DialogEval: A Benchmark for Cross-Framework Annotation Capability of Large Language Models in Classroom Dialogue Analysis

Anonymous ACL submission

Abstract

Automated Classroom Dialogue Encoding (ACDE) is critical for pedagogical improvement, yet LLMs' ability to grasp deep structures remains under-explored. We propose DialogEval, a benchmark utilizing the Three Bs framework to evaluate decoding capabilities between, behind, and beyond the words. We evaluated eight LLMs-encompassing reasoning-enhanced, general-purpose, and domain-specific families-across FIAC, IRF, and SEDA frameworks. These serve as definitive representatives of the Quantitative Interaction, Linguistic Structure, and Sociocultural schools, respectively, establishing an evolutionary gradient of cognitive demands. Using a multi-disciplinary corpus of 20 high-quality sessions, we identify a Logic Threshold for contextual dependency. Results show that logic-driven reasoning (CoT) outperforms dense rule manuals, which paradoxically trigger cognitive overload. Furthermore, domain-specific models exhibit Domain Rigidity and Proactive Inhibition, where prior schemas interfere with novel frameworks. This study offers robust diagnostic tools and theoretical insights for AI-assisted educational discourse analysis. The contributions of this paper are available at: https://acl-dialogeval.github.io/benchmark/.

1 Introduction

Language is not merely a vehicle for communication but the very medium of thought. As Vygotsky posited, higher mental functions originate on the social plane through interaction before being internalized as individual cognition (Vygotsky, 1978). In educational settings, this social genesis is operationalized through classroom discourse a process that Alexander and Mercer describe as "interthinking" where knowledge is collectively negotiated and constructed (Alexander, 2020; Mercer, 2002). Consequently, Classroom Discourse Analysis (CDA) serves as a vital lens for **unpacking the black box of teaching effectiveness, revealing how these linguistic interactions fundamentally shape learning outcomes** (Cazden, 1988).

However, the pedagogical intent of interactions is rarely explicit in the surface-level text. Drawing on Bakhtin's dialogism, we recognize that discourse is inherently situated; meaning emerges from intersubjectivity between participants (Nystrand et al., 1997). In authentic classrooms, surface-level utterances often mask underlying realities: a "question" may functionally be a disciplinary command, and "feedback" may serve as a rhythmic signal rather than a cognitive evaluation. Thus, the "ground truth" of educational dialogue does not lie in the text itself, but exists Between The Words (contextual dependencies), Behind The Words (pedagogical intent), and Beyond The Words (sociocultural norms and theoretical perspectives). Decoding these hidden dimensions has traditionally relied on expert human coding, a process that is resource-intensive and difficult to scale.

Notably, the transformative potential of LLMs in Automated Classroom Dialogue Encoding (ACDE) is increasingly recognized. While pioneering work demonstrates proficiency in surface-level linguistic features (Demszky et al., 2021; Yan et al., 2024), existing evaluations often treat discourse analysis as a generic text classification task, relying on single-turn prompts or shallow definitions. Such approaches rarely examine whether advanced strategies, like Chain-of-Thought (CoT), can bridge the gap between AI's statistical probabilities and the nuanced, theory-driven reasoning required by educational experts.

To bridge this gap, we present DialogEval, a benchmark evaluating LLMs' discourse decoding at varying cognitive depths via three research questions: **RQ1 (Between The Words):** How do

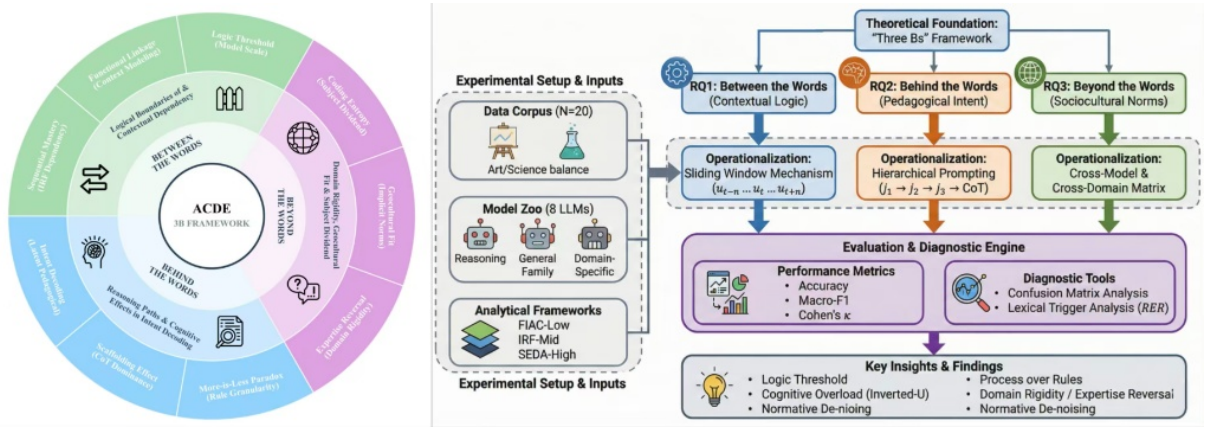

Figure 1: Overview of the ACDE 3B Framework (left) and its corresponding Experimental Methodology (right)

model scale and reasoning capabilities influence contextual dependency resolution? **RQ2 (Behind The Words):** How do rule-driven vs. logic-driven strategies affect pedagogical intent inference and cognitive load patterns? **RQ3 (Beyond The Words):** How do pre-existing domain knowledge and geocultural backgrounds reshape the interpretation of sociocultural norms?

We evaluate eight domestic and global LLMs across FIAC, IRF, and SEDA as definitive frameworks representing the quantitative, linguistic, and sociocultural schools, respectively. Spanning Science and Art disciplines, this multi-disciplinary study ensures both cognitive diversity and the generalizability of our findings.

In summary, our main contributions are: (1) **The Three Bs framework** (Fig.1) operationalizes ACDE into three cognitive dimensions: Between, Behind, and Beyond The words. (2) **A multi-disciplinary benchmark** evaluates eight representative domestic and international LLMs across diverse disciplines and geocultural contexts. (3) **Cognitive insights** reveal a Logic Threshold and provide empirical evidence of Proactive Inhibition, where granular rules paradoxically hinder domain-specific model performance.

2 Related Work

2.1 Evolution of Discourse Frameworks

Classroom discourse analysis has undergone a paradigm shift from **quantifying behavioral frequencies** to **evaluating cognitive quality**, as illustrated in Fig. 2. Historically, this evolution comprises three phases with increasing analytical depth. The **Quantitative Interaction School** (e.g., FIAC (Flanders, 1970), S-T Analysis (Fujita, 1979)) focuses on **observable event frequencies**, treating talk as explicit behavioral units. The **Linguistic Structure School** (e.g., IRF/E (Sinclair and Coulthard, 1975), TAP (Osborne et al., 2004)) shifts attention to **sequential patterns and logical scripts**. The contemporary **Sociocultural and Cognitive Dialogue School** (e.g., Mercer's Three Types (Mercer, 1995), SEDA (Hennessy et al., 2016)) targets **deep thinking and knowledge construction**, requiring high-inference interpretation of communicative functions. Crucially, this evolution implies a gradient of complexity: from identifying surface features to interpreting latent cognitive dynamics. ***This trajectory raises a fundamental question: To what extent do the varying inferential demands of these frameworks impose distinct challenges on Large Language Models (LLMs)?*** It remains to be verified whether high-inference sociocultural coding necessitates significantly more complex reasoning than low-inference quantitative tracking. **Investigating this potential disparity in coding complexity serves as the core motivation for the systematic exploration presented in this study.**

2.2 Evolution of Discourse Technologies

Classroom discourse analysis has evolved from **statistical inference** to **automated deep semantic understanding**. Early research utilized **statistical methods** like Lag Sequential Analysis on **manually coded data** (Bakeman and Gottman, 1997). While **traditional machine learning algorithms** (e.g., SVM, Naive Bayes) introduced automation, they **relied heavily on manual feature engineering** (Yang and Liu, 1999). Subsequently, **deep learning models** such as CNNs and LSTMs **enhanced sequential feature cap-**

