MECH: A Cost-Effective Multi-Task Cascade Framework for Classroom Opinion Evolution Recognition

Yancui Li[1,2,†], **Xiaoyu Zhou**[1,2,†], **Guoyi Miao**[1,2,*], **Fang Kong**[3]
[1]School of Computer and Information Engineering, Henan Normal University, Xinxiang, Henan 453007, China
[2]Key Laboratory of Artificial Intelligence and Personalized Learning in Education of Henan Province, China
[3]School of Computer Science and Technology, Soochow University, China
miaoguoyi@htu.edu.cn

Abstract

Classroom discourse analysis is critical for tracing cognitive restructuring, yet existing research predominantly focuses on Dialogue Acts (DA), overlooking the deeper dimension of Opinion Evolution (OE). In this paper, we formally define the task of Classroom Opinion Evolution Recognition and introduce the Classroom Opinion Evolution Dataset (COED). Addressing the "Accuracy-Cost-Data" trilemma in real-world educational scenarios and the "overconfidence" failure mode of traditional confidence-based cascading systems on long-tail samples, we propose the Multi-task Enhanced Cascade Hybrid (MECH) framework. Grounded in the CODA (Continuous Opinions and Discrete Actions) theory, MECH conceptually translates the "Action-Opinion" dualism into a risk-aware routing mechanism. Instead of relying solely on prediction confidence, this mechanism utilizes high-risk argumentative DA signals derived from multi-task learning to construct a "semantic safety net" effectively routing implicit or ambiguous samples to a Large Language Model for reasoning. Experimental results demonstrate that MECH achieves a state-of-the-art accuracy of 78.55% while reducing API costs by 44.4%. Furthermore, the framework exhibits robustness in few-shot scenarios (using only 20% of data), offering a cost-effective and interpretable solution for large-scale educational dialogue analysis. Our code and data are publicly available at https://github.com/ywh24284-code/MECH.

1 Introduction

Classroom discourse serves not only as a medium for knowledge transmission but also as a critical venue for the restructuring of students' cognitive structures (Resnick et al., 2015). Tracing the micro-level trajectories of learner conceptions—characterized by introduction, reinforcement, conflict, and reorganization—is essential for implementing personalized instructional interventions and precise cognitive diagnosis (Rosé and Ferschke, 2016).

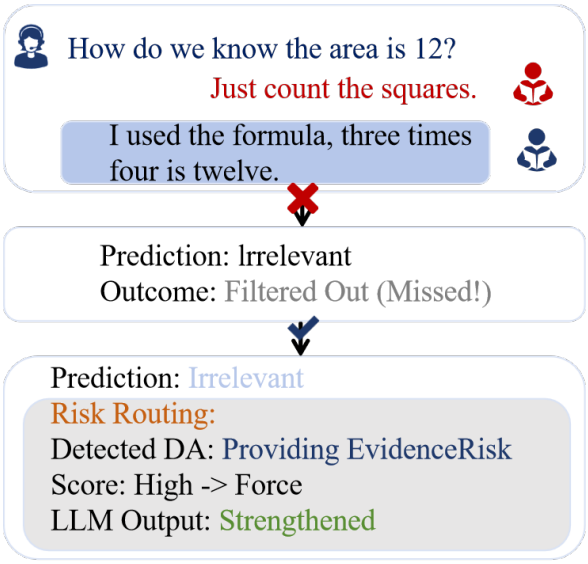

Figure 1: An illustration of the Classroom Opinion Evolution Recognition task and the "Ignorant Confidence" challenge.

However, existing datasets primarily focus on dialogue act annotation (Suresh et al., 2022), overlooking the deeper dimensions of opinion evolution. To address this, grounded in Conceptual Change Theory (Posner et al., 1982) and Argumentation Theory (Toulmin, 1958), we formally propose the task of **Classroom Opinion Evolution Recognition** with a taxonomy comprising six categories. Furthermore, we introduce the **Classroom Opinion Evolution Dataset (COED)**, a corpus containing 14,672 annotated utterances derived from the TalkMoves dataset (Suresh et al., 2022).

Implementing automated analysis in real-world educational scenarios faces an **"Accuracy-Cost-Data" trilemma**. Classroom dialogue is replete with administrative noise, while samples reflecting critical cognitive evolution are sparse, implicit,

[†]Equal contribution. [*]Corresponding author.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 22460–22473

and context-dependent. While Large Language Models (LLMs) possess the reasoning capabilities required for such tasks, deploying them on massive volumes of noisy data is cost-prohibitive (Zhao et al., 2023). Traditional cascade systems (e.g., FrugalGPT (e.g., FrugalGPT, Chen et al., 2024) attempt to mitigate this via confidence-based filtering. However, these methods suffer from the problem of **"Overconfidence"** (Guo et al., 2017) in this domain: lightweight models frequently assign high confidence to erroneous predictions on semantically complex long-tail samples (as shown in Figure 1). Relying solely on confidence scores causes high-value samples to be erroneously intercepted, leading to a sharp increase in false negative rates.

To tackle these challenges, we draw inspiration from the CODA (Continuous Opinions and Discrete Actions) model (Martins, 2008). CODA posits that an individual's internal opinion is continuous and implicit, whereas their external expressions are constrained to discrete actions. Applying this to classroom discourse, We hypothesize that explicit Dialogue Acts (DA) serve as crucial observable anchors for tracing the inherently implicit trajectories of student Opinion Evolution (OE). Based on this dualism, we propose the Multi-task Enhanced Cascade Hybrid (MECH) framework. MECH employs Multi-Task Learning to operationalize the CODA theory into a Semantically-Aware Risk Routing Mechanism. Even if the lightweight discriminative model predicts an utterance as irrelevant in the opinion dimension, the detection of high-risk argumentative DA signals (e.g., *Making a Claim*) triggers a mandatory routing to the LLM expert. This constructs a "semantic safety net" that ensures the recall of implicit but critical evolution instances.

In summary, our main contributions are as follows:

- **New Task and Dataset:** We formally define the Classroom Opinion Evolution Recognition task and release COED, the first educational dialogue corpus featuring dual annotations for both Dialogue Acts and Opinion Evolution.
- **Methodological Innovation:** We propose the MECH framework, which conceptually translates the "Action-Opinion" dualism of the CODA theory into a deep learning "DA-OE" multi-task routing mechanism. This semantic-aware router effectively mitigates the failure of confidence-based cascades on long-tail samples, and is further synergized with definition-augmented prompting strategies to ensure highly reliable LLM reasoning in complex educational contexts.
- **Experimental Validation:** Experiments demonstrate that MECH reduces API costs by 44.4% while surpassing the discriminative baseline by 3.11% in F1-score. Furthermore, it exhibits strong robustness in cold-start scenarios (e.g., with 20% data) by adaptively leveraging LLM reasoning to compensate for sparse supervision.

