Appendix: Enhancing LLM-Based Data Annotation with Error Decomposition

A Annotation Tasks and Ground Truth Dataset

A.1 Bloom

A.1.1 Dataset description: The Bloom dataset comprises 21,380 learning objectives collected from 5,558 courses offered at an Australian university in 2021. Each learning objective is treated as an individual text instance and annotated according to Bloom's Taxonomy's six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. Learning objectives were assigned single or multiple labels when they reflected more than one cognitive process.

Annotation followed a multi-stage procedure. First, a trained annotator with prior expertise in Bloom's Taxonomy manually coded the full dataset. A second independent annotator, also trained in Bloom's Taxonomy, independently annotated a random 30% subset of learning objectives without access to the first annotator's labels. Inter-rater reliability, measured using Cohen's k, was $k = 0.63$. The two annotators then discussed the disagreement cases, identified the source of disagreement,and revised the corresponding labels in the entire dataset accordingly, which increased inter-rater reliability on the 30% sample to $k = 0.80$.

A.1.2 Sampling: Since the original dataset contains a mixture of single-label and multi-label annotations, we randomly sampled single-labeled learning objectives: 150 for the evaluation set, 300 for prompt optimization, and 20 for active prompting.

A.1.3 Annotation Rubric: As the original dataset does not provide explicit documentation regarding the definition or adaptation of Bloom's Taxonomy used during annotation, we adopted the original definitions from the revised Bloom's Taxonomy classification rubric, which specifies six cognitive categories, for use in our experiment.

Label	Description
Remember	Retrieving relevant knowledge from long-term memory.
Understand	Determining the meaning of instructional messages, including oral, written, and graphic communication.
Apply	Carrying out or using a procedure in a given situation.
Analyze	Breaking material into its constituent parts and detecting how the parts relate to one another and to an overall structure or purpose.
Evaluate	Making judgments based on criteria and standards.
Create	Putting elements together to form a novel, coherent whole or make an original product.

Table 1. Annotation Rubric of Bloom Dataset

A.2 MathDial

A.2.1 Dataset Description. The MathDial dataset comprises 2,861 one-to-one tutoring dialogues grounded in multi-step math word problems. Each dialogue is treated as a sequence of teacher-student turns and is annotated according to a taxonomy of four pedagogical teacher moves: Focus (guiding progress), Probing (deepening understanding), Telling (revealing answers), and Generic (conversation management).

Annotation and data collection followed a semi-synthetic procedure where 91 professional teachers interacted synchronously with a Large Language Model simulating diverse student behaviors. Inter-rater reliability for the teacher moves taxonomy, measured using Cohen's k between two independent annotators after discrepancy resolution, was k=0.67.

A.2.2 Sampling. Each conversation contains multiple teacher utterances, with an average of approximately 4-6 teacher moves per conversation.We sampled complete conversations from the original dataset to include around 150 teacher utterances for the evaluation set, 100 for prompt optimization, and 20 for active prompting.

A.2.3 Annotation Rubric. The annotation rubric classifies teachers' intentions into four categories based on the pedagogical function of each utterance:

Label	Description
Generic	The teacher's intention includes Greeting/Farewell and General inquiry. These utterances serve social interaction and general classroom discourse without specific mathematical content.
Focus	The teacher's intention includes seeking a strategy, guiding student focus, and recalling relevant information. These scaffolding utterances direct student attention toward productive problem-solving approaches.
Probing	The teacher's intention includes Asking for Explanation, Seeking Self Correction, Perturbing the Question, and Seeking World Knowledge. These utterances prompt students to articulate their reasoning or evaluate their solutions.
Telling	The teacher's intention includes Revealing Strategy and Revealing Answer. These utterances directly communicate mathematical answers or solution methods to the student.

Table 2. Annotation Rubric of MathDial Dataset

A.3 Conversational Uptake

A.3.1 Dataset Description. The Conversational Uptake dataset comprises 2,246 student-teacher exchanges extracted from US elementary math classroom transcripts collected by the National Center for Teacher Effectiveness (NCTE) between 2010-2013. Each exchange consists of a student utterance followed by a teacher response, annotated for the degree to which the teacher demonstrates they are following what the student is saying or trying to say. The dataset represents data from 317 teachers across 4 school districts in New England.

Annotation followed a rigorous multi-stage procedure developed by experts in math instruction quality. Each exchange was annotated by three domain experts (math teachers and trained classroom observation raters). Inter-rater reliability, measured using Spearman correlation, was $\rho = 0.474$.

A.3.2 Sampling. We sampled student-teacher exchanges from the dataset: 150 exchanges for the evaluation set, 100 exchanges for prompt optimization, and 20 exchanges for active prompting. Each exchange represents a single student contribution paired with the subsequent teacher response.

A.3.3 Annotation Rubric. The annotation rubric measures the degree to which a teacher demonstrates they are following what the student is saying or trying to say, using a three-level scale:

Label	Description
Low	Teacher shows minimal or no evidence of following the student's contribution. The response may ignore, dismiss, or redirect away from what the student said, or provide only generic acknowledgment that could apply to any student input.
Mid	Teacher demonstrates moderate engagement with the student's contribution. The response shows the teacher heard the student through acknowledgment, repetition, or brief affirmation, but does not deeply explore or extend the idea.
High	Teacher actively builds upon and extends the student's contribution. The response demonstrates deep engagement by asking probing questions, requesting elaboration, or connecting the student's idea to broader concepts.

Table 3. Annotation Rubric of Conversational Uptake Dataset

A.4 GUG

A.4.1 Dataset Description. The GUG (Grammatical vs UnGrammatical) dataset comprises 3,129 sentences randomly selected from essays written by nonnative speakers of English. Each sentence is evaluated on a 4-point Likert scale measuring the degree of grammaticality, ranging from incomprehensible (Level 1) to perfect or near-perfect grammar (Level 4).