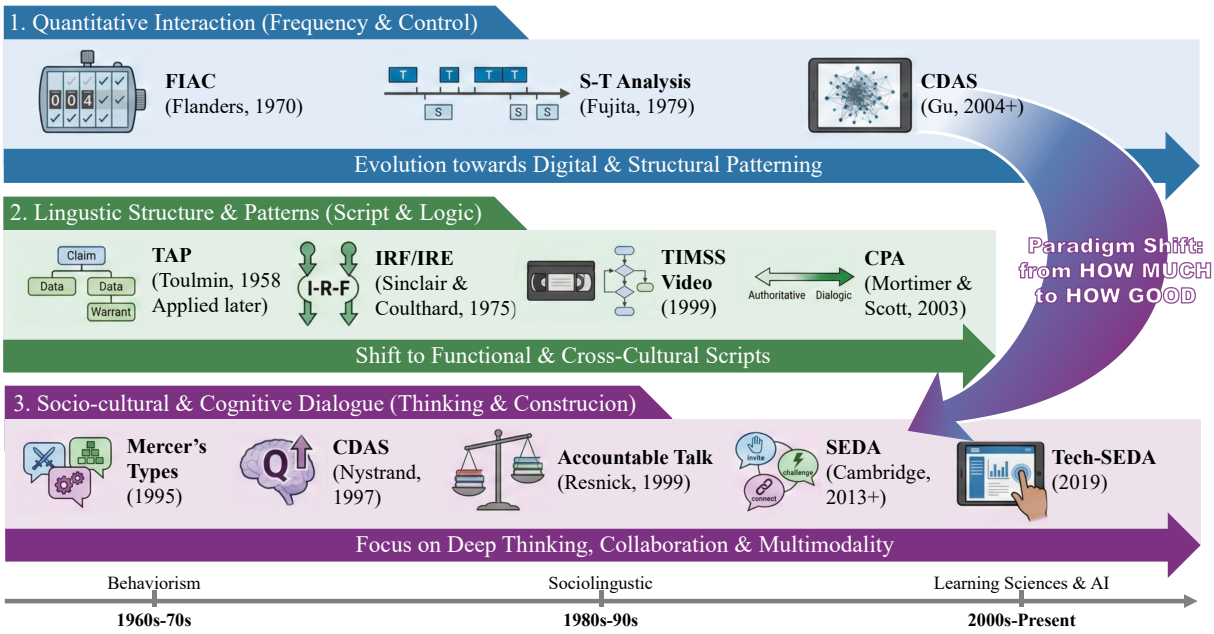

Figure 2: The evolutionary trajectory of classroom discourse analysis frameworks. The figure depicts three developmental phases: (1) Quantitative Interaction, (2) Linguistic Structure, and (3) Sociocultural & Cognitive Dialogue. The overarching trend indicates a paradigm shift from frequency-based "How Much" to quality-based "How Good," implying an increasing gradient of inferential complexity for automated coding.

ture (Kim, 2014), and **pre-trained Transformers** like BERT achieved state-of-the-art accuracy via **bidirectional contextual understanding** (Devlin et al., 2019). Furthermore, approaches including Epistemic Network Analysis (ENA) and Graph Neural Networks (GNNs) have **advanced the modeling of complex interaction structures** (Scarselli et al., 2008). Critically, supervised models remain constrained by their reliance on large-scale labeled data and difficulties in decoding high-inference structures like SEDA. LLMs mitigate these bottlenecks through emergent few-shot learning and Chain-of-Thought reasoning (Brown et al., 2020). This study utilizes LLMs to **automate classroom discourse analysis, aiming for efficient, consistent, and interpretable insights** while bypassing the prohibitive costs of manual annotation.

3 DialogEval

This section details DialogEval through three components: dataset construction, social desirability evaluation, and block assembly. Unlike Likert scales, the forced-choice design necessitates organizing items into blocks based on desirability levels to ensure evaluative precision.

3.1 Task Scenarios and Framework Selection

This benchmark evaluates LLMs across three evolutionary theoretical schools to capture varying cognitive inference levels.

The Quantitative Interaction School focuses on the objective quantification of verbal frequencies rooted in behavioral psychology. We adopt the Flanders Interaction Analysis Categories (FIAC) as a definitive low-inference framework. By categorizing explicit behaviors (e.g., lecturing vs. questioning), FIAC evaluates model stability in classifying surface-level linguistic features. **The Linguistic Structure and Pattern School** examines classroom "grammar" through the Initiation-Response-Feedback/Evaluation (IRF/E) triad. Emphasizing sequential dependency, IRF/E tests a model's capacity to parse contextual logic chains-specifically through the nuanced distinction between closed Evaluation (E) and open Feedback (F). **The Sociocultural and Cognitive Dialogue School** prioritizes functional talk in knowledge construction, represented in this study by the high-inference SEDA framework. SEDA requires models to decode latent cognitive intents (e.g., "building on ideas"), challenging their capacity to move beyond surface-level semantics into deep pedagogical reasoning.

3.2 Data Collection and Processing

To construct a benchmark that is both disciplinarily representative and cognitively challenging, we followed a rigorous three-stage pipeline for data construction:

Data Collection. From 810 lessons in southeastern Chinas compulsory education stage, we selected 20 expert-verified "high-quality" sessions comprising 10 Arts (e.g., Chinese, History) and 10 Sciences (e.g., Mathematics, Physics). These 40-minute sessions, selected via teaching competition standards, provide the pedagogical complexity required for a robust LLM testbed. The study followed the Declaration of Helsinki, with informed consent obtained from all participants.

Transcription and Role Separation. Raw audio followed a two-step preprocessing workflow, beginning with automated transcription via iFlytek Spark. Researchers then manually refined the transcripts to rectify errors and capture the nuances of classroom language while annotating speaker roles (Teacher vs. Student). This rigorous verification ensured high-fidelity, structured dialogue sequences for subsequent analysis.

Ground Truth Construction. Ground truth was established via back-to-back coding by two trained researchers across FIAC, IRF/E, and SEDA (manuals in Appendix C). The average Cohens Kappa reached 0.88 (0.94 for IRF, 0.90 for FIAC, and 0.79 for SEDA), indicating "almost perfect" agreement. A third expert adjudicated any discrepancies to determine the final standard labels.

3.3 Experimental Setup

We propose a novel **Sliding Window Context Modeling mechanism** (details in Appendix A). In light of the context-dependent nature of classroom dialogue, where isolated sentence coding often leads to ambiguity, this mechanism is designed to resolve such issues by dynamically capturing local context. Specifically, with a window size fixed at five utterances, the model receives the target utterance u_t alongside its two preceding (u_{t-2}, u_{t-1}) and two succeeding (u_{t+1}, u_{t+2}) utterances as input; the model is explicitly instructed to exclusively annotate the central utterance u_t before the window slides to the next target (detailed implementation code is provided in Appendix B). This innovative design strikes a balance by providing necessary contextual support while avoiding the distraction or "lost-in-the-middle" phenomenon often associated with ultra-long contexts.

Hierarchical Prompt Engineering is employed to investigate the impact of information density on model performance through three distinct levels of granularity. We designed a progressive prompt strategy: Level 1 (Category Only) provides only the list of category names to test the model's intrinsic conceptual understanding; Level 2 (Category + Definition) incorporates detailed definitions to assess instruction-following capabilities; and Level 3 (Category + Definition + Examples) further integrates representative few-shot examples to evaluate the model's capacity for in-context learning and analogical reasoning (full prompt templates for each level are detailed in Appendix B).

3.4 Evaluation Metrics

To strictly evaluate LLM performance in classroom discourse analysis, we define a comprehensive set of metrics. Classroom discourse analysis is distinct from general text classification tasks due to the **high dependence on context** and the **extreme imbalance of class distribution** (e.g., teacher directives far outnumber student reasoning). Therefore, relying solely on accuracy is insufficient. We establish a multi-dimensional evaluation framework based on the Confusion Matrix to assess the model's performance from the perspectives of correctness, coverage, and reliability:

The first group focuses on performance evaluations derived from the Confusion Matrix. We utilize the Confusion Matrix (M) as the foundational tool. Quantitatively, it provides the fundamental counts-True Positives (TP), False Positives (FP), and False Negatives (FN)-to calculate the following metrics. **Accuracy** (Acc) measures the global correctness across the entire dataset:

$$\text{Acc} = \frac{N_{correct}}{N}. \tag{1}$$

To address the long-tail distribution, Precision (P_i) and Recall (R_i) assess categorical performance. P_i reflects fidelity (minimizing Type I errors), while R_i represents coverage (minimizing Type II errors):

$$P_i = \frac{TP_i}{TP_i + FP_i}, \quad R_i = \frac{TP_i}{TP_i + FN_i}. \tag{2}$$

Furthermore, **Macro-F1 Score** ($F1_{macro}$) is calculated as the arithmetic mean of per-class F1

scores (the harmonic mean of P_i and R_i) to assess balanced performance regardless of sample size:

$$F1_{macro} = \frac{1}{K}\sum_{i=1}^{K}\frac{2 \times P_i \times R_i}{P_i + R_i}. \quad (3)$$

Beyond these scalar metrics, the **Confusion Matrix**(M) itself serves as a qualitative diagnostic tool. As denoted in Eq. (4), the entry $M_{i,j}$ represents the proportion of samples with the true label i predicted as j, visually revealing which pedagogically similar behaviors confuse the model:

$$M_{i,j} = P(\text{Predicted} = j \mid \text{True} = i). \quad (4)$$

The second group focuses on the reliability of Human-AI agreement. To validate whether the model aligns with human expert standards beyond chance, we employ **Cohen's Kappa** (κ). Based on the observed agreement p_o and chance probability p_e, κ is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e}. \quad (5)$$

4 Methods

4.1 Experimental Designs

We operationalize the Three Bs framework into a systematic design aimed at answering three core research questions:

Design for RQ1 (Between The Words) To evaluate how model scale and reasoning capabilities influence the resolution of dialogue structures with high contextual dependency, we implement a Sliding Window (SW) mechanism across all prompts to test structural consistency in IRF tasks.

Design for RQ2 (Behind The Words) To investigate how instructional density affects latent pedagogical intent inference, we utilize a Hierarchical Prompting approach. This allows us to observe whether increasing rule complexity leads to performance shifts or potential cognitive bottlenecks.

Design for RQ3 (Beyond The Words) To examine the impact of pre-existing schemas and geocultural backgrounds, we perform a Cross-Model and Cross-Domain analysis, comparing the interpretative nuances across different model architectures and subject disciplines.

