-reflection to improve the precision of SLM outputs, all while preserving the computational efficiency? In response to this challenge, we propose *Entrospect*, an information-theoretic framework predicated on Self-Refine that lessens the dependency on explicit semantic outputs from the model. Contrary to Self-Refine's equal consideration of all revision suggestions, Entrospect employs an unsupervised mechanism to identify the most effective revision candidate, minimizing the impact of inferior ones, as illustrated in Figure 1.

Specifically, Entrospect is implemented with an Optimal Revision Suggestion Selector (ORSS) Inspired by (Wu et al., 2024) and (Yang et al., 2024b), the ORSS intervenes between the "reflect" and the "refine" stages that are tightly connected in the Self-Refine's pipeline. It evaluates revision suggestions generated through self-reflection and identifies the one that minimizes the semantic uncertainty in the model's refinement of the prior response, where low-quality suggestions conceivably ruining the successive procedures are ruled out. This selective approach distinguishes Entrospect from its predecessors, enhancing both the quality and reliability of the refined responses.

Architecturally, Entrospect retains the simplicity and efficiency of Self-Refine, operating as a parameter-free, recurrent finite-state machine (FSM) where modules are interconnected through purpose-specific prompts. This design ensures computational efficiency while maintaining the flexibility to adapt to diverse conversational AI tasks. Figure 2 summarizes the multifaceted contributions of Entrospect, the central focus of this study.

We evaluated Entrospect on natural language reasoning tasks, including the MATH dataset (Hendrycks et al., 2021) for math reasoning and HaluEval (Li et al., 2023) for hallucination detection. The results show Entrospect outperforms baselines like zero-shot, few-shot, Chain-of-Thought (CoT), and Self-Refine. These findings underscore two critical advances:

1. **Selective Use of Self-Reflection:** We highlight that the outcomes of a model's self-reflection should not be directly or entirely relied upon as guidance for the response refinement.

2. **ORSS-Driven Optimization:** Our proposed Entrospect improves Self-Refine by introducing ORSS, an information-theoretic mechanism that unsupervisedly identifies the optimal revision from multiple candidates. Combined with our semantic similarity-based stopping condition, Entrospect allows a more robust and systematic approach to self-reflection for response refinement. Compared to its precursor, Self-Refine, Entrospect accomplishes a remarkable performance boost, delivering up to 36.2% improvement in accuracy under identical dataset and model conditions, while elevating computational efficiency by as much as 10 times.

2 Related Work

2.1 Self-Reflection of Language Models

The empirical foundation of self-reflection is that given some queries, language models may not be able to provide proper answers every time (Yan and Xu, 2023). Self-reflection assists in alleviating such problems by explicitly instructing a language model to review its generated response, providing a feedback on potential deficiencies within the current response and how they could be eliminated. The feedback is subsequently used for guiding the refinement of the previous answer. This procedure can be fully automated through a prompt-driven framework, by which a language model iteratively reflects and refines the answer to a query on its own (Lee et al., 2024).

Techniques like Self-Refine introducing mechanisms for models to improve their own responses (Madaan et al., 2024), especially in question-answering (QA) scenarios, to enhance generation quality. This approach has been further advanced in research such as Reflexion and Agent-Pro (Shinn et al., 2024; Zhang et al., 2024b), which extend self-reflection to agentic scenarios, increasing the efficiency and success rate of task execution during scenario exploration and trajectory execution. However, there remains significant room for improvement in its performance, particularly when it comes to SLMs.

Through extensive review, we found lack of report on the effectiveness of self-reflection applied to models which possess fewer than 10 billion parameters. Its success relies heavily on the context generated during the self-reflection process (Cheng et al., 2024) and is prone to overconfidence in its generated content (Zhang et al., 2024a), including biases.

We assessed the self-reflective capabilities of several SLMs across a variety of tasks, with Self-Refine chosen as a baseline approach. Our findings reveal that reflective thinking of these models fails to produce meaningful improvements in their generative performance. Entrospect is specifically designed to enhance the performance of SLMs by leveraging information theory to assist in the self-reflection process.

2.2 Enhancing the Reasoning Capabilities of Small Language Models

Recent studies have made significant strides in enhancing the reasoning capabilities of SLMs. Bi et al. introduced Solution-Guidance Fine-Tuning (Bi et al., 2024), focusing on problem understanding and decomposition to improve SLMs' generalization and reasoning abilities. Wang and Lu explored continual pre-training on a synthetic dataset to inject multi-step reasoning abilities into moderate-sized models (Wang and Lu, 2023). Fu et al. specialized small models towards multi-step reasoning through knowledge distillation from large models (Fu et al., 2023). Yu et al. developed TRIPOST, an algorithm enabling small models to self-improve via interaction with large ones (Yu et al., 2023).

However, these methods often necessitate a substantial amount of additional data, whether it is synthetically created or derived from larger models, which may not be readily accessible or easy to produce. They entail a certain degree of computational overhead, be it in data generation, pre-training, or iterative training processes. Differently, Entrospect does not require any additional data or specialized training, thus drastically reducing both overhead and resource demands, allowing broader applicability across diverse domains and use cases.

3 Methodology

3.1 Problem Definition

While frameworks like Self-Refine aim to automate response refinement in language models through self-reflection, they do not inherently ensure that such refinements are beneficial. This limitation is particularly pronounced in SLMs, where constrained semantic capabilities lead to unreliable self-reflections, resulting in *reflective contamination*. Reflective contamination occurs when the model's self-generated feedback contains biases, which can degrade rather than improve the refined response.

