TutorCaseBench: A Case-Grounded Benchmark for Evaluating LLM Tutors in Supporting Students with Challenging Behaviors

Anonymous ACL submission

Abstract

LLMs are increasingly used as AI tutors, yet existing benchmarks overlook the *challenging behaviors* common among students with learning and behavioral difficulties. We introduce TutorCaseBench, a benchmark that grounds AI-tutor evaluation in qualitative case studies from pedagogy, educational psychology, and special education. Our pipeline converts documented student categories into case-specific scenarios, tutor–student dialogues, and rubrics grounded in source evidence. The benchmark contains 621 dialogues across five challenge types, evaluated through open-ended response generation (1,080 items) and multiple-choice selection (1,727 items). Across 11 state-of-the-art LLMs, even the strongest model satisfies a rubric criterion in only 57.1% of open-ended cases, while models are better at selecting appropriate strategies than generating situated responses. These results reveal a gap between intervention recognition and context-specific adaptation, especially for open-weight models. All data are publicly available at `https://anonymous.4open.science/r/TutorCaseBench-anonymous`.

1 Introduction

Large language models (LLMs) are increasingly deployed as conversational AI tutors for scalable, individualized instruction (Jurenka et al., 2024). Yet, effective tutoring requires more than providing correct answers: a tutor who misreads emotion, fails to re-engage, or escalates refusal can demotivate the student, making affective and behavioral factors as consequential as cognitive ones (Baker et al., 2010; D'Mello and Graesser, 2012; Karumbaiah et al., 2023). This is especially important for students with learning and behavioral difficulties, whose struggles often appear as refusal, withdrawal, emotional distress, or social conflict (Sugai and Horner, 2006). We use *challenging behaviors* to refer to such disruptive or disengaged responses, which make tutoring difficult to sustain and require support tailored to the learner and context (Pepe and Addimando, 2013; Patnaik and Subban, 2023).

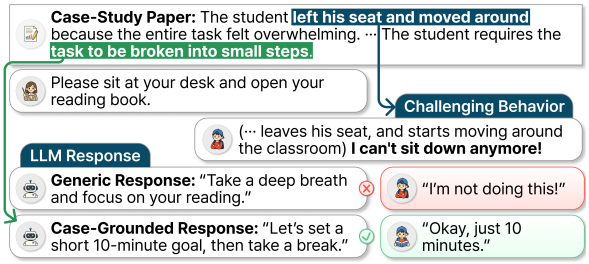

Figure 1: We introduce TutorCaseBench, a case-grounded benchmark grounded in qualitative case-study literature for evaluating whether LLM tutor responses adapt appropriately to challenging student behaviors.

Existing AI-tutor benchmarks (Macina et al., 2023, 2025; Sonkar et al., 2023) remain centered on cognitive aspects of tutoring, such as math-error detection and scaffolding. Their inputs often depict idealized student behavior rather than disengagement, resistance, or other challenging behaviors (Macina et al., 2023; Markel et al., 2023; Liu et al., 2024b), while their rubrics reward fixed pedagogical preferences (Shridhar et al., 2022; Liu et al., 2024a). As a result, they provide limited evidence on whether AI tutors can deliver case-sensitive support for students with challenging behaviors.

Our work grounds this evaluation in a resource the NLP community has so far overlooked: qualitative case literature from pedagogy, educational psychology, and special education. These studies document individual students, their specific behaviors in learning contexts, and interventions that practitioners observed to be effective or ineffective. Unlike hypothetical personas, such cases provide evidence about student backgrounds, precipitating events, and practitioner responses, allowing us to ground realistic, challenging pedagogical scenarios in empirically attested interventions.

We introduce an automated LLM pipeline that converts qualitative case-study papers into struc-

tured tutoring evaluation cases. For each documented student category, the pipeline extracts verbatim learner and instructional evidence and constructs scenarios with learning contexts, challenging behaviors, and intervention strategies. Each scenario is then converted into a grounded tutor–student dialogue ending with the student exhibiting challenging behavior, paired with a rubric derived from the documented interventions. The resulting items are evaluated in *immediate* and *long-term* settings through an LLM-as-a-judge checklist rubric (Zheng et al., 2023) and a multiple-choice task in the style of Hendrycks et al. (2021).