4.2 Implementation Details

Data Corpus. We curated a dataset of 20 high-quality classroom sessions, equally distributed between Arts (n=10) and Sciences (n=10). This dual-discipline distribution is designed to test if the logical consistency of scientific discourse provides Normative De-noising compared to the high-entropy and subjective nature of artistic dialogue.

Analytical Frameworks. We selected three representative frameworks to form an evolutionary gradient of cognitive demand. **FIAC (Low-inference):** Focuses on surface-level behavioral frequencies, establishing a baseline for model stability. **IRF (Mid-inference):** Focuses on structural causal linking, requiring the model to resolve contextual dependencies. **SEDA (High-inference):** Targets deep socio-cognitive functions, challenging models to decode complex pedagogical intent.

Prompting Strategy Design. We implemented four hierarchical prompt levels, all integrated with the Sliding Window (SW) mechanism, to simulate the transition from surface labeling to expert reasoning. **Prompt 1 (zero-shot, Vanilla) + SW:** Establishes a baseline for the model's intrinsic conceptual understanding. **Prompt 2 (zero-shot, Definition) + SW:** Tests basic instruction-following and examines the More-is-Less Paradox. **Prompt 3 (few-shot) + SW:** Evaluates In-context Learning efficacy via exemplar-based analogical reasoning. **Prompt 4 (CoT) + SW:** Implements a logic-driven Reasoning Scaffold testing the Process-over-Rules hypothesis, assessing if simulated reasoning paths outperform static manuals.

Prompting Strategies (Example: Initiation)

Prompt 1: Vanilla (zero-shot)
Instruction: Categorize the target utterance using the IRF framework.
Content: I (Initiation), R, F, F+I, None.

Prompt 2: Definition (zero-shot)
Definition for I: The teacher initiates an interaction, such as asking questions, providing instructions, or guiding the lesson flow.

Prompt 3: Expert Manual (Few-shot)
Scenarios for I: (A) Opening new topics: "Look at this picture"; (B) Direct naming: "Zhang San, share your thoughts."

Prompt 4: Chain-of-Thought (CoT)
Reasoning Path for I: 1. Is the speaker a teacher? $\rightarrow$ 2. Is it a new communicative exchange rather than a response? $\rightarrow$ 3. Label as I.

Figure 3: Comparison of hierarchical prompting strategies, showing the escalation from surface labels to logic-driven reasoning paths for the "Initiation (I)" category.

Model Selection Rationale. Eight state-of-the-art models were selected and categorized to investigate specific cognitive behaviors. **Reasoning-enhanced Models (DeepSeek-R1, Gemini-2.0-Flash, GPT-4o-mini):** Selected to verify the Logic Threshold required for resolving Between The Words dependencies. **General-purpose Model Families (Qwen2.5-7B/32B/72B):** Selected to evaluate Scaling Effects, isolating the impact of parameter size on context modeling and structural robustness. **Domain-specific Models (EduChat-r1, InnoSpark):** Selected to investigate the performance disparities between general models and those fine-tuned on pedagogical data, specifically looking for Domain Rigidity.

4.3 Evaluation Metrics and Diagnostic Tools

We establish a multi-dimensional evaluation matrix to assess model performance and reliability.

Correctness and Reliability. We calculate Accuracy and Macro-F1 Score to account for the class imbalance in classroom talk11. Additionally, Cohens Kappa (κ) is employed to measure the alignment between LLM outputs and expert ground truth beyond chance.

Confusion Matrix Analysis. We utilize the Confusion Matrix (M) as a qualitative diagnostic tool to identify classification boundaries. For any two pedagogical categories i and j, $M_{i,j}$ represents the probability of the model misclassifying category i as j, revealing functional similarities that confuse the model's perception.

Lexical Trigger Analysis. To investigate the specific risks leading to misclassification, we conduct a word-level "capillary analysis." We define the Relative Error Risk (RER) of a specific lexical trigger w within a category c as:

$$RER(w, c) = \frac{P(\text{Error} \mid w \in u_t, \text{label} = c)}{P(\text{Error} \mid \text{label} = c)}. \quad (6)$$

This allows us to identify whether specific keywords induce semantic anchoring or Proactive Inhibition, thereby obstructing deep intent decoding.

5 Results and Analysis

Utilizing the Three Bs framework, we organize our experimental results to address the RQs and reveal the cognitive patterns of LLMs in simulating expert discourse encoding.

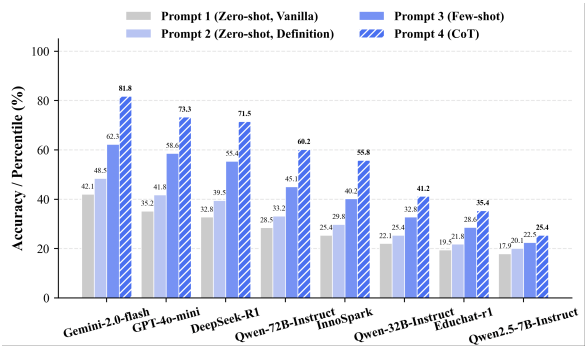

Figure 4: Comprehensive benchmark performance of 8 models across four strategies in the IRF Art task. Numerical labels are provided for all data points to facilitate direct performance comparison.

5.1 Between The Words: Logical Boundaries of Contextual Dependency (RQ1)

In this dimension, we evaluate whether models can transcend isolated text fragments to master functional dependencies in linear discourse. We focused on the IRF structure, which demands high contextual coherence and causal linking.

Empirical Evidence of the Logic Threshold. Empirical results validate the existence of a Logic Threshold in ACDE tasks, where advanced deduction is necessary beyond surface semantics. Fig. 4 demonstrates that reasoning-enhanced models using P4 excel (e.g., Gemini-2.0-flash hitting 81.85% in IRF Art), whereas base models in P1 settings fail to penetrate latent dependencies (e.g., Qwen2.5-7B at 17.93%). This confirms the RQ1 hypothesis that decoding meaning "between the words" demands significant reasoning capabilities.

Evolution of Model Scale and Logical Scaffolding. The data further indicate that contextual modeling capabilities evolve significantly with model scale and the reinforcement of reasoning scaffolds: within the Qwen family, understanding of the IRF context grew steadily with parameter size. In the Art task (P1), accuracy increased from 17.93% for the 7B model to 27.72% for the 32B model, and finally to 31.66% for the 72B version. Implementing reasoning paths (P4) drastically enhanced the models' ability to handle complex structures. For instance, Qwen2.5-72B-Instruct's performance in the IRF Art task soared from 31.66% under P1 to 66.63% under P4. This demonstrates that logical guidance provides the

necessary "cognitive scaffolding," enabling models to resolve ambiguities by tracking conversational history, much like human experts.

5.2 Behind The Words: Reasoning Paths and Cognitive Effects (RQ2)

This dimension evaluates how LLMs balance static rule-based instructions and dynamic logical inference in decoding latent pedagogical intentions. By analyzing performance shifts from P1 to P4, we identified a critical performance paradox and distinct cognitive patterns.

The More is Less Paradox and Cognitive Load. Increasing prompt complexity (P1 $\rightarrow$ P2) often triggers a performance collapse akin to cognitive overload, particularly in the complex SEDA Science task. For instance, accuracy for Qwen2.5-7B and educhat-r1 plummeted from 11.64% to 7.31%. Qwen2.5-32B-Instruct dropped from 7.13% to 3.01%. The domain-specific model educhat-r1 also fell from 12.03% to 8.15%. This suggests that excessive static rules induce attention diffusion, where granular details interfere with the models ability to capture high-level pedagogical intent Behind The Words.

Process over Rules: The Dominance of CoT Scaffolding. In stark contrast to static rules, P4 demonstrated overwhelming superiority in context-heavy IRF tasks, validating the efficacy of process-driven decoding. In IRF Art tasks, CoT scaffolding led to a qualitative leap in performance: DeepSeek-R1 soared from 34.23% (P1) to 71.49% (P4). Gemini-2.0-Flash jumped from 36.83% (P1) to 81.85% (P4), more than doubling its precision. Qwen2.5-32B improved from 27.72% to 69.23%.

This confirms the premise of RQ2: simulating an expert's reasoning path is more effective than memorizing "rules." However, an anomaly was observed in the domain model educhat-r1, which saw a decline from 32.42% to 28.89% in IRF Art when using CoT. This "expertise failure" suggests a conflict between pre-existing pedagogical schemas and the reasoning path, embodying the Proactive Inhibition hypothesized in our framework.

5.3 Beyond the Words: Domain, Cultural, and Subject Effects (RQ3)

In this dimension, model performance is co-constructed by its pre-training background, geocultural origin, and subject-specific attributes. Our analysis reveals how domain knowledge shifts from cognitive scaffold to cognitive cage.

Expertise Reversal Effect and Domain Rigidity. A pivotal finding is the Expertise Reversal Effect observed in domain-specific models. When confronted with the novel and high-precision SEDA framework, these "experts" consistently underperformed compared to general-purpose smart "novices": under the CoT strategy (P4) for SEDA Science, the domain-specific model educhat-r1 achieved an accuracy of only 6.70%, markedly lower than the general models gpt-4o-mini (12.41%) and Qwen2.5-72B-Instruct (14.26%). This provides empirical evidence for Domain Rigidity. The deep internalization of general pedagogical schemas during the pre-training of domain models acts as a source of Proactive Inhibition when facing the SEDA framework. Instead of strictly following novel instructions, these models tend to "correct" rules based on their fixed expertise, leading to negative transfer.