2 Related Work

2.1 Classroom Dialogue Analysis

Classroom dialogue analysis has long been a focal point in the fields of Educational Data Mining (EDM) and Computational Linguistics (CL). In terms of computational modeling, prevailing research concentrates on Dialogue Act Recognition (DAR). For instance, Suresh et al. (2022) constructed the renowned TalkMoves dataset, defining teacher discourse strategies in K-12 mathematics classrooms (e.g., *Revoicing*, *Pressing*) and establishing RoBERTa-based classification baselines. However, these works primarily target the communicative intent of utterances, failing to uncover the underlying "cognitive evolution." While some studies have attempted to analyze classroom argumentation structures (Lugini and Litman, 2020), they predominantly focus on static argument extraction rather than dynamic cognitive tracking.

2.2 Efficient Reasoning with Large Language Models

Although Large Language Models (LLMs) excel in complex reasoning tasks, their prohibitive inference costs constrain their application in real-time educational scenarios. To address this challenge, model cascading has been extensively investigated. Classic cascading approaches, such as CascadeBERT (Li et al., 2021) and FrugalGPT (Chen et al., 2024), employ an "easy-to-hard" strategy: lightweight models are utilized to process simple samples first, invoking the larger model only when the confidence score falls below a predefined threshold. However, such confidence-based routing strategies rely heavily on the calibration quality of the smaller model (Guo et al., 2017). This reliance is particularly problematic when handling long-tail or Out-of-Distribution (OOD) samples, where small models frequently exhibit "overconfidence."

2.3 Multi-Task Learning in Discourse Analysis

Multi-Task Learning (MTL) enhances model generalization by sharing representation layers across related tasks (Caruana, 1997). In discourse analysis, leveraging correlations between tasks is a standard approach. For instance, Stolcke et al. (2000) demonstrated a strong correlation between dialogue acts and discourse structure; similarly, in argumentation mining, Stab and Gurevych (2017) employed joint models to simultaneously identify argument components and argumentation relations.

3 Task Definition and Dataset

3.1 Task Formulation

In our educational context, an "opinion" refers to a context-dependent cognitive stance or hypothesis proposed by a student within a problem-solving thread. Accordingly, we formulate the identification of classroom opinion evolution as a context-aware sequence labeling task. Given a dialogue history $H = \{u_1, u_2, \ldots, u_{t-1}\}$, where u_t represents the current utterance, the objective is to classify u_t into one of the predefined opinion evolution categories $y \in \mathcal{Y}$.

Drawing upon a synthesis of Conceptual Change Theory (Posner et al., 1982) and Argumentation Theory (Toulmin, 1958), we define a label space $\mathcal{Y}$ consisting of six categories: *Irrelevant*, *New opinion*, *Reinforcement*, *Weakening*, *Adoption*, and *Rebuttal*. These categories capture the complete lifecycle of an opinion, spanning from its introduction to its resolution. Detailed definitions for each category and their associated cognitive mappings are provided in Appendix A.

3.2 COED Dataset

To facilitate data-driven research, we constructed the Classroom Opinion Evolution Dataset (COED) by extending the TalkMoves corpus (Suresh et al., 2022).

We recruited three annotators with backgrounds in educational psychology. To ensure high quality, we employed a double-blind annotation strategy where each document was independently labeled by two annotators. The average Cohen's Kappa reached 0.90, indicating strong inter-annotator agreement. Disagreements were resolved via an expert adjudication mechanism to establish gold-standard labels.

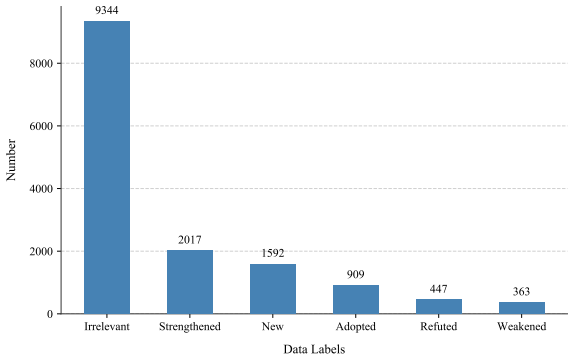

Figure 2: Category distribution of the COED dataset.

The final dataset comprises 14,672 utterances derived from 100 K-12 mathematics classrooms. We specifically chose mathematics as our testbed because its explicit reasoning chains and clear convergence criteria minimize annotation ambiguity compared to highly subjective domains. To ensure experimental rigor, we employed a session-level holdout protocol: all utterances belonging to a unique classroom session were exclusively assigned to the training (11,978), validation (1,249), or test (1,445) sets, maintaining an approximate 8:1:1 ratio.

As illustrated in Figure 2, the dataset exhibits a severe long-tailed distribution characterized by high noise and sparsity. Specifically, non-argumentative utterances dominate the corpus (63.69%), creating a high-noise environment that necessitates robust extraction of valuable cognitive cues. Conversely, high-value cognitive shifts are extremely scarce; notably, the *Rebuttal* and *Weakening* labels collectively account for merely 5.52% of the data, posing a significant challenge for the effective recall of these critical long-tail classes.

4 Methodology

This section introduces the Multi-task Enhanced Cascade Hybrid framework (MECH). The design of this framework follows a "hypothesis-driven" research paradigm, aiming to empirically verify the core hypothesis that "explicit dialogue acts serve as crucial anchors for implicit opinion evolution" via computational modeling.

4.1 Theoretical Framework and System Architecture

Our modeling approach is grounded in the CODA (Continuous Opinions and Discrete Actions) theory (Martins, 2008). We conceptualize opinion

