Emergent Wisdom at BEA 2025 Shared Task: From Lexical Understanding to Reflective Reasoning for Pedagogical Ability Assessment

Raunak Jain[1*] **and Srinivasan Rengarajan**[1]
[1]Intuit
{raunak_jain1,srinivasan_rengarajan}@intuit.com

Abstract

For the BEA 2025 shared task on pedagogical ability assessment, we introduce ***LUCERA*** (Lexical Understanding for Cue Density–Based Escalation and Reflective Assessment), a rubric-grounded evaluation framework for systematically analyzing tutor responses across configurable pedagogical dimensions. The architecture comprises three core components: (1) a rubric-guided large language model (LLM) agent that performs *lexical and dialogic cue extraction* in a self-reflective, goal-driven manner; (2) a cue-complexity assessment and routing mechanism that sends high-confidence cases to a fine-tuned T5 classifier and escalates low-confidence or ambiguous cases to a reasoning-intensive LLM judge; and (3) an *LLM-as-a-judge* module that performs structured, multi-step reasoning: (i) generating a domain-grounded reference solution, (ii) identifying conceptual, procedural and cognitive gaps in student output, (iii) inferring the tutor's instructional intent, and (iv) applying the rubric to produce justification-backed classifications. Results show that this unique combination of LLM powered feature engineering, strategic routing and rubrics for grading, enables competitive performance without sacrificing interpretability and cost effectiveness.

1 Introduction

High-quality formative feedback is a cornerstone of effective learning: timely, specific guidance helps learners close knowledge gaps, consolidate correct mental models, and sustain motivation (Anderson et al., 1995; Hattie and Timperley, 2007). Yet providing rich feedback at scale remains difficult. The BEA-2025 Shared Task (Kochmar et al., 2025) tackles this challenge by pairing a learning-science–grounded evaluation taxonomy (Maurya et al., 2025) with MRBENCH (Maurya et al., 2025), a benchmark that fuses math-centric tutoring dialogues from MATHDIAL (Macina et al., 2023) and BRIDGE (Wang et al., 2024b). The competition assesses four pedagogically salient dimensions—*Mistake Identification* (MI), *Mistake Location* (ML), *Providing Guidance* (PG), and *Actionability* (ACT)—thereby offering a unified, standardised test-bed for measuring the pedagogical competence of AI tutors.

While we participate in the shared task, our goal extends beyond leader-board performance (see 7 for more details). We introduce ***LUCERA*** (Lexical Understanding for Cue density–based Escalation and Reflective Assessment), a novel hybrid architecture that unifies fast lexical heuristics, confidence-aware routing, and reasoning capabilities of large-language-models (LLMs). *LUCERA*, positioned as a general research contribution, demonstrates how an adaptive cascade can deliver interpretable, scalable, rubric-faithful evaluation—attributes that matter both inside and outside competition settings.

Existing approaches to pedagogical-quality assessment occupy two extremes. At one end, rule-based cue extractors offer transparency and speed but falter when feedback is implicit or domain-specific (Lehman et al., 2019; Wang et al., 2020; Wollny et al., 2021; Macina et al., 2023). At the other, rubric-grounded LLM judges achieve broad coverage yet impose high computational cost and, when used indiscriminately, act as opaque monoliths that are hard to audit (Liu et al., 2023b; Maurya et al., 2025; Tack et al., 2023). Bridging these extremes, stepwise chain-of-thought (CoT) verification recovers subtle pedagogical intent but further magnifies latency and cost (Daheim et al., 2024; Wang et al., 2024b; Jain, 2025).

LUCERA orchestrates these complementary paradigms in a three-stage pipeline. A lightweight lexical-cue extractor provides instant, interpretable signals; a complexity-aware router allocates re-

[*] Corresponding author: raunak_jain1@intuit.com

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 1108–1120
July 31 - August 1, 2025 ©2025 Association for Computational Linguistics

sponses to either a heuristic XGBoost scorer, a fine-tuned T5 classifier, or a reflective CoT judge; and a final rubric-aligned verdict is produced only in low confidence scenarios. All LLM-based tasks in this work were performed using Qwen/Qwen2-1.5B-Instruct (Yang et al., 2024). This design achieves a 2.4× throughput gain over blanket LLM judging on the BEA-2025 dev set while maintaining rubric fidelity. Beyond the task, we argue that *LUCERA* offers a principled template for scaling LLM-based pedagogical quality assessment wherever feedback quality, cost, and transparency must be balanced.

The remainder of this paper is organised as follows: Section 2 surveys prior work on pedagogical-quality assessment, LLM judges, verification pipelines, and intelligent routing; Section 3 describes *LUCERA's* architecture; Section 4 and Section 5 detail the feature extraction and classification components; Section 6 explains the reflective LLM judge; Section 7 reports empirical results; and Section 7 concludes with limitations and directions for future research.

2 Related Works

Surface-level cue extraction. Early work framed pedagogical quality as a pattern-recognition problem: if a tutor turn contains directive verbs (*try*, *consider*), contrastive discourse markers (*however*, *because*), or worked-example fragments, it likely advances learning (Lehman et al., 2019; Wang et al., 2020). Rule-based and linear classifiers built on these *lexical cues* offered millisecond latency and clear rationales, and they continue to power production intelligent tutoring systems (Bringula and Basa, 2018). Nevertheless, large corpus studies show cue sparsity and STEM-specific jargon severely degrade their recall and domain transferability (Wollny et al., 2021; Macina et al., 2023).

LLM-as-a-Judge and Confidence based Cascades. The arrival of GPT-4–class models sparked a shift to *rubric-grounded* prompting: an LLM reads a turn and scores each dimension directly (Liu et al., 2023b). Frameworks such as LLM-RUBRIC formalise this practice and report sizeable gains across open-ended tasks (Xia et al., 2025). Yet *unconditional* LLM judging inflates inference cost (Jung et al., 2025; Schuster et al., 2022), produces verbose rationales of uneven quality (Saito et al., 2023; Ohi et al., 2024; Wang et al., 2024a), and can hallucinate additional rubric criteria (Li et al., 2023). Recent selective evaluation frameworks provide provable guarantees of human agreement while maintaining high coverage (Jung et al., 2024), achieving better human alignment than monolithic LLM judges while being substantially more cost-effective. These findings strongly motivate the search for *selective* depth in LLM-based evaluation.

