

000  
001  
002  
003  
004  
005  
006  
007  
008  
009  
010  
011  
012  
013  
014  
015  
016  
017  
018  
019  
020  
021  
022  
023  
024  
025  
026  
027  
028  
029  
030  
031  
032  
033  
034  
035  
036  
037  
038  
039  
040  
041  
042  
043  
044  
045  
046  
047  
048  
049  
050  
051  
052  
053  

# SID: MULTI-LLM DEBATE DRIVEN BY SELF SIGNALS

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Large Language Models (LLMs) have exhibited impressive capabilities across diverse application domains. Recent work has explored Multi-LLM Agent Debate (MAD) as a way to enhance performance by enabling multiple LLMs to discuss and refine responses iteratively. Nevertheless, existing MAD methods predominantly focus on utilizing external structures, such as debate graphs, using LLM-as-a-Judge, while neglecting the application of self signals, such as token logits and attention, that arise during generation. This omission leads to redundant computation and potential performance degradation. In this paper, we shift the focus to the self signals of multi-LLM debate and introduce a Self-Signals Driven Multi-LLM Debate (SID), which leverages two types of self-signals: model-level confidence and token-level semantic focus, to adaptively guide the debate process. Our approach enables high-confidence agents to exit early at the model level and compress the redundant debate contents based on the attention mechanism. We evaluate our method on various LLMs and Multimodal LLMs across multiple challenging benchmarks. Experimental results demonstrate that our method not only outperforms existing MAD techniques in accuracy but also reduces token consumption, highlighting the effectiveness of utilizing self signals in enhancing both the performance and efficiency of multi-agent debate systems.

## 1 INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable capabilities (Brown et al., 2020; Kojima et al., 2023) across a wide range of domains, including science, technology, engineering, mathematics (STEM) questions (Hendrycks et al., 2021; Wang et al., 2024), and complex reasoning tasks (Rein et al., 2023). The emergence of Multimodal LLMs (MLLMs) further extends the potential to the visual input domain (Lu et al., 2022; Li et al., 2023; Liu et al., 2024a). However, current models still suffer from inherent limitations such as inaccuracies and hallucinations.

Multi agent debate (MAD) offers an orthogonal approach to enhancing model performance, in which multiple agents iteratively discuss and refine their answers accordingly (Du et al., 2024; Liu et al., 2024c; Sun et al., 2025). However, a challenge arises from the prevalence of redundant content and repeated consensus points during debate, which not only waste computational resources but also introduce informational noise, potentially impairing the agents' final judgments (Du et al., 2024; Li et al., 2024b). Moreover, this iterative discussion paradigm incurs substantial token overhead, which becomes increasingly incongruent with the growing capabilities of modern foundation models (OpenAI et al., 2025). This inherent contradiction between performance gains and token consumption cost presents a central dilemma in contemporary MAD research.

To alleviate this problem, several optimization strategies have been proposed. Broadly, these methods typically fall into two categories: (i) structural optimization, such as adopting various prompting skills (Liu et al., 2024c), reducing communication via sparse debate graphs or clustering agents into local debate groups (Liu et al., 2024b); and (ii) history management, including summarization of prior discussions or introducing agent self-generated confident score (Sun et al., 2025). Whereas these approaches improve the efficiency of information flow in *external* ways (*i.e.*, restructuring agent communication or using LLM-as-a-judge to interpret history), they often suffer from secondary errors such as hallucinations in judges or summaries as evident in (Xiong et al., 2023; Zhang et al., 2024; Tian et al., 2025)). This limitation motivates us to think: *can we avoid relying on error-prone external mechanisms, and instead leverage more reliable self signals from each agent's generative process to prevent unnecessary and potentially wasteful debate?*

Motivated by the above, in this work, we present a framework that leverages self signals available during LLM inference to improve debate efficiency and performance. In this framework, two types of signals: *model-level confidence* and *token-level semantic focus*, are extracted and used to provide complementary guidance for distinguishing essential information from redundancy, thereby enhancing overall debate quality and efficiency. The model-level confidence, estimated from the probability distribution over the initially generated answer, quantifies how certain the model is about its response. We leverage this signal to design an ***early-exit mechanism*** that avoids invoking debate when the model is already sufficiently confident, thereby reducing potential noise and redundancy. The token-level semantic focus, derived from attention patterns conditioned on disagreement-oriented prompts, identifies spans in the debate content that the model considers semantically relevant to the disagreement among different agents. We extract and reconstruct these high-attention spans to form a more compact context, thereby introducing a ***compression mechanism*** that preserves critical points of contention while significantly reducing token overhead.

By integrating these two mechanisms, each leveraging a different level of self signal, we propose a unified Self Signal Driven Debate framework (SID) to enhance LLM performance. This framework enables early exit for confident agents and extracts focused context for the remaining ones, dynamically adapting the debate process based on the model’s own epistemic signals. We evaluate our method across multiple LLMs and MLLMs on diverse benchmarks, including MMLUpro, Math, GPQA, ScienceQA, and MMstar. SID consistently outperforms existing MAD approaches in most scenarios, while also achieving up to a 40% reduction in token consumption. These results demonstrate the strong effectiveness of our approach and highlight the significant potential of leveraging internal belief signals in multi-agent systems to jointly optimize performance and efficiency. Our key contributions can be summarized as follows:

- We present **SID**, a multi-agent debate framework that leverages self signals from the LLM generation process to enhance agent debate.
- We instantiate two types of LLM self signals: model-level confidence and token-level semantic focus, and leverage them to design an early-exit and a compression mechanism, respectively, effectively reducing redundancy and enhancing debate performance.
- Integrating the two proposed mechanisms, we construct an effective and efficient debate framework, SID. Experiments across multiple benchmarks, on both LLMs and MLLMs, demonstrate the significant advantages of SID over existing methods.

## 2 RELATED WORK

**Reasoning Augmentation** To enhance the reasoning capabilities of LLMs, researchers have explored various techniques. Early work primarily focused on guiding the model through step-by-step reasoning through Chain-of-Thought (CoT) prompts (Wei et al., 2023) or generating multiple reasoning paths (self-refinement) and voting for the optimal solution through self-consistency or using multi-round self-reflection (Zhang et al., 2024; Yao et al., 2023). Additionally, subsequent research has found that the model’s self-correction capabilities are limited, leading to stagnation in reasoning quality (Zhang et al., 2024). This has partially motivated the rise of multi-agent collaborative paradigms, particularly multi-agent debate (MAD) (Du et al., 2024), which introduces external perspectives and dynamic feedback among agents to overcome the limits of self-reflection. Our work differs from these studies in that they focus primarily on improving reasoning ability through context prompts, whereas we propose to use self-signals from a model to optimize the context prompt at the token level, thus improving the effectiveness of performance and token ratio.

**Uncertainty Analysis** Uncertainty in LLMs is typically categorized into aleatoric (data-related) and epistemic (model-related) uncertainty (Kiureghian & Ditlevsen, 2009; Gawlikowski et al., 2023; Hu et al., 2023; Ye et al., 2025). Given the structured nature of current tasks (e.g., QA, math, science), recent works have focused on quantifying epistemic uncertainty. **Mainstream approaches include:** (i) probability-based metrics, such as token-level entropy or negative log-likelihood (Tu et al., 2025) on the level of the attention layer (Schuster et al., 2022; Laaouach, 2025; Corallo & Papotti, 2024) and reasoning chain (Yang et al., 2025); (ii) ensemble-based methods, e.g., Monte Carlo sampling (Metropolis et al., 1953; Hastings, 1970) and Bayesian methods (Kwon et al., 2020); and (iii) verbalization-based techniques that prompt the model to self-report confidence (Tian et al., 2023).

108 Among these, probability-based methods are especially attractive due to their seamless integration  
 109 with the generation process, without requiring multiple generations, which incur significant token  
 110 overhead. Recent work, such as ReConcile and DebUnc (Xiong et al., 2023; Chen et al., 2024a;  
 111 Yoffe et al., 2025; Kirchhof et al., 2025) further explores agent-level uncertainty in interactive  
 112 settings, emphasizing the role of uncertainty as a confident signal for learning and output control. Our  
 113 method differs from theirs in model-level adaptive debate scheduling and token-level compression  
 114 instead of passing the full history. Our method aligns with the uncertainty in the interactive setting,  
 115 leveraging self-signals as dynamic indicators of agent-level uncertainty to control agent participation  
 116 during debates.

117 **Multi LLM Debate Systems** Previous multi-LLM debate employs a role-playing setup (Liang  
 118 et al., 2024), which has been demonstrated strengths in collaborative tasks. Subsequent research  
 119 has shown that it is less suited for certain types of problem-solving scenarios. Multi Agent Debate  
 120 (MAD) (Du et al., 2024) introduces external perspectives to enrich the system’s reasoning capabilities.  
 121 DMAD (Liu et al., 2024c) proposes specialized prompt strategies to diversify agent behavior.  
 122 S2-MAD (Zeng et al., 2025) introduces a selective sparsity mechanism, allowing agents to selec-  
 123 tively participate based on internal cues. CortexDebate (Sun et al., 2025) constructs a dynamic sparse  
 124 debate graph by letting agents serve as self-judges and output confidence scores. These works focus  
 125 on improving performance via external states (e.g., optimizing structures, using LLM-as-a-Judge).  
 126 **Compared to previous approaches (Zhang et al., 2025), our method can be orthogonal and comple-  
 127 mentary, which provides a new angle by integrating self-signals into the debate process instead of  
 128 optimizing communication structures.**

### 3 PRELIMINARIES

131 We first introduce the naive multi agent debate paradigm in this section. Let  $\mathcal{V}$  denote the vocabulary  
 132 and  $\text{Tok}$  the tokenizer. Given a query (e.g., natural language, image, and text)  $Q$ ,  $\mathbf{x} = \text{Tok}(Q)$   
 133 is the tokenized prompt. An casual LLM  $M$  produces a response sequence  $\mathbf{y} = (y_1, \dots, y_m)$   
 134 with per-step logits  $\ell_t \in \mathbb{R}^{|\mathcal{V}|}$  and probabilities  $\pi_t = \text{softmax}(\ell_t)$ . A debate involves  $n$  agents  
 135  $\mathcal{A} = \{1, \dots, n\}$  over rounds  $t = 0, 1, \dots, T$ . Let  $\mathbf{y}_t^{(j)}$  be agent  $j$ ’s response at round  $t$ ; round 1  
 136 is the initial answering round without debate context. The per-round input to agent  $j$  at round  $t+1$   
 137 concatenates the query, its own last response, and other agents’ last responses:

$$140 \quad \mathbf{x}_{t+1}^{(j)} = \text{Tok}\left(Q \parallel \mathbf{y}_t^{(j)} \parallel (\text{Concat}_{k \neq j} \mathbf{y}_t^{(k)})\right). \quad (1)$$

142 Here, both  $\parallel$  and  $\text{Concat}$  represent concatenation between prompt groups.

### 4 METHOD

144 The above naive framework suffers from several issues, such as excessive redundancy and low effi-  
 145 ciency. To address these challenges, as shown in Algorithm 1 and Figure 1, we propose Self Signal  
 146 Driven Debate (SID), a framework that leverages internal confidence signals readily available dur-  
 147 ing inference to adaptively guide the multi-LLM debate process. Specifically, SID utilizes two types  
 148 of self signals from the LLM: *model-level confidence* and *token-level semantic focus* (see exam-  
 149 ples in Figure 3 and the Appendix F). *Model-level confidence*, derived from the token-wise output  
 150 probability distribution (logits), reflects how confident an agent is in its initial answer. We leverage  
 151 this signal in a newly designed early-exit mechanism to enhance debate efficiency. *Token-level se-  
 152 mantic focus*, extracted from the self-attention maps conditioned on disagreement-oriented prompts,  
 153 captures regions of high variability and knowledge density throughout the debate. This signal is  
 154 incorporated into a compression mechanism to alleviate token redundancy. In the following two  
 155 sections, we introduce these two components in detail.

#### 4.1 EARLY-EXIT WITH MODEL-LEVEL CONFIDENCE

158 We first introduce an early-exit mechanism to mitigate redundant debate, motivated by the intuition  
 159 that cross-LLM discussion is more necessary when a single model lacks confidence in its response.

```

162
163 Algorithm 1 Self-Signal Driven Debate (SID)
164 Require: Query  $Q$ ; LLM agents  $\{M_j\}_{j=1}^m$ ; rounds  $N$ ; confidence threshold
165  $\theta$ ; top- $p$  ratio  $\rho$ ;
166 1:  $Y_1 \leftarrow M_1(Q)$  ▷ Model-level Confidence(Sec. 4.1)
167 2:  $u \leftarrow \phi_U(U(Y_1))$ 
168 3: if  $u \leq \theta$  then
169 4:   return  $Y_1$  ▷ early exit
170 5: end if
171 6:  $Y_1 \leftarrow \{y_1\}$ ;  $y_1^{(j)} \leftarrow M_j(Q)$  for  $j=2..m$ 
172 7: for  $t = 2$  to  $N$  do
173 8:   for  $j = 1$  to  $m$  do
174 9:      $X_t^{(j)} \leftarrow (Q \parallel y_{t-1}^{(j)} \parallel [\text{PROMPT}] \parallel \text{Concat}_{k \neq j}(Y_{t-1}^k))$ 
175 10:     $A_t^{(j)} \leftarrow \text{FORWARDATTENTION}(M_j, X_t^{(j)})$  ▷ forward only
176 11:     $\hat{C}_t^{(j)} \leftarrow \text{TopP}(A_t^{(j)}, \rho)$ 
177 12:     $S_t^{(j)} \leftarrow \text{SemanticPreserve}(\hat{C}_t^{(j)})$  ▷ Sec. 4.2
178 13:     $y_t^{(j)} \leftarrow M_j(Q \parallel y_{t-1}^{(j)} \parallel S_t^{(j)})$  ▷ generate with compressed
179 14:    context
180 15:   end for
181 16: end for
182 17: return  $Y_N$ 
183
184
185
186
187
188
189
190
191
192
193

```

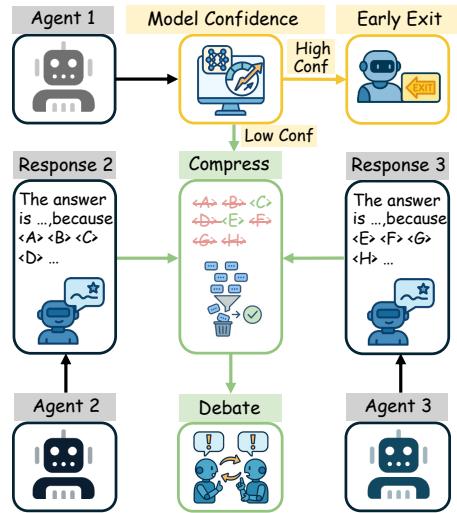


Figure 1: Overall framework of SID

The key to this mechanism lies in extracting an effective confidence score. Intuitively, the more peaked the model’s output distribution over the vocabulary (i.e., lower entropy), the more confident it is in its prediction. Based on this motivation and following conventional methods (Tu et al., 2025), we adopt two token-wise uncertainty metrics: entropy  $\text{Ent}(\pi_t) = -\sum_{v \in \mathcal{V}} \pi_{t,v} \log \pi_{t,v}$  and negative log-likelihood  $\text{NLL}_t = -\log \pi_{t,y_t}$ , to estimate the confidence for a generated answer  $y$ . To aggregate token-level metrics into a sequence-level confidence score, we explore four aggregation strategies: (1) averaging over tokens (average), (2) taking the maximum value (max), (3) using the first token’s value (first), and (4) using the penultimate token’s value (penultimate). This yields eight confidence measures (four for entropy and four for NLL). Each variant captures different facets of uncertainty: e.g., max focuses on worst-case ambiguity, while penultimate emphasizes late-stage uncertainty often aligned with final reasoning steps in autoregressive generation. We concatenate these eight measures to form a vector  $U(y)$ , which we empirically find to be statistically significant in distinguishing incorrect answers (see Figure 2 for details).

After obtaining confidence scores, the next challenge lies in leveraging them to effectively guide model-level early exits during multi-LLM debates. To this end, we propose two different drop-in strategies that convert the agent’s confidence vector into a binary decision boundary:

**Vocabulary-Adaptive Threshold** We first tried a straightforward method by directly setting a fixed threshold on the confidence metrics across different types of models. However, this naive strategy yielded suboptimal performance, likely due to the inherent dependency of entropy and NLL magnitudes on the vocabulary size  $|\mathcal{V}|$  of the underlying LLM. For example, under a uniform generation assumption, both entropy and NLL equal  $\log |\mathcal{V}|$ . Thus, larger vocabularies naturally induce higher values, while using a single threshold across models leads to unfairness and unreliability. Based on this motivation, we propose a vocabulary-adaptive threshold as follows:

$$\theta(V) = \alpha \log |\mathcal{V}|, \quad \text{and decide Terminate iff } \phi_U(U(y)) \leq \theta(V), \quad (2)$$

where  $\alpha > 0$  is a hyper-parameter,  $\phi_U$  is an operator to filter noisy metrics  $U(y)$ . This strategy ensures fair and consistent confidence evaluation across models with varying vocabulary sizes.

**Calibrated Confidence** While the aforementioned method provides a simple and robust solution, it relies on a uniformity assumption over token distributions that may not hold in practice. To capture more nuanced confidence signals, we introduce an alternative method, using a lightweight nonlinear classifier  $C : \mathbb{R}^d \rightarrow [0, 1]$  trained over a small held-out set. This model takes the confidence vector as input and outputs a scalar confidence score, calibrated against correctness labels:

$$\text{Terminate iff } C(U(y)) \geq \tau_c, \quad \tau_c \in (0, 1). \quad (3)$$

216 **Confidence-Guided Early Exit** In practice, we adopt the vocabulary-adaptive threshold for gating  
 217 due to its sufficiently strong performance and training-free simplicity. Specifically, gating is  
 218 applied in the first round: if an agent’s confidence reaches a high value, it is terminated early, sig-  
 219 nalling that the system is already sufficiently confident in its answer (see Appendix Figure 17–20  
 220 for examples). Conversely, if the model exhibits low confidence, this suggests that the question is  
 221 sufficiently challenging and unlikely to be resolved without additional reasoning, thereby motivating  
 222 the initiation of a multi-agent debate with the input described in Eq.1.

## 223 4.2 ADAPTIVE COMPRESSION WITH TOKEN-LEVEL SEMANTIC FOCUS

225 In addition to model-level confidence, we further exploit another self signal from the LLM: token-  
 226 level semantic focus, to improve debate efficiency. As a debate progresses, the accumulated context  
 227 from multiple agents often becomes repetitive and redundant. We observe that this increasing re-  
 228 dundancy can dilute the signal-to-noise ratio, potentially degrading the effectiveness and efficiency  
 229 of the debate process. A common approach to mitigate this is to use LLM-as-a-judge to sum-  
 230 marize past exchanges. However, this method is limited by the summarization capabilities of the base  
 231 models, which can be prone to hallucination or information loss (Li et al., 2024a). To address this  
 232 limitation, we instead leverage attention, an intrinsic mechanism of transformer-based models that  
 233 naturally reflects the model’s focus and the debate’s salient region, as an internal signal to implement  
 234 a token compression framework.