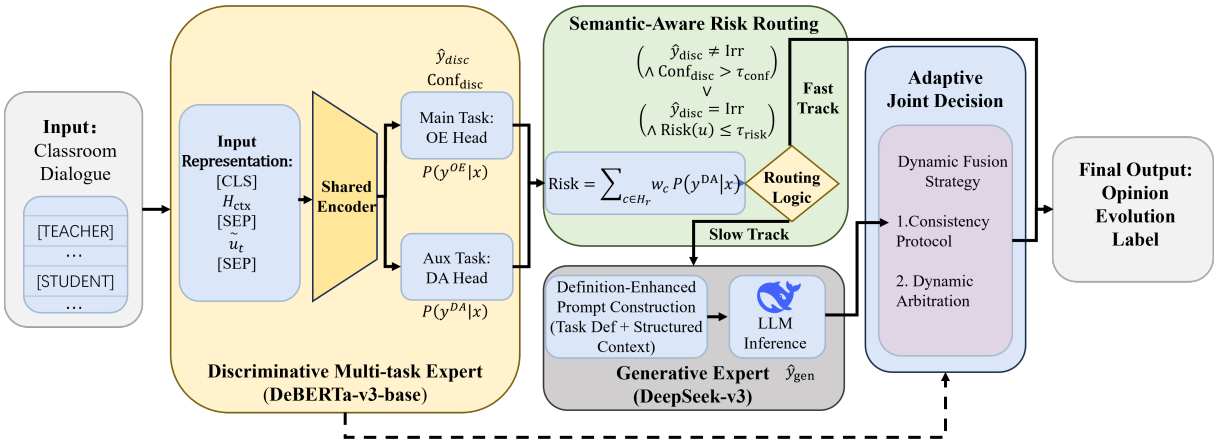

Figure 3: Overview of the Multi-task Enhanced Cascade Hybrid Framework (MECH).

evolution as a dynamic process driven by internal, continuous cognitive states, whereas externally observed "dialogue acts" (DAs) are the discrete actions that manifest this process. In the context of classroom dialogue, we hypothesize that explicit DAs function as the observable logical warrants driving implicit OE.

As illustrated in Figure 3, MECH operationalizes this theoretical intuition into a **semantic-aware cascade workflow**. Rather than a static pipeline, the architecture dynamically orchestrates a lightweight discriminative expert and a generative reasoning expert. The core mechanism involves identifying explicit dialogue acts that manifest internal cognitive states via multi-task learning, and using them as risk signals to route samples. This ensures that computationally expensive reasoning is selectively applied to complex cognitive shifts while efficiently processing administrative discourse. To ensure reproducibility, specific model hyperparameters—including loss function weights, routing thresholds, and definitions of high-risk dialogue acts—are detailed in Appendix B.

4.2 Discriminative Multi-task Expert

As the first stage, we employ DeBERTa-v3-base as a lightweight shared encoder to process the majority of samples. To explicitly model the semantic correlation between tasks, we design a multi-task architecture by connecting two parallel linear heads on top of the encoder: an **OE Head** for the primary opinion evolution task ($P(y^{OE}|x) = \text{Softmax}(W_{OE}h + b_{OE})$) and a **DA Head** for the auxiliary dialogue act task ($P(y^{DA}|x) = \text{Softmax}(W_{DA}h + b_{DA})$).

To capture the significant role asymmetry and interaction dependency in classroom discourse, we explicitly model speaker dynamics by prepending a relative role token $r_i \in \{\texttt{[TEACHER]}, \texttt{[CURRENT]}, \texttt{[OTHER]}\}$ to each utterance u_i, forming a role-augmented representation $\tilde{u}_i = r_i \oplus u_i$. Here, `[CURRENT]` and `[OTHER]` denote whether the context utterance shares the same speaker as the target turn. The input sequence for the target turn t is constructed as:

$$\mathbf{X} = [\text{CLS}] \oplus \text{H}_{\text{ctx}} \oplus [\text{SEP}] \oplus \tilde{u}_t \oplus [\text{SEP}] \quad (1)$$

where $\text{H}_{\text{ctx}} = \tilde{u}_{t-k} \oplus \cdots \oplus \tilde{u}_{t-1}$ represents the sequence of the most recent k role-augmented utterances. This design allows the model to learn specific interaction patterns.

To address the significant long-tail distribution inherent in the data, we employ a strategy combining class-weighted cross-entropy loss with weighted random sampling. The total loss function is defined as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{oe}}(y_{\text{oe}}, \hat{y}_{\text{oe}}; \mathbf{w}_{\text{oe}}) + \lambda \cdot \mathcal{L}_{\text{da}}(y_{\text{da}}, \hat{y}_{\text{da}}) \quad (2)$$

where $\mathbf{w}_{\text{oe}}$ denotes the class weight vector, and λ represents the balancing coefficient for the auxiliary task. By jointly optimizing $\mathcal{L}_{\text{total}}$, the model is compelled to explicitly learn dialogue act features. This not only enhances performance on the primary task but also endows the discriminative model with preliminary interpretability: By examining the output of the DA Head, we can discern the underlying communicative intent upon which the opinion judgment is based.

4.3 Generative Expert

To address complex contexts or ambiguous samples that challenge the discriminative model, we introduce a Large Language Model (LLM) as a

reasoning expert in the second stage, employing DeepSeek-v3 as the backbone. To maximize reasoning performance while controlling computational costs, we design a *definition-augmented prompting module* to explicitly inject cognitive frameworks from educational psychology into the LLM's reasoning space. As detailed in Appendix C, our input prompts are dynamically constructed with three synergistic components.

First, to bridge potential gaps between general LLM knowledge and educational standards, we incorporate **Cognitive Alignment Instructions** containing precise definitions. These serve as constraints for the Chain-of-Thought (CoT) process, compelling the model to classify based on specific argumentation logic (e.g., "whether new evidence is introduced") rather than pre-trained priors. Second, drawing on turn-taking theory (Sacks et al., 1974), we provide a **Structured Interaction Context** by encoding recent dialogue history with explicit speaker consistency tags (i.e., *Same* or *Switch*). This structural aid enables the model to distinguish whether an opinion change stems from "self-correction" or "social adoption." Finally, to mitigate hallucinations and facilitate automated evaluation, we enforce **Output Constraints** via a strict slot-filling format, ensuring standardized results in a zero-shot reasoning setting.

4.4 Semantic-aware Risk Router

Traditional cascading models rely solely on prediction confidence, often leading to missed detections (false negatives) due to "overconfidence" on semantically ambiguous samples. To address this, we design a *Dialogue Act-based Risk Scoring* mechanism to construct a "semantic safety net" specifically for samples predicted as irrelevant. This mechanism grounds routing decisions in explicit semantic signals rather than relying blindly on black-box confidence.

For an input u, we utilize the output of the discriminative model to execute a dual-channel routing logic.

When the discriminator predicts the "Irrelevant" class, we first check if its confidence reaches the filtering threshold τ_{irr}. In standard cascades, samples satisfying $P(\hat{y}|x) \geq \tau_{\text{irr}}$ are immediately filtered to save costs. However, to prevent missing implicit argumentative intent, we calculate a semantic risk score for these high-confidence candidates:

$$\text{Risk}(u) = \sum_{c \in \mathcal{H}_{\text{risk}}} w_c \cdot P(y^{\text{DA}} = c|x) \quad (3)$$

where $\mathcal{H}_{\text{risk}}$ denotes the set of high-risk dialogue acts (e.g., *Making a Claim*, *Providing Evidence*) derived from the auxiliary task head, and w_c represents the empirically assigned risk weight for act c. To accurately capture nuanced semantic shifts, these acts are stratified into strong and medium tiers; specific act categorizations and exact weight values are detailed in Appendix B.2. If $\text{Risk}(u) > \tau_{\text{risk}}$, the sample is deemed to have a "semantic miss risk" despite the high classification confidence. In such cases, the system overrides the discriminator's decision and forcibly routes the sample to the generative expert.