Stepwise verification and reflective reasoning. Recent studies introduce a verification stage in which an LLM first generates a reference solution and then aligns it with the learner's work before labelling (Daheim et al., 2024; Wang et al., 2024b). Such *chain-of-thought* (CoT) pipelines help identify correct pedagogical strategies by boosting understanding of student gaps (Jain, 2025). Complementary efforts build testbeds (e.g., TutorGym) and benchmarks that grade the fidelity of reasoning chains (Li et al., 2025; Jacovi et al., 2024). However, each additional reasoning step multiplies latency and cost, making blanket deployment impractical at classroom scale.

Intelligent routing and hybrid cascades. Outside education, researchers mitigate the cost–accuracy trade-off by cascading small and large models, deferring only hard instances. Contemporary confidence-tuned cascades (Xu and McAuley, 2022), cascade-aware training (Zhang et al., 2024), and calibrated ensemble policies (Wagner et al., 2024) achieve 1.5–3× speed-ups without loss of accuracy. *Educational* NLP, by contrast, has yet to embrace hybrid routing: state-of-the-art graders for assignments (Chiang et al., 2024), short-answer scoring (Chang and Ginter, 2024), and essay evaluation (Latif and Zhai, 2024; Jiang and Bosch, 2024) all deploy a single, monolithic LLM without confidence-based deferral. Bridging this gap remains an open opportunity for future assessment systems.

Summary and open gap. The literature thus presents three partially solved challenges—speed (cue extractors), depth (CoT verifiers), and transparency (rubric-grounded judging)—addressed in isolation. No prior system unifies them under a single rubric while allocating compute *proportionally* to instance difficulty. By integrating cue density, calibrated confidence, and stepwise verification into one adaptive cascade, *LUCERA* fills this gap and provides the first cost-aware, rubric-consistent pipeline for tutor-response evaluation.

3 System Overview

LUCERA is a three-stage pipeline designed to assess the pedagogical quality of tutor responses. The system processes tutor responses through the following components:

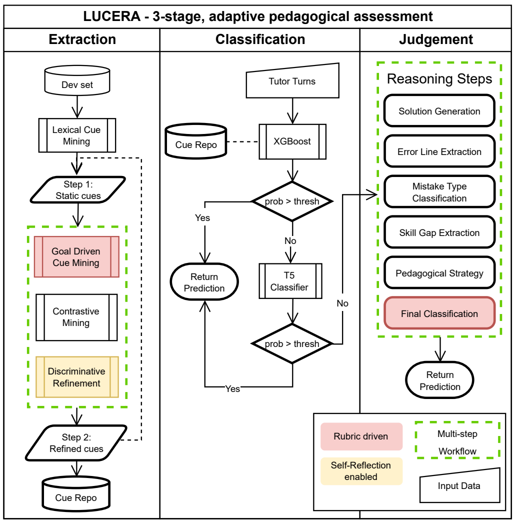

Figure 1: Overview of *LUCERA's* three-stage pipeline: (1) Rubric-Guided, goal driven lexical cue extractor identifies pedagogical features, (2) XGBoost and T5 based classifiers to solve for feature rich scenarios (3) Multi-step LLM judge for ambiguous and complex evaluation scenarios.

1. Rubric-Guided Lexical Cue **Extraction**: Identifies lexical and dialogue cues aligned with rubric dimensions via a self-reflective LLM agent. Maps directly to pedagogical criteria, maintaining interpretability and transparency throughout the feature extraction process.

2. XGBoost or Seq2Seq based **Classification**: Routes cases based on cue density and confidence. Deploys lightweight XGBoost or T5 models for efficient assessment of high-confidence cases with clear lexical patterns.

3. Step-wise LLM **Judgement**: Handles ambiguous or complex cases through multi-step reasoning. Generates reference solutions, identifies student knowledge gaps, and applies rubric criteria to deliver in-depth pedagogical analysis.

The rubric schema grounds both the cue extractor and the LLM judge, providing uniform evaluation criteria. After extracting cues from a tutor response, the system applies confidence-based routing: high-certainty cases proceed to a lightweight T5 classifier, while ambiguous ones are escalated to a reasoning-intensive LLM judge, conserving computation without sacrificing pedagogical depth.

4 Rubric-Guided Lexical Cue Extraction

We systematically developed a comprehensive feature taxonomy spanning multiple linguistic levels to classify tutor responses across pedagogical dimensions (MI, ML, PG, ACT), enabling fine-grained analysis of pedagogical signals in tutorial discourse.

4.1 Step 1: Linguistically-Grounded Feature Engineering

Lexical Cues

These cues provide shallow yet effective insights into semantic content and form of tutor responses:

- **Volumetric Features**: Basic text-level metrics including word, character, and sentence counts (Yang, 2024) serve as proxies for response depth. Low word counts may negatively correlate with PG and ACT due to insufficient detail.

- **Question Words**: Presence of interrogatives (e.g., "what," "why," "how") identified via counting pre-defined question words (Demszky et al., 2018), hypothesized to positively correlate with PG and ACT by signaling engagement and elaboration.

- **Feedback Words**: Terms like "correct," "mistake," or "however"—extracted using sentence-level sentiment or discourse tagging (Negi and Buitelaar, 2015)—expected to signal MI by indicating evaluative judgment. For example, "You're close, but remember" is a definite feedback phrase.

- **Directive Verbs**: Instructional verbs (e.g., "calculate," "explain," "solve") extracted using POS tagging and grammatical mood detection (Cohen et al., 2004), often implying actionability. For example, "Let's calculate that," "can you think of a way to calculate?" are all instructional phrases.

- **Hedging Words**: Words like "maybe," "might," or "could" introducing nuance or tentativeness, often associated with PG and ACT by softening directive tone (Deng et al., 2025). For example, "I think maybe you need one more step," "maybe we can use a hundreds chart or count up" demonstrate this.