235 **Prompt-conditioned Attention Extraction** Given the query  $Q$ , the agent  $j$ ’s previous answer  
 236  $\mathbf{y}_t^{(j)}$ , and other agents’ responses  $\{\mathbf{y}_t^{(k)} : k \neq j\}$ , we construct a concatenated input:

$$239 \quad \mathbf{x}_{t+1}^{(j)} = \text{Tok} \left( Q \parallel \mathbf{y}_t^{(j)} \parallel [\text{PROMPT}] \parallel (Concat_{k \neq j} \mathbf{y}_t^{(k)}) \right), \quad (4)$$

241 where  $[\text{PROMPT}]$  is a task instruction. Here, we use the prompt: "Identify the key  
 242 points where they disagree with your own reasoning. Concentrate  
 243 on those disagreements and decide which line of reasoning is  
 244 better.", motivated by prior work demonstrating the benefits of identifying disagreements  
 245 among agents (Du et al., 2024; Zeng et al., 2025). This prompt directs the model’s attention  
 246 toward segments of the debate that involve semantic conflict, thereby enhancing its focus on  
 247 critical reasoning divergences. We then define  $\mathcal{Q}$  as the set of token positions within the injected  
 248 prompt,  $\mathcal{C}$  as the positions corresponding to other agents’ responses (i.e.,  $Concat_{k \neq j} \mathbf{y}_t^{(k)}$ ), and  
 249  $A^{(l,h)} \in [0, 1]^{L \times L}$  as the attention score at layer  $l$  and head  $h$ . For  $c \in \mathcal{C}$ , a prompt-conditioned  
 250 semantic focus score is computed by:

$$252 \quad s(c) = \max_{l,h} \max_{q \in \mathcal{Q}} A_{q,c}^{(l,h)}, \quad (5)$$

254 which represents the maximum attention weight from any prompt token to  $c$  across all heads and  
 255 layers, capturing the extent to which  $c$  is considered relevant to the disagreement-focused instruction.

257 **Compression with Semantic Preservation** While token-level attention scores  $s(c)$  enable fine-  
 258 grained identification of salient contents, directly selecting individual tokens may result in frag-  
 259 mented phrases or broken sentence structures. Such fragments hinder the model’s ability to interpret  
 260 the compressed input coherently. To address this, we apply a semantic preservation heuristic that  
 261 extends high-attention tokens to complete sub-sentential units. Concretely, we first select the top-  
 262  $p$  fraction of context tokens, forming  $\hat{\mathcal{C}} = \text{Top}_p \{(c, s(c))\}_{c \in \mathcal{C}}$ . Using the tokenizer’s offset map  
 263  $\Psi : c \mapsto [T_a(c), T_b(c)]$ , we merge overlapping spans and then expand to sentence boundaries to  
 264 preserve semantics information. To ensure semantic completeness, we then expand each segment to  
 265 align with syntactic boundaries, such as commas, periods, or coordinating conjunctions. We denote  
 266 this process as the SemanticPreserve operation (see Appendix C.1, Figure 6 for implementation  
 267 details), which produces a minimal set of semantically coherent text spans as follows:

$$268 \quad \mathcal{S} = \text{SemanticPreserve} \left( \text{Merge} \left( \{\Psi(c) : c \in \hat{\mathcal{C}}\} \right) \right). \quad (6)$$

270 We denote the compressed textual summary for agent  $j$  as  $\text{Text}(\mathcal{S})$ , and the next-round input in the  
 271 debate process (Eq.1) becomes:  
 272

$$273 \quad \hat{\mathbf{x}}_{t+1}^{(j)} = \text{Tok}\left(Q \parallel \mathbf{y}_t^{(j)} \parallel \text{Text}(\mathcal{S})\right). \quad (7)$$

274

275 In practice, replacing full histories by  $\text{Text}(\mathcal{S})$  yields substantial token compression while preserving  
 276 points of disagreement.  
 277

### 278 4.3 OVERALL METHOD

279 Based on the aforementioned early-exit method with model-level confidence (Sec. 4.1) and adaptive  
 280 compression mechanism with token-level semantic focus (Sec. 4.2), we then present the overall SID  
 281 framework. As shown in Figure 1, after initial generation, each agent assesses its confidence using  
 282 token-level uncertainty metrics derived from output logits. If the agent is sufficiently confident, it  
 283 exits the debate early, avoiding unnecessary interaction. For less confident cases, the debate proceeds  
 284 with a compression mechanism guided by the model’s own attention dynamics. A disagreement-  
 285 oriented prompt steers the attention toward semantically relevant spans in other agents’ responses.  
 286 These spans are then selected and reconstructed into a concise context for the next round, preserving  
 287 key points of contention. By coupling generation-time uncertainty with attention-driven semantic  
 288 focus, SID adapts the debate trajectory according to each agent’s internal belief state, achieving both  
 289 high efficiency and robustness without additional training. Readers could refer to Algorithm 1 for a  
 290 more detailed illustration of the overall implementation.  
 291

## 292 5 EXPERIMENTS

### 293 5.1 EXPERIMENT SETUP

294 **Tasks and Benchmarks.** Results on both LLM and MLLM tasks are presented. For LLM tasks,  
 295 we evaluate our method on MMLUpro (Wang et al., 2024), and Math (Hendrycks et al., 2021)  
 296 datasets, as they represent a wide range of problem-solving tasks in different domains. For MLLM  
 297 tasks, we evaluate on ScienceQA (Lu et al., 2022) and MMStar (Chen et al., 2024b) datasets. **Taken**  
 298 **together, these four benchmarks span (i) text-only vs. multimodal inputs, (ii) factual, analytical and**  
 299 **symbolic reasoning, and (iii) both LLM and MLLM settings, forming a compact yet diverse testbed**  
 300 **for multi-agent debate methods. More expanded experiments can refer to Appendix D**. In consistent  
 301 with previous methods, we randomly sample 100 questions from each dataset for evaluation. For  
 302 the ScienceQA dataset, we utilize the lecture and hint as additional text information following (Liu  
 303 et al., 2024c). For all other datasets, we adopt a zero-shot prompt setting by default.  
 304

305 **Models.** To ensure representative coverage of different foundation models, we evaluate both  
 306 general-purpose and reasoning-oriented models. For LLM tasks, we test on LLaMA-3.1-Instruct-  
 307 8B (LLaMA3.1-8B) Grattafiori et al. (2024) and the recently released GPT-OSS-20B OpenAI et al.  
 308 (2025). For MLLM tasks, we evaluate LLaVA-v1.6-Vicuna-13B (Hugging Face version, LLaVA1.6-  
 309 13B) and the GLM4.1V-Thinking (GLM4.1V) reasoning model (Team et al., 2025).  
 310

311 **Implementation Details** We follow the setup of prior work (Du et al., 2024; Liu et al., 2024c)  
 312 to ensure fair comparison, using  $n = 3$  agents and  $N = 2$  debate rounds across all SID, MAD,  
 313 and DMAD settings. The number of self-consistency samples is set to 3. Additionally, we incor-  
 314 porate step-back prompting (Zheng et al., 2024) and self-contrast (Zhang et al., 2024) as reasoning  
 315 augmentation methods in complement to IO (directly output) and COT methods. For model-level  
 316 confidence, we set the NLL-max threshold  $\alpha$  to 1.0 for reasoning-oriented models, 0.5 for general-  
 317 purpose models, and 0.25 for MLLMs. To mitigate the impact of attention sinks and special tokens  
 318 on specific token logits (Xiao et al., 2024), we empirically set  $\phi(U)$  as the maximum of NLL and en-  
 319 tropy, and exclude certain position metrics when computing model-level confidence. The confidence  
 320 calibration method is trained on a held-out set of 50 samples with  $\tau_c$  as 0.9. More implementation  
 321 details are presented in Appendix C.  
 322

323 **Evaluation Metrics** For the Math dataset (Hendrycks et al., 2021), we adopt the official exact  
 324 match metric to evaluate agent responses. For all other question-answering datasets, which consist  
 325 of multiple-choice questions, we use accuracy as the evaluation metric.  
 326

324  
 325 Table 1: Performance comparison across different LLMs for various datasets (Math subsets and  
 326 MMLUpro). SID-v and SID-c denote our method using the vocabulary-adaptive threshold and cali-  
 327 brated confidence, respectively, to implement the early-exit mechanism. (see Sec.4.1 for details)

Model	Method	Alg.	C&P	Geo.	Int.A.	Num	Pre.A	Pre.C.	MMLUpro	Avg
LLaMA3.1-8B	COT	61	38	34	14	37	54	28	39	38.13
	IO	65	37	35	15	<b>46</b>	59	28	25	38.75
	SBP (Zheng et al., 2024)	46	28	21	12	33	46	24	15	28.13
	Self-Consistency (Wang et al., 2023)	58	25	32	12	40	55	25	45	36.50
	Self-Contrast (Zhang et al., 2024)	54	36	27	11	31	53	27	36	34.38
	MAD (Du et al., 2024)	61	36	36	16	37	60	29	41	39.50
	DMAD (Liu et al., 2024c)	55	36	32	13	36	58	26	39	36.88
GPT-OSS-20B	SID-v	<b>67</b>	<b>43</b>	<b>40</b>	18	41	64	<b>31</b>	<b>47</b>	43.88
	SID-c	<b>67</b>	<b>43</b>	39	<b>20</b>	41	<b>65</b>	30	<b>47</b>	<b>44.00</b>
GPT-OSS-20B	COT	85	81	56	36	70	84	44	61	64.63
	IO	85	81	60	40	74	87	42	64	66.63
	SBP (Zheng et al., 2024)	65	64	44	37	16	73	11	26	42.00
	Self-Consistency (Wang et al., 2023)	75	67	44	31	70	79	23	69	57.25
	Self-Contrast (Zhang et al., 2024)	84	75	65	36	67	88	35	65	64.38
	DMAD (Liu et al., 2024c)	91	90	73	51	66	89	47	65	71.50
	SID-v	<b>94</b>	<b>92</b>	79	<b>65</b>	87	<b>91</b>	<b>62</b>	<b>71</b>	<b>80.13</b>
	SID-c	<b>94</b>	<b>92</b>	<b>80</b>	<b>62</b>	<b>87</b>	<b>91</b>	61	70	79.63

## 5.2 MAIN RESULTS

343  
 344 **Overall Performance** Table 1 and Table 2 respectively present the overall performance across  
 345 LLMs (including LLaMA3.1-8B and GPT-OSS-20B) and MLLMs (including LLaVA1.6-13B  
 346 model and GLM4.1V) in different datasets. Our SID consistently achieves the best performance in  
 347 most scenarios, demonstrating its strong effectiveness. Additionally, we observe that MAD methods  
 348 outperform reasoning augmentation baselines such as self-consistency, which aligns with findings  
 349 reported in (Liu et al., 2024c). Another notable observation is that both the vocabulary-adaptive  
 350 threshold (SID-v) and calibrated confidence (SID-c) yield very similar performance when imple-  
 351 menting the early-exit mechanism described in Sec.4.2. This suggests that the simple thresholding  
 352 strategy can already approximate the learned decision boundary well. Given its training-free nature  
 353 and practical effectiveness, we recommend SID-v as the preferred choice in real-world applications.

354  
 355 **Accuracy and Efficiency** Figure 2(a) compares the performance and token efficiency of our SID  
 356 framework against the baseline MAD method, reporting metrics of both the accuracy and the token  
 357 consumption ratio. The token ratio is computed relative to the MAD setting (i.e., MAD has a to-  
 358 ken ratio of 1). Results show that SID achieves up to a 30% reduction in token usage on science  
 359 and reasoning datasets, while also attaining higher accuracy, demonstrating its significantly better  
 360 efficiency and effectiveness. Note that on thinking models such as GPT-OSS and GLM4.1V, our  
 361 method exhibits more significant token reduction, as their reasoning processes are inherently less  
 362 amenable to token-level compression (see Figure 21,22 for examples). We also compare the *actual*  
 363 *running times* in Figure 5 of the Appendix, where SID demonstrates substantially lower inference  
 364 time, further underscoring its efficiency advantages. Additionally, Figure 2(b) presents accuracy  
 365 curves across different debate rounds. SID consistently improves with additional rounds, highlight-  
 366 ing its strong scalability under extended deliberation.

367  
 368 **Statistical Significance Analysis** The statistical significance of our model-level confidence met-  
 369 ric is illustrated in Figure 2(c) and Figure 7–16, where results for both the LLM (GPT-OSS-20B)  
 370 and MLLM (LLaVA1.6-13B) are presented. In the figure, C and W denote correct and incorrect  
 371 responses, respectively. Across two tasks of varying difficulty: GPQA and MMLUpro, our SID  
 372 maintains a consistent confidence threshold within the correct group for the same model (e.g., NLL  
 373 max  $\approx 7.5$ ), highlighting the stability and robustness of our model-level confidence signal.

## 5.3 ABLATION AND ANALYSIS

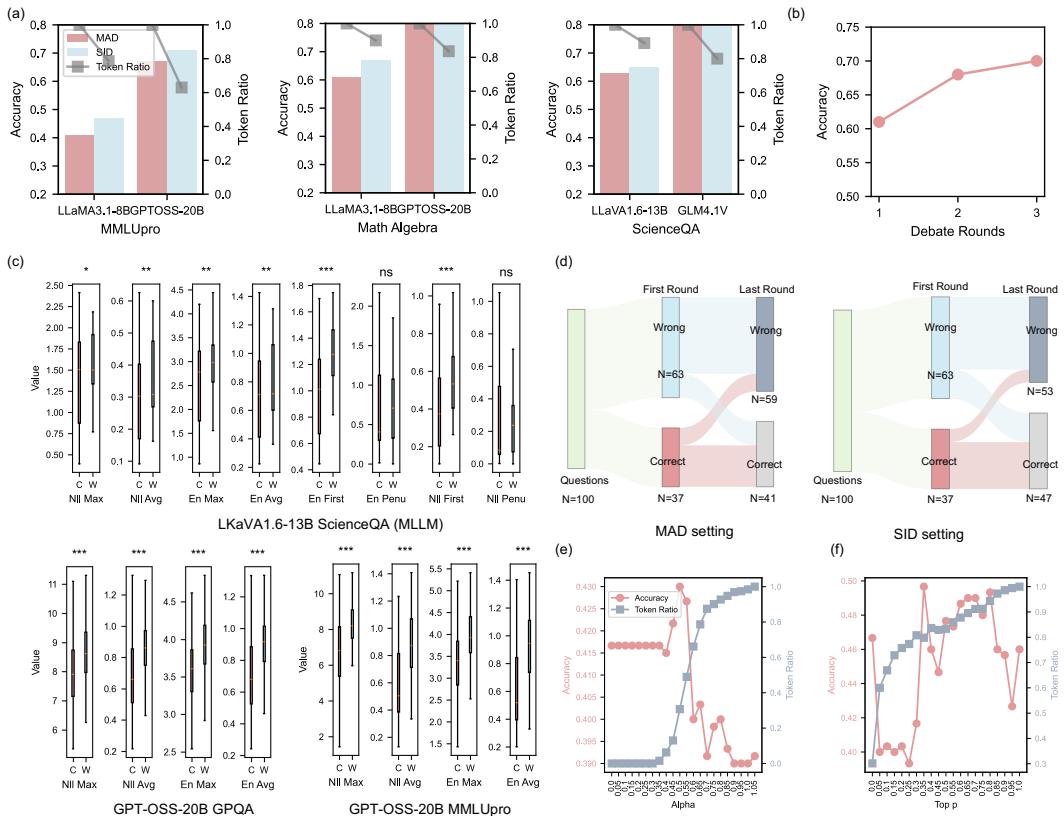
374  
 375 **Ablation of Key Components** Using the LLaMA3.1-8B model and the MMLUpro dataset, we  
 376 conduct a comprehensive ablation study to evaluate the key design components of our framework.  
 377 As shown in Table 3, the baseline MAD setup yields suboptimal performance. In contrast, incor-  
 378 porating our proposed early-exit mechanism based on model-level confidence (Section 4.1) and the  
 379 compression mechanism guided by token-level semantic focus (Section 4.2) leads to substantial im-

378  
379  
380  
Table 2: Performance on Sci.QA and MMStar based  
on MLLMs LLaVA1.6-13B and GLM4.1V.

Model	Method	Sci.QA	MMStar	Model	Method	Sci.QA	MMStar
LLaVA1.6-13B	CoT	63	11	GLM4.1V	CoT	83	29
	IO	62	9		IO	83	32
	Self-Consis	63	11		Self-Consis	84	29
	MAD	65	12		MAD	90	47
	SID-v	<b>65</b>	<b>14</b>		SID-v	<b>91</b>	<b>54</b>
	SID-c	<b>65</b>	<b>14</b>		SID-c	<b>91</b>	<b>54</b>

378  
379  
380  
Table 3: Ablation Study of SID on  
MMLUpro based on LLaMA3.1-8B.

Method	Accuracy	Token Ratio
Baseline Single-round CoT	37.67	0.17
Baseline MAD	39.50	1.00
Baseline MAD + Compression	41.67	0.73
Baseline MAD + Compression + Early Exit	46.83	0.53
SID w/o Semantic Preservation	34.50	0.46
SID w/o Early Exit w/ Token-level Summary	39.50	0.68
<b>SID</b>	<b>46.83</b>	<b>0.53</b>

414  
415  
416  
417  
418  
419  
Figure 2: (a) Accuracy and token ratio comparison across strategies in MAD vs SID. (b) Performance  
with more debate rounds in LLM and MLLM. (c) Significance tests on model-level confidence  
signals. C means the correct group, and W means the wrong group. Statistical significance  
is indicated as follows:  $p < 0.05$ (\*),  $p < 0.01$ (\*\*), and  $p < 0.001$ (\*\*\*). (d) Answer correction  
flow in the MAD vs SID setting. (e) Ablation of the early-exit confidence threshold  $\text{top-}p$  and (f) the  
semantic-preservation ratio  $\alpha$  on accuracy and token ratio.