To formalize this problem, consider the t-th refinement round, where the model $\mathcal{M}_\theta$ generates feedback F_t based on the query Q, reflection prompt P_{reflect}, and current response A_t. This feedback, represented as $\mathcal{M}_\theta\left(Q\|A_t\|P_{\text{reflect}}\right)$, consists of two components: 1) A valid portion $S_t = \rho_t F_t$, which supports effective refinement. 2) Reflective contamination $N_t = (1-\rho_t)\, F_t$, which introduces biases. Here, $\rho_t \in (0,1)$ represents the proportion of valid feedback in F_t. The refined response A_{t+1} is then generated using F_t, Q, and the refinement prompt P_{refine}, expressed as:

$$\begin{aligned} A_{t+1} &= \mathcal{M}_\theta\left(Q\|A_t\|F_t\|P_{\text{refine}}\right) \\ &= A_t + \alpha_t^S S_t - \alpha_t^N N_t \\ &= A_t + \alpha_t^S \rho_t F_t - \alpha_t^N \left(1-\rho_t\right) F_t \\ &= A_t + \left[\left(\alpha_t^S + \alpha_t^N\right)\rho_t - \alpha_t^N\right] F_t, \end{aligned} \tag{1}$$

where α_t^S and α_t^N are partial attention factors ($\alpha \in (0,1)$) applied to the valid and contaminated portions of F_t, respectively.

The Core Problem:

1. A successful refinement requires $A_{t+1} \geq A_t$, but this is not guaranteed. When ρ_t is low (i.e., the feedback contains more contamination), the refined response may degrade, as described by the condition:

$$\rho_t < \frac{\alpha_t^N}{\alpha_t^S + \alpha_t^N}. \tag{2}$$

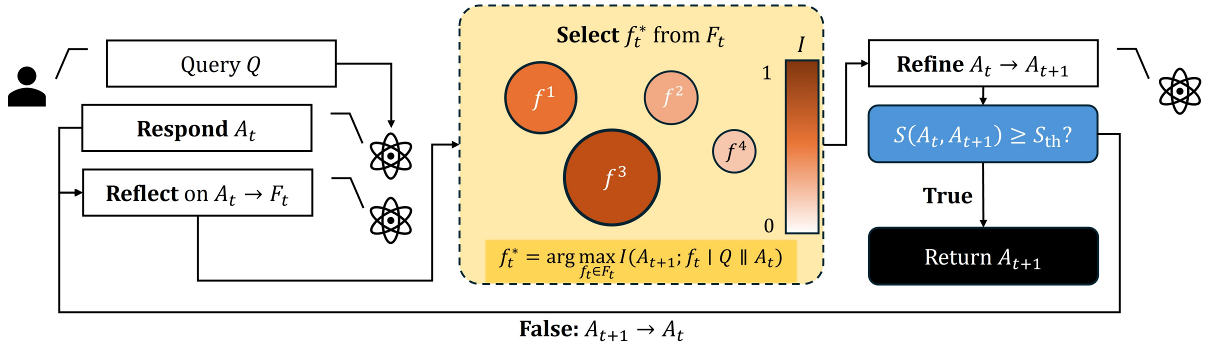

Figure 3: The pipeline of our Entrospect prompt-driven framework, extending the original Self-Refine structure with an Optimal Revision Suggestion Selector (ORSS) and a universal semantic similarity-based stopping condition. The framework requires no supervised pre-training or access to the model's internal parameters, granting it to be generalizable to various language models and reasoning tasks.

2. SLMs, with their limited semantic competence, often exhibit low ρ_t and high α_t^N (or low α_t^S), making them prone to degradation during the refinement phase of the response.

Objective: Within the realm of black-box models, α_t^S, α_t^N and ρ_t are inaccessible. This presents a significant obstacle in accurately differentiating between S_t and N_t. An alternative perspective involves concentrating exclusively on the optimal component of F_t. Entrospect proposes an unsupervised mechanism driven by information theory, providing a systematic solution to this complication.

3.2 Optimal Revision Suggestion Selector

By employing a formatting prompt, we can steer the model's self-reflective output towards a systematic arrangement of multiple revision suggestions. In this way, F_t is characterized as an ensemble of strings $\{f_t^0, f_t^1, \ldots, f_t^n\}$, framing our goal as "discerning an optimal revision suggestion from this set". However, in the absence of supervision, defining what constitutes *optimal* becomes a fundamental issue.

To address this, we propose a solution called the Optimal Revision Suggestion Selector (ORSS), which uses heuristic information-theoretic approaches for prompt selection (Wu et al., 2024; Yang et al., 2024b). These studies suggest that an optimal prompt should minimize the semantic uncertainty of a language model when processing a query, which is equivalent to maximizing the conditional mutual information (CMI) between the input and the output. Unlike recent work which assumes a manually constructed prompt pool, F_t as the candidate set in our case is constructed in an automatic fashion, where revision suggestions become prompt candidates, and the one to be selected renders the maximum CMI following Equation 3:

$$
\begin{aligned}
f_t^* &= \arg\max_{f_t \in F_t} I\left(A_{t+1}; f_t \mid Q \| A_t\right), \\
\text{where } I &= H\left(A_{t+1} \mid Q \| A_t\right) \\
&\quad - H\left(A_{t+1} \mid f_t, Q \| A_t\right).
\end{aligned} \tag{3}
$$

In Equation 3, $Q \| A_t$ stands for the prompt "Please provide a refined solution of <Q> given <A_t>", and $(f_t, Q \| A_t)$ signifies a slightly different prompt "Please provide a refined solution of <Q> given <A_t>. <f_t>". The two Hs characterize the *marginal entropy* and the *conditional entropy* in classical information theory, respectively. The value of CMI I stands for **the extent to which a revision suggestion f_t enhances the model's confidence in the refinement applied to the current answer A_t.**