- **Pronoun Ratios**: Ratios of second-person (you/your) to first-person (I/my) pronouns indicating student-centeredness or tutor-centeredness, relevant for PG and ACT (Qureshi and Strube, 2022). **Student-centeredness** refers to responses focusing on engaging students directly, guiding actions, or providing feedback, characterized by higher frequency of second-person pronouns. **Tutor-centeredness** reflects tutor's perspective, explanations, or insights, marked by higher frequency of first-person pronouns. Higher ratio of second-person to first-person pronouns suggests student-centric approach emphasizing direct instruction, while lower ratio indicates tutor-centric approach sharing tutor's reasoning. For example, "but remember your initial calculation," "but actually, you did add Kylie's 3 towels" are student-centered responses indicating PG.

4.1.1 Syntactic Complexity

Syntactic complexity is measured via average sentence length and subordinate clause density using dependency parsing (Crossley and McNamara, 2022). High complexity may hinder comprehension, potentially impacting PG and ACT despite informative content.

4.1.2 Pragmatic and Discourse Cues

These features capture pragmatic and contextual dimensions:

- **Discourse Markers**: Cues such as "however," "for example," or "but" indicating relationships between discourse units, helping differentiate between elaboration for PG and contradiction for MI (Dai and Huang, 2018).

- **Conversational Uptake**: Semantic alignment of tutor's response with preceding turns, computed using pre-trained dialogue embedding models (Demszky et al., 2021). High uptake suggests relevance and coherence, especially for PG.

- **Pedagogical Intent**: Pre-trained NLI models capturing latent pedagogical intent beyond surface features, by computing the 3-way softmax probabilities (entailment, contradiction, neutral) between tutor responses (premise) and intent descriptions (hypothesis) (Reimers and Gurevych, 2019). The entailment probability values [0-1] directly serve as feature weights, enabling nuanced quantification of pedagogical intents like supportiveness and elaboration.

- **Dialogue Act (DA) Classification**: Responses categorized into high-level DAs (e.g., 'Correction,' 'Hint,' 'Instruction') using pre-trained models (Noble and Maraev, 2021), serving as semantically rich signals—e.g., 'Correction' aligns with MI/ML while 'Instruction' relates to PG and ACT.

4.1.3 Feature Encoding Summary

Feature encoding employs a dual representation strategy: (1) numeric quantification (counts for volumetric features, pronoun ratios) and (2) TF-IDF vectorization (Salton and Buckley, 1988) with category-specific lexicons (feedback, directive, hedging words, discourse markers). Pedagogical intent features leverage NLI entailment probability values [0-1] as continuous feature weights. This complementary approach integrates statistical surface patterns with semantic-level analysis to capture both explicit and implicit pedagogical signals.

4.2 Step 2: LLM-Driven Discriminative Feature Refinement

Our approach employs a multi-stage pipeline that transforms initial lexical features into discriminative, contextually-validated pedagogical indicators. This process ensures alignment with assessment rubrics through progressive refinement, as illustrated in Figure 1 (*Extraction - Refined Cues*):

1. **Goal-Directed Feature Extraction**: LLM analyzes conversation data to identify discriminative features through an iterative, objective-oriented process guided by the initial seed features from 4. The extraction process leverages a T5-based (Raffel et al., 2020a) encoder-decoder framework fine-tuned on pedagogical conversations.

2. **Adversarial Refinement**: Features undergo validation against contradictory examples from other conversations, enabling the

LLM to eliminate spurious correlations and strengthen genuinely predictive indicators.

3. **Lexical Cue Repository Update**: Validated features are populated back into the cue repository, and steps 1-3 are repeated until no new features are found, ensuring a comprehensive and stable set of pedagogically discriminative features.

This methodology produces feature sets that transcend mere textual presence to capture pedagogical quality signals validated against both assessment criteria and challenging counterexamples.

4.3 Feature EDA Summary

We conducted systematic exploratory data analysis on development set tutor responses to quantify relationships between engineered features and pedagogical dimensions (MI, ML, PG, ACT) using Pearson correlation coefficients and distributional statistics, as detailed in Table 1. Key abbreviations include: **Vol** (Volumetric Features), **Ques** (Question Words), **Fdbk** (Feedback Words), **DV** (Directive Verbs), **Hed** (Hedging Words), **ProR** (Pronoun Ratios), **Y/N/TSE** (Yes/No/To Some Extent), **Read** (Readability score based on `flesch_reading_ease` (Flesch, 1948)), and **H/M/L** (Higher/Medium/Lower correlation trend across response categories).

4.3.1 Feature Influence based on Pearson Correlation

Table 1 summarizes key feature influences on pedagogical dimensions, where H/L/M indicates High-/Low/Medium influence for Yes/No/TSE classes respectively. Analysis of these patterns reveals several critical insights:

- **Volumetric Features** (H/L/L or H/L/H): Longer responses correlate with effective tutoring across dimensions (Chi et al., 2001; Ward et al., 2011), with verbosity particularly important for actionability where detailed guidance enables student progress (VanLehn, 2011).

- **Question Words** (variable patterns): Strong association with ML (H/L/H) shows questioning is essential for modeling learning; moderate impact on MI (M/L/H) reveals interrogatives' dual purpose in challenging misconceptions and guiding reflection (Graesser et al., 2010; VanLehn et al., 2006). TSE pattern (H) suggests questions create partial pedagogical value (Chen et al., 2011). Radar analysis (Figure 3) confirms Question Words heavily influence ACT while moderately affecting PG and MI across both TSE and "No" classifications.

- **Pronoun Ratios**: Reveals dimension-specific strategies—PG/MI benefit from tutor-centric language (L/H for "Yes"/"No") where expert explanation is valued; ACT/ML favor student-centric approaches (H/L) positioning students as active participants (Nystrand and Gamoran, 1997; Mercer and Littleton, 2009; Biber and Gray, 2006). N-gram analysis (Figure 2) shows distinctive phrases like "looks like you" and "remember that" strongly correlate with PG.

- **Feedback & Directive Words**: Inverse patterns between feedback (L/H/L) and directives (H/L/M) highlight tension between evaluation and instruction (Shute, 2008; Hattie and Timperley, 2007). Combined with hedging patterns, this suggests effective tutoring balances definitive guidance with tentative suggestion (Rowland, 2002; Mackiewicz and Thompson, 2010). Readability scores and Feedback markers most heavily influence MI classification as shown in Figure 3.