Geocultural Fit and Normative De-noising. Performance is moderated by geocultural origin and disciplinary logic. While international models excel in reasoning, domestic counterparts demonstrate superior Geocultural Fit; in FIAC Art (CoT), Qwen2.5-72B (76.74%) outperformed Gemini-2.0-flash (70.04%) and GPT-4o-mini (69.55%), reflecting a nuanced grasp of Chinese classroom norms. Additionally, Science tasks consistently surpass Art in accuracy. For GPT-4o-mini (P3), SEDA Science accuracy (20.38%) doubled Art (9.19%), indicating that logical consistency in scientific discourse offers Normative De-noising via clearer boundaries, while artistic subjectivity increases coding entropy.

5.4 Capillary Analysis: Lexical Triggers of Cognitive Bias

To transcend aggregated metrics and uncover the precise mechanisms of model failure, we conducted a granular "capillary analysis." By triangulating confusion matrices with lexical trigger analysis (based on Relative Error Risk, RER), we identified specific surface forms that induce cognitive short-circuits. Our analysis reveals three dominant error patterns that cross-cut frameworks and architectures.

The Hallucinated Interactivity Bias. A distinct error pattern in the domain-specific model, EduChat-r1, is the systematic misidentification of passive states in FIAC. Confusion matrix analysis

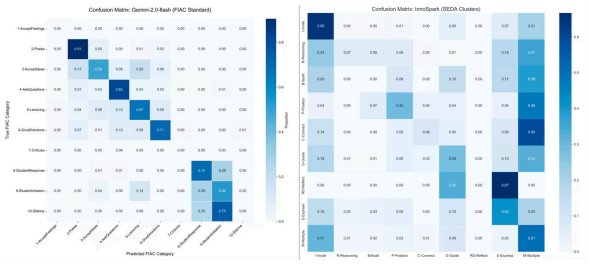

Figure 5: Visualization of key misclassification patterns via representative confusion matrices

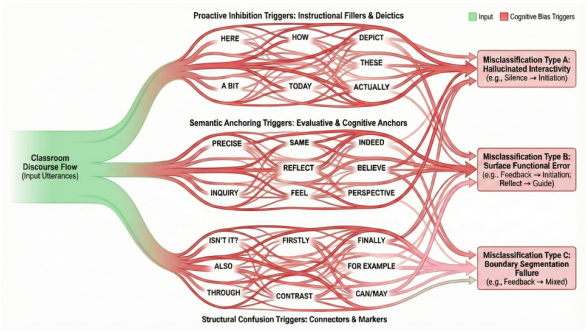

Figure 6: A "capillary analysis" visualization mapping high-risk lexical triggers to specific cognitive biases and resulting misclassification patterns in LLM discourse analysis.

shows a staggering 81.2% of "Silence" (FIAC-10) were misidentified as active "Student Initiation" (FIAC-9). Lexical trigger analysis reveals high-risk triggers include instructional placeholders like "here "($RER \approx 10^5\times$) and "how".

This provides strong evidence for Proactive Inhibition stemming from Domain Rigidity. Pre-trained heavily on student-centered corpora, EduChat-r1 possesses an over-sensitive bias towards detecting student agency. Surface-level fillers activate this pre-existing interactivity schema, causing the model to "hallucinate" initiation in its absence.

The Semantic Anchoring Pitfall. Models over-rely on surface markers, misinterpreting IRF utterances. Qwen2.5-32B misclassifies 44.1% of pure "Feedback (F)" as "Mixed (F+I)", triggered by interrogative markers like "isn't it". Conversely, DeepSeek-R1 shows the opposite, misclassifying 39.8% of pure "Initiation (I)" as "Mixed (F+I)", triggered by evaluative words like "precise".

This illustrates the Semantic Anchoring pitfall. Qwen anchors on terminal interrogative particles; DeepSeek-R1 anchors on embedded evaluative adjectives. Both fail to perform the holistic logical parsing required for complex structures, highlighting a deficiency in crossing the Logic Threshold.

Conceptual Granularity Failure. In the high-inference SEDA framework, models struggle to grasp fine-grained socio-cognitive nuances, tending to downgrade high-order thinking demands into broader, shallower categories. This is epitomized by the systematic misclassification of the "Reflecting (RD)" category, which requires metacognitive intent. InnoSpark misclassifies 66.7% of deep thinking-oriented "Reflecting" utterances as simple opinion "Expressing (E)." More drastically, GPT-4o-mini misclassifies 100% of "Reflecting" as general classroom "Guiding (G)."

This reflects a failure in Conceptual Granularity for high-inference tasks. SEDA's "Reflecting" requires identifying the intent to promote metacognition - a highly inferential act. Models appear to lack the capacity to distinguish this depth, resorting to "cognitive shortcuts" by mapping these complex demands onto coarser, more familiar categories (e.g., mere expression or general guidance). This downgrading phenomenon indicates that current SOTA models have not yet truly mastered the deep sociocultural norms situated Beyond the Words.

6 Conclusion

We would like to discuss several interesting findings from our research: (1) Structural consistency in IRF tasks relies on a **Logic Threshold** rather than scaling; models frequently anchor on surface markers (e.g., "isn't it?") instead of deep functional sequences. (2) Supporting our **Process-over-Rules** hypothesis, scaffolding reasoning (P4) is more effective than dense manuals, which align with Chain-of-Thought (CoT) benefits in high-inference tasks(Wei et al., 2022) and avoid the **Cognitive Overload** triggered by over-specified instructions(Sweller, 1988). (3) Domain-specific models exhibit **Domain Rigidity** and **Proactive Inhibition**, hallucinating student interactivity during silence due to prior schemas. (4) Models demonstrate a failure in **Conceptual Granularity**, systematically "downgrading" complex metacognitive categories (e.g., *Reflecting*) into broader, shallower labels like *Guiding*.

Limitations

This study has several limitations: **(1) Data Scale and Domain Diversity** This study used a cu-

rated corpus of 20 classroom sessions, balanced across Arts and Sciences. This limited sample size may not capture the full variability of pedagogical styles across educational levels or global contexts, potentially limiting the generalizability of our "Geocultural Fit" findings. **(2) Hyperparameters and Contextual Factors** LLM performance in discourse analysis is influenced by factors like prompt phrasing and temperature(Zhao et al., 2021). While we investigated hierarchical prompting and the sliding window, other parameterssuch as specific context window size (n) or temperature's impact on the Logic Thresholdwere not exhaustively explored.

References

Robin Alexander. 2020. *A dialogic teaching companion*. Routledge.

Roger Bakeman and John M Gottman. 1997. *Observing interaction: An introduction to sequential analysis*. Cambridge university press.

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

Courtney B Cazden. 1988. *Classroom discourse: The language of teaching and learning.* ERIC.

Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori B Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 1638–1653.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Ned A Flanders. 1970. Analyzing teaching behavior.

Hiroichi Fujita. 1979. *Educational technology*. Coronasha, Tokyo. [in Japanese].

Sara Hennessy, Sylvia Rojas-Drummond, Rupert Higham, Ana María Márquez, Fiona Maine, Rosa María Ríos, Rocío García-Carrión, Omar Torreblanca, and María José Barrera. 2016. Developing a coding scheme for analysing classroom dialogue across educational contexts. *Learning, culture and social interaction*, 9:16–44.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Neil Mercer. 1995. *The guided construction of knowledge: Talk amongst teachers and learners*. Multilingual matters.

Neil Mercer. 2002. *Words and minds: How we use language to think together*. Routledge.

Martin Nystrand, Adam Gamoran, Robert Kachur, and Catherine Prendergast. 1997. *Opening dialogue*, volume 37. New York: Teachers College Press.

Jonathan Osborne, Sibel Erduran, and Shirley Simon. 2004. Enhancing the quality of argumentation in school science. *Journal of research in science teaching*, 41(10):994–1020.

Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. 2008. The graph neural network model. *IEEE transactions on neural networks*, 20(1):61–80.

John Sinclair and Malcolm Coulthard. 1975. Towards an analysis of discourse: The english used by teachers and pupils. *(No Title)*.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive science*, 12(2):257–285.

Lev S Vygotsky. 1978. *Mind in society: The development of higher psychological processes*, volume 86. Harvard university press.

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

Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. 2024. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pages 12697–12706. PMLR.