Applying this pipeline to peer-reviewed case-study papers, we construct TutorCaseBench, a benchmark of 621 dialogues across five challenge types: Emotional Distress, Task Resistance, Social-Interpersonal Difficulty, Attention/Engagement Failure, and Academic Struggle. It contains 1,080 open-ended and 1,727 multiple-choice items.

We evaluate 11 state-of-the-art LLMs on TutorCaseBench and find that current models remain limited in providing case-sensitive support. Even the strongest model satisfies a documented intervention in only 57.1% of open-ended cases, while models perform better when selecting among provided strategies than generating situated tutor responses. This suggests a gap between recognizing appropriate interventions and adapting them to specific student contexts, especially for open-weight models, which often fall back on generic responses.

Contributions.

1. A case-grounded benchmark construction pipeline that converts qualitative pedagogical case studies into tutoring scenarios, dialogues, and case-specific evaluation criteria.

2. TutorCaseBench, a benchmark of 621 case-grounded dialogues covering five types of challenging student behavior, with both open-ended and multiple-choice evaluation settings.

3. An empirical analysis of 11 LLM tutors, showing broad difficulty with case-specific support and a tendency among open-weight models toward generic responses.

2 Related Work

2.1 Literature-Grounded Student Simulations

AI tutor evaluation requires realistic student–tutor interactions, but collecting them at scale is difficult due to privacy and logistical constraints (Macina et al., 2023). LLM-based student simulation offers a scalable alternative, typically modeling either learner knowledge and reasoning, such as misconceptions and knowledge-component states (Macina et al., 2023; Lu and Wang, 2024; Qi et al., 2025), or learner traits, such as personas and motivational profiles (Markel et al., 2023; Liu et al., 2024b; Jin et al., 2025; Marquez-Carpintero et al., 2025).

However, even fine-tuned models show limited alignment with real student behavior (Scarlatos et al., 2026). Simulated students also rarely express frustration or disengagement, remain unnaturally attentive and linguistically sophisticated, and accept tutor feedback without resistance (Martynova et al., 2025). Thus, existing frameworks mostly simulate idealized student behavior, overlooking disengaged, resistant, or off-task behaviors.

Source documents offer another way to ground dialogue synthesis. Prior work shows that document grounding improves coherence and factual consistency (Feng et al., 2020, 2021), and supports persona simulation when documents describe behavior and internal states (Wang et al., 2025). In education, textbook-grounded dialogue generation improves content accuracy (Wang et al., 2024a), but textbooks provide limited detail on student behavior. We instead ground synthesis in pedagogy, educational psychology, and special education literature, enabling dialogues with realistic challenging behaviors beyond cooperative learner settings.

2.2 Evaluation of AI Tutors

AI tutor evaluation has shifted from general-purpose NLG metrics to pedagogy-grounded frameworks. Earlier approaches relied on BLEU (Papineni et al., 2002) and BERTScore (Zhang et al., 2019), comparing tutor responses against reference answers, but such metrics poorly capture pedagogical quality because multiple valid responses may exist (Tack and Piech, 2022; Jurenka et al., 2024).

To address this, prior work has proposed pedagogy-grounded criteria. Tack and Piech (2022) found that AI responses fell short of human teachers in teacher-likeness, student understanding, and helpfulness. Later frameworks introduced criteria such as coherence, correctness, equitable tutoring (Macina et al., 2023), as well as usefulness, care, and human-likeness (Wang et al., 2024b). However, these criteria remain fragmented and lack consensus (Jurenka et al., 2024). Recent benchmarks have begun to consolidate them, including

MRBench's eight-dimensional taxonomy (Maurya et al., 2025) and MathTutorBench's focus on mathematical expertise, student understanding, and scaffolding quality (Macina et al., 2025).

Despite this progress, existing benchmarks largely use fixed criteria, often in mathematical error-remediation settings. Yet support for students exhibiting challenging behaviors often depends on the specific student context. We therefore take a case-specific approach, deriving evaluation criteria from evidence-based interventions in pedagogical case studies rather than applying a unified rubric.

3 Data Construction

We construct TutorCaseBench from case-study papers documenting diverse learning and behavioral challenges, using an automated LLM pipeline to convert them into structured tutor–student dialogues for evaluation. The benchmark includes two tasks: open-ended response generation and multiple-choice question (MCQs) selection.