- **Discourse Context**: Discourse markers strongly influence MI (H/L/L) and ML (H/L/H) (Fraser, 1999; Sanders et al., 2000). Contrasting readability patterns (MI: M/L/H vs. others: M/H/L) suggest misconception identification benefits from accessible language while model learning sometimes requires complex formulations (McNamara et al., 2010; Crossley et al., 2017). Action-oriented phrases ("closer look," "look at") strongly correlate with ACT dimension (Figure 2). TSE class presents unique classification challenges with subtle linguistic markers and mixed signals—often providing information without prompting direct action.

These patterns reflect pedagogical trade-offs: correction versus guided discovery (Hmelo-Silver et al., 2006), authoritative versus collaborative stance (Scott et al., 2002), and comprehensive explanation versus concise instruction (Wittwer and Renkl, 2010). Differential patterns validate our

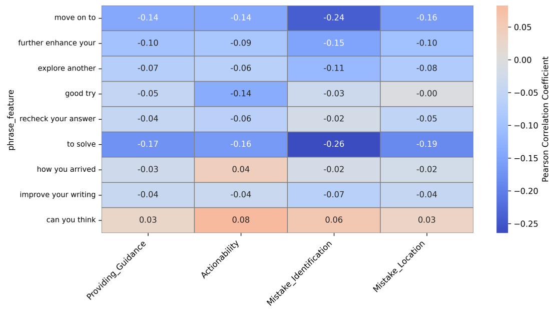

Figure 2: Word Correlation with Pedagogical Dimensions (Yes)

taxonomy's ability to capture distinct tutoring aspects, while shared patterns highlight fundamental qualities of effective pedagogical communication.

Feature	MI	ML	PG	ACT
Vol	H/L/L	H/L/H	H/L/H	H/L/H
Ques	M/L/H	H/L/H	H/L/M	H/L/M
Fdbk	M/L/L	L/H/L	L/H/L	L/H/L
DV	H/L/M	M/H/M	M/H/M	H/L/M
Hed	H/L/H	M/H/H	M/H/M	H/L/H
ProR	M/H/L	L/H/H	L/H/M	H/L/H
DM	H/L/L	H/L/H	M/L/M	M/L/M
Read	M/L/H	M/H/L	M/H/L	M/H/L

Table 1: Summary of Feature Influence Based on Correlations (H: High, L: Low, M: Medium) for classes Yes/No/TSE.

5 Model Cascade, Confidence-Based Routing, and Task Submission

Based on the extracted refined cues from Section 4, we first train an XGBoost model as a baseline to cover cases where lexical coverage is high. Once we identify lack of lexical coverage, we escalate the classification process to a T5 transformer architecture with a generative classification task instruction. We detail the baseline and T5 model architectures, training methodology etc in the following sections. The evaluation metrics are: exact macro F1 score (Ex. F1), exact accuracy (Ex. Acc), lenient macro F1 score (Len. F1), and lenient accuracy (Len. Acc).

5.1 Stage 1: XGBoost + Lexical Cues as Baseline

For a baseline, we train a multi-label (Yes/No/TSE) multi-class classification model using XGBoost (Chen and Guestrin, 2016) with a 70/30 train-val split. Hyperparameter tuning was performed using cross-validation, focusing on key parameters such as `max_depth`, `learning_rate`, `n_estimators`, `min_child_weight`, and `subsample`. Table 2 presents the performance on validation dataset of the best run (XGBoost is considerably better than all Yes (majority class for all labels) as a baseline).

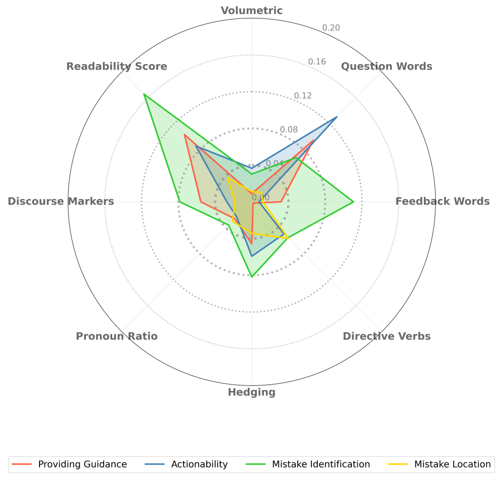

Figure 3: Normalized Combined Radar Chart by Feature Group(TSE)

Task	Ex.Acc	Ex.F1	Len.Acc	Len.F1
MI	0.71	0.64	0.81	0.73
ML	0.67	0.66	0.85	0.72
PG	0.68	0.66	0.84	0.71
ACT	0.75	0.71	0.81	0.73

Table 2: XGBoost Performance Metrics by Pedagogical Dimension

5.1.1 Feature Impact and Model Limitations:

While the XGBoost model performed well across different pedagogical dimensions, several limitations were identified for improvement in subsequent iterations:

- **Syntactic Complexity and Question-Related Features:** Complex syntactic structures, such as nested clauses or subordinate sentences, can confuse the model. For example, "While it seems correct, you might want to double-check the calculation" may be misclassified as *Yes* for MI due to ambiguous framing. Additionally, interrogative cues are essential for classifying ACT and PG, but rhetorical questions can mislead the model. For instance, "Do you think this is correct?" could be interpreted as actionable, despite expressing doubt.

- **Lexical Features and Semantic Nuance:** Lexical cues, such as keyword spotting, can lead to errors when words appear in unexpected contexts. For instance, "You're doing great! But remember" is encouraging but points to an approach to guide the student toward the correct answer.

- **Pronoun Usage and Intent Ambiguity:** Shifts between tutor-centric ("I/my") and student-centric ("you/your") language cause inconsistent classification. A statement like "You could explain it better" may be classified as a *Yes* for ACT, whereas a similar structure with "I" might be a strong *Yes* for PG. Detecting perspective shifts remains challenging.