Conversely, for non-irrelevant predictions (e.g., *New*, *Refuted*), we employ a standard confidence threshold. If the discriminator's confidence exceeds τ_{conf}, the model is considered sufficiently certain, and the prediction is output directly to ensure efficiency. Samples failing either the confidence check or the risk check are processed by the generative expert.

4.5 Adaptive Joint Decision

When a sample is routed to the generative expert, the system must resolve potential discrepancies between the predictions of the discriminative model ($\hat{y}_{\text{disc}}$, the domain-adapted expert) and the generative model ($\hat{y}_{\text{gen}}$, the general reasoning expert). To optimally balance the task-specific inductive bias of the discriminator with the robust contextual reasoning of the LLM, we propose a confidence-calibrated adaptive arbitration strategy. Let $c_{\text{disc}} \in [0, 1]$ denote the prediction confidence of the discriminator. The final decision $\hat{y}_{\text{final}}$ is determined through a hierarchical protocol:

1. Consensus Alignment: If both experts agree ($\hat{y}_{\text{disc}} = \hat{y}_{\text{gen}}$), the consensus prediction is adopted directly, representing the highest reliability state.

2. Dynamic Threshold Arbitration: In the event of conflicting predictions, we establish a hierarchy of trust gated by c_{disc}. Specifically, we define two confidence thresholds: an upper bound α_{high} for absolute override, and a relaxed bound α_{mid} for fine-grained resolution.

- *High-Confidence Override:* If $c_{\text{disc}} \geq \alpha_{\text{high}}$, the system prioritizes $\hat{y}_{\text{disc}}$, operating under the assumption that the domain expert has captured definitive, task-specific semantic signals.

- *Intra-Class Resolution:* If both models predict stance-related categories (i.e., $\hat{y}_{\text{disc}}, \hat{y}_{\text{gen}} \in \mathcal{Y}_{\text{opinion}}$) but diverge on the exact label, we prioritize the discriminator provided $c_{\text{disc}} \geq \alpha_{\text{mid}}$. This leverages the discriminator's superior boundary-learning capabilities for fine-grained opinion classification.
- *Generative Fallback:* In all remaining uncertainty intervals (e.g., $c_{\text{disc}} < \alpha_{\text{mid}}$), the system defaults to $\hat{y}_{\text{gen}}$, relying on the LLM's zero-shot reasoning capabilities to disambiguate complex semantics.

3. Data-Scale Calibration: Crucially, discriminative models trained in low-resource (few-shot) settings often suffer from under-calibration, leading to systematically lower confidence distributions. To prevent an over-reliance on the generative model in such scenarios, our arbitration thresholds are dynamically parameterized. When the empirical average confidence of the discriminator indicates a low-resource regime, we apply a decay penalty Δ to the arbitration thresholds (i.e., $\alpha \leftarrow \alpha - \Delta$). This calibration ensures that the domain expert's inductive bias remains adequately weighted across varying scales of training data.

5 Experimental Setup

5.1 Baseline Models

We compare MECH against two categories of baselines.

Discriminative PLMs We utilize **BERT-base** (Devlin et al., 2019), **RoBERTa-large** (Liu et al., 2019), and **DeBERTa-v3-base** (He et al., 2023), with the latter serving as the strong non-generative baseline.

Generative LLMs We evaluate both zero-shot and fine-tuned settings: (1) **Proprietary models**: GPT-4o (OpenAI et al., 2024) and DeepSeek-v3 (DeepSeek-AI et al., 2024) via APIs; (2) **Open-source models**: Llama-3.1-8B (Dettmers et al., 2023) and Qwen2-7B (Yang et al., 2024), fine-tuned on the full dataset using QLoRA (Dettmers et al., 2023).

5.2 Evaluation Metrics

Given the long-tail distribution, we report **Macro-F1** as the primary metric to effectively penalize misclassifications of tail samples. We also report **Cost Ratio** (relative to vanilla LLM inference) to assess cost-effectiveness.

Method	Macro F1	Accuracy		
		DeBERTa	LLM	MECH
Baseline	0.6135	0.7322	0.5385	0.7640
Proposed	**0.6828**	**0.7543**	**0.6455**	**0.7855**

Table 1: Performance comparison between baseline and proposed MECH. "LLM" corresponds to the DeepSeek-v3 accuracy on the routed subset.

5.3 Implementation Details

Experiments were conducted on a single NVIDIA RTX 4090 (24GB). Discriminative models were optimized using AdamW (lr: 1e-5 to 2e-5, batch size: 8–16) for 12–15 epochs. For Llama-3.1 and Qwen2, we employed QLoRA (4-bit NF4 quantization, $r = 16$, $\alpha = 32$) with gradient accumulation to fit memory constraints. Proprietary models were accessed with temperature set to 0.0 for reproducibility.

6 Experimental Results and Analysis

6.1 Scientific Hypothesis Verification: The Auxiliary Role of Dialogue Acts

To validate our core scientific hypothesis—that identifying dialogue acts significantly aids in capturing opinion evolution—we compare a single-task baseline (**Baseline**) against our proposed dual-task hybrid framework (**Proposed**).

- **Baseline**: A discriminative model trained in a single-task setting, supervised solely on opinion labels without risk routing. This configuration simulates baseline performance in the absence of communicative intent awareness.
- **Proposed**: Our complete method, which jointly models dialogue acts and opinion evolution while incorporating the risk routing mechanism.