Annotation followed a multi-stage procedure. First, two expert judges—native English speakers with linguistic training—independently annotated a subset of 442 sentences while viewing the previous sentence from the essay as context. Inter-rater reliability for these experts, measured using unweighted κ , was κ=0.574 (with a Pearson's $r = 0.759$). Five additional crowdsourced judgments were collected for each sentence using a small subset of "gold" expert-labeled sentences to filter out unreliable contributors. The final gold standard labels were established by averaging the six total judgments (one expert and five crowdsourced) for each item

A.4.2 Sampling. We sampled sentences from the dataset: 150 sentences for the evaluation set, 100 sentences for prompt optimization, and 20 sentences for active prompting.

A.4.3 Annotation Rubric. The annotation rubric for Grammatical Understanding of Grammar (GUG) assessment uses a 4-point scale:

Label	Description
Perfect or Near-Perfect Grammar	Sounds like native English, completely comprehensible, no grammatical errors. NOTE: There can be minor spelling errors, typos, capitalization or comma errors that don't make the sentence unclear. In this set of sentences, contractions and possessive markers will always be separated into two parts, and there may be spaces before periods and commas, so please ignore these issues.
One or More Minor Grammatical Errors	Comprehensible with minor grammatical errors, including subject-verb agreement errors, determiner/article errors (such as 'a cats' or 'those person'), minor preposition errors (such as using 'in' when 'on' would be a much more native-sounding choice), word choice errors, etc. Can also be characterized by spelling and comma errors that are more serious than those above.
One or More Serious Grammatical Errors	Somewhat comprehensible - these sentences may have multiple possible interpretations and it is not at all clear which is most likely. These more serious grammatical errors would include errors such as the lack of a verb, object, etc., verb tense errors, serious preposition errors (such as using 'to' when 'with' is the only preposition that makes sense), etc.
Incomprehensible	So many grammatical errors that it is difficult to fix. The sentence cannot be understood.

Table 4. Annotation Rubric of GUG Dataset

B Prompts, decoding parameters, and prompt optimization

B.1 Zero Shot

B.1.1 Decoding parameters.

- Temperature=0
- Maximum tokens=150

B.1.2 Bloom Prompt:

```
Your expertise lies in categorizing educational goals based on Bloom's Taxonomy.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the following learning outcome and select the ONE category that BEST describes the primary cognitive level required.
Learning Outcome: [TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Carefully analyze the main action verb and cognitive demand
2. Consider which cognitive level is the PRIMARY focus
3. Select ONLY ONE category that best fits
4. Respond with just the category name in lowercase
The 6 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this learning outcome? Respond with just the category name:
```

B.1.3 MathDial Prompt:

```
Your expertise lies in categorizing teacher moves in math tutoring based on pedagogical intent.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze ONE SPECIFIC teacher utterance and classify it into one of the four categories above. You have the full conversation context to understand the situation, but you must classify ONLY the specified utterance.
Teacher Utterance:[TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Read the full context to understand the conversation flow
2. Focus on THIS SPECIFIC teacher utterance
3. Analyze the pedagogical intent of THIS utterance
4. Classify ONLY this utterance (not the whole conversation)
5. Respond with only one word: generic, focus, probing, or telling
The 4 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this teacher move? Respond with just the category name:
```

B.1.4 Conversational Uptake Prompt.

```
Your expertise lies in evaluating teacher uptake in educational conversations.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze how well the teacher demonstrates that they are following what the student is saying
or trying to say.
Student-Teacher exchange: Student: "[STUDENT TEXT]" Teacher: "[TEACHER TEXT]"
INSTRUCTIONS:
1. Read the student's contribution carefully
2. Analyze how the teacher responds to the student
3. Respond with only one word: Low, Mid, or High
The 3 categories are: Low, Mid, High. Which ONE category best describes the teacher's uptake?
Respond with just the category name:
```

B.1.5 GUG Prompt:

```
You are an expert at evaluating English sentence grammaticality.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the grammaticality of the following sentence.
SENTENCE: "[TEXT TO BE ANNOTATED]"
INSTRUCTIONS: 1. Read the sentence carefully
2. Identify any grammatical errors
3. Determine which level (1-4) best describes the sentence's grammaticality
4. Respond with only a number: 1, 2, 3, or 4
The 4 levels are: 1 (Incomprehensible), 2 (Serious errors),3 (Minor errors), 4
(Perfect/Near-perfect).Which ONE level best describes this sentence's grammaticality? Respond
with just the number:
```

B.2 Few Shot

B.2.1 Decoding parameters.

- Temperature=0
- Maximum tokens=150

B.2.2 Bloom Prompt:

```
Your expertise lies in categorizing educational goals based on Bloom's Taxonomy.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the following learning outcome and select the ONE category that BEST describes the primary cognitive level required.
EXAMPLES: [ANNOTATED EXAMPLES]
Learning Outcome: [TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Carefully analyze the main action verb and cognitive demand
2. Consider which cognitive level is the PRIMARY focus
3. Select ONLY ONE category that best fits
4. Respond with just the category name in lowercase
The 6 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this learning outcome? Respond with just the category name:
```

B.2.3 MathDial Prompt:

```
Your expertise lies in categorizing teacher moves in math tutoring based on pedagogical intent.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze ONE SPECIFIC teacher utterance and classify it into one of the four categories above. You have the full conversation context to understand the situation, but you must classify ONLY the specified utterance.
EXAMPLES: [ANNOTATED EXAMPLES]
Teacher Utterance:[TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Read the full context to understand the conversation flow
2. Focus on THIS SPECIFIC teacher utterance
3. Analyze the pedagogical intent of THIS utterance
4. Classify ONLY this utterance (not the whole conversation)
5. Respond with only one word: generic, focus, probing, or telling
The 4 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this teacher move? Respond with just the category name:
```