5.2 Stage 2: Instruction-Based Seq2Seq Classification

Model Setup: We fine-tune T5-Base (approximately 250M parameters) for instruction-based classification across four pedagogical dimensions: MI, ML, PG, ACT. Each dimension is specified using a distinct prompt prefix. This approach leverages T5's encoder–decoder architecture, with a 512-token context and unified text-to-text pretraining, facilitating efficient and accurate classification (Raffel et al., 2020b; Qorib et al., 2024).

Prompt Template: The model generates a single-token prediction $y \in \{\texttt{yes}, \texttt{no}, \texttt{maybe}\}$. Each prompt follows this structure:

```
[PEDAGOGICAL_DIMENSION]
[LEXICAL_CUES] <lexical cues>
[TUTOR_TURNS] <concatenated tutor turns>
Output: {yes, no, maybe}
```

Example (Mistake Location):

```
[Providing Guidance]
[LEXICAL_CUES] let us use, what was
[TUTOR_TURNS] ok let us use the information
to help us what was her gross revenue this week?
Output: maybe
```

Loss Function and Evaluation Metrics: We minimize the single-token cross-entropy over our dataset $\mathcal{D}$:

$$\mathcal{L}(\theta) = -\sum_{(x,y)\in\mathcal{D}} \log p_\theta(y \mid x),$$

where y is the correct label and decoding is constrained to one step (greedy, $\text{max\_length} = 1$).

Training Details: We fine-tuned `T5-Base` via HuggingFace Transformers on an Apple M3 Mac (no GPU) using 70%/30% train/eval splits. Preprocessing involved excluding student turns and initial problem-introduction turns, concatenating remaining tutor turns, and truncating leftmost tokens when exceeding the 512-token limit. Training used batch size 8, AdamW optimizer (weight decay 0.01, LR = 3×10^{-5} with 10% steps linear warmup), and ran for 5 epochs with early stopping (patience=2, dropout=0.1). Decoding was performed greedily with evaluation via dev loss.

Why T5-Base? We selected T5-Base for its empirical and architectural advantages: superior classification performance, with Flan-T5 variants consistently outperforming decoder-only models on GLUE, SuperGLUE, and word-sense disambiguation tasks while matching GPT-3.5 in few-shot settings (Papadopoulos et al., 2024; Liu et al., 2023a); multi-task pre-training on diverse tasks equipping it with transferable NLP skills that generalize without separate heads (Raffel et al., 2020b; Liu et al., 2023a); hardware efficiency at 250M parameters, comfortably running on modest hardware with 512-token input handling and position embeddings preventing truncation issues (Scao et al., 2022; Hu et al., 2023); parameter-efficient fine-tuning via Adapter and LoRA methods matching larger models on MNLI, QNLI, and SST-2 (Hu et al., 2023); and low-data robustness requiring fewer labeled examples to achieve competitive scores compared to masked-encoder counterparts (Papadopoulos et al., 2024; Liu et al., 2023a).

5.3 Confidence-Based Routing Strategy

We implement a probability-based cascade to balance computational efficiency with classification accuracy. Many tutor utterances lack explicit lexical cues that our XGBoost baseline relies on, necessitating a dynamic routing approach.

For input x, we define probability vectors:

- XGBoost: $p_{\text{xgb}}(x) \in [0,1]^C$ (sigmoid activations)
- T5: $p_{\text{t5}}(x) \in [0,1]^C$ (softmax over {`yes`, `no`, `maybe`})

The cascade operates in three stages:

1. Route inputs through XGBoost. Accept prediction if $\max_c p_{\text{xgb}}(x)_c \geq \tau_1$.
2. If $\max_c p_{\text{xgb}}(x)_c < \tau_1$, escalate to T5. Accept if $\max_c p_{\text{t5}}(x)_c \geq \tau_2^{(c)}$ for any class c.

Algorithm 1: Learn class-specific precision cutoffs & coverage

```
Input: Validation set $V$, classes $\mathcal{C}$,
model probabilities $p_c(x)$ for each class $c$,
true labels $y(x)$, target precisions $\{\alpha_c\}$,
start threshold $\tau_0$, step size $\delta$,
upper bounds $\{U_c\}$
Output: Class-wise thresholds $\{\tau_c^*\}$ and
        coverages $\{\gamma_c^*\}$
foreach class $c \in \mathcal{C}$ do
    $N_c \leftarrow |\{x \in V : y(x) = c\}|$;
    $\tau \leftarrow \tau_0, \quad \tau_c^* \leftarrow U_c, \quad \gamma_c^* \leftarrow 0$;
    while $\tau \leq U_c$ do
        $S \leftarrow \{x \in V : p_c(x) \geq \tau\}$;
        if $|S| = 0$ then
            break
        $\text{prec} \leftarrow \frac{|\{x \in S : y(x) = c\}|}{|S|}$;
        if $\text{prec} \geq \alpha_c$ then
            $\tau_c^* \leftarrow \tau$;
            $\gamma_c^* \leftarrow |S|/N_c$;
            $\tau \leftarrow \tau + \delta$;
        else
            break;
return $\{(c, \tau_c^*, \gamma_c^*) \mid c \in \mathcal{C}\}$
```

3. If T5's confidence is insufficient, defer to an LLM judge.

Thresholds τ_1 for XGBoost and $\tau_2^{(c)}$ for T5 classes are learned on held-out data using Algorithm 1 to guarantee $\geq 95\%$ precision while maximizing coverage. In our experiments, stages 1 and 2 combined to produce 65-70% of predictions with required confidence, with the remaining 30-35% escalated to the LLM judge (discussed in the following section). See Table 3 for the learned thresholds and coverage at each stage.

Stage	Dimension	Thresholds (Yes/No/TSE)	Coverage at Stage
XGBoost	ML	0.85 / 0.45 / 0.55	0.38
	MI	0.82 / 0.48 / 0.52	0.35
	PG	0.88 / 0.42 / 0.58	0.32
	ACT	0.86 / 0.45 / 0.55	0.36
T5	ML	0.80 / 0.45 / 0.55	0.86
	MI	0.78 / 0.42 / 0.58	0.85
	PG	0.82 / 0.48 / 0.52	0.62
	ACT	0.81 / 0.45 / 0.55	0.60

Table 3: Learned thresholds for Yes/No/TSE classes and coverage percentages at each stage for each pedagogical dimension

6 Step-wise LLM-as-a-Judge

When both our lexical + XGBoost baseline and T5 classifier fall below confidence thresholds, we escalate to a multi-step "LLM-as-a-Judge" for final pedagogical-quality classification. In our dev-set evaluation, 31-34% of conversations across each dimension were escalated to the judge.

