CantoTalk: Probing Teacher Expertise From Fine-Tuned Talk Move Representations

Mayank Sharma*
Stanford University
masharma@stanford.edu
Xinman Liu*
Stanford University
xinman@stanford.edu
Gordon Wing-Leung Yeung
Stanford University
wlyeung@stanford.edu

ABSTRACT

Classroom discourse profoundly shapes student learning, yet analyzing teacher talk at scale remains challenging in non-Western and low-resource language contexts. This paper introduces CantoTalk, a dataset of 7,518 Cantonese teacher utterances from Hong Kong mathematics classrooms annotated with ten talk-move categories. We investigate whether LLMs can reliably classify these moves and encode systematic differences in teacher expertise. Fine-tuning five open-weight LLMs yields strong performance, with the best model (Qwen3-8B) achieving micro-F1 of 0.81 and macro-F1 of 0.77. Probing utterance-level embeddings reveals that teacher expertise is linearly separable with 0.79 balanced accuracy, well above chance even after controlling for surface features. Clustering analyses uncover three coherent discourse styles differing in pedagogical authority, scaffolding, and dialogic engagement, with qualitative analysis showing systematic differences in how experienced and novice teachers execute similar talk moves. These findings demonstrate that fine-tuned LLM representations capture teacher expertise, offering a new lens for analyzing classroom discourse and informing teacher feedback tools.

Keywords

Talk Move Classification, LLM Fine-tuning, Representation Learning, Teacher Expertise Detection, Low-Resource NLP, Educational Discourse Analysis

1. INTRODUCTION

Classroom discourse profoundly shapes student learning and is influenced by pedagogical traditions and cultural norms [1, 2]. Yet in typical classrooms, students contribute far less than teachers: they speak for only 27 seconds per hour of teacher talk and ask just 2 academic questions per day, compared to teachers' 142 questions [3]. Ensuring equitable classroom talk therefore depends on teachers' ability to facilitate discourse that invites reasoning, dialogue, and participation from all learners [4]. To support these goals, researchers have developed structured "talk-move" frameworks to help teachers better engage students in class and achieve their instructional objectives [5, 6]. Traditional analysis of these talk moves, however, has relied on labor-intensive manual coding methods that cannot scale to support widespread teacher development or classroom research [4].

Recent advances in large language models (LLMs) offer the potential to automate the identification of talk moves and provide immediate actionable insights to teachers [7]. However, classroom discourse remains largely an out-of-distribution domain for foundational models, which exhibit biases and misalignment to learning outcomes [8]. While there is emerging work on enhancing talk-move analysis and aligning models with pedagogical frameworks [9, 10], most studies focus on Western, English-speaking corpora and remain limited to classification tasks and frequency counts. Such analyses overlook cross-cultural adaptations of talk moves [11], the challenges of leveraging LLMs in low-resource language settings [12], as well as other contextual factors such as how teacher expertise influence discourse patterns in classroom [13].

This study therefore aims to address these gaps by examining how LLMs can identify and interpret talk moves in the authentic, low-resource setting of Hong Kong mathematics classrooms. We investigate two key research questions with CantoTalk[1]: **(1) Can LLMs be fine-tuned to perform reliable talk move classification in Cantonese classroom discourse?** and **(2) How do the distributions and linguistic realizations of talk moves differ between novice and experienced teachers, and can these patterns be uncovered from fine-tuned representations of the LLMs?**

2. RELATED WORK

*Equal contribution.

[1]GitHub Repository: https://github.com/matrix-mayank/canto-talk

2.1 Talk Moves in Mathematics Classrooms

"Talk moves" are utterance-level strategies used by teachers and students to elicit, clarify, or extend thinking in service of productive, dialogic learning [14]. Within mathematics instruction, these moves help navigate the tension between open-ended exploration and specific instructional objectives [5, 15]. Various frameworks have operationalized these strategies for analysis. [14] established a four-cluster taxonomy organized by instructional goals: encouraging students to expand their contribution, listen carefully to one another, make their reasoning explicit, and engage with others' ideas. This "Talk Moves as Tools" (TMT) framework has been widely adapted and extended. For instance, Suresh et al. [7] outline six categories: keeping everyone together, getting students to relate, restating, revoicing, pressing for accuracy, and pressing for reasoning. Research also suggests that the quality of these moves directly impacts engagement. Alic et al. [16] distinguish between "funneling" questions that guide students toward predetermined answers versus "focusing" questions that deepen mathematical thinking; while Demszky et al. [17] also highlight the importance of teacher "uptake" of student contributions, i.e., a dialogic teaching practice that acknowledges and builds upon student ideas, in ensuring students feel heard and prompted to engage more deeply with mathematical concepts.

However, these frameworks largely focus on dynamics within English-speaking classrooms; yet, the privileging of individual student talk in Western research may inadvertently undervalue other legitimate forms of mathematical communication in cultures where norms around authority, participation, and learning are different [18]. For instance, Xu and Clarke [11] demonstrate through cross-cultural analysis of Shanghai, Seoul, and Melbourne classrooms that choral responses (whole-class unison replies to a teacher's prompt) and strategic use of silence can be equally effective for developing mathematical competence, challenging assumptions that individual verbal participation is the primary indicator of engagement. Similarly, Ng et al. [19] and Ni et al. [20] highlight that in the specific context of Hong Kong mathematics classrooms, talk moves such as "inviting choral response" and "teacher evaluation" are common culturally-specific practices that put more emphasis on teacher control in advancing student's knowledge of curricular content.

2.2 Analysis of Classroom Discourse

Despite their importance, most assessments of talk moves rely on manual observation that is resource-intensive and difficult to scale [4]. Advances in computational approaches offer the potential to automate the identification and analysis of talk moves at scale [7]. With transformer-based language models, the field has progressed rapidly from early studies using small, curated datasets like TalkMoves to more robust systems trained on extensive classroom corpora such as NCTE-119 and SAGA-22 [7, 10]. The transition from Bi-LSTM architectures to BERT-based models has yielded substantial performance improvements, with reported F1-scores rising from 65% to exceeding 82% [21, 10]. More recent developments in LLMs and generative AI present new opportunities for automated feedback systems and talk move classification. While earlier studies have shown that task-specific BERT-based architectures outperform general-purpose LLMs like ChatGPT for talk move classification — a gap attributed to BERT variants learning discourse-specific cues through fine-tuning, such as turn-level patterns and move co-occurrences, that zero-shot LLMs lack [8] — the latest research suggests that fine-tuning LLMs significantly alters this landscape. For example, fine-tuned GPT-3.5-turbo models have been shown to outperform state-of-the-art RoBERTa classifiers in identifying tutor talk moves, achieving a macro F1 score of 0.87 compared to 0.78, highlighting the critical role of fine-tuning and leveraging context (such as providing long conversation segments) to adapt LLMs effectively for the discourse analysis tasks on large-scale mathematics classroom corpora [9].