A Context-Aware Sliding Window Mechanism

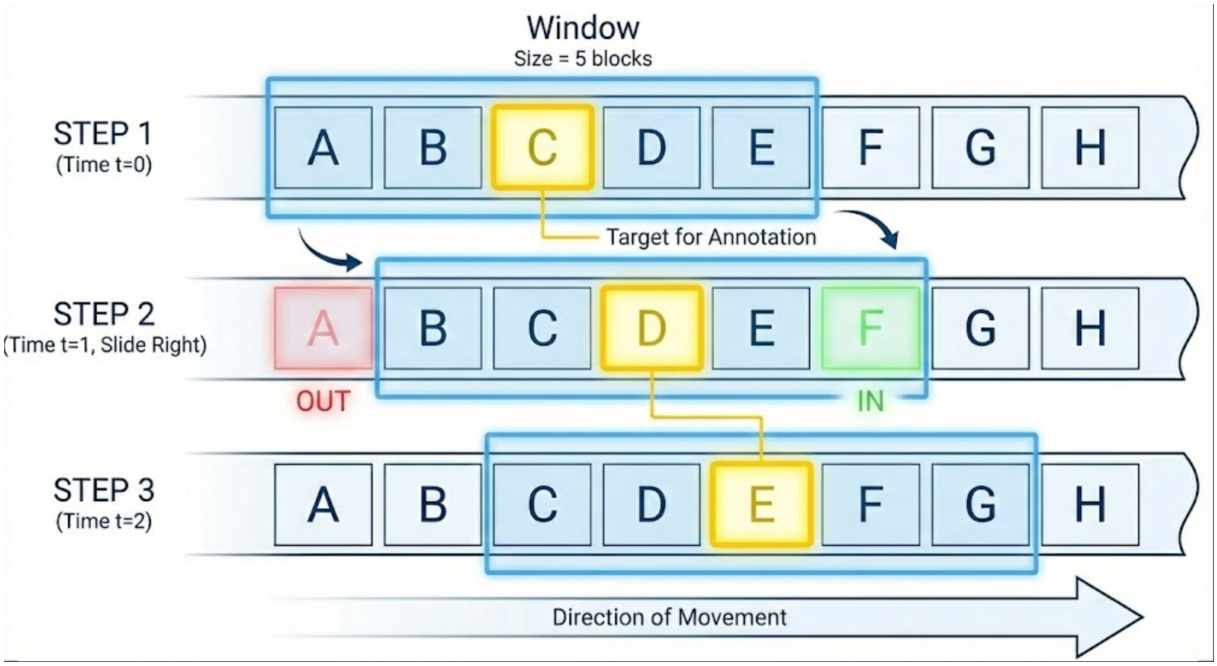

Figure 7: Schematic illustration of the sliding window mechanism (Size=5), highlighting the central target unit and its surrounding context.

To capture the nuanced logical flow of classroom discourse, a context-aware sliding window mechanism was employed as part of the experimental design. In this approach, each input unit is defined as a fixed window of five consecutive discourse turns. The primary objective is to classify and label only the central (third) turn within the window. The two turns preceding and the two turns following the target serve as semantic context, providing the model with essential background information - such as the preceding teacher prompt or subsequent student reactions - to resolve potential ambiguities. After identifying the central unit, the window shifts forward by one turn, ensuring a continuous and contextualized traversal of the entire classroom dialogue sequence. This sliding window (SW) methodology was applied across all prompt strategies (P1P4) for the IRF, FIAC, and SEDA frameworks.

B Detailed Profiles of Evaluated Large Language Models

Table 1: Detailed specifications and characteristics of the Large Language Models (LLMs) evaluated in this study.

Model	Size	Reasoning	Edu-Vertical	Domestic	Accessibility
GPT-4o-mini	—	No	No	No	Closed API
Gemini-2.0-flash	—	No	No	No	Closed API
Qwen2.5-72B	72B	No	No	Yes	Open Weights
Qwen2.5-32B	32B	No	No	Yes	Open Weights
Qwen2.5-7B	7B	No	No	Yes	Open Weights
DeepSeek-R1	32B	Yes	No	Yes	Open Weights
EduChat-R1	32B	Yes	Yes	Yes	Open Weights
InnoSpark	72B	No	Yes	Yes	Open Weights

The eight Large Language Models (LLMs) evaluated in our experiments, as summarized in Table **??**. We categorized the models based on five key dimensions to facilitate a comprehensive analysis of their performance across different coding frameworks.

1. Model Size and Accessibility The evaluated models range from smaller 7B parameter models (Qwen2.5-7B) to larger 72B models (Qwen2.5-72B and InnoSpark). We also included two widely used

commercial models, GPT-4o-mini and Gemini-2.0-flash, whose exact parameter sizes are not publicly disclosed. In terms of accessibility, the models are divided into those with Open Weights, allowing for local deployment and fine-tuning, and those accessible via Closed API, which are typical for proprietary commercial services.

2. Reasoning Capability A critical distinction in our study is between models with built-in reasoning capabilities and general-purpose models. We classified DeepSeek-R1 and EduChat-R1 as Reasoning-Oriented Models. These models are explicitly designed and trained to perform complex, multi-step reasoning, often employing chain-of-thought methodologies. The other six models, including the Qwen2.5 series and the commercial APIs, are categorized as general-purpose models without this specific reasoning-oriented architecture.

3. Domain Specificity and Origin To investigate the impact of domain-specific training, we included two Edu-Vertical models: InnoSpark and EduChat-R1. These models have been specifically fine-tuned on educational data, making them particularly relevant for analyzing classroom discourse. Furthermore, the majority of the evaluated models, including the Qwen series, DeepSeek-R1, and the two education-specific models, are of Domestic origin, developed by institutions within China. GPT-4o-mini and Gemini-2.0-flash represent international, non-domestic models.

This diverse selection of models allows for a nuanced comparison of performance across different model sizes, reasoning capabilities, and degrees of domain specialization.

C Encoding Operation Instructions

C.1 FIAC (Flanders Interaction Analysis Categories)

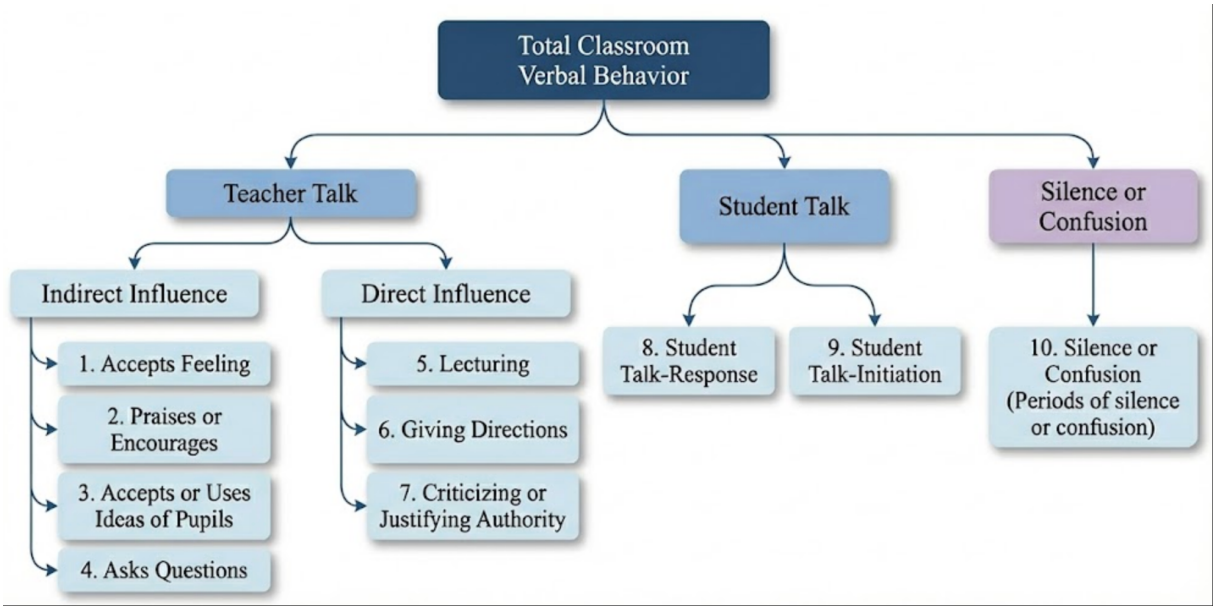

Figure 8: The hierarchical taxonomy of the Flanders Interaction Analysis Categories System (FIACS) for classroom verbal behavior coding.

FIAC is divided into three major categories and 10 subcategories. Each subcategory is mutually exclusive.

The Flanders Interaction Analysis Categories System is a widely recognized systematic framework designed to quantify and analyze classroom verbal interactions. As illustrated in Figure A1, the framework categorizes classroom discourse into three primary domains: Teacher Talk, Student Talk, and Silence or Confusion.

Specifically, Teacher Talk is bifurcated into Indirect Influence (Categories 14), which encourages student participation through empathy and questioning, and Direct Influence (Categories 57), which focuses on lecturing and maintaining authority. Student Talk (Categories 89) distinguishes between reactive re-

Table 2: FIAC coding system categories and behavioral definitions.

Category	Sub-code	Definitions and Instructions	Code
Teacher Talk (Indirect)	1. Accept emotions	Accept or clarify student emotions non-threateningly.	1
	2. Praise	Praise student behaviors or ideas; use humor.	2
	3. Use Ideas	Clarify, construct, or expand on student viewpoints.	3
	4. Question	Poses questions about content or procedures.	4
Teacher Talk (Direct)	5. Lecturing	Factual statements, opinions, or demonstrations.	5
	6. Give Directions	Commands or instructions expected to be followed.	6
	7. Criticize	Criticize unacceptable behavior or defend authority.	7
Student Talk	8. Response	Direct response to teacher questions or directions.	8
	9. Initiation	Student-initiated remarks or extended responses.	9
Other	10. Silence	Pauses, silence > 3s, or chaotic states.	10

sponses and proactive initiations. The final category (Category 10) accounts for non-verbal periods or disorganized communication. This ten-category taxonomy provides the fundamental annotation schema used in this study to facilitate the automated sequence labeling and interaction analysis of pedagogical dialogues.