1. **Solution Reasoning Pathway Generation:** The judge prompts the LLM to generate a step-by-step expert solution for the given problem, establishing a reference reasoning pathway against which to align the student's response (Wei et al., 2022; Daheim et al., 2024; Jain, 2025). This includes parsing the problem, identifying relevant concepts, applying them systematically, and verifying the final result.

2. **Error Extraction:** The judge isolates the precise span where student reasoning diverges from the expert chain—this concrete "mistake locus" anchors all downstream diagnostic steps (Daheim et al., 2024; Macina et al., 2023). The goal is solely to extract and localize the deviation.

3. **Mistake Classification:** The mistake is mapped to a structured taxonomy enabling standardized reasoning about pedagogical strategies (Macina et al., 2023). Categories include conceptual errors, procedural/arithmetic errors, misapplied formulas, comprehension errors, and logical breakdowns in multi-step reasoning (Macina et al., 2023; Wang et al., 2024b; Daheim et al., 2024).

4. **Skill Gap Mapping:** Based on the mistake classification, the judge infers the underlying cognitive skill gap (Jain, 2025), referencing Bloom's revised taxonomy (Anderson et al., 2001; Krathwohl, 2002). This includes gaps in: Remember (recalling facts), Understand (grasping concepts), Apply (executing procedures), Analyze (breaking down structure), Evaluate (judging correctness), and Create (developing alternate methods).

5. **Last Tutor Turn Strategy Identification:** Conditioned on the diagnosed cognitive gap, the judge infers the most probable pedagogically aligned instructional strategy (Macina

et al., 2023; Wang et al., 2024b) which the last tutor turn most likely employed. This may include focus questions, probing questions, worked examples, hints, or problem simplification.

6. **Final Classification:** Integrating all intermediate steps along with the inferred instructional strategy employed by the tutor's last turn, the judge produces a final classification (Yes/No/TSE) for the last tutor turn according to the BEA 2025 Shared Task dimensions (Kochmar et al., 2025). The turn is evaluated for each of the following dimensions: (1) *Mistake Identification*, (2) *Mistake Location*, (3) *Providing Guidance*, and (4) *Actionability*.

Example Judge Output for PG:

> **Conversation History:** `Tutor:` Hi, could you please provide a step-by-step solution for the question below? Tyson decided to make muffaletta sandwiches for the big game. Each sandwich required 1 pound each of meat and cheese and would serve 4 people. There would be 20 people in total watching the game. The meat cost $7.00 per pound and the cheese cost $3.00 per pound. How much money would he spend on the meat and cheese to make enough sandwiches to serve 20 people? `Student:` To serve 20 people, Tyson needs to make 20/4 = 5 sandwiches. Each sandwich requires 1+1 = 2 pounds of meat and cheese...
>
> **Extracted Error:** `Each sandwich requires 1+1 = 2 pounds of meat and cheese.`
>
> **Mistake Type:** *Right-idea.* The student has the right idea but inaccurately combines meat and cheese quantities into one, leading to a misapplied calculation.
>
> **Skill Gap:** *Analyze and decompose independent components.* The student understands facts and unit costs but fails to reason about meat and cheese as distinct cost components.
>
> **Recommended Strategy:** *Provide a hint.* The tutor asks a guiding question to nudge the student to recalculate the meat cost independently, prompting correction without explicit error labeling.
>
> **Judge Classification:** *Yes.* The tutor turn offers appropriate scaffolding to guide the next step in solving the problem.

6.1 Dev-Set Escalation Impact

Evaluating 30 examples across all rubric dimensions, the LLM Judge reduced classification errors by 50%–60% compared to our T5 baseline, achieving Macro F1 scores above 75% in three tracks and 83.3% in one (Table 4).

Extrapolating these results to hybrid system performance with judge escalation on 30%–35% of low-confidence cases (**Hypo-Full** column), projections indicate that selective escalation can substantially bridge the gap to top-performing systems.

Track	Top	T5-subm	Judge-30	Hypo-Full
1	71.81	61.0	83.3	67.69
2	59.8	47.7	76.6	56.37
3	58.3	49.0	73.3	56.29
4	70.9	56.6	76.6	62.6

Table 4: Per-track F1: **Top** = best shared-task model; **T5-subm** = T5 model submitted results; **Judge-30** = LLM on 30 escalated dev cases; **Hypo-Full** = simulated performance assuming judge intervention on 30–35% of cases.

7 Submission Results and Analysis

Our team, **Emergent Wisdom**, participated in tracks 1 to 4 based on the architecture described in 3. The metrics and ranking of our best submission, according to the official leaderboard[1], is shown in Tables 5 and 6, which also contrast our Stage 1–2 (router + encoder–decoder) results on the test set against the top shared-task systems. Δ indicates (Ours – Top).

	Top		**Ours**			**Δ**	
Tr	**Acc**	**F1**	**Acc**	**F1**	**Rank**	**Acc**	**F1**
1	94.6%	89.6	93.2%	88.0	21	−1.4	−1.6
2	86.3%	83.9	78.9%	74.4	15	−7.4	−9.5
3	81.9%	78.0	77.3%	69.2	24	−4.6	−8.8
4	88.4%	85.3	80.5%	77.8	30	−7.8	−7.5

Table 5: Lenient metrics performance (Stages 1–2 on test set; Δ = Ours – Top).

	Top		**Ours**			**Δ**	
Tr	**Acc**	**F1**	**Acc**	**F1**	**Rank**	**Acc**	**F1**
1	86.2%	71.8	85.5%	61.0	34	−0.8	−10.8
2	76.8%	59.8	71.9%	47.7	25	−4.9	−12.1
3	66.1%	58.3	61.0%	49.0	21	−5.1	−9.3
4	73.0%	70.9	66.4%	56.6	22	−6.6	−14.3

Table 6: Exact metrics performance (Stages 1–2 on test set; Δ = Ours – Top).