3.3 Eliciting the Convergence of Entrospect

We established a universal mechanism to enable Entrospect to automatically terminate its iterations. The core principle is that, at the semantic level, A_t and A_{t+1} are essentially equivalent. Consequently, when a language model employs greedy search ($temperature = 0$) for output sampling, subsequent outputs naturally converge toward consistency, rendering the increments from reflection and refinement negligible. Given these circumstances, the framework no longer introduces meaningful improvements to the response, a state we defined as "convergence". More precisely, we leverage the *cosine similarity* $S(\cdot, \cdot)$ to quantify the degree of semantic resemblance between two answers, modeled as

$$
\begin{aligned}
S\left(A_1, A_2\right) &= \frac{\mathbf{v_1} \cdot \mathbf{v_2}}{\|\mathbf{v_1}\|\|\mathbf{v_2}\|} \\
&= \frac{\sum_{i=1}^{m}(v_{1i} \cdot v_{2i})}{\sqrt{\sum_{i=1}^{m} v_{1i}^2} \cdot \sqrt{\sum_{i=1}^{m} v_{2i}^2}},
\end{aligned} \tag{4}
$$

where $\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \dots & v_m \end{bmatrix}^{\mathrm{T}}$ indicates the A's tokenized vector in a continuous, m-dimensional semantic space. The range of S is $[-1, 1]$, with a higher value referring to a stronger semantic similarity between the two entities compared. Leveraging semantic similarity as a stopping condition for the iterative refinement procedure guarantees an appropriate termination juncture, thus optimizing performance results.

3.4 Framework of Entrospect

Slightly different from the three-step process of *respond* → *reflect* → *refine* adopted by Self-Refine, Entrospect follows an extended four-step strategy: *respond* → *reflect* → ***select*** → *refine*. In the following, we detail each step sequentially; see Figure 3 for an intuitive illustration of the pipeline and Algorithm 1 for implementation guidance.

Respond: The iterations begin with the language model generating an initial answer A_0 for the input query Q.

Reflect and Select: During iteration t, the model $\mathcal{M}_\theta$, guided by the prompt P_{reflect}, the original query Q, and the current answer A_t, generates a set of candidate revision suggestions denoted as $F_t = \{f_t^0, f_t^1, \dots, f_t^n\}$. The prompt P_{reflect} serves as a directive that instructs the model on how to evaluate potential deficiencies in the current answer and construct appropriate F_t accordingly. Thereafter, the ORSS selects the optimal f_t^* that maximizes the CMI between the input and the output of the model. In practical implementation, the *Cross-Entropy Loss* $\mathcal{L}_{\text{CE}}$ output by the model for a given input can be used to calculate the marginal entropy and the conditional entropy, allowing for the straightforward computation of the CMI.

Refine: Leveraging the f_t^* as the key instruction to the refinement, the model $\mathcal{M}_\theta$ utilizes the prompt P_{refine}, in conjunction with the original query Q and the current answer A_t, to generate an updated answer A_{t+1}.

Stop Condition: Subsequent to the generation of the A_{t+1}, we exert the semantic textual similarity measure to check whether the iterative process should be terminated. When A_t and A_{t+1} exhibit a high degree of semantic resemblance, this suggests that Entrospect has entered a state of convergence from the current iteration onward. Following that, A_{t+1} is designated as the final output. To meet the requirements of long-text encoding with high representational fidelity, we opted for the *Jina Embeddings V3* (Sturua et al., 2024) with a dedicated LoRA adapter for text-matching tasks, an encoder-based model which natively supports an input sequence length of up to 8192 tokens. In our experiments, $S \geq 0.9$ is adopted as the threshold for considering A_t and A_{t+1} semantically equivalent.

We detailed the instructions involved in the operation process of Entrospect in Figure 6.

Algorithm 1 The algorithm pipeline of Entrospect

Require: query Q, model $\mathcal{M}_\theta$, prompt P_{reflect} ($:= P_{\text{f}}$), prompt P_{refine} ($:= P_{\text{r}}$), semantic similarity threshold S_{th}

```
A_0 ← M_θ(Q)                                  ▷ Respond
A_t ← A_0
while True do
    F_t ← M_θ(P_f ‖ Q ‖ A_t)                   ▷ Reflect
    {f_t^0, f_t^1, ..., f_t^n} ← list(F_t)    ▷ Itemize
    I_max ← 0
    for f_t in list(F_t) do                   ▷ Select (ORSS)
        H_t^marg ← L_CE(M_θ(P_r ‖ Q ‖ A_t))
        H_t^cond ← L_CE(M_θ(P_r ‖ Q ‖ A_t ‖ f_t))
        I_t ← H_t^marg − H_t^cond
        if I_t > I_max then
            f_t^* ← f_t
        end if
    end for
    A_{t+1} ← M_θ(P_r ‖ Q ‖ A_t ‖ f_t^*)      ▷ Refine
    if S(A_t, A_{t+1}) ≥ S_th then
        break
    end if
    A_t ← A_{t+1}
end while
return A_{t+1}
```

4 Experiments and Results

4.1 Experimental Settings

We evaluated Entrospect equipped by four of the latest SLMs, including DeepSeek-R1-distilled Qwen 2.5 1.5B (Yang et al., 2024a; Guo et al., 2025), Qwen 2.5 7B (Yang et al., 2024a), Llama 3.1 8B (AI, 2024), and GLM-4 9B (GLM et al., 2024), as compared to the baselines (see Section 4.4) on a math reasoning dataset and a hallucination detection dataset, namely MATH (Hendrycks et al., 2021) and HaluEval (Li et al., 2023). Each SLM was quantized to INT4 precision with either AutoGPTQ or BitsAndBytes (Pan, 2023; Dettmers et al., 2022).

4.2 Datasets

To comprehensively assess whether Entrospect heightens the ubiquitous reasoning performance of SLMs, we sourced our validation data from two representative datasets, MATH and HaluEval, with illustrative examples provided in Table 2.

Table 1: Accuracies (%) of various methods equipped by four of the latest SLMs on reasoning tasks MATH (The average accuracies of level 1 to level 5) and HaluEval. We highlight the best results in **bold**.