B.2.4 Conversational Uptake Prompt.

```
Your expertise lies in evaluating teacher uptake in educational conversations.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze how well the teacher demonstrates that they are following what the student is saying or trying to say.
EXAMPLES: [ANNOTATED EXAMPLES]
Student-Teacher exchange: Student: "[STUDENT TEXT]" Teacher: "[TEACHER TEXT]"
INSTRUCTIONS:
1. Read the student's contribution carefully
2. Analyze how the teacher responds to the student
3. Respond with only one word: Low, Mid, or High
The 3 categories are: Low, Mid, High. Which ONE category best describes the teacher's uptake? Respond with just the category name:
```

B.2.5 GUG Prompt:

```
You are an expert at evaluating English sentence grammaticality.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the grammaticality of the following sentence.
EXAMPLES: [ANNOTATED EXAMPLES]
SENTENCE: "[TEXT TO BE ANNOTATED]"
INSTRUCTIONS: 1. Read the sentence carefully
2. Identify any grammatical errors
3. Determine which level (1-4) best describes the sentence's grammaticality
4. Respond with only a number: 1, 2, 3, or 4
The 4 levels are: 1 (Incomprehensible), 2 (Serious errors),3 (Minor errors), 4 (Perfect/Near-perfect).Which ONE level best describes this sentence's grammaticality? Respond with just the number:
```

B.3 Auto-Chain of Thought

B.3.1 Decoding parameters.

- Temperature=0
- Maximum tokens=150

B.3.2 Bloom Prompt:

```
Your expertise lies in categorizing educational goals based on Bloom's Taxonomy.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the following learning outcome and select the ONE category that BEST describes the primary cognitive level required.
EXAMPLES: [ANNOTATED EXAMPLES]
Learning Outcome: [TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Carefully analyze the main action verb and cognitive demand
2. Consider which cognitive level is the PRIMARY focus
3. Select ONLY ONE category that best fits
4. Respond with just the category name in lowercase
Please think step by step.
The 6 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this learning outcome? Respond with just the category name:
```

B.3.3 MathDial Prompt:

```
Your expertise lies in categorizing teacher moves in math tutoring based on pedagogical intent.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze ONE SPECIFIC teacher utterance and classify it into one of the four categories above. You have the full conversation context to understand the situation, but you must classify ONLY the specified utterance.
EXAMPLES: [ANNOTATED EXAMPLES]
Teacher Utterance:[TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Read the full context to understand the conversation flow
2. Focus on THIS SPECIFIC teacher utterance
3. Analyze the pedagogical intent of THIS utterance
4. Classify ONLY this utterance (not the whole conversation)
5. Respond with only one word: generic, focus, probing, or telling
Please think step by step.
The 4 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this teacher move? Respond with just the category name:
```

B.3.4 Conversational Uptake Prompt.

```
Your expertise lies in evaluating teacher uptake in educational conversations.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze how well the teacher demonstrates that they are following what the student is saying
or trying to say.
EXAMPLES: [ANNOTATED EXAMPLES]
Student-Teacher exchange: Student: "[STUDENT TEXT]" Teacher: "[TEACHER TEXT]"
INSTRUCTIONS:
1. Read the student's contribution carefully
2. Analyze how the teacher responds to the student
3. Respond with only one word: Low, Mid, or High
Please think step by step.
The 3 categories are: Low, Mid, High. Which ONE category best describes the teacher's uptake?
Respond with just the category name:
```

B.3.5 GUG Prompt:

```
You are an expert at evaluating English sentence grammaticality.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the grammaticality of the following sentence.
EXAMPLES: [ANNOTATED EXAMPLES]
SENTENCE: "[TEXT TO BE ANNOTATED]"
INSTRUCTIONS: 1. Read the sentence carefully
2. Identify any grammatical errors
3. Determine which level (1-4) best describes the sentence's grammaticality
4. Respond with only a number: 1, 2, 3, or 4
Please think step by step.
The 4 levels are: 1 (Incomprehensible), 2 (Serious errors),3 (Minor errors), 4
(Perfect/Near-perfect).Which ONE level best describes this sentence's grammaticality? Respond
with just the number:
```

B.4 Self-Consistency

B.4.1 Decoding parameters.

- Temperature=0.7
- Maximum tokens=400
- Number of reasoning paths (n_samples)=3
- Aggregation method: Majority vote

B.4.2 Bloom Prompt:

```
Your expertise lies in categorizing educational goals based on Bloom's Taxonomy.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the following learning outcome and select the ONE category that BEST describes the primary cognitive level required.
EXAMPLES: [ANNOTATED EXAMPLES]
Learning Outcome: [TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Carefully analyze the main action verb and cognitive demand
2. Consider which cognitive level is the PRIMARY focus
3. Select ONLY ONE category that best fits
4. Respond with just the category name in lowercase
Please think step by step.
The 6 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this learning outcome? Respond with just the category name:
```

B.4.3 MathDial Prompt:

```
Your expertise lies in categorizing teacher moves in math tutoring based on pedagogical intent.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze ONE SPECIFIC teacher utterance and classify it into one of the four categories above. You have the full conversation context to understand the situation, but you must classify ONLY the specified utterance.
EXAMPLES: [ANNOTATED EXAMPLES]
Teacher Utterance:[TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Read the full context to understand the conversation flow
2. Focus on THIS SPECIFIC teacher utterance
3. Analyze the pedagogical intent of THIS utterance
4. Classify ONLY this utterance (not the whole conversation)
5. Respond with only one word: generic, focus, probing, or telling
Please think step by step.
The 4 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this teacher move? Respond with just the category name:
```