As shown in Table 1, the experimental results strongly support our hypotheses:

Reconstructing Feature Space via Multi-Task Learning (Addressing "Implicit Semantics") Even without reliance on external LLMs, merely introducing DA (Dialogue Act) as an auxiliary task improves the Accuracy of the discriminative model (DeBERTa) from 0.7322 to 0.7543. This improvement confirms the intrinsic coupling between DA and OE (Opinion Evolution). Explicit supervision signals from communicative intents optimize the feature space of the shared encoder. This enables the model to leverage explicit pragmatic features—such as "providing evidence" or "refuting"—to aid

Model	Macro F1	Accuracy	Cost (%)
Discriminative Baselines			
BERT-base-uncased	0.5461	0.6913	0
RoBERTa-large	0.6228	0.7633	0
DeBERTa-v3-base	0.5810	0.7488	0
Generative Baselines			
Qwen2-7B (FT)	0.5509	0.6837	0
Llama-3.1-8B (FT)	0.5608	0.6775	0
GPT-4o (ZS)	0.5688	0.7156	100
DeepSeek-v3 (ZS)	0.5963	0.6830	100
MECH (Ours)	**0.6828**	**0.7855**	**55.6**

Table 2: Performance and cost comparison between MECH and mainstream discriminative and generative baselines. "FT" denotes Fine-Tuned, and "ZS" denotes Zero-Shot.

inference when processing semantically ambiguous opinion evolutions, thereby mitigating the limitations of singular semantic understanding to some extent.

Validation of Risk Signal Effectiveness (Addressing "Unwarranted Confidence") When comparing performance on the subset of samples routed to the LLM (denoted as DeepSeek Accuracy), our proposed method (0.6455) significantly outperforms the baseline (0.5385). In the baseline, routing relies solely on confidence scores, which results in a large number of hard samples—where the model exhibits "confident but incorrect" behavior—being retained within the discriminative stage. Conversely, in our proposed approach, the DA-based risk score functions as a "semantic radar," successfully filtering out long-tail samples that are challenging for the discriminator but possess high-risk argumentative intents. The improvement in DeepSeek Accuracy demonstrates that the samples selected by this mechanism are indeed "genuinely hard" instances requiring strong reasoning capabilities, rather than random noise.

6.2 Main Experiment Comparison: Breaking the "Accuracy-Cost-Data" Trilemma

Table 2 presents the performance comparison between our method (MECH) and mainstream discriminative and generative models on the test set. To ensure a fair comparison, all baseline models were evaluated using a data setting identical to that of MECH. Experiments demonstrate that MECH successfully identifies the "Pareto optimal solution" between performance and cost.

Limitations of Discriminative and Generative Models The Macro-F1 of the best discriminative baseline (RoBERTa-large) is 0.6228, which is significantly lower than that of MECH (-6.0%). This indicates that merely scaling up parameters cannot fully resolve the complex contextual dependency issues in educational dialogues; logical reasoning capabilities must be introduced. Furthermore, generative models (including fine-tuned 7B/8B models and SOTA large models like GPT-4o and DeepSeek-v3) failed to surpass the strong discriminative baselines on this task. This suggests that without domain-specific structural modeling, general-purpose LLMs struggle to accurately grasp the deep pragmatic logic within educational dialogues.

Comprehensive Advantages and Efficiency of MECH MECH achieves a significant performance breakthrough, reaching a Macro-F1 of 0.6828 and an accuracy of 0.7855. This represents an improvement of 11.4% and 6.99%, respectively, compared to GPT-4o, validating the effectiveness of combining a lightweight discriminator with LLM reasoning capabilities. Meanwhile, its cost ratio is only 55.6%, as the risk routing mechanism filters out approximately 44.4% of simple samples. This substantially reduces computational and deployment costs while achieving SOTA performance, demonstrating superior practicality and scalability.

6.3 Data Efficiency and Cost Dynamics: Robustness in Cold-Start Scenarios

Addressing the challenge of scarce and expensive high-quality annotated data in the educational domain, we evaluated the performance and cost-efficiency of MECH in low-resource scenarios. Figure 4 illustrates the comparison of Macro-F1 trends between the discriminative baseline (DeBERTa) and MECH, alongside the corresponding MECH API cost (LLM routing rate).

Experimental results indicate that MECH exhibits exceptional few-shot adaptability. Using only 20% of the training data, MECH achieves a Macro-F1 of 0.6041, performing competitively with the fully-trained DeBERTa on 100% data (0.6111). Importantly, at this 20% data regime, MECH routes only 37.5% of samples to the LLM. This translates to a 62.5% cost reduction compared to a pure LLM approach, proving that the advertised cost savings hold robustly even under severe cold-start condi-

Model Variants	F1-Score by Category					Recall	Macro F1
	New	Strengthened	Weakened	Adopted	Refuted		
w/o Risk-Aware Routing	0.5108	0.6526	0.5176	**0.6989**	0.7033	0.6710	0.6622
w/o Multi-Task Learning	0.5032	0.6514	0.3614	0.6135	0.6744	0.6864	0.6135
w/o Role Features	0.3719	0.6608	0.3291	0.5714	0.6216	0.5382	0.5720
w/o Adaptive Voting	0.4458	0.6694	0.4375	0.6667	0.7160	0.6723	0.6356
w/o Consistency Prompt	0.5472	0.6542	0.4771	0.5570	0.6250	0.6289	0.6231
MECH (Ours)	**0.5653**	**0.6789**	**0.5417**	0.6811	**0.7442**	**0.6997**	**0.6828**

Table 3: Ablation study on different components of MECH.

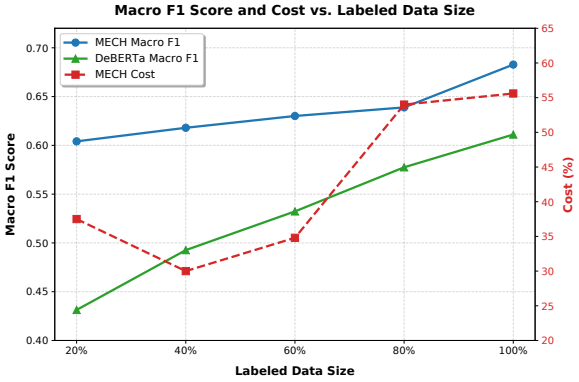

Figure 4: Cost-efficiency analysis in low-resource settings. The solid lines (left axis) compare Macro-F1 scores, while the dashed red line (right axis) depicts the MECH API Cost (percentage of samples routed to the LLM).

tions.

The mechanism behind this phenomenon highlights the adaptivity of the risk-aware router. When the discriminator is severely under-trained (20%), the router effectively leverages the LLM's zero-shot priors to compensate for information scarcity. Interestingly, as the training data scales to 100%, the routing cost does not decrease but rather rises to 55.6%, accompanied by a peak MECH F1 of 0.6828. This counter-intuitive trend occurs because a fully-supervised discriminator develops a highly sensitive risk-detector, intentionally routing more "pseudo-irrelevant" complex long-tail samples to the generative expert. Ultimately, MECH guarantees drastic cost reductions when supervision is scarce, and dynamically trades compute foraccuracy when fully optimized.

6.4 Ablation Studies and Analysis

To verify the effectiveness of each component within the MECH framework, we conducted a detailed ablation study to evaluate their impact on per-class F1 scores, Macro-F1, and overall Recall. Detailed results are presented in Table 3.