We use `gemini-3-flash-preview` (Google Deepmind, 2025) for extractive stages (Stage 1–2) and `gemini-3.1-pro-preview` (Google Deepmind, 2026) for generative stages (Stage 3–5), with default thinking level `high` and `temperature=0`. Inputs that triggered Gemini's safety policy were processed with `gpt-5.5` (OpenAI, 2026a) instead.

3.1 Paper Collection

Stage 0: Paper collection. We first collect candidate case-study papers from the ERIC database[1] by querying the ERIC API[2] for papers that include at least one student *and* one methodology descriptor. Full descriptor list used for filtering is provided in Appendix A. We restrict publications to 2010–2026 to keep the corpus contemporary while ensuring sufficient volume. Since descriptor matching alone does not guarantee that a paper qualitatively describes student classroom behavior with actionable teacher insights, we apply additional filtering based on their abstracts (see Appendix B for details).

3.2 Open-Ended Response Task

TutorCaseBench's open-ended response task evaluates whether a model produces a next tutor response aligned with documented interventions for a tutor–student dialogue grounded in case studies.

[1] https://eric.ed.gov/
[2] https://eric.ed.gov/?api

3.2.1 Tutor–Student Dialogue Construction

Stages 1–4 convert the collected papers into tutor–student dialogues paired with a hidden answer key of effective and ineffective interventions. Figure 2 provides an overview of this pipeline.

Stage 1: Student Category Extraction. Starting from the initial set of papers, the LLM rates how concretely each paper documents three dimensions essential to our benchmark—*student behavior*, *learning context*, and *teacher actions*—on a 1–5 scale (prompt in Appendix N.2). Papers scoring below 3 on any dimension are filtered out. We set this cutoff after inspecting samples and finding them too general to ground a concrete dialogue.

For papers that pass screening, the model then extracts *student categories*—distinct student profiles within the paper, each with a label, rationale, demographics, and verbatim supporting excerpts for the three dimensions above (prompt in Appendix N.3). Rather than treating each paper as a single case, we use each student category as a separate scenario-construction unit, since a paper may document multiple students or groups of students with different behaviors, contexts, or instructional needs. Categories are then filtered by the same criterion at a stricter threshold (3 or below), since downstream dialogue synthesis requires per-category evidence to be concrete on its own.

Stage 2: Evidence Extraction. Next, the LLM reads each paper from Stage 1 twice—once for learner evidence, once for instructional evidence—tagging each verbatim span with the categories it applies to. The first read extracts learner evidence along four dimensions—*observable behaviors*, *cognitive states*, *affective states*, and *difficulty triggers*—along with verbatim student quotes (prompt in Appendix N.4). These dimensions adapt Pintrich's self-regulated learning framework (Pintrich, 2000; Panadero, 2017), replacing its contextual axis with difficulty triggers to surface what specifically prompted the student's struggle.

The second read takes the paper and the prior learner spans as input, and extracts spans of instructional evidence, recording *when* the strategy applies, *how* it was enacted, and its documented *outcome* (prompt in Appendix N.5). Each instructional span is tagged as *effective* or *ineffective*, given a temporal scope of *immediate*, *long-term*, or *both*.

Stage 3: Scenario Construction. The LLM then composes one or more concrete *scenarios* per cate-