However, these advances are unevenly distributed, where developing NLP applications for low-resource languages remains challenging due to the scarcity of annotated datasets and the complexity of transferring linguistic rules from high-resource languages [22, 23]. For example, Ng et al. [24] noted that standard linguistic markers for authority (such as modal verbs) in English do not map directly to Cantonese classroom discourse, complicating the use of out-of-the-box English models. Recent benchmarks confirm that while top-tier proprietary models (e.g., GPT-4o) show emerging competence, they still exhibit significant gaps in culturally nuanced understanding and morphological generalization compared to English, with smaller open-source models lagging substantially in both fluency and accuracy when processing Cantonese colloquialisms [12]. Consequently, effective implementation requires two forms of specialization: linguistic, as Cantonese colloquialisms are under-represented in multilingual corpora [12]; and pedagogical, as culturally specific moves like *Request for Choral Response* and *Evaluation* function as authority-reinforcing practices absent from Western taxonomies [19], and thus cannot be recovered from general-purpose training alone.

2.3 Teacher Expertise and Authority

Understanding the difference between novice and experienced teachers requires examining how discourse establishes authority in the classroom: authority structures determine which talk moves are functionally available, since students are less likely to engage with reasoning demands from a teacher who has not yet established interactional authority [24]. Tong et al. [13] note that in traditional classrooms, particularly within Chinese contexts, discourse often follows an "Initiation-Response-Feedback" (IRF) pattern — where the teacher initiates a question, a student responds, and the teacher evaluates the response — which novice teachers tend to rely on to maintain control and efficiency. Experienced teachers, however, demonstrate more adaptive responsiveness, leveraging not only IRF but also "Initiation-Response-Feedback-Response" (IRFR), where the student is invited to react to the teacher's feedback, enabling a more dialogic exchange, as well as iterative initiation and response chains. Kosel et al. [25] similarly found that novice teachers tend to mostly evaluate student engagement in terms of easily observable "surface cues" (e.g., students hand-raising, digressive gaze) coupled with some cues regarding students' factual knowledge. In contrast, experienced teachers integrate a broader spectrum of not only "surface cues" but also "deep cues" (e.g., students' motivation, attentiveness, confidence, quality of verbal contributions) when diagnosing student engagement and gauging instructional effectiveness.

These findings point to the importance of accounting for not just the presence of talk moves, but how teachers configure authority and engage students within specific cultural and classroom norms.

3. DATA

The dataset used in this study is derived from classroom recordings originally collected by Ni et al. [26]. We refer to this annotated corpus as CantoTalk, which comprises 7,518 teacher utterances (M = 48.19, SD = 57.62 characters) and 7,088 student utterances (M = 7.36, SD = 10.60 characters), drawn from 64 recorded and transcribed lessons across 32 upper-primary mathematics teachers in 16 Hong Kong schools, with each teacher providing 2 lessons [26]. These utterances were labeled using a ten-category scheme designed by Ng et al. [19], including: *Ask for Expression*, *Say More*, *Revoice*, *Press for Reasoning*, *Repeat and/or Add On*, *Agree/Disagree*, *Explain Others*, *Request for Choral Response*, *Evaluation*, *Other Moves* (see Appendix A for full definition and examples).

The scheme is built upon established discourse frameworks, including Talk Moves as Tools (TMT), Classroom Discourse Analyser (CDA), authority theory, and accountability principles in productive disciplinary engagement [27, 14, 28, 29, 30], with the addition of *Request for Choral Response* and *Evaluation* to reflect the structured teacher-led nature of Hong Kong mathematics classrooms and the use of collective participation to reinforce curricular knowledge [19]. The dataset was manually annotated by the researchers of the original study, achieving inter-coder agreement ranging from 0.7 to 1.0 [26]. Appendix Table A provides a complete list of talk move labels, and their descriptions.

Teacher metadata includes years of teaching experience ranging from 1 to 23 years (see Figure 1), grade level(s) taught, and school context. We categorized teachers into novice ($\leq$ 5 years, n = 8, 2110 utterances), intermediate (6–9 years, n = 7, 1922 utterances), and experienced ($\geq$ 10 years, n = 17, 3506 utterances) based on experience thresholds established in the teacher development literature. For our expertise detection analyses, we excluded intermediate teachers to maximize contrast between groups, yielding a final dataset of 5,616 utterances with a 62.4% experienced to 37.6% novice ratio. We note that years of experience serves as a coarse proxy for instructional expertise, as experience correlates with expertise only probabilistically and nonlinearly. We adopt this operationalization pragmatically, as direct measures of instructional quality were not available in the dataset, and interpret all expertise-related findings with this caveat in mind.

This dataset, which we refer to as CantoTalk, including all teacher and student utterances with talk-move annotations, contextual windows, and teacher metadata (anonymized experience levels, grade levels taught) is hosted on the GitHub repository.

4. METHODS

We adopt a two-stage approach. First, we fine-tune a set of open-weight LLMs from different research groups and parameter scales to identify the best-performing model for talk-move classification. Second, using the top model, we extract utterance-level embeddings to compare discourse patterns between novice and experienced teachers. Teachers with 6–9 years were excluded from embedding comparison analyses. Note that we initially explored few-shot prompting using off-the-shelf LLMs, such as Deepseek-R1 and Gemini-2.5-Pro to classify teacher talk moves, but excluded this method from our final analysis as it yielded suboptimal performance ($F_1 < 0.35$) and proved computationally prohibitive, requiring over one hour per transcript.

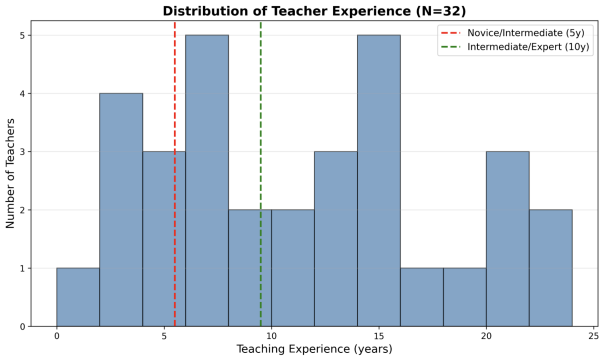

Figure 1: Distribution of teacher experience in years.

4.1 Fine-tuning

Training Data Construction. We perform a transcript-level split (70% train, 15% validation, and 15% test) to ensure that no utterance's immediate conversational context spans across splits, preventing direct context leakage between classroom sessions. Only teacher-labeled utterances are used. Because talk-move classification is strongly context-dependent, each training instance is constructed using a ±6-utterance sliding window containing six preceding turns, the focal utterance, and six following turns, formatted as `<speaker>: <content>` pairs to preserve turn-taking structure. Examples without a full context window are discarded.

Prompt Format. All models receive prompts in a standardized chat format: a Cantonese (i.e., Traditional Chinese) instruction, the full set of talk-move definitions (Cantonese + English), and the contextual dialogue grouped into preceding, current, and following blocks, followed by a directive to output all applicable talk-move categories as a comma-separated list. Chat templates for the chosen models differ only in formatting tokens, and the informational content in all of them is identical. Full prompts are presented in Appendix B.