Model Name	Method	MATH	HaluEval
DeepSeek-R1-Distilled Qwen 2.5 Instruct 1.5B	Zero-Shot	94.2	80.5
	5-Shot	90.2	29.5
	Zero-Shot CoT	91.3	91.0
	Self-Refine	88.5	80.0
	Entrospect	**98.4**	**95.5**
Qwen 2.5 Instruct 7B	Zero-Shot	78.2	94.5
	5-Shot	72.8	91.0
	Zero-Shot CoT	83.8	98.0
	Self-Refine	73.0	97.5
	Entrospect	**86.0**	**100.0**
Llama 3.1 Instruct 8B	Zero-Shot	61.7	94.5
	5-Shot	56.5	94.0
	Zero-Shot CoT	73.7	94.5
	Self-Refine	44.3	95.0
	Entrospect	**80.5**	**99.5**
GLM 4 Instruct 9B	Zero-Shot	55.0	98.5
	5-Shot	57.9	97.5
	Zero-Shot CoT	65.8	97.5
	Self-Refine	56.8	97.5
	Entrospect	**69.7**	**100.0**

MATH (Hendrycks et al., 2021)**:** a dataset designed to measure the mathematical reasoning capabilities of language models, consisting of problems sourced from high school math competitions, tagged with difficulty levels from 1 to 5 and covering a wide range of topics including algebra, geometry, number theory, and combinatorics. MATH is notable for its complexity compared to the other datasets of the same category (Frieder et al., 2024), e.g. GSM8K (Cobbe et al., 2021). Besides, the latest findings have unveiled that MATH suffers less leakage than GSM8K does from the worsening cheating on model training (Xu et al., 2024), underlining its fairness. We randomly chose 120 samples from each difficulty level to serve as our experimental dataset.

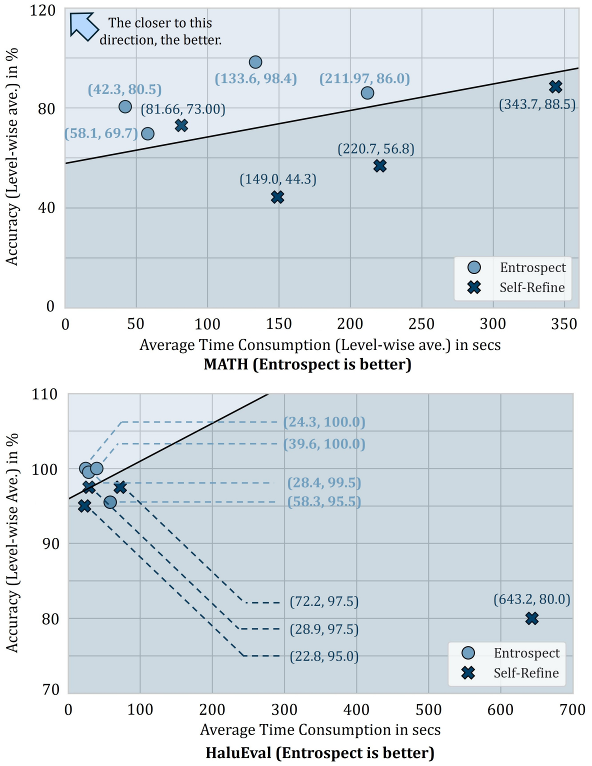

Figure 4: The Accuracy-ATC results derived from evaluating Entrospect and Self-Refine across four models and two tasks. The dividing lines in the chart correspond to the decision boundaries determined by linear SVMs fit on the data points of Entrospect and Self-Refine. Data points positioned closer to the **top-left corner** signify a more favorable trade-off between computational efficiency and reasoning accuracy, indicating superior overall performance.

HaluEval (Li et al., 2023)**:** a dataset that gauges the performance of language models in recognizing hallucinations, featuring general user queries and task-specific examples across question answering, dialogue, and text summarization. We randomly sampled 200 pairs from this dataset, providing a robust evaluation platform for analyzing the effectiveness of our framework in detecting and reducing hallucinations.

4.3 Evaluation Metrics

We selected two evaluation metrics, i.e. Accuracy and Average Time Consumption (Han et al., 2023; Xu et al., 2023; Xiao et al., 2024), to provide both qualitative and quantitative insights into the effectiveness of Entrospect.

Accuracy: a pivotal evaluation metric, is delineated as the proportion of problems correctly resolved relative to the total number of problems the model attempts, computed via $A_{\text{correct}} / (A_{\text{correct}} + A_{\text{wrong}}) \times 100\%$. A higher accuracy signifies that a prompting scheme is more effective in lifting the model's reasoning outcomes.

Average Time Consumption: We measured the Average Time Consumption (ATC) of the selected prompting schemes, spanning from the moment the input is supplied to the generation of the final output. Given the sample size N of the validation set, ATC is calculated by $\frac{1}{N}\sum_{k}^{N}(t_{k_\text{o}} - t_{k_\text{i}})$, where $t_{k_\text{o}} - t_{k_\text{i}}$ denotes the duration, counted in seconds, from the moment the k-th input is supplied to the time the k-th output is generated. A smaller ATC embodies better computational efficiency of a prompting method, which is vital for industrial implementation, notably on edge computing devices running local SLMs. In our assessments, both of the above metrics are considered for more comprehensive analysis.

4.4 Baseline Selection

We compared Entrospect against the following well-established prompting methods as well as its ablated version, functioning as robust benchmarks for appraising the performance uplift in SLMs achieved with Entrospect.

Zero-Shot and Few-Shot Prompting (Brown et al., 2020)**:** Zero-shot prompting directs a language model to perform tasks with only high-level instructions, often sacrificing accuracy for complex inputs. Conversely, few-shot prompting supplies demonstrations to improve context awareness and performance, yet its success hinges on the quality of examples, which may not fully capture task complexity and may be labor-intensive to gather in practice.

Chain-of-Thought Prompting (Wei et al., 2022)**:** An approach that guides language models to generate a structured reasoning path before arriving at the final answer, encouraging more systematic and transparent problem solving. A key downside is the increased potential for longer outputs, as irrelevant, inaccurate, and repetitive steps may appear in the generated thought chain, especially concerning SLMs, impairing the overall outcome.