420  
421  
422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432  
433  
434  
435  
436  
437  
438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448  
449  
450  
451  
452  
453  
454  
455  
456  
457  
458  
459  
460  
461  
462  
463  
464  
465  
466  
467  
468  
469  
470  
471  
472  
473  
474  
475  
476  
477  
478  
479  
480  
481  
482  
483  
484  
485  
486  
487  
488  
489  
490  
491  
492  
493  
494  
495  
496  
497  
498  
499  
500  
501  
502  
503  
504  
505  
506  
507  
508  
509  
510  
511  
512  
513  
514  
515  
516  
517  
518  
519  
520  
521  
522  
523  
524  
525  
526  
527  
528  
529  
530  
531  
532  
533  
534  
535  
536  
537  
538  
539  
540  
541  
542  
543  
544  
545  
546  
547  
548  
549  
550  
551  
552  
553  
554  
555  
556  
557  
558  
559  
560  
561  
562  
563  
564  
565  
566  
567  
568  
569  
570  
571  
572  
573  
574  
575  
576  
577  
578  
579  
580  
581  
582  
583  
584  
585  
586  
587  
588  
589  
590  
591  
592  
593  
594  
595  
596  
597  
598  
599  
600  
601  
602  
603  
604  
605  
606  
607  
608  
609  
610  
611  
612  
613  
614  
615  
616  
617  
618  
619  
620  
621  
622  
623  
624  
625  
626  
627  
628  
629  
630  
631  
632  
633  
634  
635  
636  
637  
638  
639  
640  
641  
642  
643  
644  
645  
646  
647  
648  
649  
650  
651  
652  
653  
654  
655  
656  
657  
658  
659  
660  
661  
662  
663  
664  
665  
666  
667  
668  
669  
670  
671  
672  
673  
674  
675  
676  
677  
678  
679  
680  
681  
682  
683  
684  
685  
686  
687  
688  
689  
690  
691  
692  
693  
694  
695  
696  
697  
698  
699  
700  
701  
702  
703  
704  
705  
706  
707  
708  
709  
710  
711  
712  
713  
714  
715  
716  
717  
718  
719  
720  
721  
722  
723  
724  
725  
726  
727  
728  
729  
730  
731  
732  
733  
734  
735  
736  
737  
738  
739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783  
784  
785  
786  
787  
788  
789  
790  
791  
792  
793  
794  
795  
796  
797  
798  
799  
800  
801  
802  
803  
804  
805  
806  
807  
808  
809  
810  
811  
812  
813  
814  
815  
816  
817  
818  
819  
820  
821  
822  
823  
824  
825  
826  
827  
828  
829  
830  
831  
832  
833  
834  
835  
836  
837  
838  
839  
840  
841  
842  
843  
844  
845  
846  
847  
848  
849  
850  
851  
852  
853  
854  
855  
856  
857  
858  
859  
860  
861  
862  
863  
864  
865  
866  
867  
868  
869  
870  
871  
872  
873  
874  
875  
876  
877  
878  
879  
880  
881  
882  
883  
884  
885  
886  
887  
888  
889  
890  
891  
892  
893  
894  
895  
896  
897  
898  
899  
900  
901  
902  
903  
904  
905  
906  
907  
908  
909  
910  
911  
912  
913  
914  
915  
916  
917  
918  
919  
920  
921  
922  
923  
924  
925  
926  
927  
928  
929  
930  
931  
932  
933  
934  
935  
936  
937  
938  
939  
940  
941  
942  
943  
944  
945  
946  
947  
948  
949  
950  
951  
952  
953  
954  
955  
956  
957  
958  
959  
960  
961  
962  
963  
964  
965  
966  
967  
968  
969  
970  
971  
972  
973  
974  
975  
976  
977  
978  
979  
980  
981  
982  
983  
984  
985  
986  
987  
988  
989  
990  
991  
992  
993  
994  
995  
996  
997  
998  
999  
1000  
1001  
1002  
1003  
1004  
1005  
1006  
1007  
1008  
1009  
10010  
10011  
10012  
10013  
10014  
10015  
10016  
10017  
10018  
10019  
10020  
10021  
10022  
10023  
10024  
10025  
10026  
10027  
10028  
10029  
10030  
10031  
10032  
10033  
10034  
10035  
10036  
10037  
10038  
10039  
10040  
10041  
10042  
10043  
10044  
10045  
10046  
10047  
10048  
10049  
10050  
10051  
10052  
10053  
10054  
10055  
10056  
10057  
10058  
10059  
10060  
10061  
10062  
10063  
10064  
10065  
10066  
10067  
10068  
10069  
10070  
10071  
10072  
10073  
10074  
10075  
10076  
10077  
10078  
10079  
10080  
10081  
10082  
10083  
10084  
10085  
10086  
10087  
10088  
10089  
10090  
10091  
10092  
10093  
10094  
10095  
10096  
10097  
10098  
10099  
100100  
100101  
100102  
100103  
100104  
100105  
100106  
100107  
100108  
100109  
100110  
100111  
100112  
100113  
100114  
100115  
100116  
100117  
100118  
100119  
100120  
100121  
100122  
100123  
100124  
100125  
100126  
100127  
100128  
100129  
100130  
100131  
100132  
100133  
100134  
100135  
100136  
100137  
100138  
100139  
100140  
100141  
100142  
100143  
100144  
100145  
100146  
100147  
100148  
100149  
100150  
100151  
100152  
100153  
100154  
100155  
100156  
100157  
100158  
100159  
100160  
100161  
100162  
100163  
100164  
100165  
100166  
100167  
100168  
100169  
100170  
100171  
100172  
100173  
100174  
100175  
100176  
100177  
100178  
100179  
100180  
100181  
100182  
100183  
100184  
100185  
100186  
100187  
100188  
100189  
100190  
100191  
100192  
100193  
100194  
100195  
100196  
100197  
100198  
100199  
100200  
100201  
100202  
100203  
100204  
100205  
100206  
100207  
100208  
100209  
100210  
100211  
100212  
100213  
100214  
100215  
100216  
100217  
100218  
100219  
100220  
100221  
100222  
100223  
100224  
100225  
100226  
100227  
100228  
100229  
100230  
100231  
100232  
100233  
100234  
100235  
100236  
100237  
100238  
100239  
100240  
100241  
100242  
100243  
100244  
100245  
100246  
100247  
100248  
100249  
100250  
100251  
100252  
100253  
100254  
100255  
100256  
100257  
100258  
100259  
100260  
100261  
100262  
100263  
100264  
100265  
100266  
100267  
100268  
100269  
100270  
100271  
100272  
100273  
100274  
100275  
100276  
100277  
100278  
100279  
100280  
100281  
100282  
100283  
100284  
100285  
100286  
100287  
100288  
100289  
100290  
100291  
100292  
100293  
100294  
100295  
100296  
100297  
100298  
100299  
100300  
100301  
100302  
100303  
100304  
100305  
100306  
100307  
100308  
100309  
100310  
100311  
100312  
100313  
100314  
100315  
100316  
100317  
100318  
100319  
100320  
100321  
100322  
100323  
100324  
100325  
100326  
100327  
100328  
100329  
100330  
100331  
100332  
100333  
100334  
100335  
100336  
100337  
100338  
100339  
100340  
100341  
100342  
100343  
100344  
100345  
100346  
100347  
100348  
100349  
100350  
100351  
100352  
100353  
100354  
100355  
100356  
100357  
100358  
100359  
100360  
100361  
100362  
100363  
100364  
100365  
100366  
100367  
100368  
100369  
100370  
100371  
100372  
100373  
100374  
100375  
100376  
100377  
100378  
100379  
100380  
100381  
100382  
100383  
100384  
100385  
100386  
100387  
100388  
100389  
100390  
100391  
100392  
100393  
100394  
100395  
100396  
100397  
100398  
100399  
100400  
100401  
100402  
100403  
100404  
100405  
100406  
100407  
100408  
100409  
100410  
100411  
100412  
100413  
100414  
100415  
100416  
100417  
100418  
100419  
100420  
100421  
100422  
100423  
100424  
100425  
100426  
100427  
100428  
100429  
100430  
100431  
100432  
100433  
100434  
100435  
100436  
100437  
100438  
100439  
100440  
100441  
100442  
100443  
100444  
100445  
100446  
100447  
100448  
100449  
100450  
100451  
100452  
100453  
100454  
100455  
100456  
100457  
100458  
100459  
100460  
100461  
100462  
100463  
100464  
100465  
100466  
100467  
100468  
100469  
100470  
100471  
100472  
100473  
100474  
100475  
100476  
100477  
100478  
100479  
100480  
100481  
100482  
100483  
100484  
100485  
100486  
100487  
100488  
100489  
100490  
100491  
100492  
100493  
100494  
100495  
100496  
100497  
100498  
100499  
100500  
100501  
100502  
100503  
100504  
100505  
100506  
100507  
100508  
100509  
100510  
100511  
100512  
100513  
100514  
100515  
100516  
100517  
100518  
100519  
100520  
100521  
100522  
100523  
100524  
100525  
100526  
100527  
100528  
100529  
100530  
100531  
100532  
100533  
100534  
100535  
100536  
100537  
100538  
100539  
100540  
100541  
100542  
100543  
100544  
100545  
100546  
100547  
100548  
100549  
100550  
100551  
100552  
100553  
100554  
100555  
100556  
100557  
100558  
100559  
100560  
100561  
100562  
100563  
100564  
100565  
100566  
100567  
100568  
100569  
100570  
100571  
100572  
100573  
100574  
100575  
100576  
100577  
100578  
100579  
100580  
100581  
100582  
100583  
100584  
100585  
100586  
100587  
100588  
100589  
100590  
100591  
100592  
100593  
100594  
100595  
100596  
100597  
100598  
100599  
100600  
100601  
100602  
100603  
100604  
100605  
100606  
100607  
100608  
100609  
100610  
100611  
100612  
100613  
100614  
100615  
100616  
100617  
100618  
100619  
100620  
100621  
100622  
100623  
100624  
100625  
100626  
100627  
100628  
100629  
100630  
100631  
100632  
100633  
100634  
100635  
100636  
100637  
100638  
100639  
100640  
100641  
100642  
100643  
100644  
100645  
100646  
100647  
100648  
100649  
100650  
100651  
100652  
100653  
100654  
100655  
100656  
100657  
100658  
100659  
100660  
100661  
100662  
100663  
100664  
100665  
100666  
100667  
100668  
100669  
100670  
100671  
100672  
100673  
100674  
100675  
100676  
100677  
100678  
100679  
100680  
100681  
100682  
100683  
100684  
100685  
100686  
100687  
100688  
100689  
100690  
100691  
100692  
100693  
100694  
100695  
100696  
100697  
100698  
100699  
100700  
100701  
100702  
100703  
100704  
100705  
100706  
100707  
100708  
100709  
100710  
100711  
100712  
100713  
100714  
100715  
100716  
100717  
100718  
100719  
100720  
100721  
100722  
100723  
100724  
100725  
100726  
100727  
100728  
100729  
100730  
100731  
100732  
100733  
100734  
100735  
100736  
100737  
100738  
100739  
100740  
100741  
100742  
100743  
100744  
100745  
100746  
100747  
100748  
100749  
100750  
100751  
100752  
100753  
100754  
100755  
100756  
100757  
100758  
100759  
100760  
100761  
100762  
100763  
100764  
100765  
100766  
100767  
100768  
100769  
100770  
100771  
100772  
100773  
100774  
100775  
100776  
100777  
100778  
100779  
100780  
100781  
100782  
100783  
100784  
100785  
100786  
100787  
100788  
100789  
100790  
100791  
100792  
100793  
100794  
100795  
100796  
100797  
100798  
100799  
100800  
100801  
100802  
100803  
100804  
100805  
100806  
100807  
100808  
100809  
100810  
100811  
100812  
100813  
100814  
100815  
100816  
100817  
100818  
100819  
100820  
100821  
100822  
10082

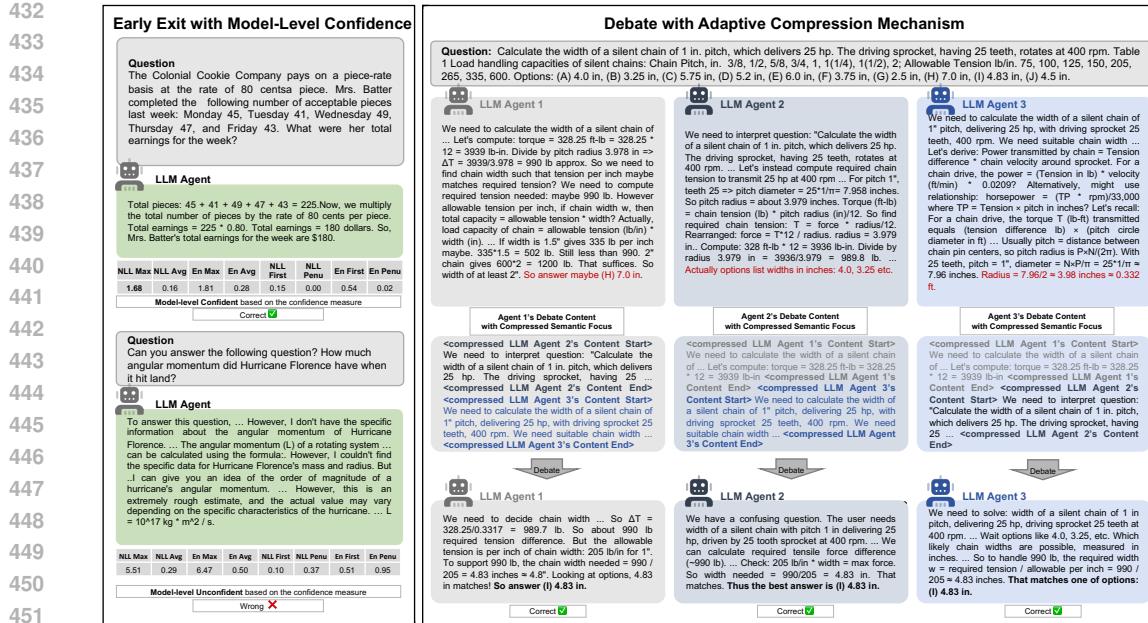


Figure 3: Case study of SID’s debate process. (Left) On MMLUpro, SID exits early for a simple arithmetic question with high confidence but fails on a complex physics question with low confidence. (Right) Three agents initially err but converge to the correct answer through debate guided by token-level semantic focus from adaptively compressed content.

produce a concise summary, which is used as the context for the next round instead of our attention-based compressed history. Empirically, this self-summarization approach leads to a 7.3% absolute accuracy drop and about a 15% increase in token usage compared to our full SID configuration. As shown in Table 3, the superior performance of our SID method validates the design choices of our framework and highlight the contribution of each component to the overall performance.

**Ablation of Vocabulary Adaptive Threshold  $\alpha$**  We further conduct an ablation study on the vocabulary adaptive threshold  $\alpha$  and early exit ratio based on the LLaMA3.1-8B. The results are presented in Figure 2 (e). Small  $\alpha$  means all questions are unconfident, thus the exit ratio is 0, equivalent to traditional MAD, whereas large  $\alpha$  means all questions are confident, thus the system stops at the first round, equivalent to only one LLM model. Our results show that  $\alpha = 0.5$  is an optimal value for this LLaMA3.1-8B model.

**Ablation of Semantic Preservation Ratio  $P$**  In our semantic preservation framework, we select the top- $p$  fraction of context tokens for further processing. The ablation study results for varying  $p$  are shown in Figure 2 (f). We observe that selecting the top tokens with  $p$  around 0.35 or 0.4 yields the best performance. It is interesting to find that when  $p > 0$  but very small, performance can degrade compared to the case where no additional context is included. Conversely, when  $p$  is too large, which means retaining a broader range of content, including potentially redundant agreement, the performance also drops. These findings suggest that both incomplete and overly redundant context can negatively impact multi-LLM debate effectiveness.

#### 5.4 VISUALIZATION RESULTS

To illustrate the mechanisms of our framework more intuitively, we present the visualizations of SID’s workflow in Figure 3. The left branch showcases the early-exit mechanism on a real-world economics question. After generating an answer, the model is assessed as highly confident (e.g., NLL Max = 1.68) by the model-level confidence module and exits early with a correct prediction. In contrast, for a more complex physics question, the model is flagged as low confidence (e.g., NLL Max = 5.51), thus prompting further debate. The right branch illustrates the debate process guided by our adaptive compression mechanism. When facing a challenging physics problem, all

486 three agents initially fail. However, by engaging in a debate using token-level compressed content  
487 driven by semantic focus, the agents collaboratively refine their reasoning and successfully converge  
488 on the correct answer. More case studies can be found in Appendix F, Figure 22. Furthermore,  
489 Figure 2(d) compares the corrections made by the MAD and SID. Our method significantly reduces  
490 the number of cases where debates drift from correct to incorrect answers, while increasing the  
491 number of beneficial corrections, i.e., debates that shift from wrong to correct outcomes, further  
492 demonstrating the high effectiveness of our method.

493

## 494 6 CONCLUSION

495

496 This work introduces SID, a multi-LLM debate framework that leverages self signals from the LLM  
497 generation process to improve both performance and efficiency. SID integrates two types of inter-  
498 nal signals: model-level confidence, which enables early exit for confident agents, and token-level  
499 semantic focus, which compresses debate history by using attention scores to retain key points of  
500 disagreement. Experiments across diverse benchmarks with various LLMs and MLLMs demon-  
501 strate the high performance and efficiency of SID, underscoring the strong potential of leveraging  
502 internal model states as effective signals for guiding collaborative problem-solving. These findings  
503 point toward a promising direction for developing new paradigms in multi-agent systems.

504

505

506

507

508

509

510

511

512

513

514

515

516

517

518

519

520

521

522

523

524

525

526

527

528

529

530

531

532

533

534

535

536

537

538

539

540 REPRODUCIBILITY STATEMENT  
541

542 Significant efforts have been made to ensure the reproducibility of our results. The implementation  
543 details of our framework are described in the main manuscript (Section 4, Algorithm 1, and 5.1), in-  
544 cluding methods, baselines, benchmarks, model configurations, and evaluation settings. Additional  
545 implementation details and the full algorithm are provided in Appendix C. To facilitate faithful repli-  
546 cation of our method, we include detailed descriptions of the key prompts and instruction formats in  
547 Table 4, Table 5, and Figure 4. We believe these materials are sufficient to enable reproducibility of  
548 our study.

549  
550 REFERENCES  
551

552 Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal,  
553 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and et al. Language Models  
554 are Few-Shot Learners, July 2020.

555 Justin Chen, Swarnadeep Saha, and Mohit Bansal. ReConcile: Round-Table Conference Improves  
556 Reasoning via Consensus among Diverse LLMs. In *Proceedings of the 62nd Annual Meet-  
557 ing of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7066–7085,  
558 Bangkok, Thailand, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.  
559 acl-long.381.

560 Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi  
561 Wang, Yu Qiao, Dahu Lin, and et al. Are We on the Right Way for Evaluating Large Vision-  
562 Language Models?, April 2024b.

563 Giulio Corallo and Paolo Papotti. FINCH: Prompt-guided Key-Value Cache Compression for Large  
564 Language Models. *Transactions of the Association for Computational Linguistics*, 12:1517–1532,  
565 November 2024. ISSN 2307-387X. doi: 10.1162/tacl\_a\_00716.

566 Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving  
567 factuality and reasoning in language models through multiagent debate. In *Proceedings of the  
568 41st International Conference on Machine Learning*, volume 235 of *ICML’24*, pp. 11733–11763,  
569 Vienna, Austria, July 2024. JMLR.org.

570 Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt,  
571 Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, and et al. A survey  
572 of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56(S1):1513–1589, October  
573 2023. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-023-10562-9.

574 Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad  
575 Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. The Llama 3  
576 Herd of Models, November 2024.

577 W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications.  
578 *Biometrika*, 57(1):97–109, April 1970. ISSN 0006-3444. doi: 10.1093/biomet/57.1.97.

579 Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn  
580 Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset,  
581 November 2021.

582 Mengting Hu, Zhen Zhang, Shiwan Zhao, Minlie Huang, and Bingzhe Wu. Uncertainty in Natural  
583 Language Processing: Sources, Quantification, and Applications, June 2023.

584 Michael Kirchhof, Gjergji Kasneci, and Enkelejda Kasneci. Position: Uncertainty Quantification  
585 Needs Reassessment for Large-language Model Agents, May 2025.

586 Armen Der Kiureghian and Ove Ditlevsen. Aleatory or epistemic? Does it matter? *Structural  
587 Safety*, 31(2):105–112, March 2009. ISSN 0167-4730. doi: 10.1016/j.strusafe.2008.06.020.

588 Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large  
589 Language Models are Zero-Shot Reasoners, January 2023.

594 Yongchan Kwon, Joong-Ho Won, Beom Joon Kim, and Myunghee Cho Paik. Uncertainty quantifi-  
 595 cation using Bayesian neural networks in classification: Application to biomedical image segmen-  
 596 tation. *Computational Statistics & Data Analysis*, 142:106816, February 2020. ISSN 0167-9473.  
 597 doi: 10.1016/j.csda.2019.106816.

598 Yassir Laaouach. HALT-CoT: Model-Agnostic Early Stopping for Chain-of-Thought Reasoning via  
 599 Answer Entropy. In *4th Muslims in ML Workshop Co-Located with ICML 2025*, June 2025.