Models and Training Procedure. We fine-tune five open-weight LLMs, including: `Qwen3-4B-Instruct-2507`, `Qwen3-8B`, `Llama-3.1-8B-Instruct`, `DeepSeek-V3.1`, and `GPT-OSS-20B` using Tinker API [31]. All models are trained using LoRA adapters (rank = 32) with a batch size of 8 and a maximum input length of 4096 tokens. Adam optimization was used with a 300-step linear warm-up. The model trained for up to 3000 steps with early stopping on validation loss. Training hyperparameters are listed in Appendix C. To address severe skew in talk-move frequencies, we apply inverse-frequency loss weighting, assigning each label i

weight $w_i = (1/f_i)^{0.5}$, normalized to have mean 1.0. For multi-label examples, we use the maximum weight among the associated labels. This weighting scheme amplifies the contribution of rare moves such as *Explain Others* and *Evaluation*.

Evaluation. Evaluation is conducted on held-out teacher utterances. Models generate label sequences using deterministic decoding (temperature = 0). Predictions are canonicalized by splitting on commas, trimming whitespace, and matching against the valid label inventory (see full details in Appendix D). We report: micro-F_1, micro precision, micro recall, macro-F_1 (averaged across labels) and per-label F_1 and support.

4.2 Embedding Analysis

Embedding Extraction. We analyze internal representations from our best-performing model to examine whether it encodes systematic differences between novice and experienced teachers. A well-known challenge in probing LMs is that pooling token embeddings over the entire prompt can entangle properties of the target utterance with artifacts from the surrounding instructions and definitions [32]. To avoid this, we extract target-only embeddings using character-level offset mappings.

For each example, we tokenize the full prompt (instruction + talk-move definitions + before/after context + target utterance) with `return_offsets_mapping=True`. We identify the character span of the target utterance using fixed delimiters (當前發言： ... 後文語境：) and select only the tokens whose offsets overlap this region. The embedding of the utterance is computed by mean-pooling the final-layer hidden states of these tokens. We use the last transformer layer because it typically captures task-specific refinements acquired during fine-tuning [33, 34]. This procedure yields a 4,096-dimensional embedding for all 7,538 utterances in the corpus. Extraction is performed via Colab Pro on an NVIDIA A100 using 8-bit quantization, enabling full processing in approximately 40 minutes.

Dimension Validation and Feature Selection. For expertise analyses, we retained only novice and experienced teachers, resulting in an embedding matrix of size $5{,}616 \times 4{,}096$. To identify dimensions associated with teacher expertise while avoiding overfitting, we divide them into a stratified 80/20 train-test split. On the training set, we conduct two-sample t-tests comparing experienced versus novice values across all 4,096 dimensions, applying a Bonferroni-corrected threshold ($\alpha = 1.22 \times 10^{-5}$). This yields 1,254 significant dimensions (30.6%). We re-test these dimensions on the held-out set at the same threshold; 348 dimensions (27.8%) remain significant. This two-stage procedure separates dimension validation from classifier training, enabling us to probe what the model encodes rather than just maximize classification accuracy, while guarding against dimensions that are spuriously significant in the training set alone. All downstream analyses use only these validated dimensions.

Expertise Classification. Overall, we train a linear probe implemented as an ℓ_2-regularized logistic regression classifier (`solver = lbfgs`, $C = 1.0$, `max_iter = 1000`) with balanced class weights to correct for class imbalance, using the validated 348-dimensional embeddings as features. All inputs are standardized (z-scored). We conduct 5-fold stratified cross-validation on the training set ($n = 4{,}492$) and evaluate the final model on the test set ($n = 1{,}124$). We report accuracy, balanced accuracy, precision, recall, and F_1. This analysis identifies whether the model's internal representations reliably encode distinctions between novice and experienced teachers. On a talk-move level, to test whether expertise is uniformly encoded across discourse practices, we repeat the probe analysis for each of the nine primary talk-move categories (excluding *Other Moves* as it contains varied moves). For each move, we subset the data to utterances containing that move, then train the same logistic regression classifier with 5-fold CV. We evaluate significance by one-sample t-tests against chance ($\mu = 0.5$) using balanced accuracy as the metric. This analysis identifies which pedagogical moves are most diagnostic of teacher expertise.

Clustering. To explore latent discourse patterns beyond binary expertise classification, we perform k-means clustering on the 348-dimensional validated embeddings. We compare solutions for k = 2–10 using silhouette [35], Calinski-Harabasz [36], and Davies-Bouldin [37] scores to choose k = 3. The association between cluster membership and expertise is tested with a χ^2 test, with structure visualized using UMAP [38] (`n_neighbors = 15`, `min_dist = 0.1`). To interpret the discovered clusters, we compute descriptive properties of each cluster such as proportion of experienced teachers, mean utterance length, and talk moves in it.

We further assess linguistic and pedagogical differences across clusters through several analyses. First, we compare utterance lengths using a one-way ANOVA and report η^2 effect sizes; post-hoc differences are examined with Bonferroni-corrected Mann-Whitney U tests. To evaluate whether cluster structure is driven by verbosity rather than discourse function, we train a shallow decision tree (`max_depth = 2`) to predict cluster membership using only utterance length, quantifying how much representational structure remains after accounting for this surface feature. Next, we conduct a qualitative comparison of experienced and novice utterances within each cluster to examine differences in execution quality beyond move frequency or length. These combined analyses triangulate the linguistic and pedagogical properties that define the discovered embedding clusters.

5. RESULTS

5.1 Fine-tuning

Table 1 reports test-set performance for all five fine-tuned LLMs. `Qwen3-8B` achieves the strongest results, obtaining the highest micro-F_1 (0.8145) and macro-F_1 (0.7687). Although `Qwen3-8B` outperforms `DeepSeek-V3.1` narrowly despite comparable size, we select `Qwen3-8B` for subsequent embedding analyses due to: (a) its slightly superior accuracy and (b) lower computational cost owing to its smaller size. `Qwen3-4B-Instruct-2507` and `Llama-3.1-8B-Instruct` form a middle tier, outperforming `GPT-OSS-20B` but trailing the top models by several points. `GPT-OSS-20B` underperforms across all metrics despite its larger parameter count.