Self-Refine (Madaan et al., 2024)**:** The framework allows a model to iteratively revise its own outputs with identified errors from the self-reflection's feedback. Despite its potential, such a strategy

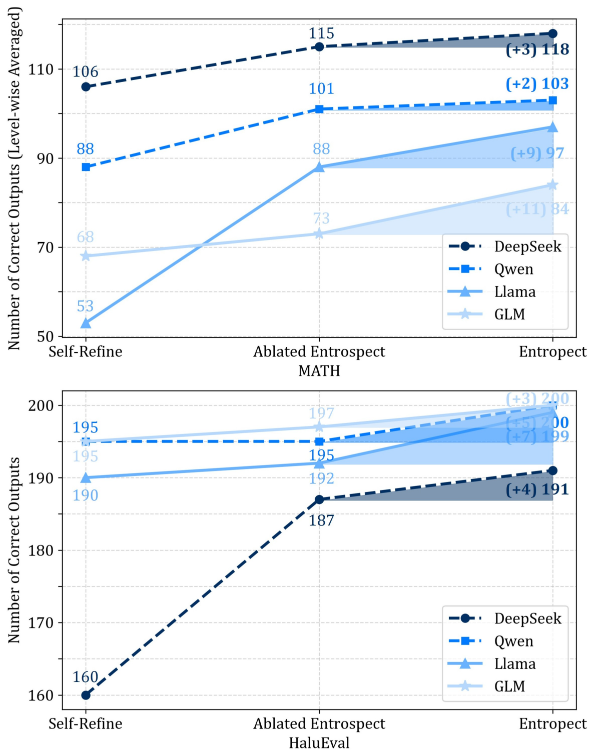

Figure 5: (A higher Number of Correct Outputs is better) Constrained on a fixed 5 rounds of refinement rather than the stopping condition, the ablated Entrospect falls into suboptimal performance in contrast to the complete version across both tasks and all involved models. This highlights the significance and efficacy of importing semantic similarity comparison as the stopping condition for our framework.

may introduce unnecessary or incorrect changes during the refinement cycles, especially for SLMs, as mentioned in Section 1.

Ablated Entrospect: The variant of Entrospect without the semantic similarity-based stopping condition. Instead, a manual setting of 5 fixed iterations is assigned. This baseline serves as the *ablation study* that verifies the efficacy of our nominated convergence policy.

4.5 Results

We report the Entrospect's state-of-the-art competences versus the baseline prompting approaches, especially Self-Refine, in augmenting the SLMs' semantic reasoning across two validation tasks.

Entrospect improves reasoning accuracies: Displayed in Table 1 and 3, SLMs armed with Entrospect outshines all other baselines pertaining to the reasoning accuracies across both MATH and HaluEval validation sets. In contrast specifically to Self-Refine, Entrospect yields a maximum improvement of 36.2% *with Llama 3.1 Instruct*

8B (44.3% → 80.5%) on the MATH dataset and 15.5% with *DeepSeek-R1-Distilled Qwen 2.5 Instruct 1.5B* (80.0% → 95.5%) on the HaluEval dataset. Moreover, Figure 7 highlights Entrospect's robustness beyond handling math problems with a fixed complexity. When set against Self-Refine, Entrospect consistently offers more substantial mitigation against the overall degradation of reasoning accuracy as the problem difficulty rises, securing a reduced decay rate as much as 52.8%.

The exceptional computational efficiency: As depicted in Figure 4, Entrospect reaches convergence faster than Self-Refine across most instances. on the MATH dataset, Entrospect reduces runtime by an average factor of up to 2.8 (e.g., *Llama 3.1 8B + Entrospect*), meanwhile demonstrating even more pronounced efficiency gains on the HaluEval dataset, with runtime reductions reaching up to 10-fold (e.g., *DeepSeek R1-Distill Qwen 2.5 1.5B + Entrospect*). Beyond its efficiency advantages, Figure 4 highlights Entrospect's ability to strike a superior balance between computational efficiency and accuracy, driving substantial overall performance enhancements in SLMs.

To investigate potential correlations between model parameter sizes and the ATC outcomes achieved by Entrospect, we employed Spearman's rank correlation coefficient alongside corresponding p-values (Spearman, 2010). However, no statistically significant relationship was observed within the scope of our experiments (MATH: $\mathrm{corr} = -0.600$, $p = 0.400$; HaluEval: $\mathrm{corr} = -0.200$, $p = 0.800$).

Ablation study: To validate whether the semantic similarity-based stopping condition is crucial for propelling a higher reasoning accuracy of Entrospect, we conducted an ablation study by removing this mechanism and fixing the number of refinement cycles to 5. Figure 5 illustrates that the ablated Entrospect constantly underperforms compared to the complete implementation, witnessing performance deficits of 1.8 → 8.9% on the MATH dataset and 1.5% → 3.5% on the HaluEval dataset across all tested SLMs. The results solidify the role of the semantic similarity-guided stopping condition as a cornerstone for enhancing Entrospect's overall performance.

5 Conclusion

This paper introduces Entrospect, an optimized Self-Refine framework that leverages an information-theoretic Optimal Revision Suggestion Selector to provide optimal revision suggestions during the self-reflection stage while eliminating ineffective ones for efficient refinement of initial responses from SLMs. Besides, the convergence of Entrospect is made possible with a dedicated semantic similarity-determined stopping condition. Through our holistic evaluations, Entrospect claimed state-of-the-art relative to the baseline methods on both of our reasoning tasks across four SLMs of diverse parameter sizes, obtaining a maximum 36.2% reasoning accuracy uplift and at most 10 times the computational efficiency exclusively over its antecedent, Self-Refine.

We aspire for this study to inspire further advancements in small language models research and furnishes new perspectives for information-theoretic prompt engineering.

Limitations

There remains much room for promoting Entrospect, and our future studies shall prioritize the following key limitations:

More solid definition of an *optimal* revision suggestion: The ORSS of Entrospect, grounded in maximizing the conditional mutual information, operates as an approximate selection technique in unsupervised settings. This approach gauges the quality of a revision suggestion by leveraging the model's intrinsic output uncertainty as a pivotal determinant. However, its reliability is compromised when the model demonstrates undue confidence in erroneous outputs. As a result, it is imperative to pursue a more precise and theoretically grounded definition of what constitutes an *optimal* revision suggestion in our future studies.