600 Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun  
 601 Liu. LLMs-as-Judges: A Comprehensive Survey on LLM-based Evaluation Methods, Decem-  
 602 ber 2024a.

603 Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image  
 604 Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the*  
 605 *40th International Conference on Machine Learning*, pp. 19730–19742. PMLR, July 2023.

606 Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. Im-  
 607 proving Multi-Agent Debate with Sparse Communication Topology, June 2024b.

608 Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming  
 609 Shi, and Zhaopeng Tu. Encouraging Divergent Thinking in Large Language Models through  
 610 Multi-Agent Debate, October 2024.

611 Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual In-  
 612 struction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*  
 613 *Recognition*, pp. 26296–26306, 2024a.

614 Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang,  
 615 and Jing Li. GroupDebate: Enhancing the Efficiency of Multi-Agent Debate Using Group Dis-  
 616 cussion, September 2024b.

617 Yexiang Liu, Jie Cao, Zekun Li, Ran He, and Tieniu Tan. Breaking Mental Set to Improve Reasoning  
 618 through Diverse Multi-Agent Debate. In *The Thirteenth International Conference on Learning*  
 619 *Representations*, October 2024c.

620 Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord,  
 621 Peter Clark, and Ashwin Kalyan. Learn to Explain: Multimodal Reasoning via Thought Chains  
 622 for Science Question Answering, October 2022.

623 Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Ed-  
 624 ward Teller. Equation of State Calculations by Fast Computing Machines. *The Journal of Chemical*  
 625 *Physics*, 21(6):1087–1092, June 1953. ISSN 0021-9606. doi: 10.1063/1.1699114.

626 OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus,  
 627 Rahul K. Arora, Yu Bai, Bowen Baker, and et al. Gpt-oss-120b & gpt-oss-20b Model Card,  
 628 August 2025.

629 David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien  
 630 Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A  
 631 Benchmark, November 2023.

632 Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and  
 633 Donald Metzler. Confident Adaptive Language Modeling, October 2022.

634 Yiliu Sun, Zicheng Zhao, Sheng Wan, and Chen Gong. CortexDebate: Debating Sparsely and  
 635 Equally for Multi-Agent Debate. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mo-  
 636 hammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL*  
 637 2025, pp. 9503–9523, Vienna, Austria, July 2025. Association for Computational Linguistics.  
 638 ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.495.

639 GLM-V. Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang,  
 640 Jiale Cheng, Ji Qi, Junhui Ji, and et al. GLM-4.5V and GLM-4.1V-Thinking: Towards Versatile  
 641 Multimodal Reasoning with Scalable Reinforcement Learning, August 2025.

648 Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea  
 649 Finn, and Christopher D. Manning. Just Ask for Calibration: Strategies for Eliciting Calibrated  
 650 Confidence Scores from Language Models Fine-Tuned with Human Feedback, October 2023.  
 651

652 Zailong Tian, Zhuoheng Han, Yanzhe Chen, Haozhe Xu, Xi Yang, Richeng Xuan, Houfeng Wang,  
 653 and Lizi Liao. Overconfidence in LLM-as-a-Judge: Diagnosis and Confidence-Driven Solution,  
 654 August 2025.

655 Weijie Tu, Weijian Deng, Dylan Campbell, Yu Yao, Jiayang Zheng, Tom Gedeon, and Tongliang  
 656 Liu. Ranked from Within: Ranking Large Multimodal Models Without Labels. In *Forty-Second*  
 657 *International Conference on Machine Learning*, June 2025.

658 Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowd-  
 659 heray, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language  
 660 Models, March 2023.

661 Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming  
 662 Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and et al. MMLU-Pro: A More Robust and Challeng-  
 663 ing Multi-Task Language Understanding Benchmark, November 2024.

664 Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc  
 665 Le, and Denny Zhou. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models,  
 666 January 2023.

667 Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient Streaming  
 668 Language Models with Attention Sinks, April 2024.

669 Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs  
 670 Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. In *The*  
 671 *Twelfth International Conference on Learning Representations*, October 2023.

672 Chenxu Yang, Qingyi Si, Yongjie Duan, Zheliang Zhu, Chenyu Zhu, Qiaowei Li, Minghui Chen,  
 673 Zheng Lin, and Weiping Wang. Dynamic Early Exit in Reasoning Models, September 2025.

674 Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik  
 675 Narasimhan. Tree of Thoughts: Deliberate Problem Solving with Large Language Models, De-  
 676 cember 2023.

677 Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shum-  
 678 ing Shi, and Zhaopeng Tu. Benchmarking LLMs via uncertainty quantification. In *Proceedings*  
 679 *of the 38th International Conference on Neural Information Processing Systems*, volume 37 of  
*NIPS '24*, pp. 15356–15385, Red Hook, NY, USA, June 2025. Curran Associates Inc. ISBN  
 979-8-3313-1438-5.

680 Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuanjing Huang, and Xipeng  
 681 Qiu. Exchange-of-Thought: Enhancing Large Language Model Capabilities through Cross-  
 682 Model Communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*  
*683 Language Processing*, pp. 15135–15153, Singapore, 2023. Association for Computational Lin-  
 684 guistics. doi: 10.18653/v1/2023.emnlp-main.936.

685 Luke Yoffe, Alfonso Amayuelas, and William Yang Wang. DebUnc: Improving Large Language  
 686 Model Agent Communication With Uncertainty Metrics, February 2025.

687 Yuting Zeng, Weizhe Huang, Lei Jiang, Tongxuan Liu, XiTai Jin, Chen Tianying Tiana, Jing Li,  
 688 Xiaohua Xu, and et al. S2-MAD: Breaking the Token Barrier to Enhance Multi-Agent Debate  
 689 Efficiency. In Luis Chiruzzo and Alan Ritter (eds.), *Proceedings of the 2025 Conference of*  
*690 the Nations of the Americas Chapter of the Association for Computational Linguistics: Human*  
*691 Language Technologies (Volume 1: Long Papers)*, pp. 9393–9408, Albuquerque, New Mexico,  
 692 April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/  
 693 v1/2025.naacl-long.475.

694 Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Wang Zhen, Dinghao Wu, and  
 695 Shuyue Hu. *If Multi-Agent Debate Is the Answer, What Is the Question?* February 2025. doi:  
 696 10.48550/arXiv.2502.08788.

702 Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weim-  
703 ing Lu. Self-Contrast: Better Reflection Through Inconsistent Solving Perspectives, June 2024.  
704

705 Huaixiu Steven Zheng, Swaroop Mishra, Xinyun Chen, Heng-Tze Cheng, Ed H. Chi, Quoc V. Le,  
706 and Denny Zhou. Take a Step Back: Evoking Reasoning via Abstraction in Large Language  
707 Models, March 2024.

708

709

710

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750

751

752

753

754

755

756 **A USE OF LARGE LANGUAGE MODELS (LLMs)**  
757758 Large Language Models (LLMs) were used solely for language refinement and proofreading pur-  
759 poses. They were not involved in research ideation and methodology design. All scientific contribu-  
760 tions and conceptual developments were carried out entirely by the authors. The LLM did not play  
761 a substantive role in shaping the research content and should not be considered a contributor.  
762763 **B LIMITATION**  
764765 Our method relies on internal model signals such as logits and attention maps, which limit direct  
766 applicability to public closed-source APIs. However, it remains well-suited for internal deployments  
767 of proprietary models, especially in multi-agent systems, and can serve as an intermediate reasoning  
768 layer prior to externalized API serving. Notably, many modern systems (e.g., GPT-5) already adopt  
769 multi-agent or tool-augmented architectures, making our approach broadly applicable and increas-  
770 ingly relevant.  
771772 **C ALGORITHM AND IMPLEMENTATION**  
773774 **C.1 SEMANTIC PRESERVATION**  
775776 The semantic preservation module plays a key role in restoring the semantic cohesion from the top-  
777  $p$  selected sparse tokens based on a model’s self-signals (*i.e.*, attention mechanism). Specifically,  
778 the method selects the most relevant textual spans based on attention distribution, but ensures that  
779 these selections are semantically coherent when mapped back to natural language. The algorithm  
780 below (Algorithm 2) shows the main pipeline, and the example (Figure 6) illustrates the comparison  
781 between without and with semantic preservation.  
782783 **Algorithm 2** Semantic-Preserving Compression  
784785 **Require:** Prompt text  $x$  with marked spans `FOCUS`, `DISCUSSION`; offset map  $\mathcal{O}$ ; Top- $p$  selected  
786 attention score  $\mathcal{C}$   
787 **Ensure:** Compressed prompt  $x'$   
788 1:  $\mathcal{U} \leftarrow \text{EXTRACTUNITS}(x, \text{DISCUSSION})$  ▷ Sentence/clause-level segments  
789 2:  $T \leftarrow \text{TOKENIZER}(x)$   
790 3:  $\mathcal{S} \leftarrow \text{MAPTOKENSTOUNITS}(\mathcal{C}, \mathcal{U}, \mathcal{O})$   
791 4:  $x' \leftarrow \text{REPLACESPAN}(x, \text{DISCUSSION}, \mathcal{S})$   
792 5: **return**  $x'$   
793794 We begin by extracting semantically coherent units (e.g., sentences or clauses) from the  
795 `DISCUSSION` span using lightweight parsing heuristics, including punctuation or newline segmen-  
796 tation. This yields a set of candidate text fragments  $\mathcal{U}$ .  
797798 We then use the Top- $p$  selected attention score  $\mathcal{C}$  (from Algorithm 1) to select the top- $p$  most relevant  
799 tokens from  $T$ . To preserve semantic interpretability, we map these selected tokens back to their  
800 enclosing segments in  $\mathcal{U}$  using the token-to-text offset map  $\mathcal{O}$ . The resulting set of informative  
801 fragments  $\mathcal{S}$  is used to replace the original `DISCUSSION` span, yielding a compressed prompt  $x'$   
802 that retains critical disagreement signals while discarding redundant or low-relevance content.  
803804 In the multi-modal setting (e.g., MLLMs), token offsets may shift due to image-text fusion. We  
805 mitigate this by anchoring to stable textual markers in the `FOCUS` span to adjust  $\mathcal{O}$  and maintain  
806 alignment.  
807808 This compression module is integrated into the overall SID framework to support efficient and in-  
809 terpretable multi-agent reasoning under token or latency constraints

810 C.2 PROMPT TEMPLATE  
811

812 In multi-task evaluation settings, especially those involving factual or multiple-choice benchmarks,  
813 we observe that models frequently generate semantically correct answers but fail to conform to  
814 the expected output format. This discrepancy is particularly pronounced in open-ended LLMs,  
815 where prior supervised fine-tuning (SFT) phases may introduce implicit formatting preferences (e.g.,  
816 `\boxed` in math domains).

817 To mitigate this, we prepend a task-specific *system prompt* that explicitly enforces the desired answer  
818 format. Our full prompting format is:

819 `<system prompt> + <question content> + <output instruction>`

820 This method proves especially helpful for models with weaker instruction-following capabilities  
821 (e.g., LLaMA3.1-8B) and significantly reduces post-hoc answer parsing failures. Another example is the GLM4.1-V thinking model. The default multiple choice response uses  
822 a special boxed token, such as `<|begin_of_box|>B<|end_of_box|>`. By emphasizing  
823 the answer returning with brackets in the system prompt, GLM4.1V thinking yields  
824 `<|begin_of_box|>(B)<|end_of_box|>`. This allows us to extract the result using brackets  
825 in a unified way. Table 4 lists the dataset-specific system prompts and the enforced answer formats  
826 used in our experiments.

827 Table 4: Dataset-specific system prompts and enforced output formats for answer extraction.  
828

829 <b>Dataset</b>	830 <b>System Prompt (Instruction)</b>	831 <b>Expected Output Format (for answer parsing)</b>
832 MMLUpro	833 You are a trivia expert who knows everything. You 834 are tasked to answer the following question. Give 835 your final answer in the format of (X), e.g., (A).	(A), (B), etc.
836 Math	837 You are a math expert. You are tasked to 838 determine the answer to the following question. Give 839 your final answer in the form of 840 <code>\boxed{answer}</code> in the last sentence of your response, e.g., <code>\boxed{[1, 3]}</code> .	<code>\boxed{...}</code>
841 GPQA	842 You are an expert in graduate-level science and 843 mathematics. You will be presented with chal- 844 lenging questions designed to test your reasoning 845 abilities. Your last sentence should be “The cor- 846 rect answer is (insert answer here).”	“The correct answer is (A).”
847 ScienceQA	848 You are a trivia expert who knows everything. You 849 are tasked to answer the following question. Give 850 your final answer in the format of (X), e.g., (A).	(A), (B), etc.
851 MMStar	852 You are an expert in multimodal task under- 853 standing, and your task is to answer the following ques- 854 tions. Give your final answer in the format of (X), 855 e.g., (A)	(A), (B), etc.

856 Table 5: Dataset-specific output instruction prompts.  
857

858 <b>Dataset</b>	859 <b>Output Instruction</b>
MMLUpro	Give your final answer in the format of '(X)'
Math	Give your final answer in the form of <code>\boxed{answer}</code> at the end of your response, e.g., <code>\boxed{[1, 3]}</code> .
GPQA	Your last sentence should be ‘The correct answer is (insert an- swer here).’ e.g., The correct answer is (A).
ScienceQA	Give your final answer in the format of '(X)'. You should only give one answer. For example, the answer is (A).
MMStar	Give your final answer in the format of '(X)'. You should only give one answer.

864  
865  
866  
867

## Reasoning Augmentation Prompt

```

868 <COT (zero-shot)>
869 <Generation 1/1>
870 Let's think step by step.

871 <IO>
872 <Generation 1/1>
873 Please directly give your answer.

874 <Self-Consistency>
875 <Generation N/N, i.e., after N-rounds generation>
876 <majority vote among multiple responses>

877 <Step-back prompting, SBP>
878 <Generation 1/2>
879 You are an expert at structured reasoning. Your task is to extract the subject concepts and
880 principles involved in solving the problem. In this step you don't need to give you final answer,
881 just extract the concepts and principles.
882 <Get Phase 1 Response>

883 <Generation 2/2>
884 Learned concepts and principles:
885 {Phase 1 Response}
886 Solve the problem step by step with your reasoning path, according to the concepts and principles
887 you have learned.

888 <Self-Contrast>
889 <Generation 1/4>
890 Let's think step by step

891 <Generation 2/4>
892 Please generate an alternative solution to this problem using a different approach or reasoning
893 method.

894 <Generation 3/4>
895 Now compare your original solution with the alternative solution:
896 1. What are the key differences between the two approaches?
897 2. Which approach seems more reliable and why?
898 3. Can you identify any weaknesses in either approach?
899 4. Based on this comparison, what is your final answer?

900 <Generation 4/4>
901 Based on your comparison of the different approaches, provide your final answer.

902
903
904
905
906
907
908
909
910
911

```

Figure 4: Details of reasoning augmentation prompt.

912 In terms of question content, we strictly follow the previous work (Du et al., 2024; Liu et al., 2024c)  
913 in parsing the question to the chat template.914 Moreover, we list the reasoning augmentation prompt (Figure 4 used in our experiments. Notably,  
915 Output Instructions should still be used after those prompts to enhance the ability to follow instruc-  
916 tions.  
917

## 918 D EXPANDED EXPERIMENTS

920 To further broaden our empirical coverage, we additionally evaluate SID on GPQA (Rein et al.,  
 921 2023), a challenging benchmark targeting advanced science knowledge and reasoning. We follow  
 922 the same evaluation protocol as in the main experiments and use GPT-OSS-20B as the base model,  
 923 with SID-v as the default debate configuration.

925 Table 6: Results on the GPQA benchmark using GPT-OSS-20B under the same evaluation protocol  
 926 as in the main text.

928 Method (GPQA)	929 CoT	929 IO	929 MAD	929 SID
930 Accuracy (%) $\pm$ std	930 $43.2 \pm 1.7$	930 $41.4 \pm 1.6$	930 $52.9 \pm 2.0$	930 <b>54.8 <math>\pm 1.5</math></b>

931 As shown in Table 6, SID achieves the best performance among all compared methods on GPQA,  
 932 demonstrating that our SID framework generalizes to challenging scientific reasoning tasks beyond  
 933 the benchmarks reported in the main paper.

935 Exchange-of-Thought (EoT) (Yin et al., 2023) derives a model-level confidence score post hoc from  
 936 the final discrete answers (e.g., options A/B/C/D) across rounds and/or agents, for example, by  
 937 measuring how frequently the most common answer appears. This effectively treats consistency  
 938 of verbalized outputs as a proxy for confidence. To more concretely compare SID with EoT, we  
 939 run both methods with LLaMA3.1-8B on the MMLUpro dataset. We follow the same evaluation  
 940 protocol as in our main experiments and configure three agents and two debate rounds, computing  
 941 the EoT confidence exactly as described in the original paper. The results are summarized in Table 7.

942 Table 7: Comparison with EoT on MMLUpro using LLaMA3.1-8B.

945 Method	946 CoT	946 MAD	946 EoT	946 SID
946 Accuracy (%)	946 39.5	946 41.5	946 42.7	946 <b>46.8</b>

948 As shown in Table 7, EoT improves over standard MAD, indicating that consistency-based aggre-  
 949 gation can indeed be beneficial. SID, however, still yields a clear additional gain.

951 We further examine whether SID extends naturally to *heterogeneous* multi-model debate. To this  
 952 end, we consider two base models on MMLUpro: LLaMA-3.1-8B (denoted as model A) and GPT-  
 953 OSS-20B (denoted as model B). We instantiate homogeneous pairs (A+A and B+B) as well as  
 954 a heterogeneous pair (A+B), and measure the accuracy of each agent’s final-round answer under  
 955 MAD and SID.

956 Table 8: Heterogeneous multi-model debate on MMLUpro with LLaMA-3.1-8B (A) and GPT-OSS-  
 957 20B (B). For the heterogeneous A+B setting, we report the final accuracy of each agent (A/B).

959 Setting	960 A+A MAD	960 A+A SID	960 B+B MAD	960 B+B SID	960 A+B MAD (A/B)	960 A+B SID (A/B)
961 Acc. (%)	961 41	961 47	961 65	961 71	961 65 / 65	961 64 / 72

963 As shown in Table 8, SID consistently improves performance over MAD in the homogeneous set-  
 964 tings: the accuracy of A+A increases from 41% to 47%, and B+B from 65% to 71%. This mirrors  
 965 the trends observed in the main experiments and indicates that SID’s model-level early exit and  
 966 token-level semantic-focus mechanisms remain beneficial across different base models.

968 The heterogeneous configuration (A+B) provides additional insight. Under MAD, both A and B  
 969 achieve roughly the same final accuracy (65%), suggesting that the weaker model A benefits from  
 970 debating with the stronger model B, while the stronger model does not fully realize its potential  
 971 within an unstructured debate protocol. Under SID, the stronger model B reaches 72% accuracy,  
 which exceeds all homogeneous baselines (including B+B with MAD at 65% and even B+B with

972 SID at 71%), while model A remains competitive at 64%. These results indicate that SID can better  
 973 exploit model complementarity in heterogeneous groups: the weaker model can still benefit from  
 974 interaction, but the stronger model is less likely to be “dragged down” by unnecessary debate and  
 975 can more reliably achieve (or slightly exceed) its best homogeneous performance.

976 Overall, this study suggests that SID is not only compatible with heterogeneous multi-model debate,  
 977 but also capable of leveraging internal self-signals to coordinate agents of different strengths more  
 978 effectively than standard MAD.