Per-Label Performance. Table 2 shows per-label F_1 scores for `Qwen3-8B`. The model performs best on moves with strong contextual cues and larger support, such as *Say More*, *Revoice*,

Table 1: Performance of FT models on talk-move classification

Rank	Model	F_1^{μ}	F_1^{M}	Prec.	Rec.
1	Qwen3-8B	**0.8145**	**0.7687**	0.8048	**0.8245**
2	DeepSeek-V3.1	0.8134	0.7401	0.8077	0.8193
3	Qwen3-4B-Instruct-2507	0.7959	0.7221	**0.8097**	0.7827
4	Llama-3.1-8B-Instruct	0.7927	0.7432	0.8020	0.7837
5	GPT-OSS-20B	0.5683	0.4164	0.7624	0.4529

Table 2: Per-label F_1 scores for talk-move classification by Qwen3-8B. Support indicates the number of test examples

Talk Move Label	F_1	Support
表達自己 (Ask for Expression)	0.823	620
對自己發言作補充 (Say More)	0.858	219
複述 (Revoice)	0.860	625
解釋自己 (Press for Reasoning)	0.742	353
重述及/或補充他人發言 (Repeat and/or Add On)	0.739	253
同意/否定 (Agree/Disagree)	0.788	90
解釋他人 (Explain Others)	0.545	5
集體作答 (Request for Choral Response)	0.884	444
評價 (Evaluation)	0.678	55
其他 (Other Moves)	0.768	396
Macro Average	**0.7687**	**3,060**

and *Request for Choral Response*, each reaching F_1 scores above 0.85. In contrast, the lowest F_1 scores correspond to categories with extremely sparse support. Noticeably, *Explain Others* has $F_1 = 0.545$ (5 instances in the test set), and *Evaluation* has $F_1 = 0.678$ (55 instances in the test set), showing reduced performance. These patterns reflect the long-tail distribution of pedagogical moves as well as their inherent semantic difficulty, which can make some moves harder to distinguish than others.

5.2 Embedding Analysis

We present results in two phases: first, expertise classification demonstrates that teacher expertise is linearly recoverable from the model's internal representations; second, cluster analysis reveals the pedagogical structure underlying these representations.

Expertise Classification. For overall set, a linear probe trained on 348 validated dimensions achieved 0.733 ± 0.016 balanced accuracy (5-fold CV) and 0.790 on held-out test data, significantly above chance (0.50). The model distinguished both experienced ($F_1 = 0.83$) and novice ($F_1 = 0.74$) discourse effectively, indicating that performance was not driven by class imbalance exploitation. These results establish that teacher expertise is linearly encoded within the model's representational geometry, a key finding suggesting that teacher experience leaves distinctive signatures in fine-tuned embeddings. We repeated probe analysis separately for each talk-move category to determine whether expertise signals vary by pedagogical function. All categories showed above-chance balanced accuracy (mean = 0.754; all > 0.70), but with meaningful variation. *Revoice* (0.783), which involves deeper interpretation of student thinking to paraphrase contributions, had the highest differentiation. In contrast, *Press for Reasoning* had the weakest differentiation (0.723). This pattern suggests the model captures expertise-sensitive nuances within all talk moves.

Cluster Discovery. K-means clustering on the 348 dimensions revealed three stable discourse clusters. A χ^2 test showed strong association with expertise ($\chi^2 = 256.7$, $p < 10^{-56}$), indicating that representational regions in the embedding space differ by teacher expertise. UMAP projections (Figure 2) reveal that the three clusters occupy largely distinct regions of the embedding space, with overlap concentrated at cluster boundaries, consistent with the natural co-occurrence of talk moves across discourse styles. To ensure that clusters truly reflected pedagogical function rather than surface-level verbosity, we conducted three validation tests. First, while clusters differed significantly in average length ($p < 10^{-145}$, $\eta^2 = 0.112$), a decision tree using only length as a predictor achieved just 65% accuracy in cluster assignment, misclassifying 35% of utterances and showing particular confusion between Clusters 2 and 3 despite their 38-character mean length difference (29.1 vs. 67.3 characters). Second, within each cluster, experienced and novice utterances showed no significant length differences (Cluster 1: $p = 0.23$; Cluster 2: $p = 0.67$; Cluster 3: $p = 0.18$), yet maintained distinct embedding patterns as evidenced by the linear probe's within-cluster classification performance. Third, experienced-novice differences in pedagogical moves persisted within each cluster even when length was held constant. These findings collectively demonstrate that embedding geometry captures deeper distinctions in discourse rather than mere stylistic artifacts.

Cluster Characterization. Through qualitative analysis of utterances within clusters, we found they capture qualitatively different instructional styles. Critically, these clusters are defined not just by which moves are used, but by *how* they are executed. The embedding space appears to align with well-established pedagogical constructs, including authority and accountability roles in classroom discourse, cognitive-demand frameworks, and distinctions between monologic, authoritative, and dialogic instructional patterns.

Cluster 1: Procedural Management Integrated with Conceptual Scaffolding ($n = 118$; experienced 78%; $M = 97.8$ chars). This cluster represents long, multi-functional turns where teachers intertwine classroom management and procedural explanations with mathematical content. We note that Cluster 1 is notably smaller than the other clusters and appears less stable in the UMAP projection, which likely reflects the relative rarity of long multi-functional turns in the corpus rather than a poorly-defined cluster; internal validation metrics (silhouette, Calinski-Harabasz, Davies-Bouldin) consistently supported $k = 3$ over alternative solutions. Experienced teachers frequently turn routine instructions into opportunities for conceptual elaboration. For instance, while managing a hands-on activity in which students were cutting paper shapes to explore symmetry, one experienced teacher turned a management moment into a mathematical one: *"No, this is not a symmetrical figure... don't be like Student 4 who cut it—once it's cut, it's no longer a parallelogram."* Rather than separating the procedural warning from the mathematical concept, the teacher delivers both within a single turn. Even in explaining parti-

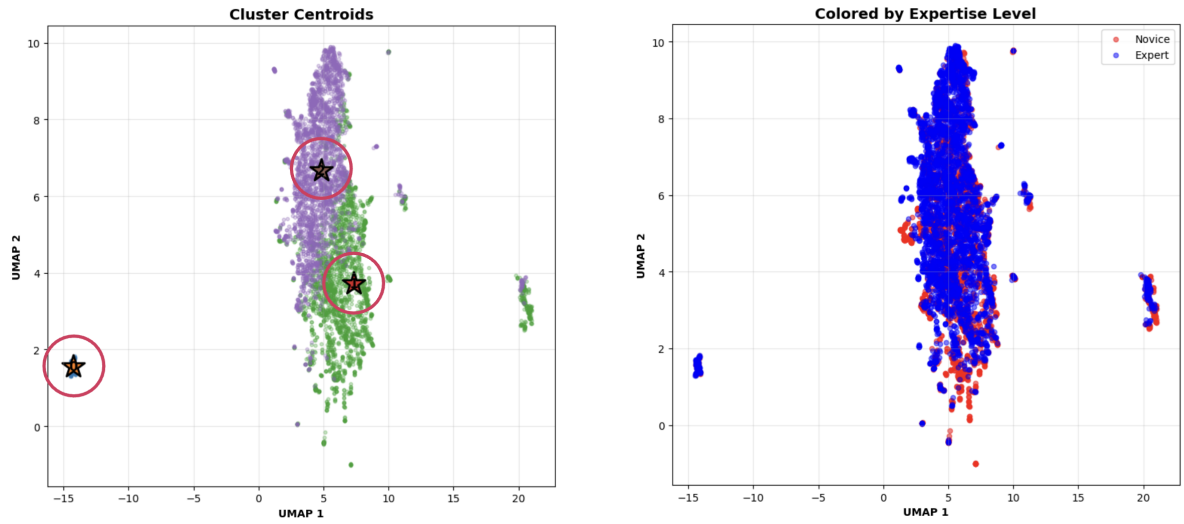

Figure 2: UMAP projection of utterance embeddings showing three discourse clusters ($k = 3$). Colors indicate cluster membership; clusters differ in expertise concentration (78%, 51%, 72%).

tioning procedures, experts weave in conceptual reframing: *"If I split this into two equal parts, this is no longer just 'one'—what is it then?"*, inviting reconceptualization of unit size. By contrast, novice teachers issue procedural instructions as distinct steps separate from conceptual elaboration: *"Mixed numbers... I'm just going to transform one of the parts... how many equal parts should I cut it into?"* Cluster 1 represents a mode where instructional guidance and content delivery occur simultaneously, with experts more likely to weave these functions within single turns while novices sequentialize them.