B.4.4 Conversational Uptake Prompt.

```
Your expertise lies in evaluating teacher uptake in educational conversations.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze how well the teacher demonstrates that they are following what the student is saying or trying to say.
EXAMPLES: [ANNOTATED EXAMPLES]
Student-Teacher exchange: Student: "[STUDENT TEXT]" Teacher: "[TEACHER TEXT]"
INSTRUCTIONS:
1. Read the student's contribution carefully
2. Analyze how the teacher responds to the student
3. Respond with only one word: Low, Mid, or High
Please think step by step.
The 3 categories are: Low, Mid, High. Which ONE category best describes the teacher's uptake? Respond with just the category name:
```

B.4.5 GUG Prompt:

```
You are an expert at evaluating English sentence grammaticality.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the grammaticality of the following sentence.
EXAMPLES: [ANNOTATED EXAMPLES]
SENTENCE: "[TEXT TO BE ANNOTATED]"
INSTRUCTIONS: 1. Read the sentence carefully
2. Identify any grammatical errors
3. Determine which level (1-4) best describes the sentence's grammaticality
4. Respond with only a number: 1, 2, 3, or 4
Please think step by step.
The 4 levels are: 1 (Incomprehensible), 2 (Serious errors),3 (Minor errors), 4 (Perfect/Near-perfect).Which ONE level best describes this sentence's grammaticality? Respond with just the number:
```

B.5 Active Prompting

B.5.1 Decoding parameters.

- Temperature=0.8 (for uncertainty estimation), Temperature=0 (for final prediction)
- Maximum tokens=150
- Pool size=20 (questions for uncertainty estimation)
- K samples=5 (predictions per question for uncertainty)
- Selection criteria: Top uncertain + top wrong examples

B.5.2 Bloom Prompt:

```
Your expertise lies in categorizing educational goals based on Bloom's Taxonomy.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the following learning outcome and select the ONE category that BEST describes the primary cognitive level required.
EXAMPLES:[SELECTED EXAMPLES WITH THE HIGHEST UNCERTAINTY AND INCORRECT PREDICTIONS]
Learning Outcome: [TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Carefully analyze the main action verb and cognitive demand
2. Consider which cognitive level is the PRIMARY focus
3. Select ONLY ONE category that best fits
4. Respond with just the category name in lowercase
Please think step by step.
The 6 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this learning outcome? Respond with just the category name:
```

B.5.3 MathDial Prompt:

```
Your expertise lies in categorizing teacher moves in math tutoring based on pedagogical intent.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze ONE SPECIFIC teacher utterance and classify it into one of the four categories above. You have the full conversation context to understand the situation, but you must classify ONLY the specified utterance.
EXAMPLES:[SELECTED EXAMPLES WITH THE HIGHEST UNCERTAINTY AND INCORRECT PREDICTIONS]
Teacher Utterance:[TEXT TO BE ANNOTATED]
INSTRUCTIONS:
1. Read the full context to understand the conversation flow
2. Focus on THIS SPECIFIC teacher utterance
3. Analyze the pedagogical intent of THIS utterance
4. Classify ONLY this utterance (not the whole conversation)
5. Respond with only one word: generic, focus, probing, or telling
Please think step by step.
The 4 categories are: [RUBRIC CATEGORIES NAME LIST], Which ONE category best describes this teacher move? Respond with just the category name:
```

B.5.4 Conversational Uptake Prompt.

Your expertise lies in evaluating teacher uptake in educational conversations.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze how well the teacher demonstrates that they are following what the student is saying or trying to say.
EXAMPLES: [SELECTED EXAMPLES WITH THE HIGHEST UNCERTAINTY AND INCORRECT PREDICTIONS]
Student-Teacher exchange: Student: "[STUDENT TEXT]" Teacher: "[TEACHER TEXT]"
INSTRUCTIONS:
1. Read the student's contribution carefully
2. Analyze how the teacher responds to the student
3. Respond with only one word: Low, Mid, or High
The 3 categories are: Low, Mid, High. Which ONE category best describes the teacher's uptake? Respond with just the category name:

B.5.5 GUG Prompt:

You are an expert at evaluating English sentence grammaticality.
RUBRIC: [FULL ANNOTATION RUBRIC]
TASK: Analyze the grammaticality of the following sentence.
EXAMPLES:[SELECTED EXAMPLES WITH THE HIGHEST UNCERTAINTY AND INCORRECT PREDICTIONS]
SENTENCE: "[TEXT TO BE ANNOTATED]"
INSTRUCTIONS: 1. Read the sentence carefully
2. Identify any grammatical errors
3. Determine which level (1-4) best describes the sentence's grammaticality
4. Respond with only a number: 1, 2, 3, or 4
Please think step by step.
The 4 levels are: 1 (Incomprehensible), 2 (Serious errors),3 (Minor errors), 4 (Perfect/Near-perfect).Which ONE level best describes this sentence's grammaticality? Respond with just the number:

B.6 Dspy Optimization

B.6.1 Pipeline Configuration.

- **Training sample sizes**: 100, 200, 300 examples
- **Optimization method**: BootstrapFewShot
- **Optimization parameters**:
  - max_bootstrapped_demos: 3
  - max_labeled_demos: 4
  - optimization_rounds: 1
- **Models tested**: GPT-3.5-turbo-0125 (temperature=0, max_tokens=500) and GPT-5 (temperature=1.0, max_tokens=20000)

B.6.2 Task Signatures. DSPy uses signature classes to define the input-output structure for each task.