## 981 E MODEL-LEVEL CONFIDENCE ANALYSIS

984 In this section, we provide additional analyses of the model-level confidence signal used by SID,  
 985 in order to clarify (i) how well it separates correct and wrong predictions, and (ii) what happens in  
 986 over-confident failure cases.

987 Recall that for each agent, we compute a model-level confidence score from internal self-signals during  
 988 generation. Concretely, given a sequence of output tokens  $\{y_t\}_{t=1}^T$  and model log-probabilities  
 989  $\log p_\theta(y_t \mid y_{<t}, x)$ , we construct confidence metrics based on aggregated negative log-likelihood  
 990 (NLL) and entropy over different token positions (e.g., all tokens, answer tokens, or reasoning  
 991 tokens). These metrics are then calibrated via the vocabulary-adaptive threshold described in Sec-  
 992 tion 4.1 to decide whether an agent should exit the debate early.

995 **Separation between correct and wrong predictions.** Figures 7–16 in the appendix, for multiple  
 996 datasets and models, the empirical distributions of our confidence metrics for *correct* (C) vs. *wrong*  
 997 (W) answers, together with significance tests. Across MMLU-Pro, MATH subsets, ScienceQA,  
 998 MMStar, and GPQA, we consistently observe that the C group exhibits noticeably *lower* NLL /  
 999 entropy than the W group, and these differences are statistically significant in most settings. In  
 1000 parallel, the correction-flow plots in the same figures show that, under SID, the number of correct-to-  
 1001 wrong (C→W) transitions across rounds is reduced, while wrong-to-correct (W→C) transitions are  
 1002 maintained or increased compared to MAD. Taken together, these results indicate that our confidence  
 1003 estimate, although not perfectly calibrated, is a useful ranking signal for debate scheduling: high-  
 1004 confidence states are more likely to be correct and more likely to remain correct under SID.

1005  
 1006 **Over-confident errors and failure cases.** We also explicitly examine cases where the model-level  
 1007 confidence is high but the final answer may be incorrect. Figures 17–20 provide qualitative exam-  
 1008 ples of such failure cases under our early-exit policy. In most of these examples, the intermediate  
 1009 reasoning trajectory is largely sensible (e.g., correctly recalling definitions or setting up equations),  
 1010 but the model makes a local slip in the last step, such as an arithmetic mistake or an incorrect option  
 1011 mapping. Because our confidence is derived from token-level log-probabilities aggregated over the  
 1012 entire reasoning sequence, these mostly plausible trajectories can still yield low NLL / entropy even  
 1013 when the final box answer is wrong. These examples illustrate the limitations of our signal: it is not  
 1014 an oracle and over-confident errors do occur.

1015 However, our quantitative analyses show that such over-confident wrong cases are relatively rare  
 1016 compared to the large mass of high-confidence correct predictions. In addition, the correction-flow  
 1017 statistics indicate that SID reduces harmful C→W transitions overall, while preserving beneficial  
 1018 W→C transitions. Thus, even though the confidence signal can occasionally fail, on balance it  
 1019 enables the early-exit mechanism to (i) protect many high-confidence correct answers from being  
 1020 overturned by noisy debate and (ii) avoid spending additional tokens on debates that are unlikely to  
 1021 change the outcome.

1022 In summary, our model-level confidence should be viewed as a ranking heuristic derived from in-  
 1023 ternal self-signals. Empirically, it exhibits a clear and statistically meaningful separation between  
 1024 correct and wrong predictions across benchmarks, and it leads to fewer C→W transitions and more  
 1025 efficient use of debate rounds when integrated into SID. This supports its use as a practical gating  
 signal for deciding when to continue or terminate multi-agent debate.

1026 **F MORE RESULTS**  
10271028 In this section, we list many visualization results to illustrate the effectiveness of our SID methods.  
10291030 Figure 6 Example comparison between w/o semantic preservation (red, brute-force token selec-  
1031 tion) and w/ semantic preservation (green, semantically coherent expansion). It can be observed  
1032 that token-level semantic focus deletes irrelevant points (for example, point 2 in the solution was  
1033 deleted), and our semantic preservation retains the semantic cohesion from the selected tokens.1034 A series of model-level confidence examples below can demonstrate the stable early exit threshold  
1035 in the same model, and the statistical significance between the correct and wrong groups. Moreover,  
1036 we also provide the correction flow from the first round to the last round.1037 Figure 7 Top: Model-level Confidence result on the Math Algebra dataset with LLaMA3.1-8B.  
1038 Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged  
1039 between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.1040 Figure 8 Model-level Confidence result on the Math Counting and Probability dataset with  
1041 LLaMA3.1-8B.

1042 Figure 9 Model-level Confidence result on the Math Geometry dataset with LLaMA3.1-8B.

1043 Figure 10 Model-level Confidence result on the Math Intermediate Algebra dataset with LLaMA3.1-  
1044 8B.

1045 Figure 11 Model-level Confidence result on the Math Number Theory dataset with LLaMA3.1-8B.

1046 Figure 12 Model-level Confidence result on the Math Prealgebra dataset with LLaMA3.1-8B.

1047 Figure 13 Model-level Confidence result on the Math Precalculus dataset with LLaMA3.1-8B.

1048 Figure 14 Model-level Confidence result on the MMStar dataset with GLM4.1V.

1049 Figure 15 Model-level Confidence result on the MMLUpro dataset with GPT-OSS-20B. Bottom:  
1050 Correction flow with 8 deltas of confidence metrics.1051 In addition, a number of model-level early exit cases are provided here to show the confident and  
1052 overconfident cases. It can be observed that the model partially analyzes the problem in overconfi-  
1053 dent cases.

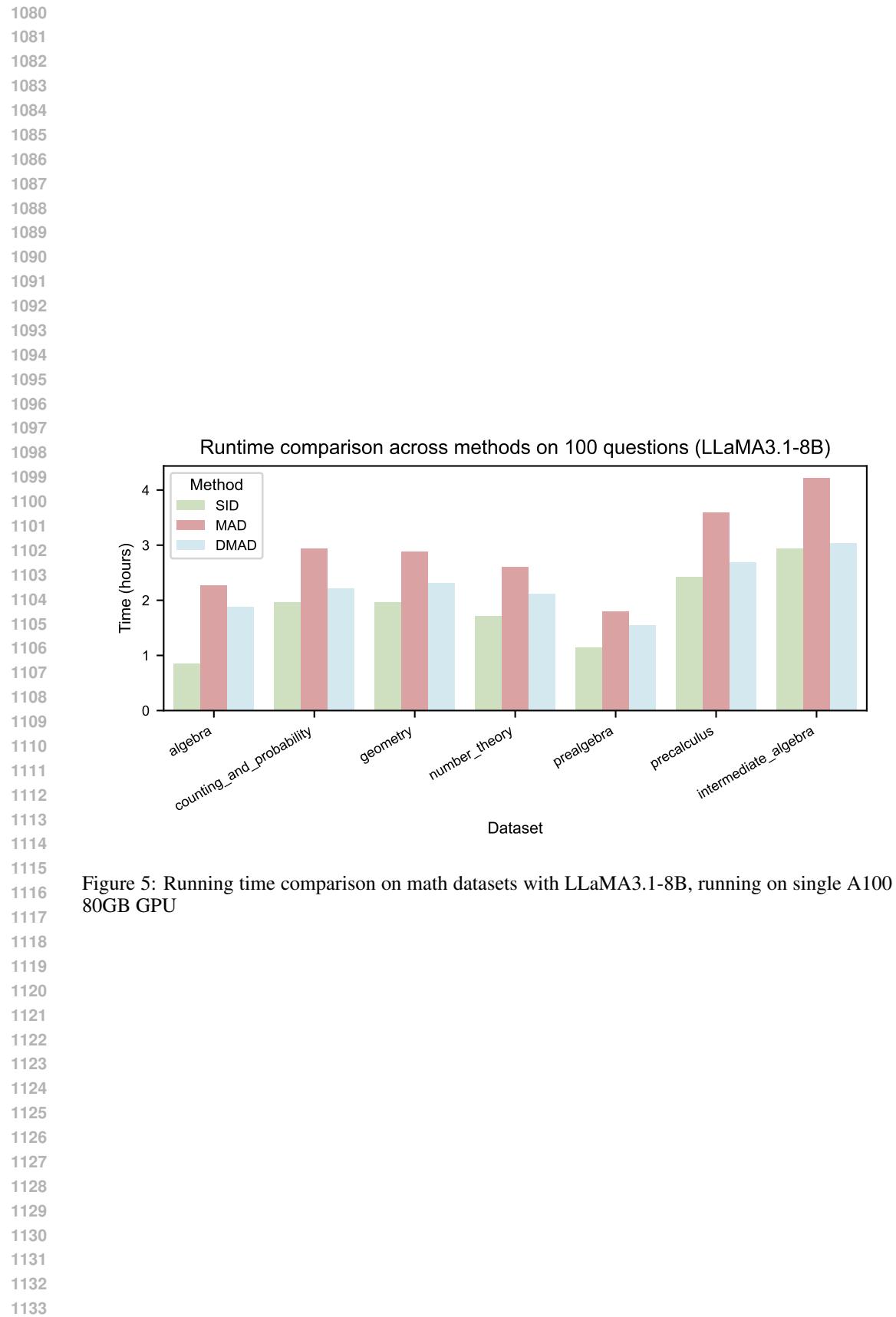
1054 Figure 17 Examples of model-level early exit cases in the MMLUpro dataset.

1055 Figure 18 Examples of model-level early exit cases in the ScienceQA dataset.

1056 Figure 19 Examples of model-level early exit cases in the Math dataset.

1057 Figure 20 Examples of model-level early exit cases in the GPQA dataset.

1058 Subsequently, Figure 21 and Figure 22 display that the token-level semantic focus module com-  
1059 presses the contents and assists agents in correcting their answers.1060  
1061  
1062  
1063  
1064  
1065  
1066  
1067  
1068  
1069  
1070  
1071  
1072  
1073  
1074  
1075  
1076  
1077  
1078  
1079



1134  
 1135  
 1136  
 1137  
 1138

## Examples of Semantic Preservation

1139  
 1140  
 1141     **<Selected Text: w/o Semantic Preservation>**  
 1142     One agent solution:  
 1143         ```To solve this, let's break it down by step:  
 1144         1. Both parents have the same meaning they are both they can only that also zy they can only).  
 1144         3. Since are but not).  
 1145         4 the probability that each of children will bezy is1or %), because they will only from both parents.  
 1146         So, the correct answer is (F) 1.  
 1147         ````  
 1148     One agent solution:  
 1149         ```To solve this let's consider the possibleotypes andotypes of the and both parents are we can  
 1150         their the a  
 1151         2 Theotypes of )  
 1151         3 parents the possible can  
 1152         4 The probability of each child is the of which is because each has of probability of2).  
 1153         So, the probability that each children will be affected equals 4 or 25%. Therefore, the answer is  
 1154         (I).  
 1155         ````  
 1156     **<Selected Text: w/ Semantic Preservation>**  
 1157     One agent solution:  
 1158         ```To solve this, let's break it down step by step:  
 1159         1. Both parents have the same autosomal recessive disorder, meaning they are both homozygous  
 1160         recessive (rr) for the disorder. When two homozygous recessive parents (rr) have children, they  
 1160         can only produce offspring that are also homozygous recessive (rr) because they can only  
 1161         contribute their recessive alleles (r).  
 1162         3. Since the disorder is autosomal recessive, the parents are carriers of the recessive allele  
 1162         (r) but do not express the disorder themselves because they are homozygous recessive (rr).  
 1163         4. Therefore, the probability that each of their children will be affected (homozygous recessive,  
 1164         rr) is 1 (or 100%), because they will only inherit the recessive allele from both parents. So,  
 1165         the correct answer is (F) 1.````  
 1166     One agent solution: ```To solve this, let's consider the possible genotypes and phenotypes of the  
 1167     parents and their offspring. Since both parents are affected by an autosomal recessive disorder,  
 1167     we can represent their genotype as 'rr' (assuming the disorder is caused by a recessive allele  
 1168     'r' and the dominant allele is 'R').  
 1169         2. The possible genotypes of their offspring are: - Rr (carrier) - Rr (carrier) 3. Since both  
 1170         parents are affected, the possible genotypes of the parents can be rr.  
 1171         4. The probability of each child being affected (rr) is the probability of inheriting two  
 1171         recessive alleles, which is 1/4, because each parent has a 1/2 chance of contributing a recessive  
 1172         allele, and the probability of both contributing a recessive allele is (1/2) \* (1/2) = 1/4. So,  
 1173         the probability that each of their children will be affected equals 1/4 or 25%. Therefore, the  
 1174         answer is (I).````  
 1175     **Legend: The red and green shaded text is an example pair for the w/o Semantic Preservation and w/  
 1176         Semantic Preservation comparison.**  
 1177  
 1178  
 1179  
 1180  
 1181  
 1182  
 1183  
 1184     Figure 6: Example comparison between w/o semantic preservation (red, brute-force token selection)  
 1185         and w/ semantic preservation (green, semantically coherent expansion)  
 1186  
 1187

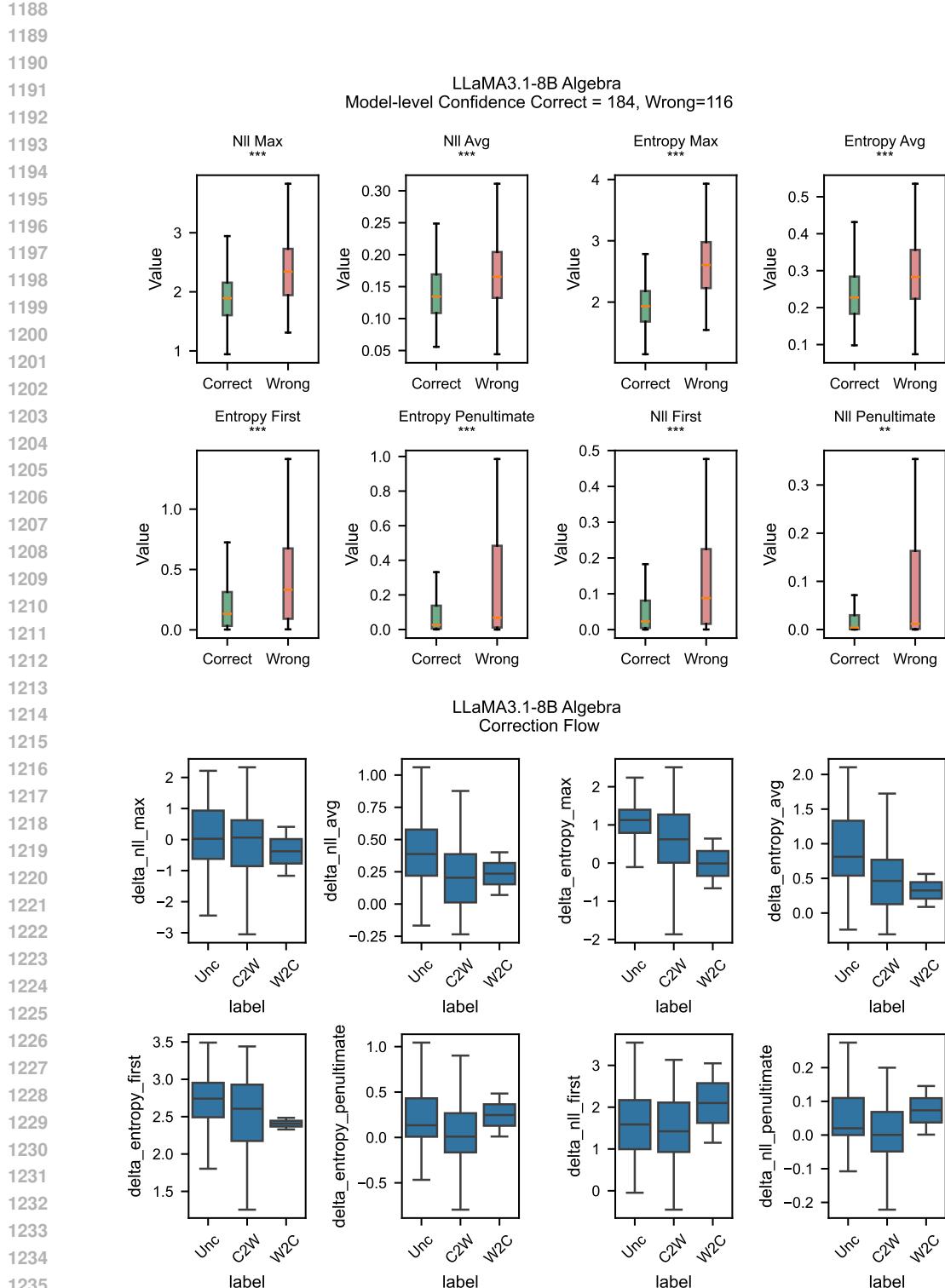


Figure 7: Top: Model-level Confidence result on the Math Algebra dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

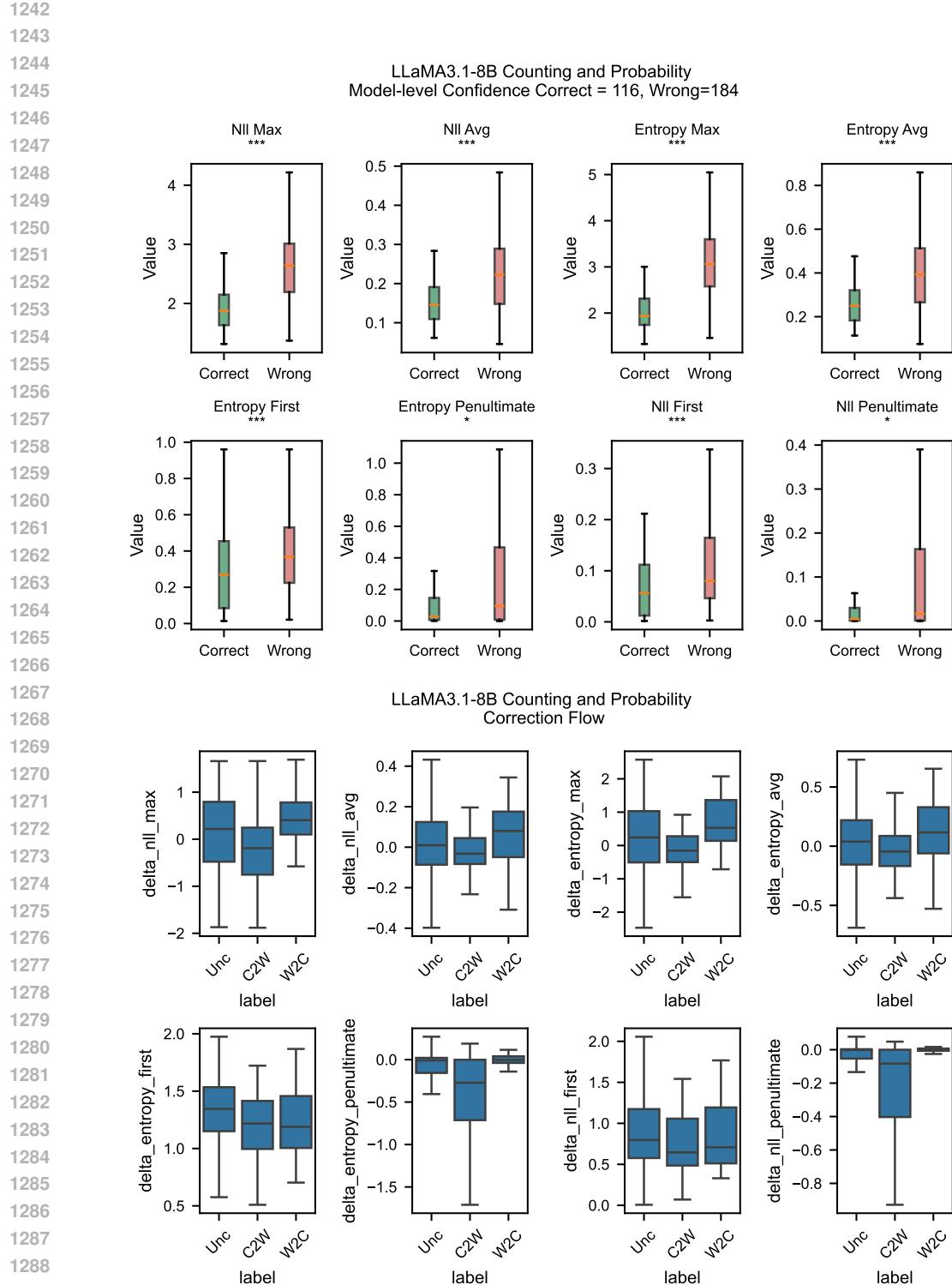


Figure 8: Top: Model-level Confidence result on the Math Counting and Probability dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