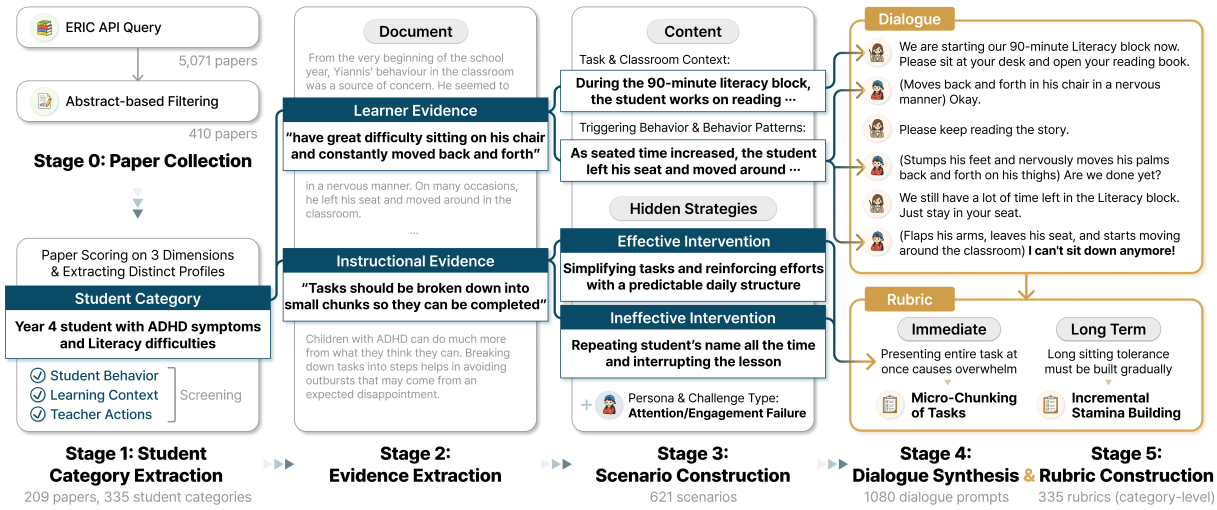

Figure 2: Overview of the TutorCaseBench data construction pipeline. Based on extracted learner and instructional evidence, we compose case-grounded tutor–student dialogues. Examples are adapted from Pantaleon (2016).

gory (prompt in Appendix N.6) with the evidence. Since the evidence within a single category often spans different tasks and classroom contexts, we construct multiple distinct scenarios per category each mapped with their evidence, rather than collapsing them into one generic scene. Scenarios are dropped when the source paper records only ineffective or no documented strategies.

Each scenario is organized as two parallel blocks: a *content* block (task, classroom context, triggering behavior, related behavior patterns) that frames what a tutor sees in the moment, paired with a hidden strategies block (immediate effective, long-term effective, and ineffective) that the model must rediscover from the content alone. Every field points back to Stage 2 spans, keeping the entire scenario grounded in the source case study.

At the category level, the model produces two additional artifacts. The *student persona* captures the learner's recurring patterns, triggers, and what tends to work or fail, ensuring consistent student turns across all scenarios in the category. The *primary challenge type* (task resistance, emotional distress, attention/engagement failure, social-interpersonal difficulty, or academic struggle) is assigned from a five-type dialogue-oriented taxonomy (see Appendix C for details), allowing us to test whether tutor models differ in handling qualitatively different types of learner difficulty.

Stage 4: Dialogue Synthesis. In the final stage, the LLM converts each scenario into a 6-turn tutor–student dialogue that ends with the student exhibiting the challenging behavior (prompt in Appendix N.7). To simulate a coherent and evidence-grounded student, the model is given the student *category* and *persona*, the scenario *content*, and the verbatim excerpts, and generates a dialogue in which every student turn is grounded in one or more excerpts. The *hidden strategies* block is excluded and tutor turns are constrained to be neutral, keeping the subsequent tutor's response open for evaluation. The model also produces a one-sentence *minimal context* containing only the task and setting, giving the responder the context needed to interpret the dialogue while leaving the student's needs to be inferred from the dialogue itself. In parallel, the full scenario content block is preserved as a *full context* field for downstream grading.

3.2.2 Open-Ended Response Evaluation

Stage 5: Rubric Construction. To evaluate open-ended tutor responses, we create case-specific rubrics by student category and response type (*immediate* or *long-term*) (prompt in Appendix N.8). Rather than scoring every response on fixed dimensions, each rubric is built from effective intervention strategies documented for that student in the source paper. We define rubrics at the category level because scenarios within a category share the same learner profile and intervention evidence.

Given the student persona, primary challenge type, synthesized dialogues, and hidden strategies from the category's scenarios, the model produces a checklist of criteria, each capturing a distinct evidence-supported response principle with a short rationale and check condition. Ineffective strategies are included as negative anchors, and a brief

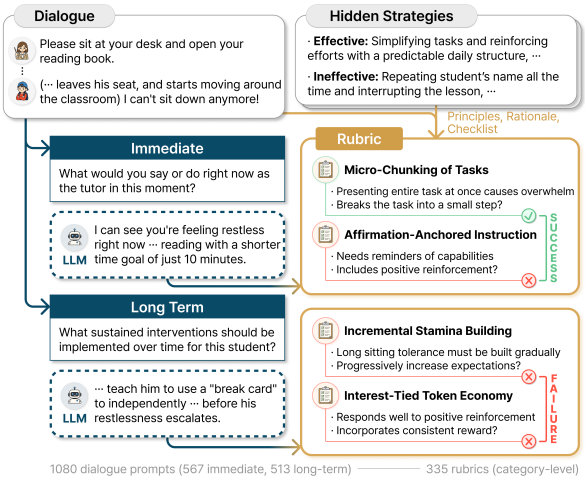

Figure 3: Overview of the rubric construction and evaluation process. Case-specific rubrics are built from intervention strategies documented per student category.