Bloom Taxonomy Classification:

```
class BloomSignature(dspy.Signature):
    learning_outcome = InputField(
        desc="The educational learning outcome to classify")
    remember = OutputField(
        desc="1 if requires remembering/recalling, else 0")
    understand = OutputField(
        desc="1 if requires understanding/explaining, else 0")
    apply = OutputField(
        desc="1 if requires applying knowledge, else 0")
    analyze = OutputField(
        desc="1 if requires analyzing/comparing, else 0")
    evaluate = OutputField(
        desc="1 if requires evaluating/judging, else 0")
    create = OutputField(
        desc="1 if requires creating/designing, else 0")
```

MathDial Teacher Move Classification:

```
class MathDialSignature(dspy.Signature):
    problem = InputField(
        desc="The math problem being discussed")
    student_solution = InputField(
        desc="The student's incorrect solution attempt")
    conversation_history = InputField(
        desc="Previous exchanges in the conversation")
    teacher_utterance = InputField(
        desc="The current teacher utterance to classify")
    move_type = OutputField(
        desc="Teacher move: 'focus', 'probing', 'telling', or 'generic'")
    reasoning = OutputField(
```

```
        desc="Brief explanation for the classification")
```

Conversational Uptake Classification:

```
class UptakeSignature(dspy.Signature):
    student_text = InputField(
        desc="The student's utterance or contribution")
    teacher_text = InputField(
        desc="The teacher's response to the student")
    uptake_level = OutputField(
        desc="One of: Low, Mid, High")
    reasoning = OutputField(
        desc="Explanation of why this uptake level was chosen")
```

GUG Grammaticality Evaluation:

```
class GrammaticalitySignature(dspy.Signature):
    sentence = InputField(
        desc="The English sentence to evaluate for grammaticality")
    grammaticality_level = OutputField(
        desc="Integer from 1 to 4, where 1=incomprehensible,
              2=serious errors, 3=minor errors, 4=perfect/near-perfect")
    reasoning = OutputField(
        desc="Brief explanation of the grammatical errors found (if any)")
```

C Human Annotation Tasks and Survey

C.1 Annotation Instruction

The annotation process was conducted in two rounds. Each annotator completed one set of 20 samples at a time, with no time limit, and was encouraged to take breaks between tasks. Annotations were recorded in Google Sheets (one sheet per task), while detailed annotation guidelines were provided in Google Docs for continuous reference. The annotation interface is shown in Fig 1.

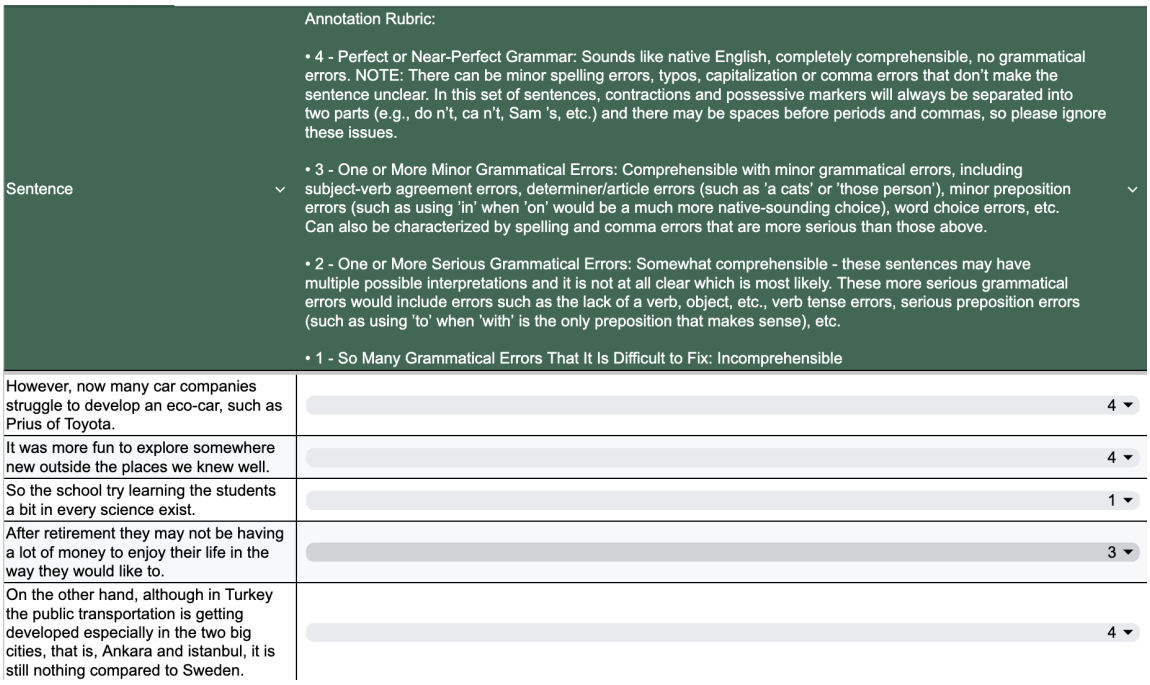

Fig. 1. Human Annotation interface: example from GUG task

C.2 Post Annotation Survey

After completing the annotation task, annotators individually completed a survey to assess their perception of the boundary and conceptual clarity of the annotation tasks. For each dimension, we included a direct measurement item(the first question in Table5) and five supplementary items . Two annotators were asked to rank the four tasks from the most difficult to the easiest to annotate. The survey interface is shown in Fig2