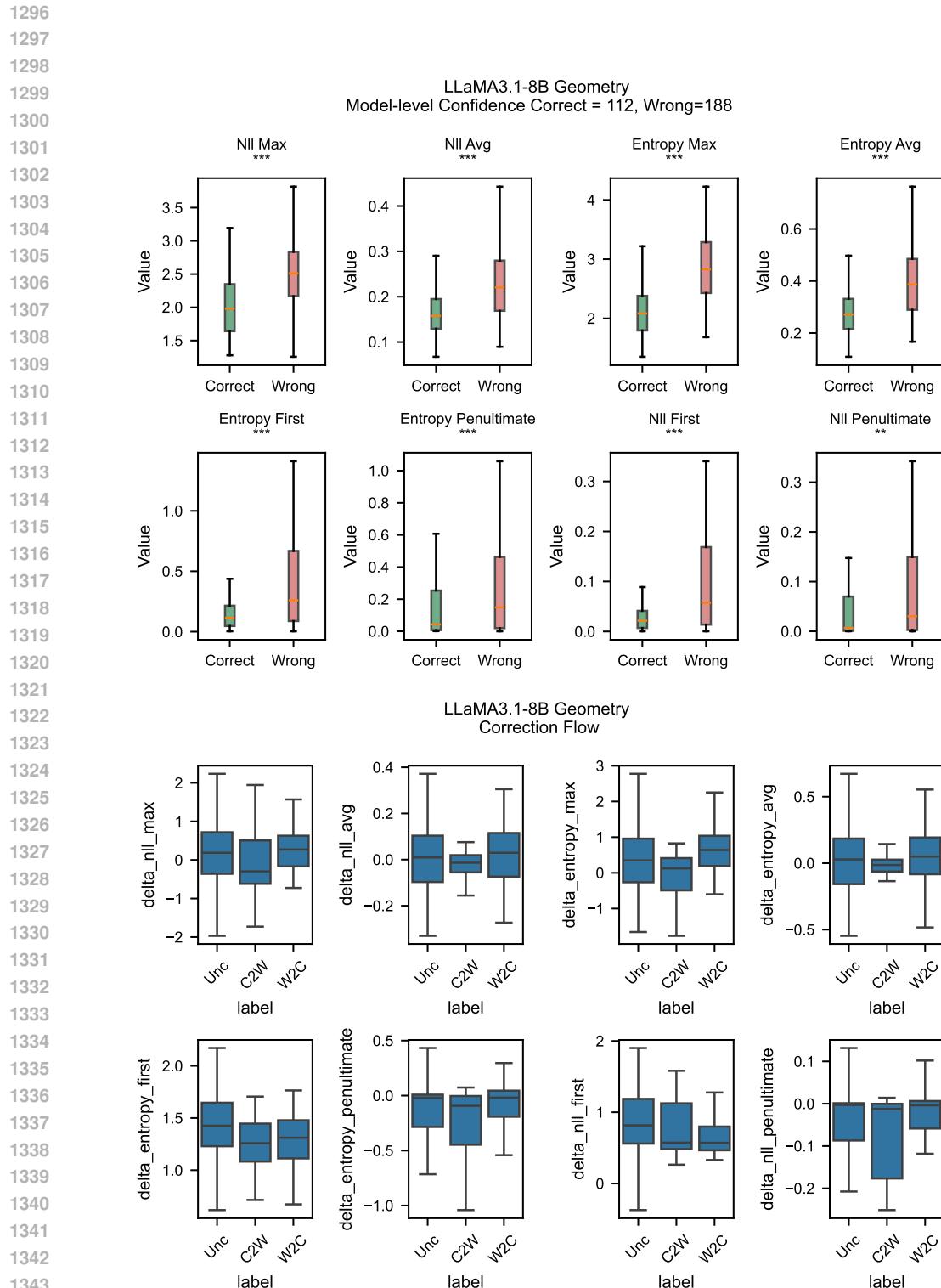


Figure 9: Top: Model-level Confidence result on the Math Geometry dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

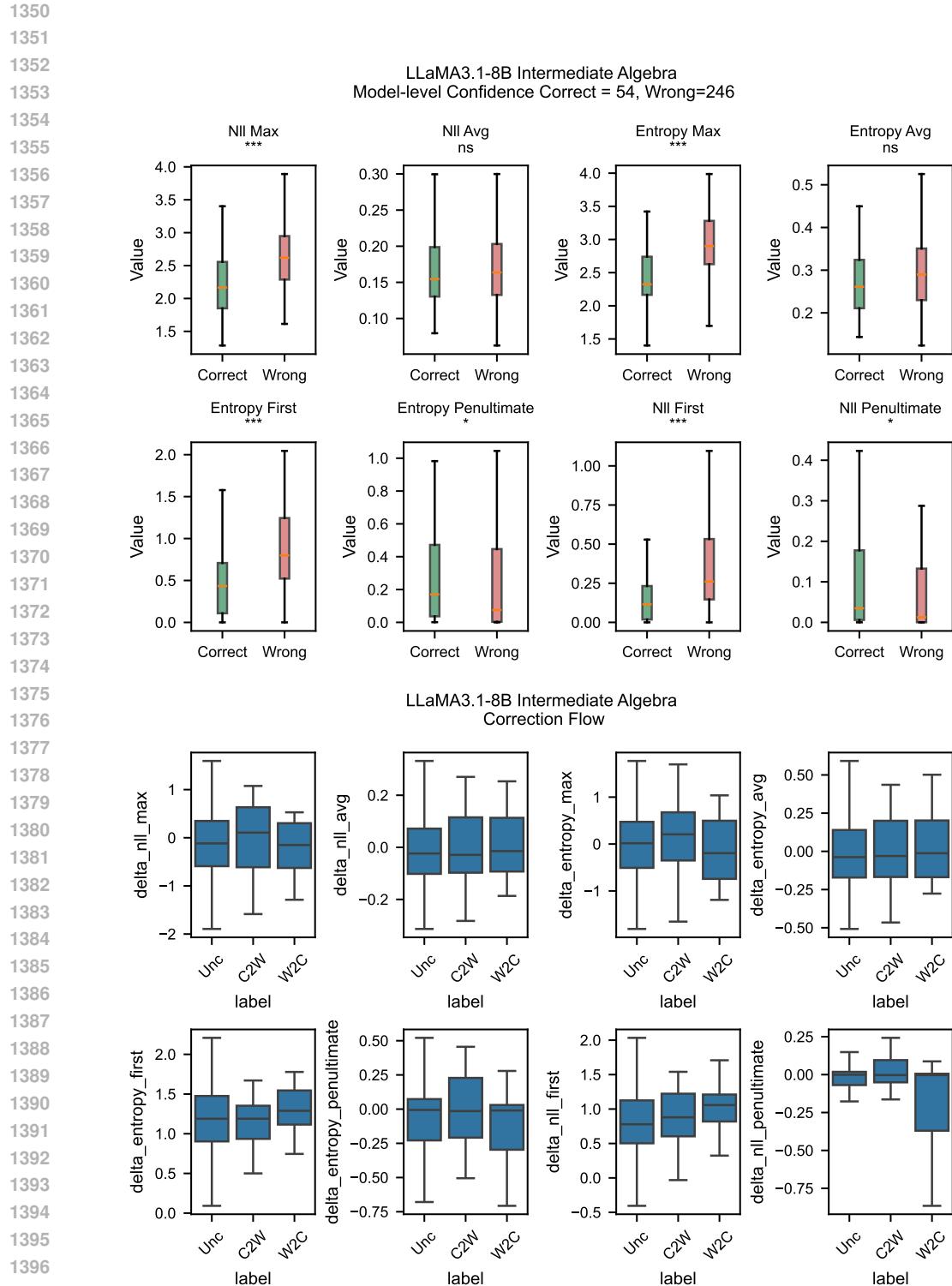


Figure 10: Top: Model-level Confidence result on the Math Intermediate Algebra dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

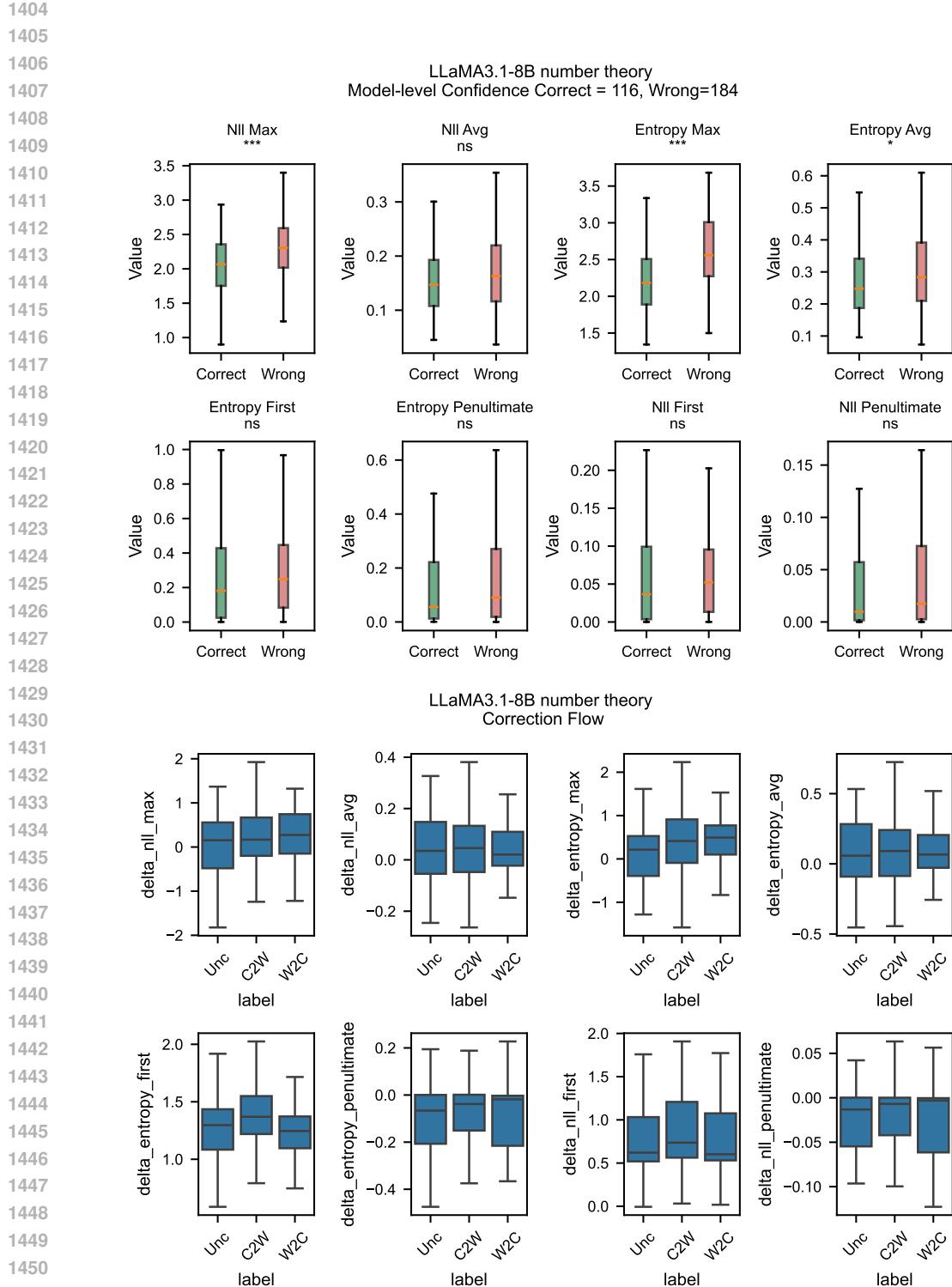


Figure 11: Top: Model-level Confidence result on the Math Number Theory dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

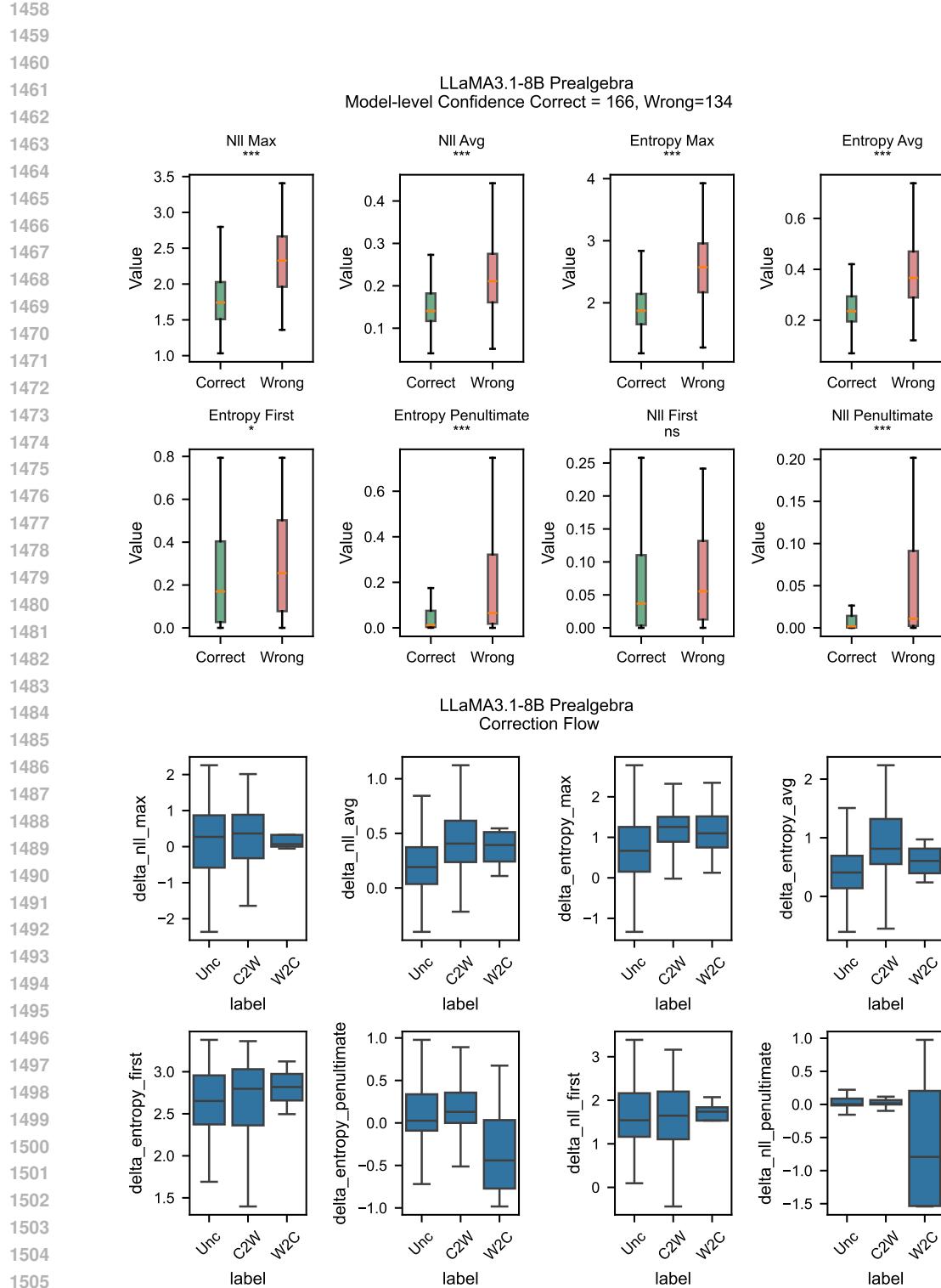


Figure 12: Top: Model-level Confidence result on the Math Prealgebra dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

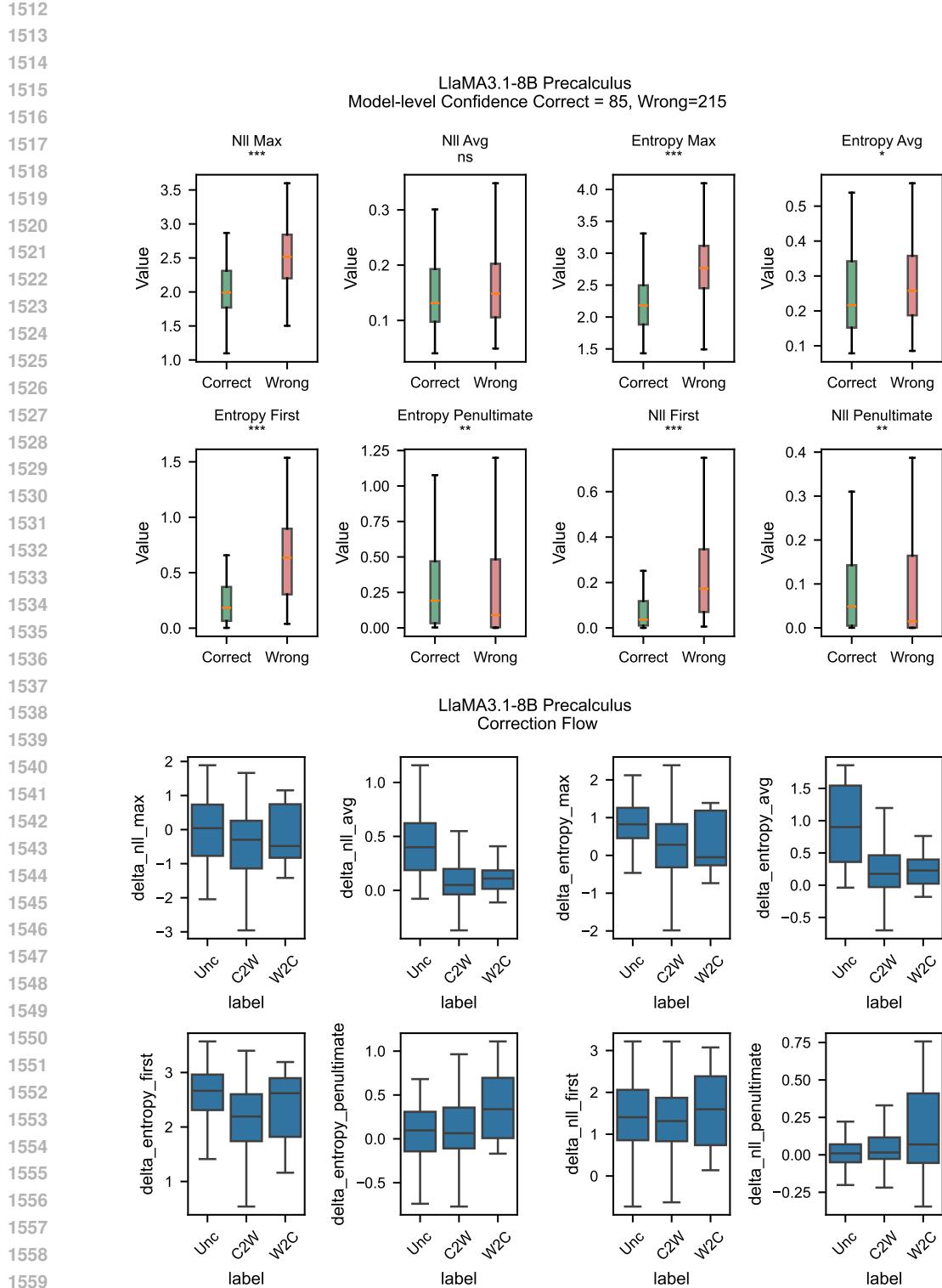


Figure 13: Top: Model-level Confidence result on the Math Precalculus dataset with LLaMA3.1-8B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