Cluster 2: Rapid-Fire Elicitation and Checking (n = 2,598; experienced 51%; M = 29.1 chars). This cluster contains short prompts dominated by *Ask for Expression* and *Request for Choral Response*, aligning with IRE (Initiate-Response-Evaluate) patterns observed in the regional context [26]. Both experienced and novice groups use this mode frequently, but for different functions. Experts often reinforce structural relationships: *"The denominators are the same—meaning the addend and augend denominators are both... ?"* While novices also employ short questioning sequences to elicit mathematical reasoning (*"Three over six. Why is it three over six?"*), they more often emphasize reinforcing recall or ensuring procedural correctness. Prompts like *"Say it again completely"* or *"Read it carefully. Read it carefully"* focus on the form or mechanics of participation [25]. Questions such as *"That's right—what is the question actually asking you to use?"* verify task comprehension. While both groups leverage rapid-fire turns to probe concepts, novices pay more attention to ensuring students follow procedures correctly.

Cluster 3: Dialogic Facilitation and Inquiry (n = 2,900; experienced 72%; M = 67.3 chars). This cluster represents medium-length turns where teachers orchestrate mathematical discourse by positioning students as sources of knowledge and facilitating peer reasoning. Teachers frequently employ *Press for Reasoning*, *Agree/Disagree*, and *Repeat and/or Add On* moves that redistribute mathematical authority across the classroom. While both groups encourage student expression and peer explanation, discernible differences emerge in their approaches. Experts probe underlying mathematical logic (*"Why did you do it so fast? Tell me, when you were measuring, how did you measure so quickly?"*), while novices often couple explanation-seeking with classroom management (*"Student 21 stand up, what is he saying?"*, *"When classmates are talking, you need to listen carefully"*). Experts frame inquiry to elevate student status (*"What's the secret to making you count so fast? Let me interview the fastest group first"*) and persist in understanding reasoning (*"I don't really understand what Student 22 is saying. Can you say it again?"*), while some novices redirect when students struggle (*"I don't know what you're talking about!"*). In short, Cluster 3 reflects attempts to democratize mathematical authority, however, teachers may vary in their approaches to sustain productive student-centered dialogue.

6. DISCUSSION

This study shows that fine-tuned LLMs can reliably classify talk moves in Cantonese mathematics classrooms and that their internal representations encode systematic differences in teacher expertise. A key contribution of our approach is its ability to surface pedagogical structure that is difficult to specify in advance. Traditional linguistic analyses require researchers to pre-select features such as question types, turn lengths, or lexical indicators, and may therefore miss diffuse or non-obvious cues to expertise. Embedding analysis instead reveals the latent distinctions the model uses to separate discourse from experienced and novice teachers, capturing how ideas are composed, how teachers respond to unfolding interaction, and how the same move is executed differently. Analysing representations thus provides a complementary lens on classroom instruction. We discuss three key contributions of our study and their implications:

Fine-Tuned Models Enable Cross-Cultural Discourse Analysis. Domain-specific fine-tuning yielded strong talk-move classification performance. Qwen3-8B achieved micro-F_1 = 0.814 and macro-F_1 = 0.768, with particularly high accuracy on culturally characteristic moves such as *Request for Choral Response* (F_1 = 0.884) and *Revoice* (F_1 = 0.860). These results indicate that open-weight models, when properly adapted, can learn context-specific pedagogical practices that differ from Western instructional norms, an encouraging outcome given that Cantonese remains severely under-represented in NLP resources. Notably, models larger than 8B parameters did not yield meaningful improvements, suggesting that performance gains came primarily from adaptation rather than scale alone; this is further supported by our few-shot prompting baseline using off-the-shelf LLMs, which yielded $F_1 < 0.35$ without fine-tuning, compared to 0.81 after adaptation. We did not evaluate larger multilingual LLMs such as Qwen3-235B due to computational constraints. Nonetheless, the underperformance of `GPT-OSS-20B` relative to smaller fine-tuned models hints that scale alone may not fully compensate for domain adaptation in low-resource pedagogical settings. Evaluating larger multilingual models as zero-shot baselines remains an important direction for future work.

LLM Embeddings Encode Expertise Beyond Surface Features. Before interpreting these results, we note that our operationalization of expertise as years of experience is a known limitation. Experience correlates with instructional quality only probabilistically, and the relationship is nonlinear: some novices exhibit expert-like discourse and vice versa. The patterns we report should therefore be understood as differences between less and more experienced teachers, not as validated markers of instructional quality per se. Embedding analysis showed that teacher expertise is linearly recoverable from the model's internal representations (balanced accuracy = 0.790 on held-out data). Expertise signals were strongest for moves involving deeper reasoning, such as *Revoice* (0.783), and weaker for procedurally simpler moves. K-means clustering revealed three coherent discourse styles: (1) procedural management and scaffolding (78% expert), (2) rapid elicitation (51% expert), and (3) dialogic facilitation (72% expert). Although lengths of utterances within the clusters differed, our results show that length is not driving the structure, and deeper pedagogical signals underlie the observed geometry. Qualitative inspection supports this interpretation. Experts and novices often use similar structures, but experts integrate conceptual connections. In Cluster 2, for instance, experts extend student thinking ("The denominators are the same…"), whereas novices default to repetitive prompting ("Say it again completely"). This computational pattern mirrors educational research showing that experienced teachers exhibit stronger adaptive responsiveness and conceptual coherence.

Implications for Automated Teacher Feedback. The linear separability of teacher expertise in embedding space provides a foundation for formative feedback tools that move beyond counting talk-move frequencies. Rather than treating discourse categories as checklists, automated systems could surface differences in how similar moves are executed, identify low-quality realizations of otherwise appropriate moves, highlight overreliance on novice-dominant discourse patterns, or suggest examples from experienced-teacher clusters as models for improvement. For instance, our Cluster 2 analysis shows that both novice and experienced teachers frequently use rapid questioning, but differ in pedagogical function: novices tend to emphasize procedural compliance, while experienced teachers use similar moves to surface underlying mathematical structure. This illustrates that the instructional value of a talk move lies in its realization, not its label.