Salvaging Recall via Risk-Aware Routing The *w/o Risk-Aware Routing* variant removes the interception mechanism based on dialogue act risk scores. Experimental results show that while performance remains acceptable for some high-confidence categories (such as *Adopted*), the overall Recall drops significantly from 0.6997 to 0.6710, accompanied by a substantial decline in the F1 score for the *New* category. This shift in data directly corroborates the core value of risk routing: intercepting false negatives. In the full model, when the discriminator misclassifies an implicit opinion expression as *Irrelevant*, high-risk DA signals forcibly trigger a routing review. Removing this mechanism causes these hard samples to be discarded directly, resulting in an immediate loss of recall. This proves that this module is a critical line of defense against the "missed detection" problem.

Capturing Implicit Intents via Multi-Task Learning Having previously compared the single-task baseline with our method in terms of accuracy, we now perform a fine-grained class analysis. The single-task baseline (*w/o Multi-Task Learning*) exhibits significant performance degradation when processing implicit opinions. The most notable gap appears in the *Weakened* category (F1 drops from 0.5417 to 0.3614). This result reaffirms our scientific hypothesis: dialogue acts are explicit precursors to opinion evolution. Without the supervision of the DA task, the model's sensitivity to these pragmatic features is significantly reduced.

Role Features as the Cornerstone of Interaction Modeling The *w/o Role Features* variant leads to the most severe deterioration, with Recall plummeting to 0.5382. This indicates that distinguishing "who said what" is crucial. Without role embeddings, the model fails to construct speaker consistency links, causing recognition failure in interaction-dependent categories (e.g., *New*, *Weak-*

ened) and degenerating into a simple text classifier.

Constraints of the Consistency Prompt on LLM Removing consistency markers from the Generative Expert's prompt (*w/o Consistency Prompt*) causes the Recall to drop to 0.6289. This suggests that zero-shot LLMs struggle with opinion attribution. Explicit *Consistency* tags serve as critical reasoning anchors, substantially improving the generator's accuracy.

Optimization via Adaptive Joint Decision Removing the adaptive strategy (*w/o Adaptive Voting*) results in a drop in Macro F1 to 0.6356, proving that simple hard voting cannot address model divergences. The adaptive strategy successfully balances the discriminator's high-confidence domain expertise with the LLM's generalized reasoning. Furthermore, this hierarchical arbitration provides a traceable decision path—distinguishing between empirical judgment and logical reasoning—which is vital for trust in educational scenarios.

The ablation experiments above quantify the contribution of each module from a statistical perspective. To further visually demonstrate how the MECH framework corrects discriminative model errors and recovers overlooked opinions in real-world contexts, we provide detailed qualitative case studies in Appendix D.

7 Conclusion

In this work, we have filled the gap regarding the dimension of "opinion evolution" in classroom discourse analysis and released **COED**, the first benchmark dataset for this domain. To address the severe "Accuracy-Cost-Data" trilemma in real-world educational scenarios, we proposed the **MECH** framework. This framework innovatively integrates multi-task learning objectives to explicitly model the intrinsic indicative role of dialogue acts in opinion evolution. Furthermore, we designed a semantic-aware risk routing mechanism to construct a "semantic safety net" specifically for long-tail implicit samples. Coupled with a definition-augmented prompting strategy and an adaptive joint decision-making mechanism, MECH effectively elicits the reasoning potential of Large Language Models (LLMs) in complex contexts. Our experimental results demonstrate that MECH not only outperforms GPT-4o on the full dataset with a significant cost reduction of 44.4% but also exhibits stronger robustness in data-scarce cold-start scenarios. Overall, this work validates the efficacy of the "Small Model Perception + Large Model Reasoning" paradigm for complex cognitive analysis tasks, paving a new path for large-scale, cost-effective, and precise instructional diagnosis.

Limitations

While MECH achieves promising results, we acknowledge limitations in two aspects pointing toward future research. First, methodologically, the current set of "high-risk dialogue acts" relies on static empirical observations, and the underlying causal mechanism of the DA–OE correlation remains unclear. Future work will conduct data-driven experiments to dynamically define risk sets and construct interpretable transition graphs. Second, regarding domain generalizability, our evaluation is confined to educational dialogues (TalkMoves). Validating the framework in other contexts, such as open-domain debates or social media deliberations, remains an important direction for future research.

Ethical Considerations

The Classroom Opinion Evolution Dataset (COED) is strictly built upon the publicly available TalkMoves dataset. This research did not conduct any new original dialogue collection involving human subjects. Our newly annotated opinion evolution information does not contain any Personally Identifiable Information (PII). Furthermore, the public release of COED will strictly adhere to the original terms of use and licensing agreements of the TalkMoves project, ensuring it is used solely for non-commercial educational research.

Acknowledgements

The authors would like to thank the anonymous reviewers and the Area Chairs for their insightful and constructive reviews. We are grateful to Xianan Wang, Yinghui Li, and Enxi Dou for their dedicated work in annotating the dataset. This research was supported by the National Natural Science Foundation of China (Grant No. 62276178), the Henan Provincial Natural Science Foundation (Grant No. 262300421797), the Henan Provincial Science and Technology Research Project (Grant Nos. 252102210102 and 262102210084), and the Jiangsu Provincial Higher Education Teaching Reform Research Project (Grant No. 2025JGYB594).

References

Rich Caruana. 1997. Multitask learning. *Machine Learning*, 28(1):41–75.

Lingjiao Chen, Matei Zaharia, and James Zou. 2024. FrugalGPT: How to use large language models while reducing cost and improving performance. *Transactions on Machine Learning Research*. Published: Dec 2024.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, and 1 others. 2024. DeepSeek-V3 technical report. *Computing Research Repository*, arXiv:2412.19437.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

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.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

Lei Li, Yankai Lin, Deli Chen, Shuhuai Ren, Peng Li, Jie Zhou, and Xu Sun. 2021. CascadeBERT: Accelerating inference of pre-trained language models via calibrated complete models cascade. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 475–486, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *Computing Research Repository*, arXiv:1907.11692.

Luca Lugini and Diane Litman. 2020. Contextual argument component classification for class discussions. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1475–1480. International Committee on Computational Linguistics.

André C. R. Martins. 2008. Continuous opinions and discrete actions in opinion dynamics problems. *International Journal of Modern Physics C*, 19(4):617–624.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. GPT-4o system card. *Computing Research Repository*, arXiv:2410.21276.

George J. Posner, Kenneth A. Strike, Peter W. Hewson, and William A. Gertzog. 1982. Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education*, 66(2):211–227.