Error Type	Survey Question (4 = Most difficult to annotate, 1 = Easiest to annotate)
Boundary ambiguity	1. The boundary between adjacent categories (e.g., scores 2 and 3) felt clear and distinct. 2. The guidelines provided sufficient and clear rules for instances that were right on the border between two scores. 3. The scale felt too fine-grained, forcing me to make a difficult choice between two closely-spaced numbers (e.g., 3 vs. 4). 4. I could clearly articulate the precise "tipping point" or feature that pushed a text from category N to category N+1. 5. When assigning a score, the adjacent categories often felt like a plausible alternative. 6. The text required me to focus on very subtle differences in wording or context to distinguish between adjacent scores.
Conceptual Misidentification	1. It was difficult to distinguish between categories that are not adjacent on the scale (e.g., confusing a low category with a high one)? 2. The labeling categories were well-defined and did not overlap conceptually across the whole scale (e.g., 1 vs. 5). 3. I was completely clear about the single, underlying concept (e.g., toxicity, relevance) the ordinal scale was measuring. 4. I felt the task was conceptually straightforward and easy to grasp. 5. The texts themselves were frequently so ambiguous that I had to guess the general magnitude of the score (e.g., low vs. high). 6. I felt very confident about the general magnitude of the score (e.g., "it's definitely high" or "it's definitely low").

Table 5. Post Survey Annotation Questions

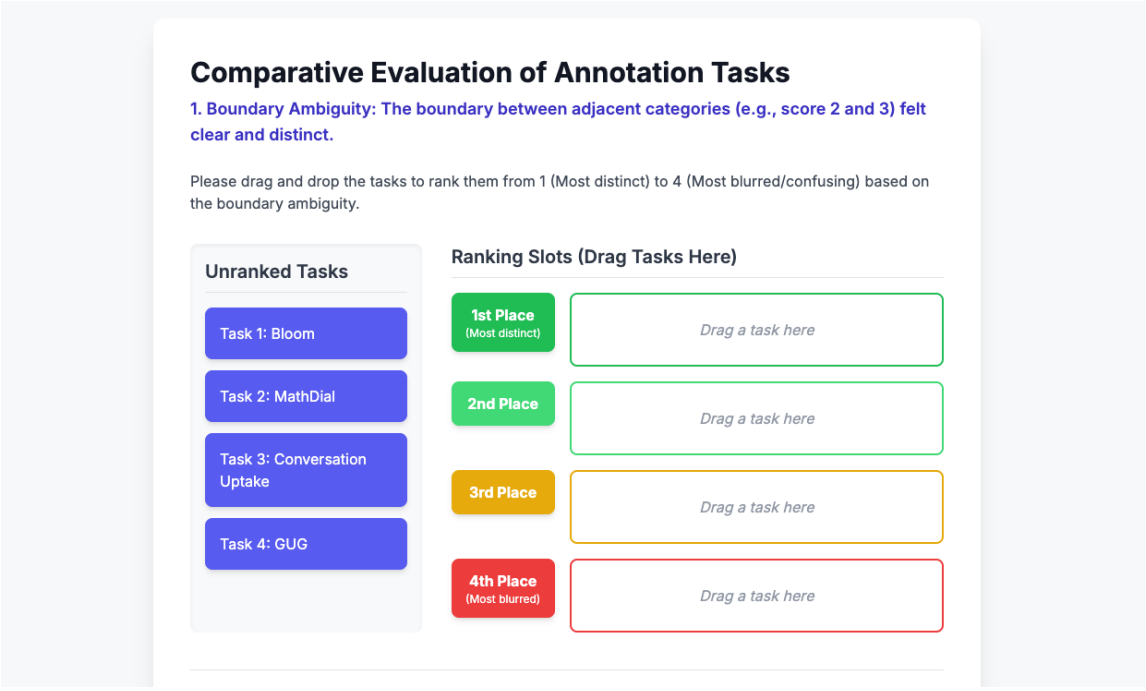

Fig. 2. Post Annotation Survey Interface

D Additional results

Here we present additional results from error decomposition based on the 30 and 40 item human annotation tasks.