Beyond semantic similarity comparison as the stopping condition: A high semantic similarity between consecutive refinement iterations as a sign of convergence is logically aligned with language models adopting greedy search sampling. In conversational situations, however, sampling methods such as Top-K and nucleus sampling are more regularly used to ensure generative variability. Our future work will seek to modify the current convergence mechanism tailored to these sampling configurations.

Ethics Statement

This study strictly adheres to the Ethical Policy of the Association for Computational Linguistics. We conducted a thorough assessment of the potential impacts of our research and did not identify any evident ethical concerns. All datasets utilized in this study were sourced from publicly available resources and were handled strictly in accordance with their respective terms of use. Nevertheless, we acknowledge the possibility of unforeseen impacts in any research and invite readers to share feedback on any potential ethical concerns they may identify.

References

Meta AI. 2024. Llama 3.1: The most capable open foundation models. https://ai.meta.com/blog/meta-llama-3-1/.

Jing Bi, Yuting Wu, Weiwei Xing, and Zhenjie Wei. 2024. Enhancing the reasoning capabilities of small language models via solution guidance fine-tuning. *arXiv preprint arXiv:2412.09906*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Ruoxi Cheng, Haoxuan Ma, and Shuirong Cao. 2024. Deceiving to enlighten: Coaxing llms to self-reflection for enhanced bias detection and mitigation. *arXiv preprint arXiv:2404.10160*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332.

Simon Frieder, Mirek Olsák, Julius Berner, and Thomas Lukasiewicz. 2024. The imo small challenge: Not-too-hard olympiad math datasets for llms. In *Tiny Papers@ ICLR*.

Yao Fu, Hao Peng, Litu Ou, Ashish Sabharwal, and Tushar Khot. 2023. Specializing smaller language models towards multi-step reasoning. In *International Conference on Machine Learning*, pages 10421–10430. PMLR.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chengcheng Han, Xiaowei Du, Che Zhang, Yixin Lian, Xiang Li, Ming Gao, and Baoyuan Wang. 2023. Dialcot meets ppo: Decomposing and exploring reasoning paths in smaller language models. *arXiv preprint arXiv:2310.05074*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*.

Dongyub Lee, Eunhwan Park, Hodong Lee, and Heui-Seok Lim. 2024. Ask, assess, and refine: Rectifying factual consistency and hallucination in llms with metric-guided feedback learning. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2422–2433.

Beibin Li, Yi Zhang, Sébastien Bubeck, Jeevan Pathuri, and Ishai Menache. 2024. Small language models for application interactions: A case study. *arXiv preprint arXiv:2405.20347*.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Qiwei Pan. 2023. Autogptq: An easy-to-use llms quantization package with user-friendly apis, based on gptq algorithm.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2024. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36.

C Spearman. 2010. The proof and measurement of association between two things. *International Journal of Epidemiology*, 39(5):1137–1150.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, et al. 2024. jina-embeddings-v3: Multilingual embeddings with task lora. *arXiv preprint arXiv:2409.10173*.

Fali Wang, Zhiwei Zhang, Xianren Zhang, Zongyu Wu, Tzuhao Mo, Qiuhao Lu, Wanjing Wang, Rui Li, Junjie Xu, Xianfeng Tang, et al. 2024. A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness. *arXiv preprint arXiv:2411.03350*.

Tianduo Wang and Wei Lu. 2023. Learning multi-step reasoning by solving arithmetic tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1229–1238.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359*.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, et al. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 214–229.

Junda Wu, Tong Yu, Rui Wang, Zhao Song, Ruiyi Zhang, Handong Zhao, Chaochao Lu, Shuai Li, and Ricardo Henao. 2024. Infoprompt: Information-theoretic soft prompt tuning for natural language understanding. *Advances in Neural Information Processing Systems*, 36.

Bin Xiao, Burak Kantarci, Jiawen Kang, Dusit Niyato, and Mohsen Guizani. 2024. Efficient prompting for llm-based generative internet of things. *arXiv preprint arXiv:2406.10382*.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Zhaozhuo Xu, Zirui Liu, Beidi Chen, Yuxin Tang, Jue Wang, Kaixiong Zhou, Xia Hu, and Anshumali Shrivastava. 2023. Compress, then prompt: Improving accuracy-efficiency trade-off of llm inference with transferable prompt. *arXiv preprint arXiv:2305.11186*.

Tianqiang Yan and Tiansheng Xu. 2023. Refining the responses of llms by themselves. *arXiv preprint arXiv:2305.04039*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024a. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Sohee Yang, Jonghyeon Kim, Joel Jang, Seonghyeon Ye, Hyunji Lee, and Minjoon Seo. 2024b. Improving probability-based prompt selection through unified evaluation and analysis. *Transactions of the Association for Computational Linguistics*, 12:758–774.

Xiao Yu, Baolin Peng, Michel Galley, Jianfeng Gao, and Zhou Yu. 2023. Teaching language models to self-improve through interactive demonstrations. *arXiv preprint arXiv:2310.13522*.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024a. Self-contrast: Better reflection through inconsistent solving perspectives. *arXiv preprint arXiv:2401.02009*.