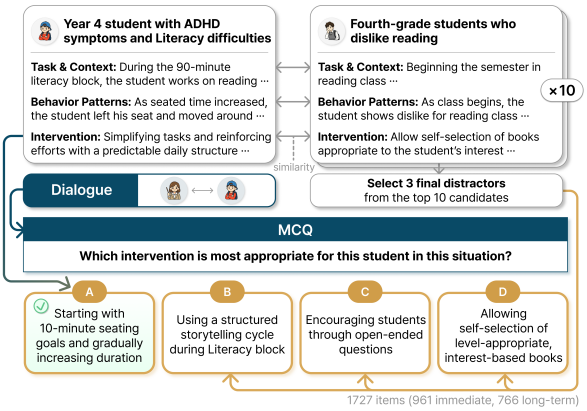

Figure 4: Each effective strategy yields a 4-option question with cross-scenario distractors. Examples from Pantaleon (2016), Lee (2012), Delony et al. (2013).

student description is produced as auxiliary context. Figure 3 provides an overview of the rubric construction and evaluation process.

Each criterion is judged independently against its check condition and scored as satisfied or not (0–1). Because criteria represent alternative evidence-supported principles rather than additive elements of a single ideal answer, satisfying any one criterion counts as success. Thus, a response receives credit for applying one supported principle well, without having to cover every strategy in the rubric.

3.3 MCQ Selection Task

To complement open-ended evaluation, we convert scenarios from §3.2.1 into 4-option MCQs. Each question uses one effective strategy from the target scenario as the correct answer and three effective strategies from other scenarios as distractors. Since all options are plausible interventions, the task tests whether a model can select the strategy that best fits the specific student and dialogue.

Figure 4 illustrates the construction process: we retrieve candidate distractors by embedding similarity, select three with an LLM judge, and rewrite all four options into a uniform surface form so the correct answer cannot be identified by style. We additionally construct a binary variant pairing one effective and one ineffective option from the same scenario as a sanity check. Full construction details for both variants are in Appendix D.

3.4 Data Statistics

Overall. Starting from 5,071 ERIC candidates, the pipeline produces 621 scenarios from 209 source papers and 335 student categories ($M = 1.85$ scenarios per category, $SD = 0.74$). The final benchmark includes 621 dialogues with 1,080 open-ended items: 567 immediate and 513 long-term. Category-level rubrics contain 2.83 immediate criteria ($SD = 0.84$) and 2.44 long-term criteria ($SD = 0.88$) on average. The MCQ set has 1,727 items: 961 immediate and 766 long-term.

Challenge-type distribution. Our taxonomy of student challenges (Appendix C) covers the dataset scenarios as follows: *Task Resistance* (28.5%), *Attention/Engagement Failure* (20.6%), *Academic Struggle* (17.4%), *Emotional Distress* (17.2%), and *Social-Interpersonal Difficulty* (16.3%).

3.5 Data Validation

We validated the grounding of the two generated artifacts used in downstream evaluation: dialogues as model inputs and rubrics as answer keys. Two authors independently scored a stratified 10% sample of dialogues ($N = 68$) and rubrics ($N = 36$): dialogue grounding was rated as whether the final student turn was supported by learner excerpts (0–1), while rubric grounding was rated by how fully the criteria were supported by instructional excerpts (1–3). Dialogue grounding reached 0.87 out of 1 with substantial agreement ($\kappa = 0.74$), and rubric grounding reached 2.86 out of 3 ($\kappa = 0.77$), with 86.1% of rubric ratings fully grounded and none unsupported. Details are provided in Appendix E.

4 Experiments

We evaluate 11 state-of-the-art models on TutorCaseBench across three tasks: open-ended response generation, MCQ selection, and reference-

based response generation. Experimental settings and model details are provided in Appendix F.