D.1 Human annotations: 30 items

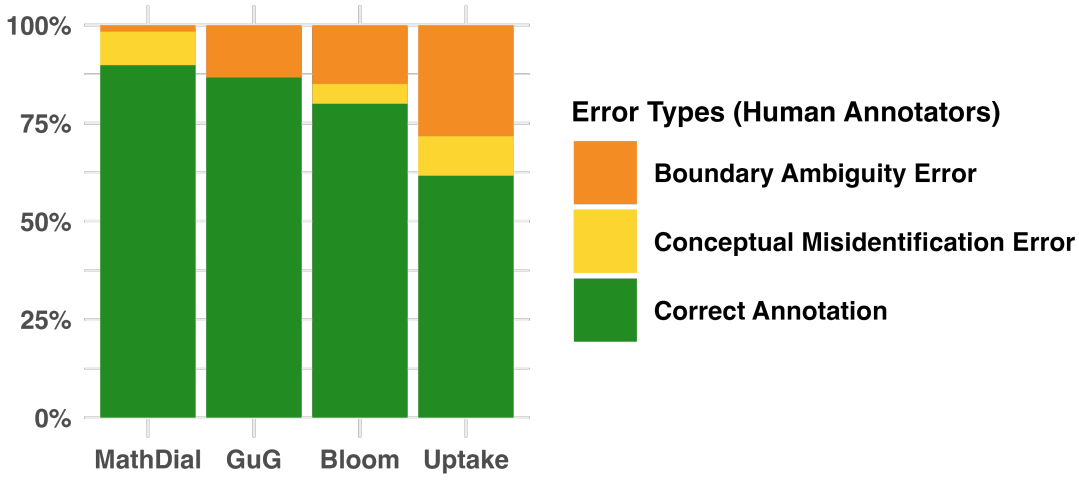

Fig. 3. **Distribution of error types observed in human annotations**

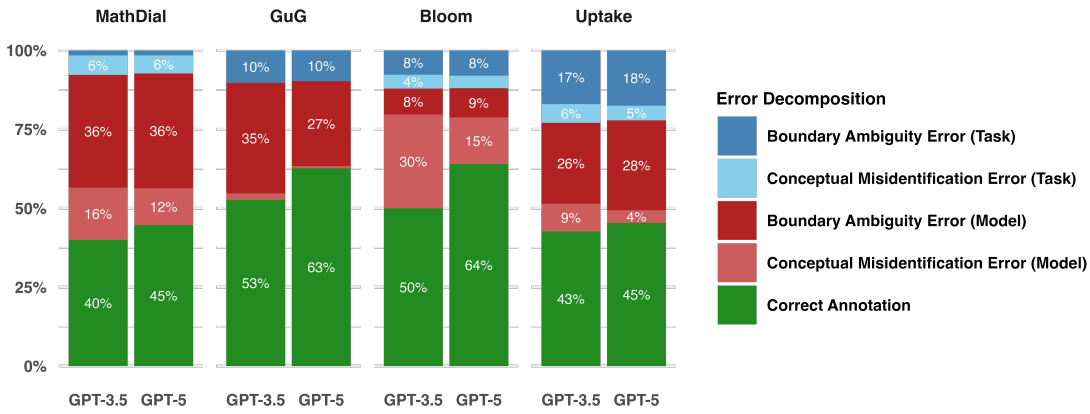

Fig. 4. **Comparison of Error Decomposition Between Two Models**

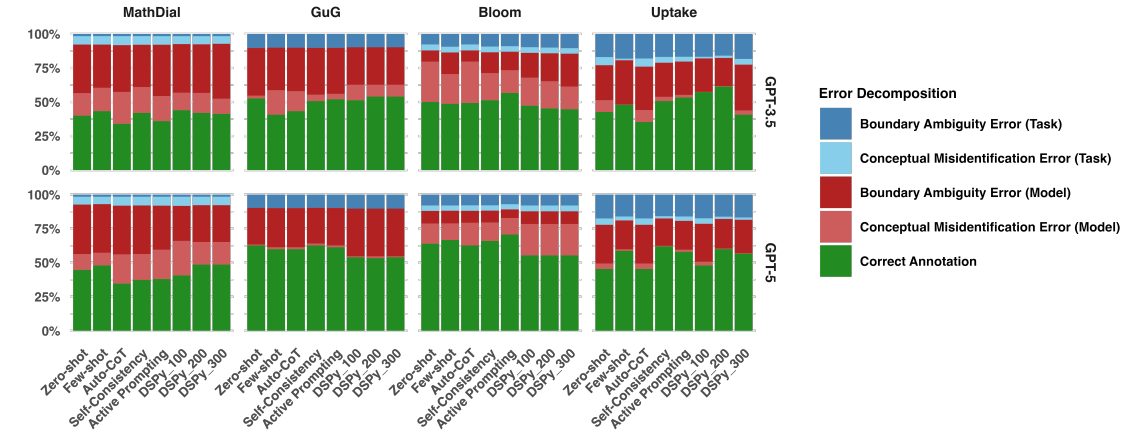

Fig. 5. **Error Decomposition Across Prompting Strategies**

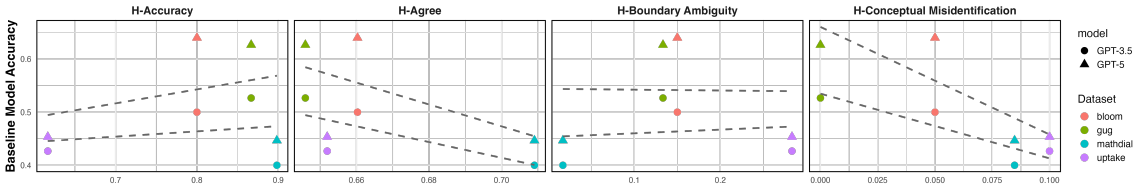

Fig. 6. **Baseline (Zero-Shot) Model Accuracy vs. Task Inherent Errors**

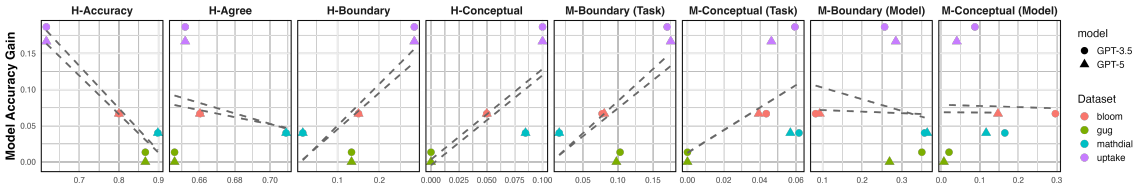

Fig. 7. **Prompting Strategies Effectiveness vs. Task Inherent Errors**