As demonstrated in section 6.1, our analysis reveals that strategic reliance on the judge component for complex cases enables performance within 1-2% macro F1-score of the top-performing systems without increasing computational needs for the whole dataset, suggesting the potential for competitive results based on intelligent routing.

[1] https://sig-edu.org/sharedtask/2025#results

Limitations

Despite strong performance, our cascade approach faces several limitations: T5-Base's 512-token context window restricts processing of longer tutoring sessions; both models struggle with ambiguous utterances serving multiple pedagogical functions; performance suffers on underrepresented classes like the "maybe" classification; confidence-based routing relies on carefully tuned thresholds; and analyzing only tutor turns misses important student context. Future work should explore larger context windows, multi-label classification, and more sophisticated conversational modeling.

References

John R. Anderson, Albert T. Corbett, Kenneth R. Koedinger, and Sharon Pelletier. 1995. Intelligent tutoring systems. In *Computer Science and Artificial Intelligence Conference*. Springer.

Lorin W. Anderson, David R. Krathwohl, Peter W. Airasian, Kathleen A. Cruikshank, Richard E. Mayer, Paul R. Pintrich, James Raths, and Merlin C. Wittrock. 2001. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman.

Douglas Biber and Bethany Gray. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. John Benjamins Publishing Company, Amsterdam.

Rex P. Bringula and Ryan S. Basa. 2018. Effects of prior knowledge in mathematics on learner–interface interactions in a learning-by-teaching intelligent tutoring system. In *Proceedings of the 26th International Conference on Computers in Education*, pages 25–30. Asia-Pacific Society for Computers in Education.

Li-Hsin Chang and Filip Ginter. 2024. Automatic short answer grading for finnish with ChatGPT. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181, Vancouver, Canada. AAAI Press.

Guanliang Chen, Jie Yang, and Claudia Hauff. 2011. Studying effective tutoring strategies in programming moocs. In *Proceedings of the 6th ACM Conference on Learning at Scale*, pages 1–10. ACM.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM.

Michelene T. H. Chi, Stephanie A. Siler, Heisawn Jeong, Takashi Yamauchi, and Robert G. Hausmann. 2001. Learning from human tutoring. *Cognitive Science*, 25(4):471–533.

Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2489–2513, Miami, Florida, USA. Association for Computational Linguistics.

William W. Cohen, Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech acts". In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona, Spain. Association for Computational Linguistics.

Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. 2017. The tool for the automatic analysis of text cohesion (taaco): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, 49(4):1227–1237.

Scott A. Crossley and Danielle S. McNamara. 2022. Computational assessment of text readability and comprehension. *Journal of Research in Reading*, 45(2):223–246.

Nico Daheim, Jakub Macina, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2024. Stepwise verification and remediation of student reasoning errors with large language model tutors. *arXiv preprint arXiv:2407.09136*.

Zeyu Dai and Ruihong Huang. 2018. Improving implicit discourse relation classification by modeling inter-dependencies of discourse units in a paragraph. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 141–151, New Orleans, Louisiana. Association for Computational Linguistics.

Dorottya Demszky, Kelvin Guu, and Percy Liang. 2018. Transforming question answering datasets into natural language inference datasets. *arXiv preprint arXiv:1809.02922*.

Dorottya Demszky, Devyani Sharma, Jonathan Clark, Vinodkumar Prabhakaran, and Jacob Eisenstein. 2021. Learning to recognize dialect features. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2315–2338, Online. Association for Computational Linguistics.

Zhujun Deng, Afida Mohamad Ali, and Zaid Bin Mohd Zin. 2025. Investigating methodological trends of hedging strategies in academic discourse: A systematic literature review. *World Journal of English Language*, 15(5):322.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

Bruce Fraser. 1999. What are discourse markers? *Journal of Pragmatics*, 31(7):931–952.

Arthur C. Graesser, Sidney K. D'Mello, and Natalie K. Person. 2010. Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Journal of Educational Psychology*, 102(4):805–820.

John Hattie and Helen Timperley. 2007. The power of feedback. *Review of Educational Research*, 77(1):81–112.

Cindy E. Hmelo-Silver, Ravit Golan Duncan, and Clark A. Chinn. 2006. Scaffolding and achievement in problem-based and inquiry learning: A response to kirschner, sweller, and clark (2006). *Educational Psychologist*, 42(2):99–107.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2023. Lora: Low-rank adaptation of large language models. *Transactions on Machine Learning Research*.

Alon Jacovi, Yoav Goldberg, Aishwarya Kamath, Matthew Peters, and Roy Schwartz. 2024. Weak-link: Uncovering reasoning chain failures of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2341–2357. Association for Computational Linguistics.

Raunak Jain. 2025. Emergent wisdom: Empowering constructivism by proxying human reasoning with llm thought traces. *OSF Preprints*. Preprint. `https://osf.io/preprints/osf/dkst7_v1`.

Lan Jiang and Nigel Bosch. 2024. Short answer scoring with GPT-4. In *Proceedings of the 11th ACM Conference on Learning @ Scale (L@S '24)*, pages 1–5, Atlanta, GA, USA. Association for Computing Machinery.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2024. Trust or escalate: LLM judges with provable guarantees for human agreement. *arXiv preprint arXiv:2407.18370*.

Jaehun Jung, Faeze Brahman, and Yejin Choi. 2025. Trust or escalate: LLM judges with provable guarantees for human agreement. In *Proceedings of the 13th International Conference on Learning Representations (ICLR)*, Addis Ababa, Ethiopia. Oral paper.

Ekaterina Kochmar, Kaushal Kumar Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 shared task on pedagogical ability assessment of AI-powered tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

David R. Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory Into Practice*, 41(4):212–218.

Ehsan Latif and Xiaoming Zhai. 2024. Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, 6:100210.

Blair Lehman, Sidney D'Mello, Amber Strain, Caitlin Mills, Melissa Gross, Allyson Dobbins, Patricia Wallace, Keith Millis, and Arthur Graesser. 2019. Automated analysis of tutorial dialogues: Unsupervised modeling of student and tutor behaviors. In *International Conference on Artificial Intelligence in Education*, pages 117–127.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. HaluEval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6449–6464, Singapore.