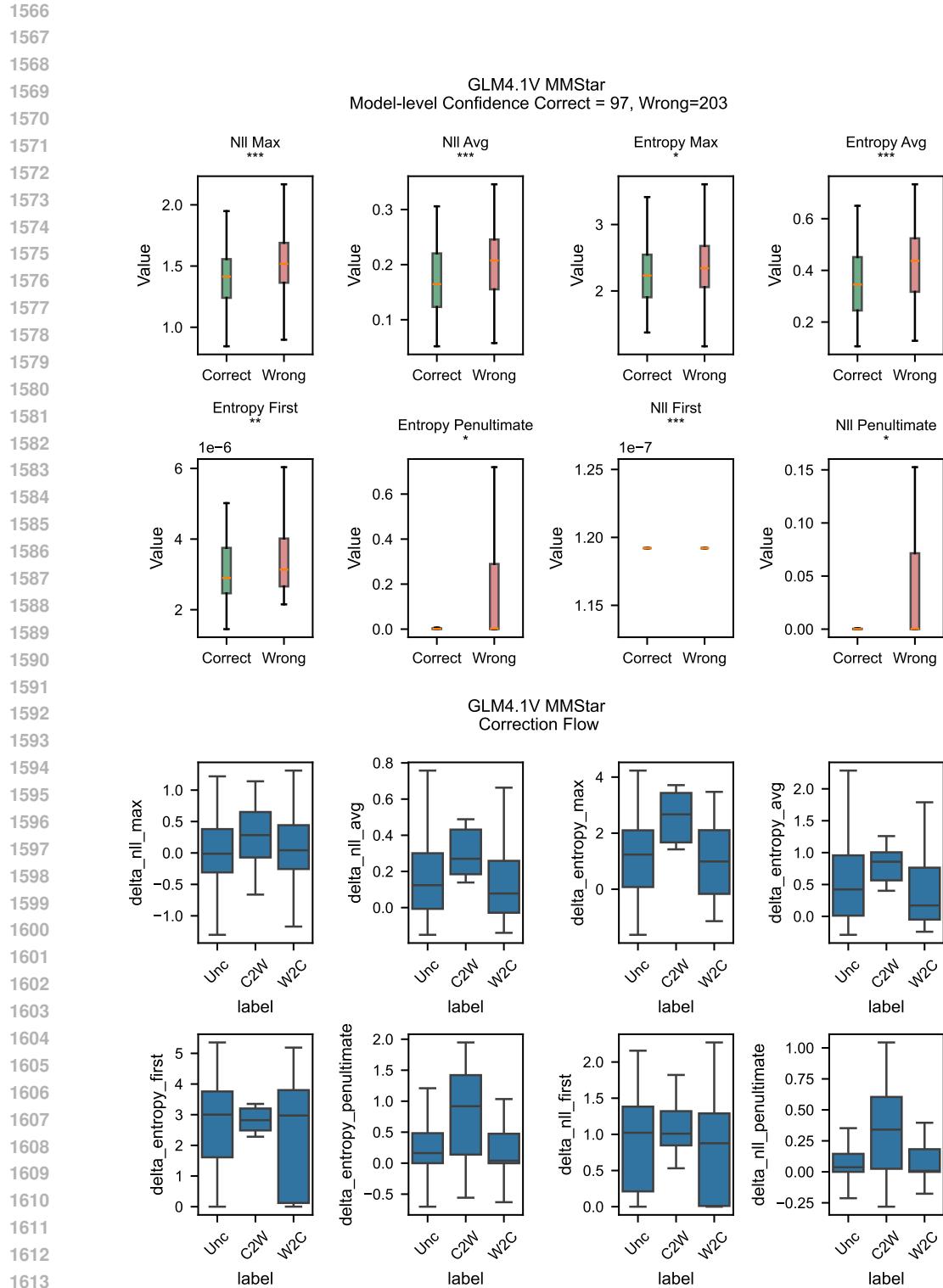


Figure 14: Top: Model-level Confidence result on the MMStar dataset with GLM4.1V. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

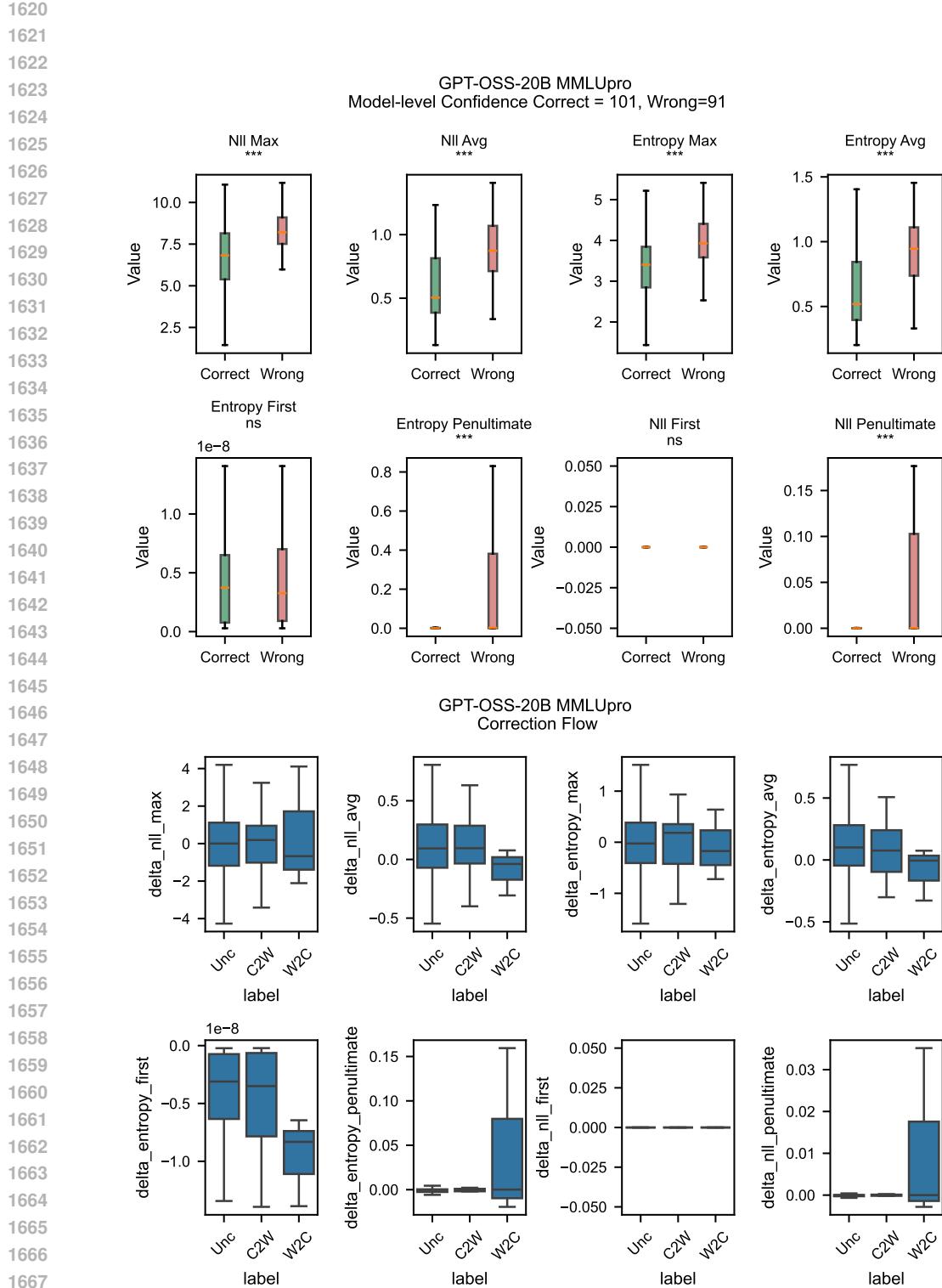


Figure 15: Top: Model-level Confidence result on the MMLUpro dataset with GPT-OSS-20B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

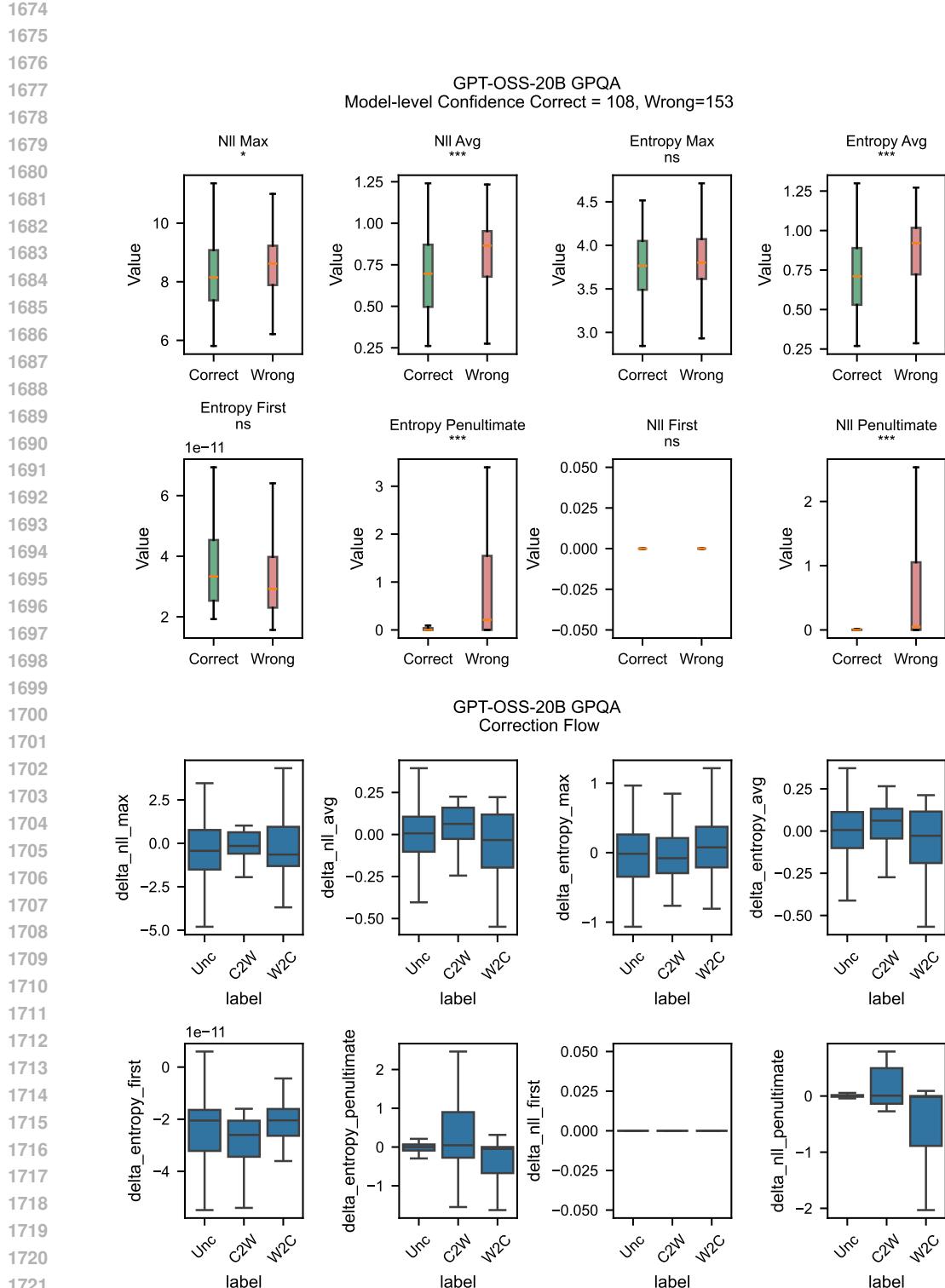


Figure 16: Top: Model-level Confidence result on the GPQA dataset with GPT-OSS-20B. Bottom: Correction flow with 8 deltas of confidence metrics. Unc: the answer remains unchanged between the first round and the last round, C2W: correct to wrong, W2C: wrong to correct.

1728  
1729  
1730  
1731  
1732  
1733

## Examples of Model-Level Confidence Exit on MMLUpro

1734  
1735  
1736  
1737  
1738  
1739  
1740  
1741  
1742  
1743  
1744  
1745  
1746  
1747  
1748  
1749  
1750  
1751  
1752  
1753  
1754  
1755  
1756  
1757  
1758  
1759  
1760  
1761  
1762  
1763  
1764  
1765  
1766  
1767  
1768  
1769  
1770  
1771  
1772  
1773  
1774  
1775  
1776  
1777  
1778  
1779  
1780  
1781

**<Question, Correct Answer: B>**

Can you answer the following question? If a freight train travels at a speed of 20 miles per hour for 6 hours, how far will it travel?: (A) 90 miles, (B) 120 miles, (C) 150 miles, (D) 140 miles, (E) 26 miles, (F) 100 miles, (G) 80 miles, (H) 110 miles, (I) 160 miles, (J) 12 miles.

**<Agent Answer: B> **

To find the distance traveled by the freight train, we need to multiply its speed by the time it traveled.  
Speed = 20 miles per hour  
Time = 6 hours  
Distance = Speed x Time  
= 20 miles/hour x 6 hours  
= 120 miles  
So, the freight train will travel 120 miles.  
The correct answer is (B).

**<Question, Correct Answer: E>**

Can you answer the following question? A driver and passenger were driving to work in the former's automobile one morning. As they were traveling at a speed of 20 m. p. h. (which was within the posted speed limit), the passenger suddenly pointed to an overturned vehicle along the side of the highway and said, "Look at that car upside down." The driver turned to look at the overturned vehicle. As he was looking toward the side of the road, the driver failed to see an abandoned vehicle with a flat tire in the highway about 200 feet in front of his approaching auto. Seconds later, the driver crashed into the rear of the abandoned auto and was injured. The jurisdiction has a relevant comparative negligence statute in effect. If the driver asserts a claim against the owner of the abandoned auto, the most likely result is that the driver will: (A) recover nothing, because the passenger's act was a supervening cause., (B) recover only a portion of damages, because both the driver and the owner of the abandoned auto were negligent., (C) recover nothing, because he was distracted while driving., (D) recover all of his damages, because the owner of the abandoned auto was negligent., (E) recover only a portion of damages, because the abandoned auto was in plain view., (F) recover all of his damages, because the abandoned auto was a traffic hazard., (G) recover all of his damages, because the defendant created a dangerous condition., (H) recover nothing, because he was driving at a speed that didn't allow him to stop in time., (I) recover nothing, because he had the last clear chance to avoid the collision..

**<Agent Answer: B>  [But Judged by Gemini and GPT, both B and E are correct.]**

(B) To determine the most likely result, let's break down the situation step by step:  
1. The driver was driving at a speed of 20 m.p.h. within the posted speed limit.  
2. The passenger pointed out the overturned vehicle, distracting the driver and causing him to take his eyes off the road.  
3. As a result, the driver failed to see the abandoned vehicle with a flat tire in front of him.  
4. The driver crashed into the abandoned vehicle.  
Given the comparative negligence statute, the jurisdiction likely follows a modified comparative negligence system. In this system, a plaintiff can recover damages even if they were partially at fault, but their recovery is reduced by the percentage of their own negligence.  
The driver's distraction caused by the passenger's comment likely contributed to the accident, making him partially responsible for the collision. However, the owner of the abandoned auto also created a hazard by leaving the vehicle on the road. Therefore, both parties contributed to the accident.  
The most likely result is that the driver will recover only a portion of damages, as both the driver and the owner of the abandoned auto were negligent.

Figure 17: Examples of model-level early exit cases on MMLUpro datasets. The exit questions are correctly answered in principle.

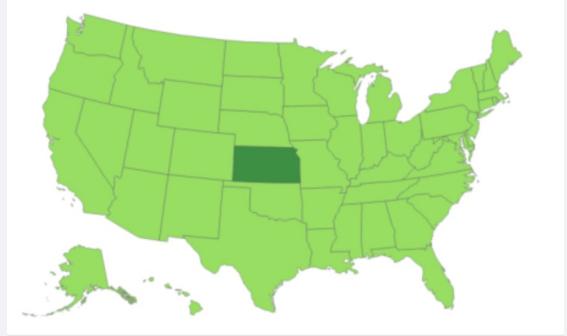
1782  
1783  
1784  
1785  
1786  
1787

## Examples of Model-Level Confidence Exit on ScienceQA

1788  
1789  
1790  
1791  
1792  
1793  
1794  
1795  
1796  
1797  
1798  
1799  
1800  
1801  
1802  
1803  
1804  
1805  
1806  
1807  
1808  
1809  
1810  
1811  
1812  
1813  
1814  
1815  
1816  
1817  
1818  
1819  
1820  
1821  
1822  
1823  
1824  
1825  
1826  
1827  
1828  
1829  
1830

<Question, Correct Answer: C>  
Question:  
What is the capital of Kansas?  
Options:  
(A) Des Moines  
(B) Wichita  
(C) Topeka  
(D) Lincoln  
Lecture:

<Agent Answer: C>   
The answer is (C) Topeka.



1831  
1832  
1833  
1834  
1835

<Question, Correct Answer: C>  
Question:  
Which country is highlighted?  
Options:  
(A) Barbados  
(B) the Dominican Republic  
(C) Saint Lucia  
(D) Saint Vincent and the Grenadines  
Lecture:

<Agent Answer: C>   
The answer is (C) Saint Lucia



Figure 18: Examples of model-level early exit cases on ScienceQA datasets. The exit questions are correctly answered in principle.