These differences suggest concrete opportunities for targeted coaching in teacher education. Contrasting examples drawn from our discourse clusters could help novice teachers recognize how small shifts in phrasing, sequencing, or follow-up change the function of a move. Rather than generic advice such as "ask more open-ended questions," professional development can focus on specific discourse strategies, such as sustaining student reasoning through clarification and revoicing when confusion arises. Grounding coaching in authentic classroom utterances may make otherwise tacit aspects of instructional expertise more visible and learnable.

At the classroom level, these insights point toward personalized, low-stakes feedback tools that support reflection over time. Session-level summaries showing how a teacher's utterances distribute across discourse clusters, paired with illustrative examples from experienced-teacher patterns, could help teachers identify missed opportunities for conceptual scaffolding or overreliance on rapid elicitation. When framed as developmental reference points rather than evaluative benchmarks, such feedback can complement self-reflection and support growth across a semester.

Finally, by making explicit the discourse norms embedded in Hong Kong mathematics classrooms, this work opens space for cross-context professional dialogue rather than prescriptive standardization. Educators in different settings can use these representations to examine their own assumptions about participation, questioning, and authority, such as the role of choral response in collective sense-making. When carefully contextualized, automated discourse feedback can support reflection and learning across instructional contexts.

7. LIMITATIONS

We acknowledge several limitations of this study. First, our embedding analysis was conducted exclusively on Qwen3-8B. While this choice was methodologically justified, it means that our findings may be model dependent. Models with different capacities or architectures may encode talk-move patterns in different ways, which limits the generalizability of our results across LLM families. Additionally, our embedding extraction approach relies on delimiter-based span detection and token-to-character offset mapping, which assumes consistent tokenizer behavior and accurate recovery of span boundaries. In practice, any misalignment, particularly in non-space-delimited languages, may result in partial contamination from surrounding context. Disentangling whether low F1 on rare categories such as *Explain Others* and *Evaluation* reflects data sparsity or intrinsic semantic difficulty also remains an open question, and controlled subsampling experiments matching support across labels would be a valuable direction for future work.

Second, this study operationalizes instructional expertise as years of teaching experience, a coarse proxy whose limitations are discussed in Sections 3 and 6. A related concern involves the effective sample size of the expertise analyses: although CantoTalk contains over 7,000 utterances, these come from only 32 teachers, with expertise analyses contrasting just 25 teachers after excluding the intermediate group. The observed embedding-based distinctions may therefore partly reflect teacher-specific idiosyncrasies rather than robust markers of expertise; replication with a larger and more diverse teacher sample remains an important next step. Additionally, intermediate teachers were excluded from classification analyses to maximize contrast between novices and experts; whether embeddings order all three experience levels approximately linearly is also an open question for future work.

Finally, the teacher utterances analyzed in this work come from a narrow instructional domain, namely mathematics classrooms in Hong Kong. As a result, the discourse patterns that the model learns to associate with expertise may reflect local or context-specific norms rather than generalizable features of high-quality instruction. Findings may not generalize to other subjects, cultures, or languages. Additionally, although we identify embedding clusters that correlate with differences in teaching experience or talk-move styles, we do not establish a direct connection to student learning outcomes. While our results show that LLM embeddings encode systematic differences between novice and experienced teachers in ways that align with theoretical constructs from the teacher development literature, we do not evaluate whether identified expertise patterns actually support student learning, engagement, or equitable participation. It is possible that some of the discourse patterns we label as "expert" reflect institutional norms, curriculum alignment, or classroom management strategies rather than practices that promote deeper learning. For example, frequent use of *Request for Choral Response* may create an appearance of engagement without supporting individual sense-making. Future work should incorporate behavioral measures or expert-based validation and extend the analysis to a wider range of instructional contexts.

8. ETHICAL IMPLICATIONS

Some important ethical implications of this research include:

Privacy and Consent. This work relies on classroom discourse data, which raises important privacy concerns. Although all transcripts were anonymized and stripped of identifying information, teaching styles and discourse patterns may still be recognizable to colleagues within the same context. We mitigated this risk through anonymized analysis, and any future deployment of similar models should adopt strict data governance practices and explicitly prohibit individual teacher identification.

Potential Benefits. This work has several positive implications for education research and practice. Automated analysis of talk moves could support scalable, low-cost professional development by enabling teachers to reflect on their discourse patterns over time. By demonstrating that LLMs can be adapted to non-Western, non-English classroom contexts, this study also contributes toward more inclusive educational technologies and reduces reliance on English-centric tools. For researchers, automated classification substantially lowers the cost of large-scale discourse analysis, opening opportunities for longitudinal and cross-context studies that are infeasible with manual coding.

Risks and Misuse. At the same time, this work carries meaningful risks. Discourse patterns should not be treated as a complete or universal measure of teaching quality, yet there is a risk that such tools could be misused for high-stakes evaluation. Teaching quality is multidimensional and context-dependent, and many critical aspects of instruction are not captured in transcripts alone. There is also a risk of reinforcing narrow pedagogical norms learned from specific frameworks and contexts, which may not generalize or align with local instructional values. Additionally, because we do not validate our findings against student learning outcomes, identified expertise patterns may reflect institutional norms or surface-level engagement rather than deep learning.

Fairness and Generalization. Our use of teaching experience as a proxy for expertise introduces fairness concerns, as experience does not always align with instructional quality. The model may misclassify effective novice teachers or disadvantage teachers working in more challenging contexts. Furthermore, because the data come from Cantonese-medium mathematics classrooms in Hong Kong, the learned patterns reflect local discourse norms and linguistic features that may not transfer to other subjects, cultures, or languages.

Responsible Use. We emphasize that models like ours should be used only for low-stakes, teacher-initiated reflection, not for administrative evaluation or accountability. Future work should validate discourse signals against student outcomes, involve educators as co-designers in new contexts, and treat automated analysis as a complement to, rather than a replacement for, human professional judgment.

9. REFERENCES

[1] C. Howe, S. Hennessy, N. Mercer, M. Vrikki, and L. Wheatley. Teacher–student dialogue during classroom teaching: Does it really impact on student outcomes? *Journal of the Learning Sciences*, 28(4-5):462–512, 2019.

[2] D. Wagner and B. Herbel-Eisenmann. Identifying authority structures in mathematics classroom discourse: A case of a teacher's early experience in a new context. *ZDM*, 46:871–882, 2014.

[3] Daniel Lee and Jamie Poskin. Teachfx app (student–teacher talk measurement tool). Math Practical Measurement, WestEd, 2025. Accessed: 2025-11-06; URL: https://mpm.wested.org/measure/teachfx-app-student-teacher-talk-measurement-tool/.

[4] Bryant Jensen, Guadalupe Valdés, and Ronald Gallimore. Teachers Learning to Implement Equitable Classroom Talk. *Educational Researcher*, 50(8):546–556, November 2021. Publisher: American Educational Research Association.