Open-ended response. Each model produces a free-form response to each open-ended item, either an immediate next-turn response or a long-term intervention plan (prompts in Appendix N.11). Immediate items ask what the tutor should say or do next, while long-term items ask what sustained interventions should be implemented over several weeks. In both cases, the model is shown only the scenario's minimal context and the dialogue up to the challenging student utterance, requiring it to infer the student's needs and respond appropriately.

We evaluate each response with `gemini-3.1-pro-preview` using the rubric from §3.2.2 (prompt in Appendix N.12). The evaluator receives the rubric, student description, dialogue, the scenario's full context, and the model response, and assigns each criterion a binary score (1 if satisfied, 0 otherwise). A response is marked successful if it satisfies at least one criterion.

MCQ selection. For each MCQ from §3.3, we provide the same minimal context and dialogue used in the open-ended setting, along with four answer options, and ask the model to select one (prompt in Appendix N.13). A response counts as correct if it matches the correct option, which is grounded in the scenario's effective strategy.

Reference-based response. We additionally create a reference-based variant to test whether models can produce better open-ended responses when relevant guidance is available. For each test dialogue, we retrieve the most similar scenarios from a separate pool and add their intervention strategies to the prompt as a reference (Figure 5). We evaluate the resulting responses with the same rubrics as the dialogue-only setting, enabling a paired comparison. Full construction details are in Appendix G.

5 Evaluation Results

5.1 Open-Ended Response

Overall Performance. Table 1 reports the success rate of open-ended responses (i.e., the proportion satisfying at least one criterion in the rubric). Overall performance remains limited even for the strongest models. Proprietary models occupy the top ranks, with Claude Opus 4.7 performing best at 57.1% and GPT-5.5 following at 52.5%. Among open-weight models, DeepSeek V4 Pro performs

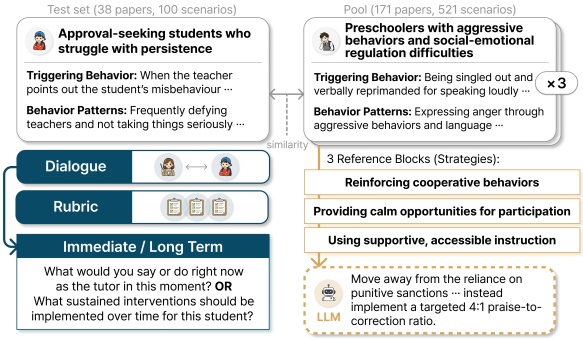

Figure 5: We use strategies from similar cases as references. Examples from Kwon et al. (2017), Bayat et al. (2010), Cassidy et al. (2018), Poon-McBrayer (2016).

best at 47.2%, while Llama 4 Maverick is lowest at 31.5%. These results indicate that generating pedagogically grounded responses for students with challenging behaviors remains difficult.

Table 1: Success rate of open-ended response (%)

Model	Immediate (N=567)	Long-Term (N=513)	Overall (N=1,080)
Claude Opus 4.7	68.3	44.8	57.1
Claude Haiku 4.5	54.9	41.9	48.7
Gemini 3.1 Pro	64.9	32.7	49.6
Gemini 3.0 Flash	56.8	32.2	45.1
GPT-5.5	64.0	39.8	52.5
GPT-5.4 Mini	47.8	32.2	40.4
Llama 4 Maverick	42.9	18.9	31.5
Gemma 4 31B	49.7	26.9	38.9
DeepSeek V4 Pro	53.8	40.0	47.2
GPT-OSS 120B	49.4	32.7	41.5
Qwen 3.5-397B	49.9	30.4	40.6

A consistent gap appears between immediate responses and long-term intervention plans. All models perform better on immediate responses, with the largest gap for Gemini 3.1 Pro (64.9% vs. 32.7%) and the smallest for Claude Haiku 4.5 (54.9% vs. 41.9%). This suggests that models are more capable of producing plausible in-the-moment responses, but struggle in formulating sustained interventions tailored to the student's needs.

Figure 6 breaks down open-ended performance by challenge type. Social-interpersonal difficulty has the lowest average performance and is the lowest-performing category for 9 of 11 models, suggesting that socially oriented student needs are especially difficult for models to address.