1836  
 1837  
 1838  
 1839  
 1840  
 1841  
 1842  
 1843

## Examples of Model-Level Confidence Exit on Math

1844 <Question, Correct Answer: [1, 4]>

1845 Question:

1846 Given a mathematics problem, determine the answer.

1847 Question: Compute the domain of the real-valued function  $f(x) = \sqrt{1 - \sqrt{2 - \sqrt{x}}}$ .

1848 Your final answer should be in the form `boxed{answer}`, at the end of your response.

1849 <Agent Answer: [1, 4]> 

1850 To find the domain of the function  $f(x) = \sqrt{1 - \sqrt{2 - \sqrt{x}}}$ , we need to ensure that the expression under the  
 1851 square roots is non-negative.  
 1852 First, we start with the innermost expression,  $\sqrt{x}$ . Since the square root of a  
 1853 number is non-negative, we know that  $x \geq 0$ .  
 1854 Next, we consider the expression  $\sqrt{2 - \sqrt{x}}$ . For this to be non-negative, we need  $2 - \sqrt{x} \geq 0$ , which implies  $x \leq 4$ .  
 1855 Finally, we consider the expression  $\sqrt{1 - \sqrt{2 - \sqrt{x}}}$ . For this to be non-negative, we need  $1 - \sqrt{2 - \sqrt{x}} \geq 0$ .  
 1856 Solving for  $\sqrt{2 - \sqrt{x}}$ , we get  $\sqrt{2 - \sqrt{x}} \leq 1$ , which implies  $2 - \sqrt{x} \leq 1$ . Simplifying, we get  $\sqrt{x} \geq 1$ , which implies  $x \geq 1$ .  
 1857 Combining the conditions, we have  $1 \leq x \leq 4$ . Therefore, the domain of the function is `boxed{[1, 4]}`.

1859 <Question, Correct Answer: 7>

1860 Given a mathematics problem, determine the answer.

1861 Question: Evaluate  $\left\lceil \left( \frac{3}{2} \right)^2 + \left\lceil \left( \frac{3}{2} \right)^2 \right\rceil \right\rceil$ .

1862 Your final answer should be in the form `boxed{answer}`, at the end of your response.

1863 <Agent Answer: 9> 

1864 To evaluate this expression, we need to follow the order of operations (PEMDAS):  
 1865 Evaluate the expressions inside the parentheses:  $\left( \frac{3}{2} \right)^2 = \frac{9}{4} = 2.25$ . Evaluate the ceiling function of the result:  $\lceil 2.25 \rceil = 3$ .  
 1866 Evaluate the expression inside the other set of parentheses:  $\lceil 3 \rceil = 3$ .  
 1867 Evaluate the ceiling function of the result:  $\lceil 3 \rceil = 3$ .  
 1868 Square the result of step 3:  $3^2 = 9$ .  
 1869 Add the results of steps 2 and 5:  $3 + 9 = 12$ .  
 1870 Therefore, the final answer is `boxed{12}`.

1885 Figure 19: Examples of model-level early exit cases on Math datasets. The bottom overconfident  
 1886 case shows the model partially analyzes the question, potentially leading to a high confidence.

1887  
 1888  
 1889

1890  
1891  
1892  
1893  
1894  
1895  
1896  
1897

## Examples of Model-Level Confidence Exit on GPQA

1898 <Question, Correct Answer: B>

Can you answer the following question? In the vacuum, we have the following mixture  

$$\left| \nu_{\text{e}} \right\rangle = e^{i p_1 x} \cos \theta \left| \nu_{\text{e}} \right\rangle + e^{i p_2 x} \sin \theta \left| \nu_{\text{e}} \right\rangle$$
where  $\theta$  is the mixing angle, and  $\left| \nu_{\text{e}} \right\rangle$  and  $\left| \nu_{\text{e}} \right\rangle$  are the basis of mass eigenstates.  
At what value of the mixing angle we will obtain the transition probability  $P(\left| \nu_{\text{e}} \right\rangle \rightarrow \left| \nu_{\mu} \right\rangle) = 1$ .  
Choices: (A)  $\pi/2$  (B)  $\pi/4$  (C)  $\pi$  (D)  $\pi/3$ .

1903 <Agent Answer: B>

analysis: We need to analyze neutrino flavor mixing: initial state is electron neutrino? They give general state  $|\nu_{\text{e}}(x)\rangle = e^{i p_1 x} \cos \theta |\nu_{\text{e}}\rangle + e^{i p_2 x} \sin \theta |\nu_{\text{e}}\rangle$ . The probability of transition from  $\nu_{\text{e}}$  to  $\nu_{\mu} = \sin^2 2\theta \sin^2(\Delta m^2 L / (4E))$ . For maximum probability 1, we need  $\sin^2 2\theta = 1$  AND  $\sin^2$  phase term = 1 (i.e., argument =  $\pi/2 + n\pi$ ). But question asks at what mixing angle we will obtain the probability 1 independent of oscillation? If you want probability 1, you need perfect mixing:  $0 = \pi/4$  gives maximal mixing so  $\sin^2 2\theta = \sin^2(\pi/2) = 1$ . So answer (B)  $\pi/4$ . So "The correct answer is (B)". assistantFinalIn a two-flavor framework the transition probability is  $\left| \langle \nu_{\text{e}} | \nu_{\mu} \rangle \right|^2$  where  $\theta$  is the mixing angle and the second sine factor contains the usual oscillation phase. To make the probability equal to unity for some distance  $L$  we need both factors to be unity. The phase factor can become unity for suitable  $L$  (e.g. when  $\Delta m^2 L / (4E) = \pi/2$ ). The remaining requirement is therefore  $\sin 2\theta = 1$ ;  $\sin 2\theta = \pi/2$ . The number of stars per unit distance is  $\propto \sin^2 2\theta$ . Among the provided options, the value that gives complete mixing is  $\pi/4$ . Therefore the correct choice is (B). The correct answer is (B).'

1915 <Question, Correct Answer: B>

Can you answer the following question? In a specific region of the sky, astronomers have observed that the number of stars varies with parallax as  $1/\text{plx}^5$ . How does the number of stars in that region of the sky change with distance (per unit range of distance,  $r$ )? Choices: (A)  $\sim r^4$  (B)  $\sim r^3$  (C)  $\sim r^2$  (D)  $\sim r^5$ .

1919 <Agent Answer: B>

analysis: We need to understand relationship between parallax and distance: parallax  $p$  (in arcseconds) approx  $1/d(\text{pc})$  maybe? For small parallax:  $p = 1/d$  in arcsec if distance in parsecs. So  $d = 1/p$ . Number of stars varying as  $1/p^5$ . Number of stars per parallax interval? They say number of stars varies with parallax as  $1/\text{plx}^5$ . That implies count  $N(p) \sim 1/p^5$ . We want how number per unit range of distance  $r$  behaves:  $N(r)$  per unit distance? The distribution of stars per unit distance range:  $dN/dr$ . The relation between  $p$  and  $r$ :  $p = 1/r$  (assuming  $p$  in arcsec,  $r$  in parsecs). So  $r = 1/p$ . So  $p = 1/r$ . So  $N(p) \sim 1/p^5$ . But  $N(p)$  is number in certain parallax interval maybe? We need convert to  $N(r)$ . Suppose number per parallax interval  $dp$  is proportional to  $1/p^5$ . The number per unit  $r$ :  $dN/dr = dN/dp * dp/dr$ .  $dN/dp = k * 1/p^5$ .  $dp/dr = d(1/r)/dr = -1/r^2$ . We take magnitude:  $|dp/dr| = 1/r^2$ . Express  $p$  in terms of  $r$ :  $p = 1/r$ . So  $(1/p^5) = 1/(1/r)^5 = r^5$ . So  $dN/dp = k * r^5$ . Multiply by  $|dp/dr| = 1/r^2$  gives  $dN/dr = k * r^5 * (1/r^2) = k * r^3$ . So number per unit distance  $\sim r^3$ . So answer (B). That matches typical uniform space density:  $N(r) \sim r^3$ . So the correct answer is (B). assistantFinalThe relationship between parallax  $p$  and distance  $r$  is  $p = 1/r$  (with  $p$  in arcseconds and  $r$  in parsecs). The number of stars as a function of parallax is given as  $N(p) \propto p^{-5}$ . To find the differential number per unit distance we use the chain rule:  $dN/dr = \frac{dN}{dp} \frac{dp}{dr} = \frac{dN}{dp} \frac{1}{r^2} = k * r^3$ . From  $(p = 1/r)$ , we have  $(dp/dr) = -1/r^2$ ; taking the magnitude,  $|dp/dr| = 1/r^2$ . Substitute  $(p = 1/r)$  into  $dN/dr = k * r^3$ . Hence, the number of stars per unit range of distance  $r$  scales as  $r^3$ . The correct answer is (B). (We did not find negative cases when using model-level early exit.)

1933  
1934  
1935  
1936  
1937  
1938

1939 Figure 20: Examples of model-level early exit cases on GPQA datasets, where we did not find  
1940 negative cases (i.e., the model is overconfident).

1941  
1942  
1943

1944  
1945  
1946  
1947  
1948  
1949  
1950

## Examples of Token-Level Compression

1951 <Token-Level Semantic Focus Analysis>  
1952 Below are the key points where other agents \*\*disagree\*\* with your own reasoning.\*  
1953 Concentrate on those disagreements and Keep or revise your answer accordingly.  
1954  
1955 <Copy the first agent message as the content is determined by internal attention that would be useful for current agent>  
1956 **One agent solution:** \*\*\*analysisWe need interpret reaction: 9-fluorenone + 2 equivalents of Selectfluor under photochemical? 400nm? Then product  
1957 observed via 19F NMR with two fluorine signals: +29.8 ppm and -115.7 ppm. Which product matches? We need consider structure of products.  
1958 9-fluorenone has no fluorine. Fluorine signals appear from added fluorine atoms (from Selectfluor) presumably fluorination of the aromatic system. 19F  
1959 NMR chemical shifts: +29.8 ppm is typical for aromatic fluorine attached to sp<sup>2</sup> carbon in aromatic ring (like fluorobenzene ~7-8 ppm? wait  
1960 fluorobenzene is 7 ppm. Actually 19F NMR for fluoride attached to benzene ring: about -113 to -115 ppm. For CF<sub>3</sub> group ~ -65 ppm. For fluorines  
1961 bound to carbonyl? Let's recall typical shifts: In 1,2-difluoro compounds, one signal can be upfield and one downfield depending on environment. For  
1962 example, 2,4-difluoroaniline maybe at ~110 and -45? Hard.  
1963 Let's analyze options:  
1964 (A) 1,5-difluoro-9H-fluoren-9-one. This would be fluorine at positions 1 and 5 of the fluorenone core. Aromatic fluorine signals for 1,5-difluoro;  
1965 symmetrical? Two identical signals? But 19F NMR would show one signal (if equivalent)? Actually if two fluorines are meta to each other? Might be  
1966 equivalent? 1,5-difluoro: two positions are equivalent if molecule symmetrical (flipped). For 1,5-difluoro-9-fluorenone, the two fluorines would be in  
1967 symmetric positions. Would give a single 19F signal. But question says two fluorine signals, so (A) unlikely.  
1968 (C) 2,6-difluoro-9H-fluoren-9-one: also symmetrical? 2,6 are para on each phenyl ring? The two fluorines are in equivalent positions (mirror plane  
1969 through center). Might again give single signal. So (C) unlikely.  
1970 (B) 2-fluoro-[1,1'-biphenyl]-2-carbonyl fluoride. This is a different product: a biphenyl with fluorine on one ring and a carbonyl fluoride on other. This  
1971 would have a single fluorine (the one at 2' position). Therefore only one 19F signal. So also unlikely.  
1972 (D) 4'-fluoro-[1,1'-biphenyl]-2-carbonyl fluoride. Also single fluorine.  
1973 ...  
1974 **One agent solution:** \*\*\*analysisWe need to analyze the described reaction: 9-fluorenone (structure: a fluorenone: a fused ring (fluorene) with a  
1975 carbonyl at center? Actually 9-fluorenone is a 9-carbonyl in fluorene. 2 equivalents of Selectfluor used in acetonitrile, irradiated with 400 nm light.  
1976 Reaction product has two fluorine signals: +29.8 ppm and -115.7 ppm in 19F NMR. Need to identify which structure among the options is correct  
1977 product. Let's parse the options:  
1978 (A) 1,5-difluoro-9H-fluoren-9-one  
1979 (B) 2'-fluoro-[1,1'-biphenyl]-2-carbonyl fluoride  
1980 (C) 2,6-difluoro-9H-fluoren-9-one  
1981 (D) 4'-fluoro-[1,1'-biphenyl]-2-carbonyl fluoride  
1982  
1983 <Delete useless background information>  
1984 We need to consider the reaction conditions. Selectfluor is used as a fluorination agent. Selectfluor is known to be a strong oxidant and fluorinating  
1985 reagent, can generate radical cation. 9-Fluorenone has a central carbonyl. But 2 equivalents of Selectfluor might produce two fluorine atoms.  
1986  
1987 <Keep useful induction>  
1988 Looking at the product options: Both (B) and (D) are biphenyl derivatives with a carbonyl fluoride group at the 2-position of the biphenyl core. The  
1989 difference is the position of a fluorine substituent on the biphenyl ring: 2' vs 4'.  
1990  
1991 <Cut out useless reasoning paths>  
1992 Options (A) and (C) are difluoro-substituted fluorene-9-one. 1,5-difluoro vs 2,6-difluoro arrangement on the fluorene core. Those are fluorinated at the  
1993 aromatic positions adjacent to the carbonyl (perhaps positions that are 1 and 5, or 2 and 6 referencing fluorene numbering). But do we expect  
1994 difluorination of the aromatic ring? Maybe the reaction could be electrophilic aromatic substitution with electrophilic fluoride; but Selectfluor is a  
1995 fluorinating agent, providing F<sup>-</sup>. With 9-fluorenone, electrophilic aromatic substitution would likely occur at positions ortho/para to the carbonyl? etc)  
1996 perhaps the substituents on the ring positions adjacent to the central carbonyl. I need to draw structure:  
1997 The numbering scheme: positions 1, 3, 5, 7 correspond to the outer ring carbons; Let's quickly sketch: The central carbonyl is at position 9. I need to  
be precise. Let's step aside: We might not need exact numbering\*\*\*  
1998 Using the reasoning from other agents as additional advice, can you give an updated answer? Examine your solution and that other agents step by  
1999 step. Let's think step by step. Your last sentence should be 'The correct answer is (insert answer here).' e.g., 'The correct answer is (A).'  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009  
2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024  
2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036  
2037  
2038  
2039  
2040  
2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051  
2052  
2053  
2054  
2055  
2056  
2057  
2058  
2059  
2060  
2061  
2062  
2063  
2064  
2065  
2066  
2067  
2068  
2069  
2070  
2071  
2072  
2073  
2074  
2075  
2076  
2077  
2078  
2079  
2080  
2081  
2082  
2083  
2084  
2085  
2086  
2087  
2088  
2089  
2090  
2091  
2092  
2093  
2094  
2095  
2096  
2097

Figure 21: Detailed example of token-level semantic focus. There is content from two agents in the figure. The first agent's answer was largely adopted, whereas the second agent's answer was largely cropped. <comments> shows places that were adaptively compressed by the model's attention mechanism.

1998  
1999  
2000  
2001  
2002  
2003  
2004  
2005  
2006  
2007  
2008  
2009

2010  
2011  
2012  
2013  
2014  
2015  
2016  
2017  
2018  
2019  
2020  
2021  
2022  
2023  
2024

2025  
2026  
2027  
2028  
2029  
2030  
2031  
2032  
2033  
2034  
2035  
2036

**Token-Level Semantic Focus Cases**

Question: If Real GDP = \$200 billion and the price index = 200 Nominal GDP is:

**LLM Agent 1**

To find the Nominal GDP, we need to multiply the Real GDP by the price index. Real GDP = \$200 billion. Price Index = 200. Nominal GDP = Real GDP \* Price Index. Nominal GDP = \$200 billion \* 200. Nominal GDP = \$40,000 billion. However, none of the options match this calculation. This might be due to the fact that the question is based on a typical assumption that the price index is a percentage increase. In this case, the correct answer should be the Real GDP multiplied by  $(1 + (\text{Price Index} / 100))$ . Price Index = 200. Price Index Percentage =  $(200 / 100) = 2$ . Nominal GDP = Real GDP \*  $(1 + \text{Price Index Percentage})$ . Nominal GDP = \$200 billion \*  $(1 + 2)$ . Nominal GDP = \$200 billion \* 3. Nominal GDP = \$600 billion.

Wrong

**LLM Agent 2**

To solve this, we need to understand the relationship between Real GDP and Nominal GDP. Real GDP is the value of goods and services produced in a country, adjusted for inflation. Nominal GDP, on the other hand, is the value of goods and services produced in a country, without adjusting for inflation. The formula to find Nominal GDP from Real GDP is: Nominal GDP = Real GDP \* (Price Index / 100). Given that Real GDP = \$200 billion and the price index = 200, we can plug in the values: Nominal GDP = \$200 billion \*  $(200 / 100)$ . Nominal GDP = \$200 billion \* 2. Nominal GDP = \$400 billion  
So, the correct answer is (I) \$400 billion.

Correct

**LLM Agent 3**

To find the nominal GDP, we need to multiply the real GDP by the price index. Nominal GDP = Real GDP \* Price Index. Nominal GDP = \$200 billion \* 200.  
Nominal GDP = \$40 trillion \* 5 (since  $200/5 = 40$ )  
Nominal GDP = \$40 trillion \* 10 (since  $200/20 = 10$ ) is not correct but  $200/5 = 40$  is correct.  
Nominal GDP = \$40 trillion \* 5 (since  $200/5 = 40$ )  
Nominal GDP = \$40 trillion \* 10 (since  $200/20 = 10$ ) is not correct but  $200/5 = 40$  is correct.  
...

Wrong

**Token-level Semantic Focus | Agent 1's Debate Content**

<compressed LLM Agent 2's Content Start> To solve this, we need to understand the relationship between Real GDP and Nominal GDP. Real GDP is the value of goods and services produced in a country, ... <compressed LLM Agent 2's Content End> <compressed LLM Agent 3's Content Start> To find the nominal GDP, we need to multiply the real GDP by the price index. ... Nominal GDP = Real GDP \* Price Index. Nominal GDP = \$200 billion \* 200 <compressed LLM Agent 3's Content End>

**Token-level Semantic Focus | Agent 3's Debate Content**

<compressed LLM Agent 1's Content Start> To find the Nominal GDP, we need to multiply the Real GDP by the price index. ... Real GDP = \$200 billion. Price Index = 200. <compressed LLM Agent 1's Content End>  
<compressed LLM Agent 2's Content Start> To solve this, we need to understand the relationship between Real GDP and Nominal GDP. Real GDP is the value of goods and services produced in a country, adjusted for inflation. ... <compressed LLM Agent 2's Content End>

**LLM Agent 1**

I see that they are all calculating the Nominal GDP by multiplying the Real GDP by the price index. However, one agent is using the correct formula to adjust for the price index, which is Nominal GDP = Real GDP \* (Price Index / 100). Real GDP = \$200 billion. Price Index = 200 Nominal GDP = Real GDP \* (Price Index / 100) Nominal GDP = \$200 billion \*  $(200 / 100)$ . Nominal GDP = \$200 billion \* 2. Nominal GDP = \$400 billion

Correct

**LLM Agent 3**

Upon reviewing my previous solution and the additional advice from other agents, I realize that I made an incorrect assumption about the price index. Let's re-evaluate the problem with the correct approach. ... Nominal GDP = Real GDP \* (Price Index / 100). Given that Real GDP = \$200 billion and the price index = 200, we can plug in the values: Nominal GDP = \$200 billion \*  $(200 / 100)$ . Nominal GDP = \$200 billion \* 2. Nominal GDP = \$400 billion.

Correct

Figure 22: Visualization of token-level semantic focus helped agents correct their answers during the SID. Specifically, SID invokes a compressed debate round, highlighting disagreement-relevant spans across debate contents. Agent 1 and Agent 3 iteratively revise their reasoning based on focused inputs, ultimately correcting earlier errors and converging on the correct answer.

2041  
2042  
2043  
2044  
2045  
2046  
2047  
2048  
2049  
2050  
2051