Wenqi Zhang, Ke Tang, Hai Wu, Mengna Wang, Yongliang Shen, Guiyang Hou, Zeqi Tan, Peng Li, Yueting Zhuang, and Weiming Lu. 2024b. Agent-pro: Learning to evolve via policy-level reflection and optimization. *arXiv preprint arXiv:2402.17574*.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

A Appendix

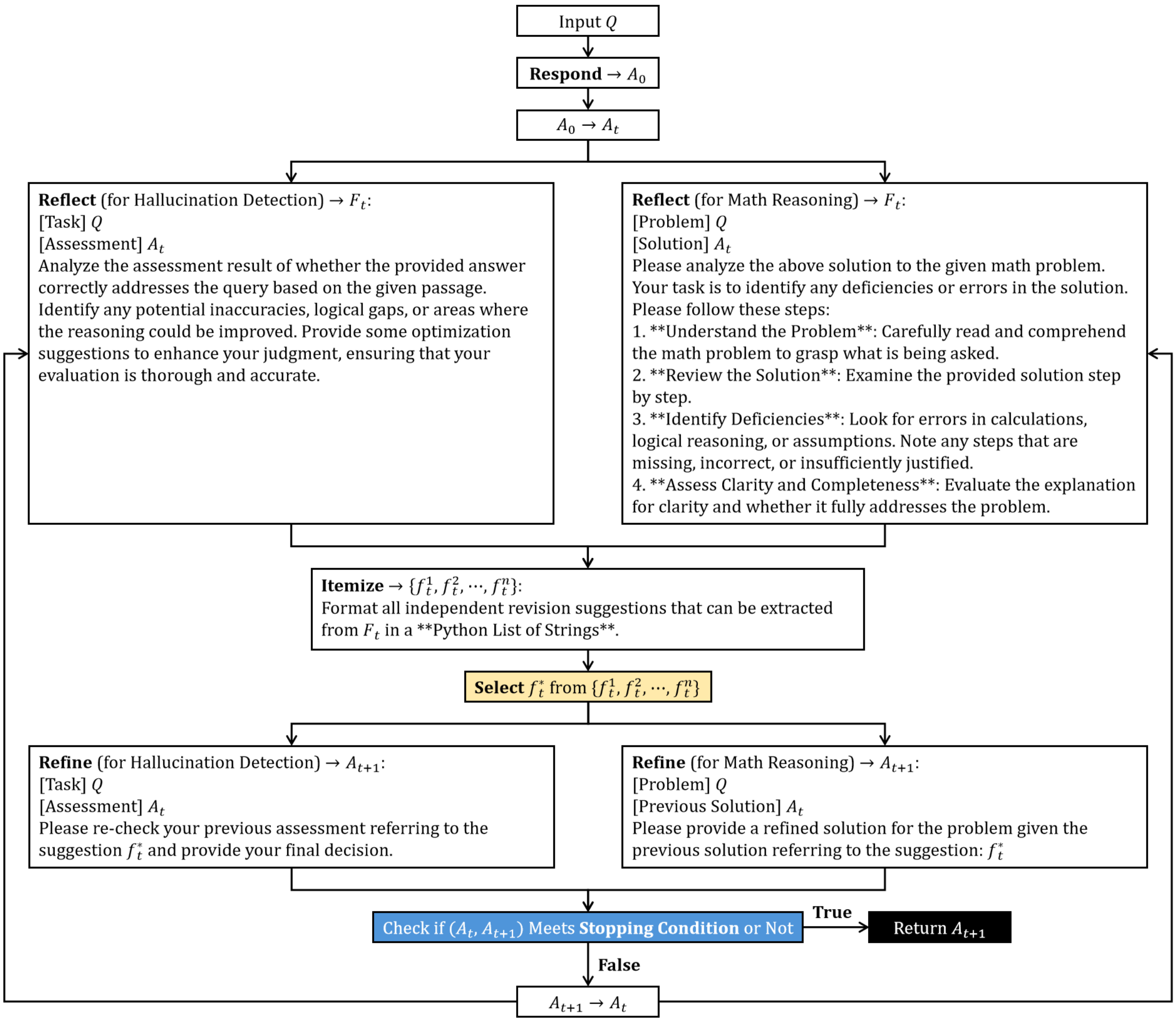

Figure 6: (Referred in Section 3.4) The detailed instructions used for all prompting nodes (modules) within the Entrospect framework during the evaluation phases. These instructions guide the SLMs through the process of generating an initial response, reflecting on its deficiencies, selecting the optimal revision, and refining the response based on the selected suggestion.

Table 2: (Referred in Section 4.2) Representative data samples from the MATH and HaluEval datasets, demonstrating a mathematical reasoning problem and a reading comprehension task.

Dataset	Query	Label
MATH	What is the simplified numerical value of $\frac{a+11b}{a-b}$ if $\frac{4a+3b}{a-2b}=5$?	Let's play with the given condition a little. Clearing out the denominator gives $4a+3b=5(a-2b)=5a-10b$. Selectively combine like terms by adding $9b-4a$ to both sides to get $12b=a-b$. This gives $\dfrac{12b}{a-b}=1$. Now, we want to find $\dfrac{a+11b}{a-b}$. Rewrite this as $\dfrac{a-b+12b}{a-b}=\dfrac{a-b}{a-b}+\dfrac{12b}{a-b}=1+1=\boxed{2}$, and we are done.
HaluEval	The following is a reading comprehension task, which provides a passage, a question related to the passage, and an answer to the question: [Passage] The ValleyCats play at Joseph L. Bruno Stadium which opened in 2002 on the campus of Hudson Valley Community College located in Troy. Joseph Bruno Stadium is a stadium located on the campus of Hudson Valley Community College in Troy, New York. [Question] The Tri-City ValleyCats play at which stadium located on the campus of Hudson Valley Community College in Troy, New York? [Answer] Troy Community Stadium, located on Hudson Valley Community College campus. Please determine whether the given answer is correct. If it is correct, output 'PASS'; if it is incorrect, output 'FAIL'.	FAIL

Table 3: (Referred in Section 4.5) The extended table of accuracies(%) on the MATH dataset, providing a detailed breakdown of all results across Level 1 to Level 5, where Entrospect performs the best with all SLMs relative to the baseline prompting methods across all difficulty levels. We highlight the best results in **bold**.