[5] Y. Ni, D. Zhou, X. Li, and Q. Li. Relations of instructional tasks to teacher–student discourse in mathematics classrooms of chinese primary schools. *Cognition and Instruction*, 32(1):2–43, 2014.

[6] Klara Sedova, Martin Sedlacek, Zuzana Salamounova, Tomas Lintner, Roman Svaricek, Jakub Vlcek, Karolina Malikova, and Ivo Rozmahel. Let them all talk: equitable participation in classroom dialogue as a result of an intervention programme. *Language and Education*, 39(6):1471–1489, November 2025. Publisher: Routledge _eprint: https://doi.org/10.1080/09500782.2025.2454637.

[7] A. Suresh, T. Sumner, J. Jacobs, B. Foland, and W. Ward. Automating analysis and feedback to improve mathematics teachers' classroom discourse. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9721–9728, 2019.

[8] D. Wang, D. Shan, Y. Zheng, K. Guo, G. Chen, and Y. Lu. Can chatgpt detect student talk moves in classroom discourse? a preliminary comparison with bert. In *Proceedings of the 16th International Conference on Educational Data Mining*, pages 515–519, 2023.

[9] Baptiste Moreau-Pernet, Yu Tian, Sandra Sawaya, Peter Foltz, Jie Cao, Brent Milne, and Thomas Christie. Classifying Tutor Discursive Moves at Scale in Mathematics Classrooms with Large Language Models. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, pages 361–365, New York, NY, USA, July 2024. Association for Computing Machinery.

[10] J. Cao, A. Suresh, J. Jacobs, C. Clevenger, A. Howard, C. Brown, and J. H. Martin. Enhancing talk moves analysis in mathematics tutoring through classroom teaching discourse. *arXiv preprint arXiv:2412.13395*, 2024.

[11] Lihua Xu and David Clarke. Speaking or not speaking as a cultural practice: analysis of mathematics classroom discourse in Shanghai, Seoul, and Melbourne. *Educational Studies in Mathematics*, 102(1):127–146, June 2019. Publisher: Springer.

[12] Chengxuan Xia, Qianye Wu, Hongbin Guan, Sixuan Tian, Yilun Hao, and Xiaoyu Wu. Evaluating Modern Large Language Models on Low-Resource and Morphologically Rich Languages:A Cross-Lingual Benchmark Across Cantonese, Japanese, and Turkish, November 2025. arXiv:2511.10664 [cs].

[13] Zhiyue Tong, Fengcun An, and Yanji Cui. Exploring teacher discourse patterns: Comparative insights from novice and expert teachers in junior high school EFL contexts. *Heliyon*, 10(16):e36435, August 2024.

[14] S. Michaels and C. O'Connor. Conceptualizing talk moves as tools: Professional development approaches for academically productive discussion. In *Socializing Intelligence Through Talk and Dialogue*, pages 347–362. 2015.

[15] C. Howe and M. Abedin. Classroom dialogue: A systematic review across four decades of research. *Cambridge Journal of Education*, 43(3):325–356, 2013.

[16] Sterling Alic, Dorottya Demszky, Zid Mancenido, Jing Liu, Heather Hill, and Dan Jurafsky. Computationally Identifying Funneling and Focusing Questions in Classroom Discourse, July 2022. arXiv:2208.04715 [cs].

[17] Dorottya Demszky, Jing Liu, Heather C. Hill, Dan Jurafsky, and Chris Piech. Can Automated Feedback Improve Teachers' Uptake of Student Ideas? Evidence From a Randomized Controlled Trial in a Large-Scale Online Course. *Educational Evaluation and Policy Analysis*, 46(3):483–505, September 2024. Publisher: American Educational Research Association.

[18] David Clarke, Li Hua Xu, and May Ee Vivien Wan. Students Speaking Mathematics: Practices and Consequences for Mathematics Classrooms in Different Countries. In *Student Voice in Mathematics Classrooms around the World*. Brill, January 2013. Section: Student Voice in Mathematics Classrooms around the World.

[19] Oi-Lam Ng, Yujing Ni, Lian Shi, Gaowei Chen, and Zhihao Cui. Designing and Validating a Coding Scheme for Analysis of Teacher Discourse Behaviours in Mathematics Classrooms. *Journal of Education for Teaching*, 47(3):337–352, May 2021. Publisher: Routledge _eprint: https://doi.org/10.1080/02607476.2021.1896340.

[20] Yujing Ni, Lian Shi, Alan Cheung, Gaowei Chen, Oi-Lam Ng, and Jinfa Cai. Implementation and efficacy of a teacher intervention in dialogic mathematics classroom discourse in Hong Kong primary schools. *International Journal of Educational Research*, 107:101758, 2021.

[21] A. Suresh, J. Jacobs, V. Lai, C. Tan, W. Ward, J. H. Martin, and T. Sumner. Using transformers to provide teachers with personalized feedback on their classroom discourse: The talkmoves application. *arXiv preprint arXiv:2105.07949*, 2021.

[22] Juan Pava, Caroline Meinhardt, Haifa Badi Uz Zaman, Toni Friedman, Sang T. Truong, Daniel Zhang, Vukosi Marivate, and Sanmi Koyejo. Mind the (Language) Gap: Mapping the Challenges of LLM Development in Low-Resource Language Contexts, 2025.

[23] Partha Pakray, Alexander Gelbukh, and Sivaji Bandyopadhyay. Natural language processing applications for low-resource languages. *Natural Language Processing*, 31(2):183–197, March 2025.

[24] Oi-Lam Ng, Wing Kin Cheng, Yujing Ni, and Lian Shi. How linguistic features and patterns of discourse moves influence authority structures in the mathematics classroom. *Journal of Mathematics Teacher Education*, 24(6):587–612, July 2020. Publisher: Springer.

[25] Christian Kosel, Elisabeth Bauer, and Tina Seidel. Where experience makes a difference: teachers' judgment accuracy and diagnostic reasoning regarding student learning characteristics. *Frontiers in Psychology*, 15:1278472, March 2024.

[26] Y. Ni, L. Shi, A. Cheung, G. Chen, O. L. Ng, and J. Cai. Implementation and efficacy of a teacher intervention in dialogic mathematics classroom discourse in hong kong primary schools. *International Journal of Educational Research*, 107:101758, 2021.

[27] Gaowei Chen, Sherice N. Clarke, and Lauren B. Resnick. Classroom Discourse Analyzer (CDA): A Discourse Analytic Tool for Teachers. *Technology, Instruction, Cognition & Learning*, 10(2):85–105, April 2015.

[28] Randi A. Engle and Faith R. Conant. Guiding principles for fostering productive disciplinary

engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and Instruction*, 20(4):399–483, 2002.

[29] M. Stein, J. Remillard, and S. Smith. How curriculum influences student learning. In *Second handbook of research on mathematics teaching and learning*, volume 1, pages 319–370. 2007.

[30] E. Yackel and P. Cobb. Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education*, 27(4):458–477, 1996.