Diversity and Case-Specificity. To better understand model-level differences, we additionally test whether stronger models produce less templated, more case-specific responses. We measure lexical

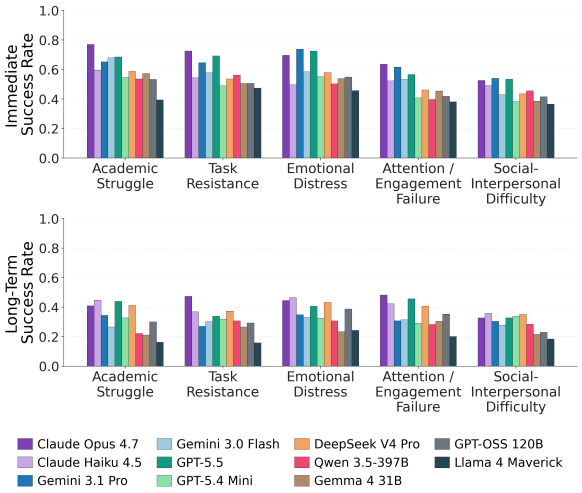

Figure 6: Success rate for open-ended response across models broken down by challenge type.

diversity using the distinct-n ratio—the fraction of unique n-grams in each model's responses—and compute Spearman correlations with success rates.

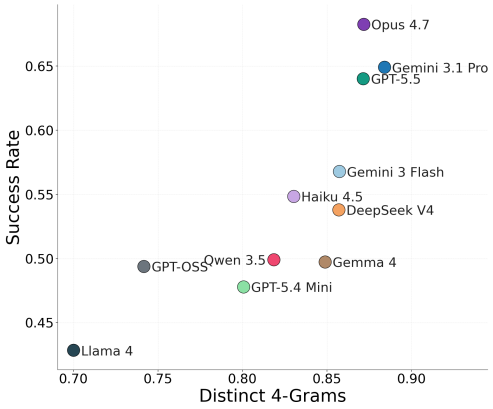

Figure 7: Correlation between distinct 4-grams ratio and immediate response success rate across models.

For immediate responses, diversity is strongly correlated with success across all tested phrase lengths, with $\rho = 0.918$ for 3-grams, $\rho = 0.936$ for 4-grams, and $\rho = 0.918$ for 5-grams (all $p < 0.001$). Figure 7 visualizes the 4-gram result, with the remaining results reported in Appendix H. These results suggest that higher-performing models are not merely relying on generic templates, but vary their responses more across cases.

For example, Llama 4 Maverick's most-used 4-gram ("take a deep breath") appears in 52.2% of its immediate responses, whereas the most-used 4-grams of Claude Opus 4.7, GPT-5.5, and Gemini 3.1 Pro appear in only 9–10% of their outputs, reflecting a more dispersed phrase distribution.

This pattern does not extend to long-term intervention plans. Across 3-, 4-, and 5-grams, distinct-n ratio is not significantly correlated with long-term success ($\rho = 0.475$, 0.406, and 0.470, respectively; all $p > 0.05$), suggesting that greater lexical diversity is not reliably associated with better performance for long-term intervention plans.

We next complement lexical diversity with an embedding-based measure of case-specificity. We embed each model response with `text-embedding-3-large` (OpenAI, 2024), retrieve its nearest non-self response, and check whether that response comes from the same dialogue case. The case-recovery rate is the fraction of responses for which this nearest neighbor comes from the same case, with higher values indicating more case-specific responses.

For immediate responses, case-recovery is strongly correlated with success rate across models (Spearman's $\rho = 0.800$, $p < 0.01$, $N = 11$; see Appendix I), suggesting that stronger models are better at tailoring their replies to the current student. However, this relationship consistently weakens for long-term plans ($\rho = 0.356$, $p > 0.05$).

Failure Analysis We qualitatively code a stratified 10% sample of open-ended cases where all models failed (25 immediate, 25 long-term), summarized in Table 2. We focus on the dominant pattern for each response type and describe the remaining failure modes in Appendix J.

The dominant immediate-response failure mode was *empathy/praise*. In one case (Thomas, 2017), an English-language learner stops revising her paragraph and says, "I afraid I can't." The case-prescribed response is to counter this expression of inability and affirm that she can do the work. All 11 models instead say variants of "it is completely normal to feel this way" and suggest working on one paragraph together. Although supportive on the surface, this leaves the self-doubt intact because the model never directly affirms her capability.