Lauren B. Resnick, Christa S. C. Asterhan, and Sherice N. Clarke, editors. 2015. *Socializing Intelligence Through Academic Talk and Dialogue*. American Educational Research Association, Washington, DC.

Carolyn P. Rosé and Oliver Ferschke. 2016. Technology support for discussion based learning: From computer supported collaborative learning to the future of massive open online courses. *International Journal of Artificial Intelligence in Education*, 26(2):660–678.

Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–374.

Abhijit Suresh, Jennifer Jacobs, Charis Harty, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. The TalkMoves dataset: K-12 mathematics lesson transcripts annotated for teacher and student discursive moves. In *Proceedings of the 13th Language Resources and Evaluation Conference*, pages 4654–4662, Marseille, France. European Language Resources Association.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press, Cambridge, UK.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2 Technical Report. *Computing Research Repository*, arXiv:2407.10671.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen

Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2023. A survey of Large Language Models. *Computing Research Repository*, arXiv:2303.18223.

A Detailed Definitions of Opinion Evolution

This appendix elaborates on the taxonomy of classroom opinion evolution introduced in Section 3.1. This taxonomy is a theory-driven framework constructed via a systematic synthesis of Conceptual Change Theory and Argumentation Theory.

A.1 Theoretical Foundations and Design Principles

The construction of this taxonomy aims to address the challenge of the "invisibility of implicit cognitive processes." Its core logic is founded on the mapping of the following two theoretical pillars:

Conceptual Change Theory (Internal Cognitive Mechanisms) This theory elucidates the mechanisms underlying deep learning. It posits that when learners encounter new information, their knowledge structures evolve primarily through two pathways:

- **Assimilation**: Integrating new information into existing mental schemas (incremental learning).
- **Accommodation**: Restructuring or revising existing schemas when new information triggers cognitive conflict (transformative learning).

While this theory describes how cognitive states change, the process is typically implicit.

Argumentation Theory (External Linguistic Representation) This theory provides tools for analyzing explicit discourse functions. It defines interactional relationships within dialogue as argumentation moves, such as support or attack. This characterizes the reasoning process observable in social interactions.

We establish a one-to-one mapping: "supportive" utterances by students (or teachers) are interpreted as external signals of internal "assimilation" or opinion reinforcement, while "attacking" utterances are regarded as precursors that trigger cognitive conflict and potentially lead to "accommodation." Consequently, each label in this taxonomy serves not merely as a classification of surface-level discourse acts but as a diagnostic indicator of a specific stage of cognitive transformation.

A.2 Taxonomy Overview

Table 4 details the six categories of opinion evolution, accompanied by authentic classroom dialogue examples drawn from our constructed COED dataset. These examples illustrate how to identify the specific evolution category based on the function of an utterance within its dialogue context.

A.3 Case Study: Tracing Opinion Evolution in Context

While Table 4 provides isolated, utterance-level definitions, an "opinion" in our framework is fundamentally a **context-dependent stance within a problem-solving thread**, rather than an isolated proposition.

To further elucidate the dynamic application of OE labels within a continuous discourse, we present a representative dialogue excerpt regarding the geometric definition of a "face." This case study observes the interaction from the perspective of explicit Dialogue Acts (DA) to demonstrate how they underpin and drive the multi-turn evolution of implicit Opinion Evolution (OE) states:

Teacher: What is a "face"?
[DA: Press Reasoning → OE: Irrelevant]
Toby (Student 1): It is the front of the shape.
*[DA: Making a Claim → OE: **New**]*
(Context: A novel hypothesis is introduced.)
Teacher: It can be the front, but not [entirely]...
*[DA: Keeping Together → OE: **Weakened**]*
(Context: Toby's specific hypothesis is challenged, creating cognitive conflict.)
Guy (Student 2): Is it its surface?
*[DA: Making a Claim → OE: **New**]*
Teacher: Its surface.
*[DA: Revoicing → OE: **Adopted**]*
Guy (Student 2): A face is the flat surface of a polygon.
*[DA: Making a Claim → OE: **Strengthened**]*
(Context: Guy consolidates and refines the concept from "surface" to "flat surface".)
Teacher: A flat surface, so it is...
*[DA: Revoicing → OE: **Adopted**]*

As demonstrated above, this process of *Proposing Hypothesis → Facing Conflict → Refining Definition* represents the essence of mathematical argumentation and opinion evolution. It clarifies that our OE labels do not target isolated sentences, but rather track the current status of the active hypothesis within the collaborative reasoning chain.

Label	Definition	Example (from COED Dataset)
Irrelevant	Utterances unrelated to the discussion topic, meta-comments regarding the dialogue state, or procedural remarks.	T: Okay. T: Make up a problem for me. *(Procedural instruction)*
New	Utterances that introduce a novel claim or concept into the dialogue.	Erik: What makes thirds? Alan: Thirds, thirds out of a, thirds out of this? *(Introducing the new problem of "thirds")*
Strengthened	Utterances that consolidate an existing opinion through new evidence or elaboration.	Erik: Okay, ten. Erik: So that's ten, this must be nine. *(Reinforcing the judgment on rod length through measurement reasoning)*
Weakened	Utterances that challenge or question an existing opinion.	Erik: But nine can. Alan: Nine can, but there is no nine rod. *(Alan questions Erik's hypothesis using the factual evidence that "there is no rod of length 9")*
Adopted	Utterances where the speaker explicitly accepts a previously disputed opinion.	Erik: Light green. Alan: Light green would make thirds out of the orange. *(Alan adopts Erik's suggestion and provides a rationale)*
Refuted	Utterances where the speaker explicitly and directly denies an opinion present in the context.	Erik: Eleven, this is twelve though. Alan: No, it isn't, look. *(Alan directly refutes Erik's judgment)*

Table 4: Taxonomy of Classroom Opinion Evolution.

Component	Parameter	Symbol	Value
Loss Function (Eq. 2)	Weakened Weight	w_{weak}	3.0
	Refuted Weight	w_{refut}	2.0
	Base Weight	w_{base}	1.0
	Aux. Task Coeff.	λ	0.67
Risk Router (Sec. 4.4)	Sem. Risk Thresh.	τ_{risk}	0.25
	Confidence Thresh.	τ_{conf}	0.96
	Irrelevant Thresh.	τ_{irr}	0.80
	Strong Risk Wt.	w_{strong}	1.0
	Medium Risk Wt.	w_{med}	0.6
Adaptive Decision (Sec. 4.5)	Expert Trust Thresh.	α_{high}	0.90
	Fine-grained Correct.	α_{mid}	0.80
	Decay Penalty	Δ	0.20
Generative & Opt.	LLM Temperature	T	0.0
	Learning Rate	lr	2e-5
	Batch Size / Dropout	B / dp	16 / 0.6

Table 5: Detailed Hyperparameter Configurations for the MECH framework.