Model Name	Method	MATH-L1	MATH-L2	MATH-L3	MATH-L4	MATH-L5
DeepSeek-R1-Distilled Qwen 2.5 Instruct 1.5B	Zero-Shot	97.5	95.0	96.7	91.7	90.0
	5-Shot	96.7	92.5	94.2	91.7	75.8
	Zero-Shot CoT	75.0	98.3	98.3	93.3	91.7
	Self-Refine	90.0	89.2	90.8	88.3	84.2
	Entrospect	**99.2**	**99.2**	**99.2**	**96.7**	**97.5**
Qwen 2.5 Instruct 7B	Zero-Shot	91.7	92.5	85.8	73.3	47.5
	5-Shot	90.8	92.5	81.7	62.5	36.7
	Zero-Shot CoT	95.0	93.3	91.7	79.2	60.0
	Self-Refine	85.0	89.2	82.5	65.8	42.5
	Entrospect	**95.0**	**95.8**	**91.7**	**84.2**	**63.3**
Llama 3.1 Instruct 8B	Zero-Shot	87.5	74.2	60.8	48.3	37.5
	5-Shot	88.3	70.0	62.5	41.7	20.0
	Zero-Shot CoT	91.7	83.3	77.5	65.8	50.0
	Self-Refine	72.5	54.2	43.3	28.3	23.3
	Entrospect	**95.0**	**88.3**	**84.2**	**74.2**	**60.8**
GLM 4 Instruct 9B	Zero-Shot	82.5	66.7	55.0	46.7	24.2
	5-Shot	86.7	66.7	66.7	44.2	25.0
	Zero-Shot CoT	90.0	81.7	75.8	51.7	30.0
	Self-Refine	85.0	69.2	63.3	45.0	21.7
	Entrospect	**92.5**	**85.8**	**79.2**	**56.7**	**34.2**

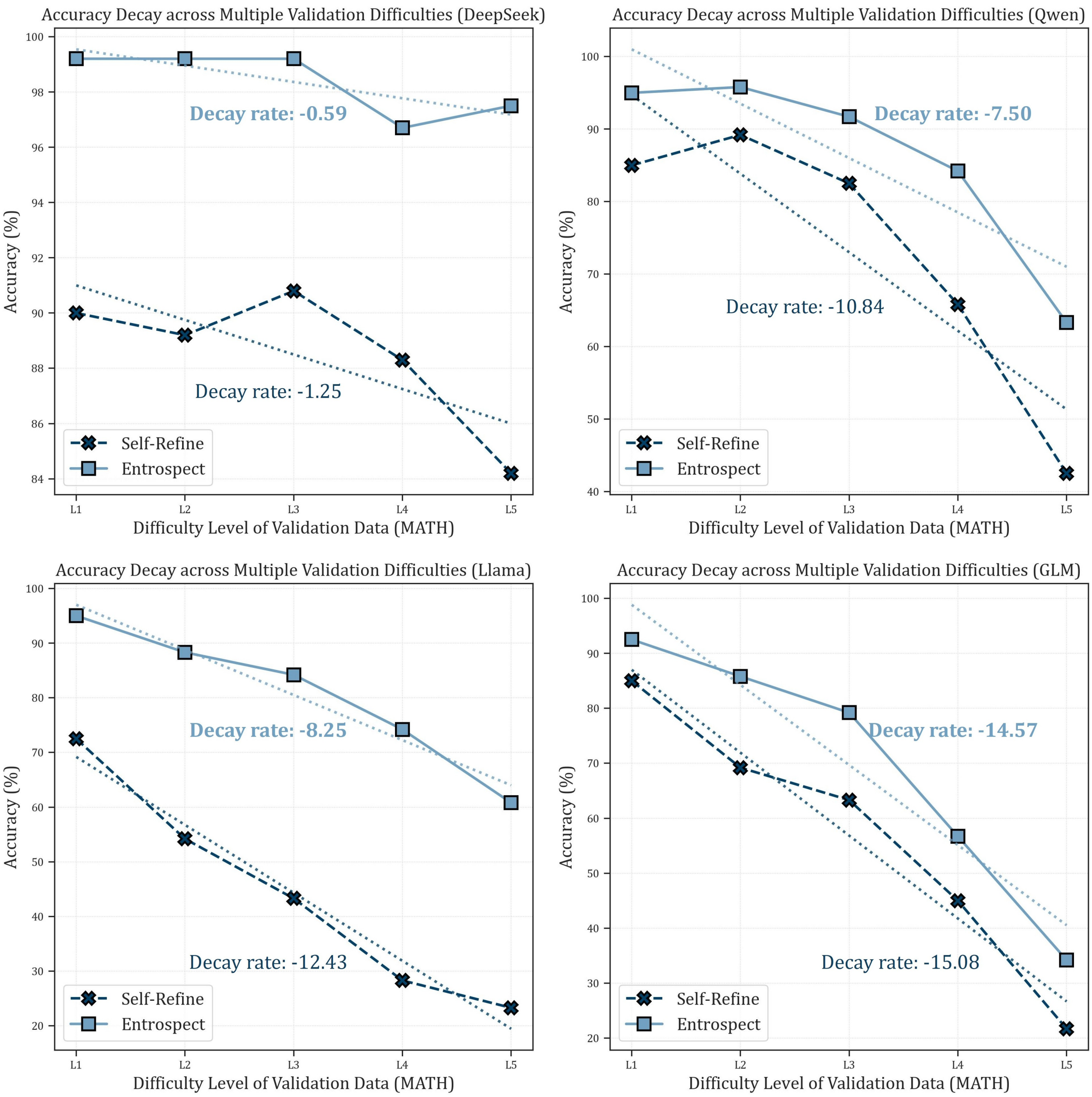

Figure 7: (Referred in Section 4.5) We employed linear regression to model the decline in reasoning accuracy, as measured by Entrospect and Self-Refine on the MATH validation set with increasing difficulty levels. The four charts correspond to the four distinct SLMs we evaluated, where the *decay rate* equals the slope of each fitted decay line. A decay rate with a larger absolute value indicates a more rapid deterioration in reasoning accuracy as the difficulty level rises. Across all tested models, observations indicate that as the difficulty level of the test data increases, the performance degradation exhibited by Entrospect is, overall, less pronounced than that of Self-Refine.