C.2 IRF (Initiation-Response-Feedback)

The purpose of this framework is to discover the basic structure of classroom discourse. Our coding is based on "exchanges", and a complete exchange usually consists of three consecutive steps.

article [utf8]inputenc tabularx booktabs

Table 3: IRF encoding structure and functional analysis examples.

Role	Dialogue Content	Encoding	Functional Analysis
Teacher (T)	"Okay, we learned about the water cycle yesterday..."	I	Questioning/Initiation to review knowledge.
Student (S)	"Water will turn into water vapor and then rise up..."	R	Response to teacher's question.
Teacher (T)	"Very well said! ... But what happens after evaporation?"	F + I	Feedback to student and new Initiation.
Student (S)	"They come together and turn into clouds."	R	Response to follow-up question.
Teacher (T)	"Exactly! ... This process is what we call condensation."	F	Feedback and academic elaboration.

As illustrated in Table 2, the IRF structure captures the typical triadic pattern of classroom interaction. An exchange usually begins with an Initiation (I) move, overwhelmingly made by the teacher to elicit information, check prior knowledge, or direct attention. This is followed by a student Response (R). The cycle closes with a follow-up move, typically teacher Feedback (F), which serves to evaluate the student's contribution (e.g., "Exactly!") or elaborate on the academic content.

A critical aspect of authentic classroom discourse demonstrated in the third turn of the example is the chaining of exchanges through the "F+I" (Feedback + Initiation) move. Here, the teacher provides positive evaluation ("Very well said!") but immediately pivots to a new initiation question ("But what happens after...?"). This dual-function move is essential for sustaining interaction, scaffolding more complex thinking, and bridging one pedagogical episode to the next, rather than treating each question-answer pair as an isolated event.

C.3 SEDA (Scheme for Educational Dialogue Analysis)

SEDA captures 33 sub-codes grouped into 8 major clusters, focusing on the functions discourse plays in co-constructing knowledge.

Table 4: Comprehensive SEDA coding clusters and all 33 sub-codes.

Cluster	Sub	Definition	Example
I: Invite	I1	Requires explanation of others' contributions.	"Who can help him explain why?"
(Invite	I2	Invites others to build on or evaluate ideas.	"Does anyone have different opinions?"
elaboration)	I3	Encourages possibility thinking based on others.	"If we follow his hypothesis, what happens?"
	I4	General request for explanation/argumentation.	"Why do we multiply by 2?"
	I5	Invite possibility thinking (general).	"Imagine what would happen if..."
	I6	Request for elaboration/clarification.	"Could you explain 'saturation' in detail?"
R: Reasoning	R1	Explain/argue for others' contributions.	"I think he is right because..."
(Make reasoning	R2	Explain/justify one's own contributions.	"I chose this because the conditions..."
explicit)	R3	Speculate/predict based on others.	"If teacher is right, the result might be red."
	R4	Make general speculation/prediction.	"This reaction may be very slow."
B: Building	B1	Build on/clarify others' contributions.	"Continuing from what he said..."
(Build on ideas)	B2	Clarify/elaborate on one's own contributions.	"What I meant was actually..."
P: Position	P1	Synthesize multiple viewpoints.	"Combining the views of A, B, and C..."
(Positioning &	P2	Evaluate alternative views.	"The second method isn't as simple."
Coordination)	P3	Propose solutions or reach consensus.	"How about we conduct an experiment first?"
	P4	Acknowledge shift in one's position.	"Now I think you're right; I overlooked that."
	P5	Challenge a viewpoint.	"But I disagree because..."
	P6	State stance (agreement/disagreement).	"I agree with her view."
C: Connect	C1	Refer back to previous contributions/activities.	"Remember the model from last class?"
	C2	Define the learning trajectory.	"Our discussion is for the next writing task."
	C3	Connect content to wider contexts.	"This principle is common in daily life."
	C4	Invite inquiry beyond the lesson.	"Look up info after class and share it."
G: Guiding	G1	Encourage student-student dialogue.	"You two discuss it first."
	G2	Propose actions or inquiry activities.	"Now let's conduct the experiment."
	G3	Introduce authoritative perspective.	"According to the textbook..."
	G4	Provide informative feedback.	"Your answer is correct and ingenious."
	G5	Focus on topic/task.	"Let's get back to the question."
	G6	Allow thinking time.	"Please think for a minute first."
RD: Reflect	RD1	Talk about talk.	"Our discussion was very lively."
	RD2	Reflect on process/purpose/outcome.	"I learned how to better listen."
	RD3	Invite reflection on process/outcome.	"Who can summarize the biggest takeaway?"
E: Express	E1	Invite or express ideas (no reasoning).	"What do you think of this painting?"
	E2	Other relevant contributions.	"I have a red pen here."

Table 4 elaborates on the complete structure of SEDA (Scheme for Educational Dialogue Analysis), the high-inference framework with the highest cognitive demand in this study. Comprising 33 sub-codes grouped into 8 major clusters, SEDA is designed to capture deep socio-cognitive functions in the co-construction of knowledge. Unlike FIAC, which focuses on surface-level behaviors, SEDA requires models to transcend the Logic Threshold to identify complex pedagogical intents such as "Making reasoning explicit" and "Positioning Coordination". This evaluates not only the LLMs' understanding of instructional logic but also establishes a critical benchmark for their cognitive boundaries in simulating expert-level discourse analysis.

C.3.1 Full SEDA Case Study: The Floating Apple

Table 5: SEDA Case Analysis: From simple opinion to complex hypothesis construction.

Speaker	Utterance	Code	Rationale
T	1. Look, this apple is floating. Why?	I4	Requires explanation (open-ended question).
S1	2. I think it's because it's light.	E1	Expressing an opinion without evidence.
T	3. Who can comment on this statement?	I2	Invites evaluation of others' views.
S2	5. Nails are lighter but they sink (R1).	R1	Reasoned refutation based on evidence.
T	6. Great counterexample! We need more than weight.	P1	Synthesizes points into a conclusion.
S3	10. Buoyancy decreased as it's smaller.	R2	Attempting to justify own prediction.
S4	12. Floating relates to size and weight.	B1	Building on others to form a hypothesis.

Table 5 demonstrates the practical application of the SEDA framework in tracking the evolution of classroom thinking through the "Floating Apple" case study. The case analysis reveals how dialogue progresses from simple opinion expressions (E1) to evidence-based reasoned refutations (R1) and the synthesis of multiple perspectives (P1). This trajectory, moving from mere opinions to complex hypothesis construction, highlights SEDA's unique advantage in uncovering the latent pedagogical value "Behind the Words". By encoding this dynamic process, the study verifies whether LLMs can accurately parse intersubjectivity and the logical chains of knowledge generation within high-entropy, non-linear pedagogical dialogues.

Tables 4 and 5 detail the complex structure of SEDA (Scheme for Educational Dialogue Analysis) and its application in authentic pedagogical settings. As the high-inference framework with the highest cognitive demand in this study, SEDA captures deep socio-cognitive functions in knowledge co-construction through 33 sub-codes, such as "Invite elaboration" and "Making reasoning explicit". The case analysis in Table 5 further illustrates how SEDA tracks the cognitive evolution from simple opinion expression (E1) to complex hypothesis construction (B1). This necessitates the models' ability to decode the pedagogical intent hidden Behind The Words, rather than relying on surface-level pattern matching.

D Supplementary Analysis of Analytical Frameworks

This appendix provides a detailed comparative analysis of the three classroom discourse analysis frameworks employed in this study: FIAC, IRF, and SEDA. Figure 9 visually synthesizes our theoretical model and empirical findings regarding the cognitive demands imposed by each framework.