B Implementation Details and Hyperparameters

B.1 Hyperparameter Settings

Table 5 presents all key hyperparameters used in the final experiments.

B.2 Definition of High-Risk Dialogue Acts

Within the semantic-aware routing mechanism, empirical observations on the validation set indicate that certain dialogue acts exhibit varying degrees of correlation with implicit opinion evolution. To maximize the recall of the semantic safety net, we expand the high-risk set into two tiers with corresponding discount weights:

Strong High-Risk Acts (Weight = 1.0): These act as direct precursors to opinion shifts (particularly *New*, *Strengthened*, and *Refuted*).

- **Making a Claim**, **Providing Evidence /Reasoning**
- **Press Reasoning**, **Relating to Another Student**
- **Revoicing**, **Restating**

Medium High-Risk Acts (Weight = 0.6): These acts signal potential cognitive dissonance but with lower certainty.

- **Press Accuracy**, **Asking for Info**

Accordingly, the semantic risk score computation in Eq. 3 is expanded into a weighted sum of probabilities derived from the discriminative model's DA Head:

$$\begin{aligned}\text{Risk}(u) = &\sum_{c \in \mathcal{H}_{\text{strong}}} P(y^{\text{DA}} = c|x) \\ &+ 0.6 \sum_{c \in \mathcal{H}_{\text{medium}}} P(y^{\text{DA}} = c|x)\end{aligned} \tag{4}$$

Category	Cognitive Definition & Annotation Criteria
Irrelevant	The utterance is unrelated to the topic, meta-comments (e.g., "I don't understand"), or procedural management.
New	Introduces a new claim, argument, or evidence not present in the preceding context.
Strengthened	Provides support, agreement, evidence, or elaboration for an existing opinion.
Weakened	Raises questions, counter-examples, or limitations regarding an existing opinion without fully negating it.
Adopted	Explicitly agrees with or accepts an opinion from another speaker (typically in *Switch* state).
Refuted	Explicitly and directly negates an opinion (typically in *Switch* state, e.g., "No...").

Table 6: Cognitive definitions for Opinion Evolution categories injected into the System Prompt.

Component	Template Content

System	`You are a classroom dialogue analysis expert. Classify the "Current Utterance" based on the "Context" into one of six categories:`
	`1. Irrelevant: [Def. from Table 6]`
	`...`
	`6. Refuted: [Def. from Table 6]`
	`Output Constraint: Output only the label word.`

User	`{context_str}`
	`Current Utterance:`
	`- Speaker: {speaker}`
	`- Consistency: {state}`
	`- Sentence: {sentence}`
	`Please classify:`

Table 7: The structured prompt template used for the Generative Expert.

C Prompt Design for the Generative Expert

To ensure the alignment of the Large Language Model's reasoning with our educational taxonomy, we designed a definition-augmented prompt. This prompt comprises two components: Cognitive Alignment Definitions (System Prompt) and a Structured Input Template (User Prompt).

C.1 Cognitive Alignment Definitions

We formalized the six categories of opinion evolution into explicit discrimination criteria and injected them into the model as system instructions. Table 6 presents these definitions, which directly constrain the model's reasoning space.

C.2 Structured Input Template

To capture the interaction flow within multi-turn dialogues, we constructed the input in a structured slot-filling format. Specifically, we introduced a `[Consistency]` slot to explicitly prompt the turn-taking status. The complete prompt template is shown in Table 7.

D Effectiveness Analysis of the Hybrid Mechanism

To qualitatively understand how MECH enhances performance, we analyze two representative cases, as detailed in Table 8, that respectively highlight the effectiveness of the "Adaptive Joint Decision" and "Risk-Aware Routing" mechanisms.

Case 1: Correcting Granularity Errors via Adaptive Decision-Making In this instance, the discriminative model, misled by surface-level negation markers (e.g., "no"), misclassifies the utterance as *Refuted*. However, its confidence score (0.7382) falls below the high-confidence threshold, indicating inherent model uncertainty. MECH's Adaptive Joint Decision mechanism detects this uncertainty and delegates the sample to the generative expert (LLM). Leveraging its semantic reasoning capabilities, the LLM accurately discerns the speaker's implicit logic—accepting the mathematical premise while questioning physical feasibility—thereby correcting the prediction to *Weakened*. This demonstrates that the confidence-based cascading strategy effectively functions as a "second opinion," mitigating the limitations of lightweight models in comprehending complex logic.

Case 2: Recalling Overlooked Opinions via Risk-Aware Routing This case illustrates the susceptibility of lightweight models to "false negative" errors in scenarios involving implicit argumentation. The discriminative model failed to establish the mathematical reasoning link between the utterance and the preceding context, misclassifying it as *Irrelevant* with high confidence (0.7728). In a conventional workflow, this sample would have been directly filtered out. MECH, however, detected the implicit intent of "Providing Evidence"

Component	Case 1 (Adaptive Decision)	Case 2 (Risk Routing)
Context	**Alan**: No, ten can't be divided into thirds. *(OE: Refuted) (DA: Providing Evidence)* **Erik**: But nine can. *(OE: New) (DA: Relating to Another Student)*	**Jamie**: How many chickens does Grandpa have? *(OE: Irrelevant) (DA: None)* **Jeff**: Six. *(OE: New) (DA: Providing Evidence)*
Target Utterance	**Alan**: Nine can, but there is no nine rod. *(OE: Weakened) (DA: Relating to Another Student)*	**Jeff**: He sells three, he will have as many as Dad. *(OE: Strengthened) (DA: Providing Evidence)*
Discriminative Expert	Prediction: Refuted Confidence: 0.73	Prediction: Irrelevant Confidence: 0.77
Generative Expert	Prediction: Weakened	Prediction: Strengthened
Reason	Insufficient confidence in the discriminative model triggered the LLM for deep logical reasoning, thereby correcting the granularity error.	The auxiliary task detected the "Providing Evidence" intent, triggering risk-aware routing and enforcing LLM verification.

Table 8: Qualitative comparison of two representative cases demonstrating MECH's error correction mechanisms.

via the auxiliary dialogue act recognition module. This generated a high-risk score that triggered the risk-aware routing mechanism to enforce an LLM re-evaluation. The LLM successfully reconstructed the underlying mathematical reasoning chain and correctly reclassified the utterance as *Strengthened*. This process validates the role of dialogue act information as a "semantic safety net," significantly enhancing the system's capability to recall high-value implicit argumentation.