D.2 Human annotations: 40 items

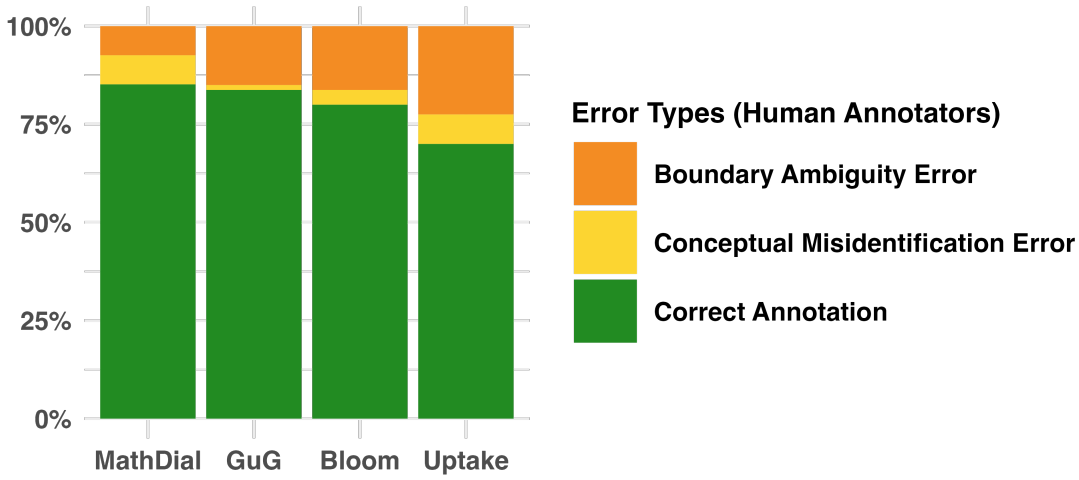

Fig. 8. **Distribution of error types observed in human annotations**

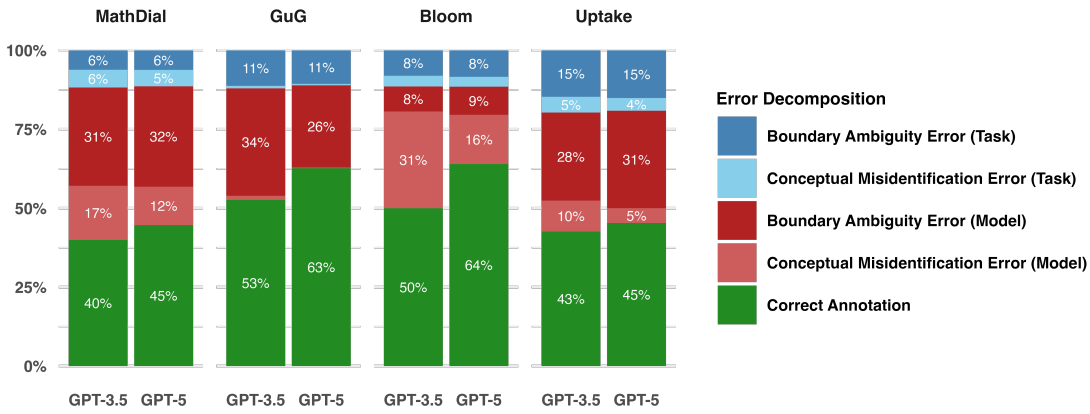

Fig. 9. **Comparison of Error Decomposition Between Two Models**

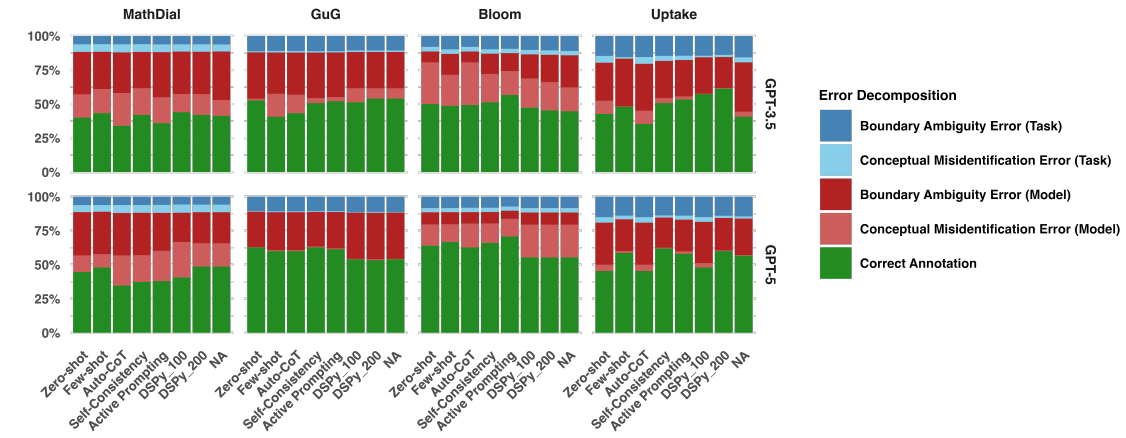

Fig. 10. **Error Decomposition Across Prompting Strategies**

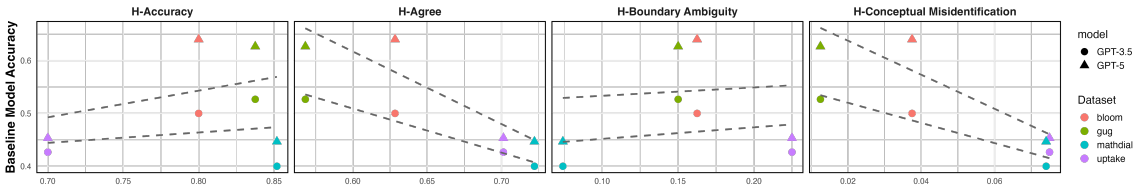

Fig. 11. **Baseline (Zero-Shot) Model Accuracy vs. Task Inherent Errors**

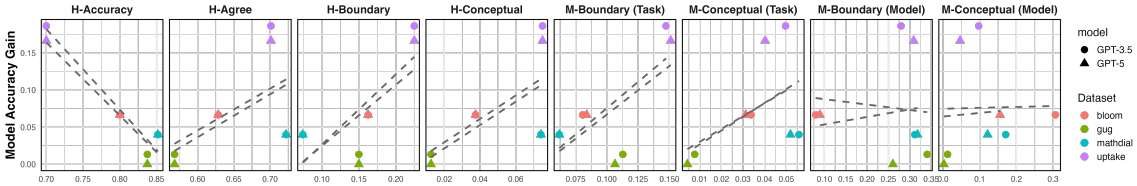

Fig. 12. **Prompting Strategies Effectiveness vs. Task Inherent Errors**