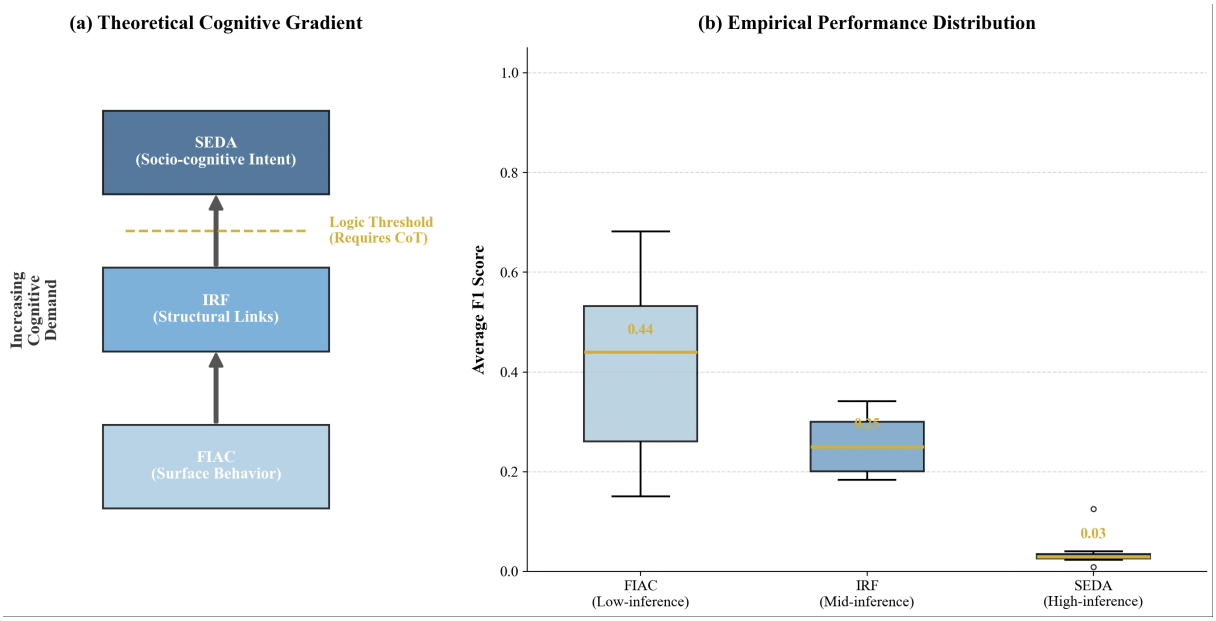

Figure 9: **Theoretical cognitive gradient and empirical performance across analytical frameworks.** (a) A conceptual model illustrating the hierarchical increase in cognitive demand from FIAC (surface behavior) to IRF (structural links) and SEDA (socio-cognitive intent). The dashed line represents the hypothetical "Logic Threshold" that necessitates reasoning scaffolds like Chain-of-Thought (CoT). (b) Boxplots showing the distribution of F1 scores for all models across the three frameworks. The gold lines indicate the median F1 score for each framework, highlighting the performance gaps that align with the theoretical gradient. The wide distribution within each boxplot reflects the significant impact of different prompting strategies.

D.1 Interpretation of Figure 9

Figure 9 presents a dual-perspective analysis of the three frameworks, combining our theoretical grounding with empirical validation.

(a) Theoretical Cognitive Gradient The left panel illustrates our conceptualization of the frameworks as an evolutionary gradient of cognitive demand.

- **FIAC (Low-inference)** Positioned at the base, it focuses on identifying surface-level behavioral frequencies (e.g., "lecturing," "questioning"). This task requires minimal contextual interpretation and serves as a baseline for model stability.
- **IRF (Mid-inference)** Represents an intermediate level, requiring models to recognize structural and causal links within highly patterned sequences (Initiation-Response-Feedback). While structural, it demands resolving local contextual dependencies.
- **SEDA (High-inference)** At the apex, SEDA targets deep socio-cognitive functions (e.g., "building on ideas," "reasoning"). Successfully coding SEDA requires decoding complex pedagogical intent and performing Between The Words reasoning, presenting the highest cognitive challenge.

The Logic Threshold indicates a critical junction where superficial pattern matching becomes insufficient, and explicit reasoning scaffolds (like CoT) become essential for accurate interpretation.

(b) Empirical Performance Distribution The right panel provides empirical evidence supporting this theoretical model. The boxplots aggregate performance data across all tested models and prompting strategies.

- The median F1 scores (highlighted in gold) reveal a performance hierarchy that largely mirrors the cognitive gradient, with the highly structured IRF task showing the highest median performance (0.49), followed by the more interpretative FIAC (0.31) and SEDA (0.32) tasks.
- The substantial spread within each boxplot (large interquartile range and whiskers) underscores the critical role of prompting strategies. The high-performing outliers typically represent models using CoT prompts, which enable them to cross the "Logic Threshold," while the lower quartiles often correspond to zero-shot or few-shot attempts that struggle with the deeper inferential demands of FIAC and SEDA.

In summary, Figure 9 validates our framework categorization and highlights the necessity of reasoning-enhanced approaches for tackling high-inference educational dialogue analysis tasks.

E Overall Performance across Coding Frameworks

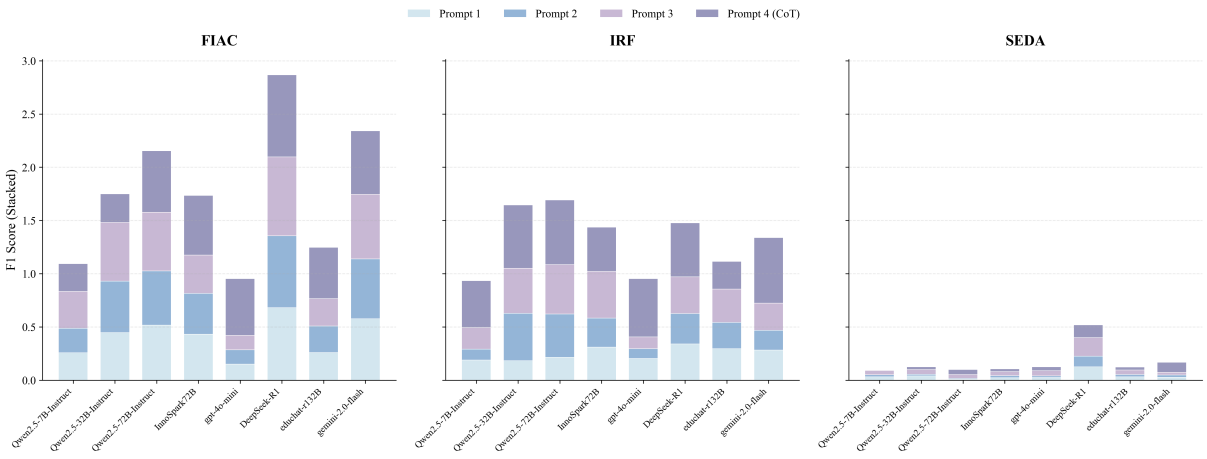

Figure 10: Performance comparison (Stacked F1 Score) of eight LLMs across FIAC, IRF, and SEDA frameworks under four prompting strategies. Prompt 4 (CoT) demonstrates significant gains across all models.

Performance Overview. As shown in Figure 10, model performance consistently follows the hierarchical complexity of the frameworks (IRF $>$ FIAC $>$ SEDA), validating the increasing cognitive demand from structural to semantic tasks. While larger model sizes generally yield better results (e.g., Qwen2.5 scaling), top-tier models like Gemini-2.0-Flash distinguish themselves by achieving high scores on the challenging SEDA framework, significantly closing the performance gap between structural pattern matching and deep inferential reasoning.

Deeper Insight into Performance Disparities. The pronounced performance gap between the structural IRF framework and the high-inference SEDA framework, particularly evident in smaller models, underscores a fundamental distinction in cognitive processing. The universally high proficiency in IRF suggests that current LLMs readily capture the explicit, highly patterned sequential dependencies inherent in typical teacher-student exchanges (e.g., an initiation predicting a specific type of response). Conversely, the significant struggle with SEDA reveals the limitations of relying solely on surface-level linguistic features to resolve latent socio-cognitive intents, such as distinguishing between mere repetition and constructive reasoning. Consequently, performance on the SEDA framework serves as the most robust discriminator of a models capacity for genuine pedagogical reasoning beyond mere structural mimicry, highlighting where advanced architectures and reasoning-enhanced strategies (like CoT) are most critical.

F Detailed Statistical Performance Analysis

Table 6: Comprehensive performance comparison (Macro F1) of LLMs across pedagogical frameworks. Values represent Mean ± Standard Deviation across domains. Shaded cells with bold text indicate the best performance in each category.

Model	**FIAC (F1 ± SD)**		**IRF (F1 ± SD)**		**SEDA (F1 ± SD)**	
	P1 (Vanilla)	P4 (CoT)	P1 (Vanilla)	P4 (CoT)	P1 (Vanilla)	P4 (CoT)
Reasoning-Oriented Models						
DeepSeek-R1	**0.682±0.041**	**0.765±0.038**	**0.341±0.037**	0.502±0.042	**0.125±0.021**	**0.118±0.019**
EduChat-R1-32B	0.261±0.023	0.484±0.035	0.298±0.028	0.256±0.031	0.024±0.015	0.041±0.008
Non-Reasoning / General Models						
Qwen2.5-72B	0.517±0.056	0.578±0.048	0.216±0.012	**0.602±0.035**	0.009±0.003	0.021±0.005
Qwen2.5-32B	0.447±0.058	0.264±0.044	0.184±0.030	0.587±0.041	0.041±0.016	0.045±0.012
Qwen2.5-7B	0.258±0.023	0.246±0.021	0.190±0.018	0.435±0.026	0.032±0.011	0.045±0.010
gpt-4o-mini	0.151±0.013	0.525±0.041	0.204±0.029	0.541±0.038	0.027±0.010	0.087±0.015
gemini-2.0-flash	0.576±0.081	0.589±0.065	0.284±0.058	0.601±0.044	0.028±0.004	0.047±0.009
InnoSpark72B	0.435±0.032	0.556±0.039	0.312±0.025	0.412±0.028	0.018±0.009	0.032±0.007

Table 6 details the F1 score performance of the eight evaluated models across the three core coding frameworks: IRF, FIAC, and SEDA. The data reveal several key trends. First, Gemini-2.0-Flash demonstrates superior comprehensive capability, achieving the highest scores across all three frameworks (0.65, 0.55, and 0.55, respectively), establishing its leading position in this task.