Justin Li, Jaemin Choi, Samir Yitzhak Gadre, Ansh Kaul, Louis-Philippe Morency, Rada Mihalcea, Minsu Seo, Darsh Shah, Luke Zettlemoyer, Amanda Stent, Jihie Hwang, Kyunghyun Cho, and Aniruddha Kembhavi. 2025. Tutorgym: A benchmark for tutoring with large language models. *arXiv preprint arXiv:2410.11895*.

Ting Liu, Yiming Chen, Daniel Brown, Sebastian Riedel, and Hoifung Poon. 2023a. Pre-trained language models can be fully zero-shot learners. *Transactions of the Association for Computational Linguistics*, 11:1032–1049.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 7321–7335.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Jo Mackiewicz and Isabelle Thompson. 2010. Assertions of expertise in online tutoring sessions. *Journal of Business and Technical Communication*, 24(1):3–28.

Kaushal Kumar Maurya, Kv Aditya Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2025. Unifying AI tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of LLM-powered AI tutors. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1234–1251, Albuquerque, New Mexico. Association for Computational Linguistics.

Danielle S. McNamara, Arthur C. Graesser, Philip M. McCarthy, and Zhiqiang Cai. 2010. Linguistic features of writing quality. *Written Communication*, 27(1):57–86.

Neil Mercer and Karen Littleton. 2009. Dialogue and the development of children's thinking. *Educational Psychology in Practice*, 25(4):365–379.

Sapna Negi and Paul Buitelaar. 2015. Towards the extraction of customer-to-customer suggestions from reviews. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2159–2167, Lisbon, Portugal. Association for Computational Linguistics.

Bill Noble and Vladislav Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 166–172, Groningen, The Netherlands (online). Association for Computational Linguistics.

Martin Nystrand and Adam Gamoran. 1997. *Opening Dialogue: Understanding the Dynamics of Language and Learning in the English Classroom*. Teachers College Press, New York.

Masanari Ohi, Masahiro Kaneko, Ryuto Koike, Mengsay Loem, and Naoaki Okazaki. 2024. Likelihood–based mitigation of evaluation bias in large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3237–3245, Bangkok, Thailand.

Theodoros Papadopoulos, Jiasheng Du, Claudiu Musat, Marija Bjelogrlic, and Robert West. 2024. Language models for text classification: Is in-context learning enough? *arXiv preprint arXiv:2403.17661*.

Muhammad Qorib, Geonsik Moon, and Hwee Tou Ng. 2024. Are decoder-only language models better than encoder-only language models in understanding word meaning? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16339–16347, Bangkok, Thailand. Association for Computational Linguistics.

Muhammad Abdul-Mageed Qureshi and Michael Strube. 2022. Linguistically motivated features for classifying shorter text into fiction and non-fiction genres. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*, pages 924–934, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020a. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020b. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tim Rowland. 2002. Being mathematically assertive: The role of hedges in mathematical discourse. *For the Learning of Mathematics*, 22(3):12–18.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. In *Proceedings of the NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Ted J. M. Sanders, Wilbert P. M. Spooren, and Leo G. M. Noordman. 2000. Toward a taxonomy of coherence relations. *Discourse Processes*, 15(1):1–35.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. In *arXiv preprint arXiv:2211.05100*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems 35*, pages 17456–17472.

Philip H. Scott, Eduardo F. Mortimer, and Orlando G. Aguiar. 2002. The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Science Education*, 90(4):605–631.

Valerie J. Shute. 2008. Focus on formative feedback. *Review of Educational Research*, 78(1):153–189.

Anaïs Tack, Ekaterina Kochmar, Zheng Yuan, Serge Bibauw, and Chris Piech. 2023. The BEA 2023 shared task on generating AI teacher responses in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 785–795, Toronto, Canada. Association for Computational Linguistics.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist*, 46(4):197–221.

Kurt VanLehn, Arthur C. Graesser, G. Tanner Jackson, Pamela Jordan, Andrew Olney, and Carolyn P. Rosé. 2006. When are tutorial dialogues more effective than reading? *Cognitive Science*, 30(1):3–62.

Joshua Wagner, Tianyi Zhang, Stephen Bach, Ankit Jha, Michael Wornow, Avery Adler, Besmira Nushi, and Titus Glaunsinger. 2024. Label with confidence: Advancing selective prediction through ensemble agreement. In *Proceedings of the 41st International Conference on Machine Learning*, pages 22478–22501.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. Interpretable automated feedback for student writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4841–4852.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 9440–9450, Bangkok, Thailand.

Rose Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2024b. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2174–2199, Mexico City, Mexico. Association for Computational Linguistics.

Wayne Ward, Ron Cole, Daniel Bolaños, Cindy Buchenroth-Martin, Edward Svirsky, and Tim Weston. 2011. My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing*, 7(4):1–29.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837.

Jörg Wittwer and Alexander Renkl. 2010. How effective are instructional explanations in example-based learning? a meta-analytic review. *Educational Psychology Review*, 22(4):393–409.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet? a systematic literature review on chatbots in education. *Frontiers in Artificial Intelligence*, 4:654924.

Mengzhou Xia, Abdullah Ali, Angela Fan, Lilian Zhong, Allyson Ettinger, Ekin Akyurek, Margaret Mitchell, and Jacob Andreas. 2025. Llm-rubric: Evaluating machine-generated text with customized rubrics. *arXiv preprint arXiv:2403.06929*.

Canwen Xu and Julian McAuley. 2022. A survey on dynamic neural networks for natural language processing. In *Journal of Machine Learning Systems*. Comprehensive review of early-exit and cascade methods in NLP.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 40 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zijiang Yang. 2024. Improving the natural language inference robustness to hard dataset by data augmentation and preprocessing. *arXiv preprint arXiv:2412.07108*.

Tianyu Zhang, Sayak Basu, Douwe Kiela, and Oriol Vinyals. 2024. Cascade-aware training and inference for more resource-efficient language models. *arXiv preprint arXiv:2405.12345*.