[31] Thinking-Machines-Lab. Tinker: a training api for large-scale model fine-tuning. https://thinkingmachines.ai/tinker/, 2025. Accessed: 2025-11-25.

[32] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting, 2024.

[33] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, July 2019. Association for Computational Linguistics.

[34] Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In Afra Alishahi, Yonatan Belinkov, Grzegorz Chrupała, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad, editors, *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online, November 2020. Association for Computational Linguistics.

[35] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987.

[36] T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974.

[37] David L. Davies and Donald W. Bouldin. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979.

[38] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2020.

[39] Yujing Ni, Gloria Ho, Jinfa Cai, Alan Cheung, Gaowei Chen, and Oi-Lam Ng. Research protocol: Teacher interventions aimed at engaging students in dialogic mathematics classroom discourse. *International Journal of Educational Research*, 86:23–35, January 2017.

APPENDIX

A. TALK MOVES FRAMEWORK

Teacher talk-move labels, their definition, function and examples, presented in Appendix Tables A. We would like to note that teacher talk move definitions, functions, and examples are synthesized from [39], [24] and [19].

Table A: Teacher Talk Moves with Definitions and Illustrative Examples

Talk Move	Definition	Examples
表達自己 (Ask for Expression)	明確邀請學生把想法說出來（不限內容深度）。 *Teacher invites a student to state their thinking or idea (no reasoning required).*	"Could you tell me what features a quadrilateral has?" "If they are the same, this procedure is what we call...?"
對自己發言作補充 (Say More)	要求同一位學生把剛才的發言延伸或補充細節。 *Teacher asks the same student to elaborate on their own prior statement.*	"Talk a little bit more." "Could you Say More about the idea you've just expressed?" "Is there anything you want to add about your example?"
複述 (Revoice)	老師用自己的語言重述學生想法以確認理解或放大訊息。 *Teacher restates a student's idea in their own words to confirm understanding or clarify for others.*	"So, let me see if I've got your thinking right. Are you saying...?" "Student X told us a square should have four equal angles." [Student: "Three parts."] "Three parts—so which one is three-eighths now?"
解釋自己 (Press for Reasoning)	追問同一學生說明「為什麼／怎麼知道」。 *Teacher asks a student to explain the reasoning behind their answer.*	"Why do you think that?" "What's your evidence?" "Could you describe the procedure she used?"
重述及/或補充他人發言 (Repeat and/or Add On)	邀請學生重述同伴想法，或在其上加以補充。 *Teacher invites students to restate or add to another student's idea.*	"Could you provide an example to illustrate the definition he is giving?" "Anything wrong with that answer, Student X?"
同意/否定 (Agree/Disagree)	老師表態同意或否定學生想法。 *Teacher expresses agreement or disagreement with a student's idea.*	"Do you agree/disagree—and why?" "Do you accept his example as a quadrilateral figure?"
解釋他人 (Explain Others)	要求學生解釋「另一位」的推理或方法。 *Teacher asks a student to explain the reasoning of another student.*	"Why do you think he said that?" "Who can explain what she means when she says that?"
集體作答 (Request for Choral Response)	要求全班齊答。 *Teacher prompts the whole class to respond together.*	"Do we all agree with that?" "The first multiple of a number is... say it together!"
評價 (Evaluation)	老師給出正誤判斷、稱讚或等第式回饋。 *Teacher provides correctness judgment, praise, or evaluative feedback.*	"You're right." "Good idea." "Not exactly correct."
其他 (Other Moves)	不屬上述類別的管理、過渡、後勤或其他話語。 *Classroom management, procedural talk, transitions, etc.*	"Please do not talk to your neighbor." "Open your textbook to page 25 for homework." "Louder, I couldn't hear just now."

B. FULL PROMPT TEMPLATE

Each training example is formatted using the model's native chat template, with identical informational content across architectures. Below is the canonical prompt used for all models (Cantonese + English definitions).

B.1 User Message

分析以下香港高小數學課堂對話，識別當前發言的所有talk move類別。
Talk Move 類別定義：

- `{label_1}`: `{definition_1}`
- `{label_2}`: `{definition_2}`
- ...
- `{label_k}`: `{definition_k}`

前文語境：
`{speaker_before_1}`: `{utterance_before_1}`
`{speaker_before_2}`: `{utterance_before_2}`
`{speaker_before_3}`: `{utterance_before_3}`
`{speaker_before_4}`: `{utterance_before_4}`
`{speaker_before_5}`: `{utterance_before_5}`
`{speaker_before_6}`: `{utterance_before_6}`
當前發言：
`{current_speaker}`: `{current_utterance}`
後文語境：
`{speaker_after_1}`: `{utterance_after_1}`
`{speaker_after_2}`: `{utterance_after_2}`
`{speaker_after_3}`: `{utterance_after_3}`
`{speaker_after_4}`: `{utterance_after_4}`
`{speaker_after_5}`: `{utterance_after_5}`
`{speaker_after_6}`: `{utterance_after_6}`
任務：根據上述定義，識別當前發言的所有適用talk move類別（用逗號分隔多個類別）。

B.2 Assistant Message

Gold multi-label target:

`<label_1>, <label_2>, ...`

B.3 Model-Specific Renderers

- `Qwen3-4B-Instruct-2507, Qwen3-8B` → `qwen3_instruct`
- `Llama-3.1-8B-Instruct` → `llama3_instruct`
- `DeepSeek-V3.1` → `deepseek`
- `GPT-OSS-20B` → `gpt_oss_no_sysprompt`

Note: Renderers differ only in tokenization wrappers; content is identical.

C. TRAINING HYPERPARAMETERS

All models were fine-tuned using LoRA adapters with the following configuration:

Component	Setting
LoRA Configuration	rank = 32; α = model default; dropout = 0.0
Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.95$, ϵ = 1e–8)
Learning Rate	base LR = 2e–5; LoRA LR scaled via `get_lora_lr_over_full_finetune_lr()`
Scheduling	300-step linear warm-up; constant LR thereafter
Training Steps	max 3000 steps; early stopping after 5 validation checks
Validation	every 100 steps (100-sample subset)
Batch Size	8
Sequence Length	max 4096 tokens
Checkpointing	every 50 steps

Table A: Hyperparameter configuration for LoRA fine-tuning across all models.

D. IMPLEMENTATION AND EVALUATION DETAILS

Tokenization and Truncation. We use each model's native tokenizer (via Tinker's `get_tokenizer()`), truncating sequences to a maximum length of 4096 tokens. Truncation is applied only to the earliest portions of the preceding context, ensuring that the focal utterance and its immediate surroundings are always preserved.

Decoding. During evaluation, models generate predictions using greedy decoding (temperature = 0) with a 50-token limit and model-specific stop sequences. Outputs are stripped of special tokens before post-processing.

Label Canonicalization. Generated text is split on commas, normalized (whitespace removed), and matched to the valid talk-move inventory. Unknown strings are discarded. Multi-label predictions are converted into binary indicator vectors for evaluation.