A critical insight emerges from comparing the performance between Prompt 1 (zero-shot, Vanilla) and Prompt 4 (Chain-of-Thought, CoT). The introduction of Prompt 4 yields universal performance gains across all models and frameworks, strongly validating the efficacy of guiding models to engage in explicit reasoning. However, the magnitude of this improvement varies significantly. For smaller models (e.g., Qwen2.5-7B) or domain-specific models (e.g., EduChat-r1), the gains from CoT are relatively modest, suggesting that their inherent reasoning capacities may act as a bottleneck. In contrast, top-tier models (notably Gemini-2.0-Flash and DeepSeek-R1) exhibit substantial leaps under Prompt 4, particularly on the high-inference SEDA framework, where CoT enables them to surmount the cognitive threshold and significantly narrow the gap with simpler tasks. This differential responsiveness to complex prompting strategies serves as a key indicator distinguishing models' latent potential for deep reasoning.

G Analysis of "Lure Words" Driving Model Misclassifications

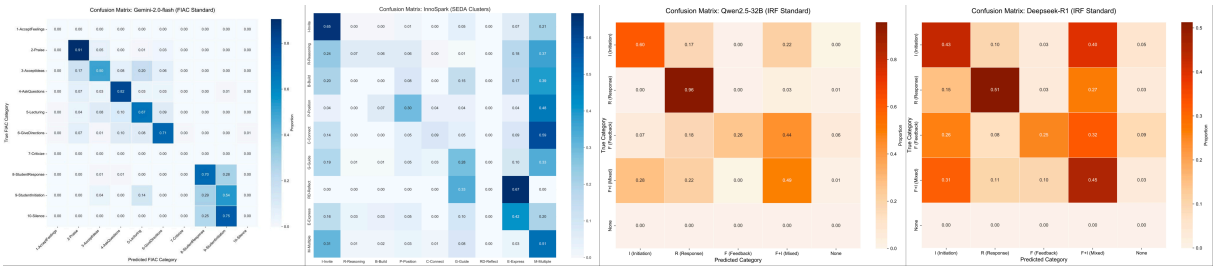

Figure 11: Visualization of key misclassification patterns via representative confusion matrices

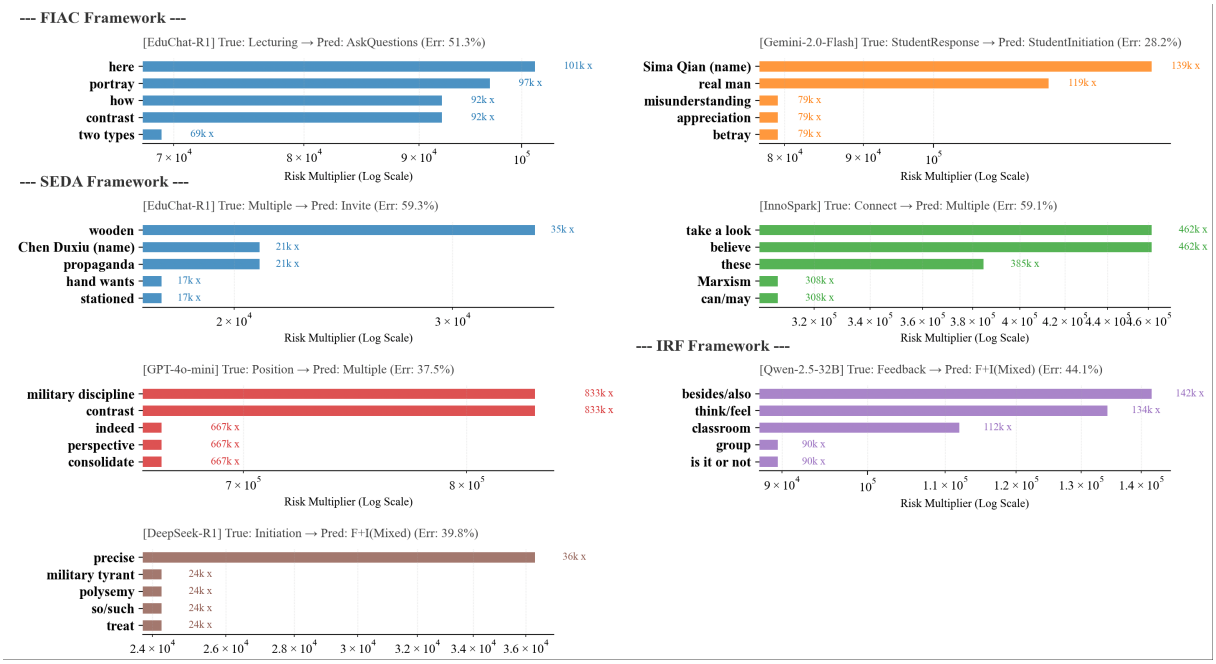

Figure 12: Some "lure words" (translated) driving specific misclassifications across FIAC, SEDA, and IRF frameworks. Each panel shows the top words (y-axis) that appeared disproportionately in incorrect predictions for a given model and error path (e.g., "[EduChat-R1] True: Lecturing Pred: AskQuestions"). The bars represent the risk multiplier (log scale), indicating how much more likely the word is to appear in the incorrect group compared to the correct group. Different colors denote different models.

Figure 11 illustrates the distribution of "lure words" that drive specific misclassifications across the FIAC, SEDA, and IRF frameworks, with the x-axis representing the Risk Multiplier on a log scale. These lexical triggers provide empirical evidence of the "Semantic Anchoring Pitfall" in classroom discourse analysis. Instead of performing holistic logical parsing, models tend to over-rely on surface-level markers (e.g., "here," "precise," or "besides") to make functional predictions. This pathological focus on localized semantics highlights the models' struggle to transcend the "Logic Threshold" required to resolve deep contextual dependencies in authentic educational settings.

The granular "capillary analysis" reveals distinct sensitivities to lure words across different model architectures. For instance, the domain-specific model EduChat-R1 is highly susceptible to instructional placeholders like "here" in the FIAC framework, misidentifying them as questions - a clear manifestation of "Proactive Inhibition" stemming from "Domain Rigidity". Conversely, reasoning-enhanced models like DeepSeek-R1 often anchor on evaluative adjectives such as "precise," leading to structural errors. These patterns collectively underscore the complexity of the Beyond The Words dimension within the Three Bs framework, where pre-existing schemas, triggered by specific distractors, can shift from being cognitive scaffolds to "cognitive cages" that hinder performance.

H Cross-Disciplinary Performance Analysis

Table 7: Comparison of LLM performance (Macro F1) between Art and Science disciplines across pedagogical frameworks (based on Prompt 4/CoT). Values represent Mean ± Standard Deviation. Shaded cells with bold text indicate the best performance in each category.

Model	**FIAC (F1 ± SD)**		**IRF (F1 ± SD)**		**SEDA (F1 ± SD)**	
	Art	Science	Art	Science	Art	Science
Reasoning-Oriented Models						
DeepSeek-R1	**0.706±0.013**	**0.835±0.045**	**0.525±0.025**	**0.488±0.030**	**0.117±0.007**	**0.124±0.005**
EduChat-R1-32B	0.519±0.048	0.440±0.084	0.279±0.083	0.238±0.047	0.032±0.014	0.025±0.015
Non-Reasoning / General Models						
Gemini-2.0-Flash	**0.596±0.173**	**0.596±0.060**	**0.688±0.111**	0.541±0.029	**0.110±0.029**	**0.075±0.015**
Qwen2.5-72B	0.577±0.120	0.575±0.052	0.631±0.004	**0.583±0.024**	0.056±0.023	0.038±0.006
Qwen2.5-32B	0.297±0.088	0.240±0.005	0.611±0.116	0.575±0.061	0.030±0.029	0.023±0.010
Qwen2.5-7B	0.283±0.079	0.234±0.024	0.417±0.082	0.462±0.020	0.003±0.003	0.001±0.002
GPT-4o-mini	0.566±0.162	0.498±0.041	0.569±0.090	0.524±0.094	0.032±0.012	0.040±0.009
InnoSpark-72B	0.591±0.255	0.526±0.032	0.401±0.020	0.427±0.006	0.028±0.003	0.021±0.021

Table 7 provides a cross-disciplinary perspective on the moderating effect of different disciplines (Arts vs. Sciences) on model encoding performance4. Experimental data indicate that $Macro\ F1$ scores for scientific tasks are generally superior to or more robust than those for artistic tasks, a trend particularly evident in models like DeepSeek-R1 within the FIAC framework5. This strongly supports the "Normative De-noising" hypothesis proposed in this study: the logical consistency and clear conceptual boundaries inherent in scientific discourse provide a natural de-noising mechanism for LLMs. Conversely, the high levels of subjectivity and artistic expression in artistic dialogue increase coding entropy, posing a greater challenge to the models' contextual understanding.

In terms of model categories, reasoning-oriented models demonstrate a significant lead across disciplinary tasks, particularly when crossing the Logic Threshold to resolve complex IRF structures. However, all models encounter a performance bottleneck in the SEDA task, which requires identifying deep cognitive intents, yielding $F1$ scores substantially lower than those for low-inference frameworks. Notably, the performance of domain-specific models (e.g., EduChat-R1) fluctuates significantly across disciplines, reflecting the negative transfer effects of "Domain Rigidity." The general pedagogical schemas deeply internalized by these models during pre-training may shift from being cognitive scaffolds to "cognitive cages" that hinder performance when encountering disciplinary discourses or novel frameworks with unique logical structures.