The dominant long-term failure mode was *generic scaffolding*. In one case (Hawkins et al., 2011), a student with autism grabs classmates' hair, especially when they wear sparkly hair grips. The case-prescribed response is to train on this trigger by having staff wear sparkly grips and rewarding looking without touching. All 11 models instead offer generic support, such as social-skills lessons, social stories, sensory diets, and peer buddies. These plans miss the central difficulty because the trigger—sparkly hair grips—never appears.

Table 2: Failure modes of open-ended cases missed by all models, by immediate and long-term responses ($N = 50$)

Failure Mode	**Description**	Imm.	LT
Empathy/Praise	Comforts the student when the response should challenge their belief or behavior.	36%	0%
Generic Scaffolding	Applies a standard support plan without identifying the student's central difficulty.	28%	64%
Over-Intervention	Steps in to correct or redirect when the tutor should step back instead.	20%	16%
Task Redesign Missed	Helps the student cope with a task that itself should be changed.	16%	20%

5.2 MCQ Selection

Table 3 reports accuracy on the MCQ task. Scores are higher and less variable than in open-ended generation, suggesting that models can recognize an appropriate intervention when it is provided as an option, even when they struggle to formulate one from scratch. Claude Opus 4.7 performs best at 66.0%, followed by Gemini 3.1 Pro at 64.1% and GPT-5.5 at 63.8%. The top open-weight model, Gemma 4 31B, reaches 59.9%, while GPT-OSS 120B is lowest at 53.3%, still above a 25% chance.

Table 3: Accuracy on 4-option MCQ cases (%).

Model	Immediate (N=961)	Long-Term (N=766)	Overall (N=1,727)
Claude Opus 4.7	66.7	65.1	66.0
Claude Haiku 4.5	62.0	54.3	58.6
Gemini 3.1 Pro	66.8	60.7	64.1
Gemini 3.0 Flash	59.8	52.3	56.5
GPT-5.5	67.7	58.9	63.8
GPT-5.4 Mini	60.5	53.3	57.3
Llama 4 Maverick	57.0	50.4	54.1
Gemma 4 31B	62.2	57.0	59.9
DeepSeek V4 Pro	57.6	54.6	56.3
GPT-OSS 120B	56.6	49.2	53.3
Qwen 3.5-397B	61.0	52.9	57.4

Challenge-type trends also mirror the open-ended setting, with social-interpersonal difficulty the weakest category for 10 of 11 models and academic struggle the strongest (see Appendix K).

These MCQ errors appear systematic rather than stochastic. 15.4% of questions are missed by all 11 models, a rate highest for Social-Interpersonal Difficulty. On these questions, models also often converge on the same distractor rather than failing independently (see Appendix L for details).

5.3 Reference-Based Response

We compare the dialogue-only open-ended prompt with a reference-based version supplemented with strategies from similar prior cases. Figure 8 shows the effect on immediate responses. Most models benefit from the references, with 10 of 11 models improving by an average of 6.7 percentage points. Qwen 3.5-397B improves the most, from 45.2% to 63.4%, while Claude Haiku 4.5 is the only model that decreases, from 62.4% to 59.1%.

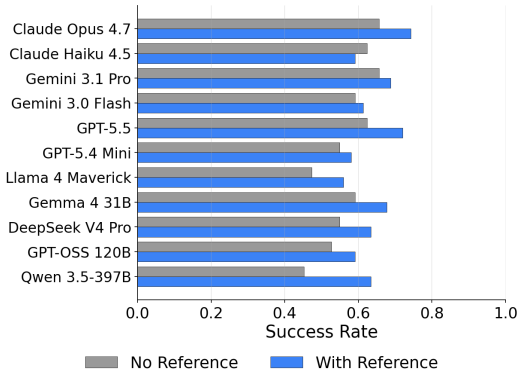

Figure 8: Success rate across models for immediate response, with and without references.

The reference effect is weaker for long-term plans, where only five models improve, with an average gain of 1.1 points (Appendix M). The contrast suggests that transfer works mainly when strategies map directly onto the next tutor move.

6 Conclusion

We introduce TutorCaseBench, a case-grounded benchmark for evaluating whether LLM tutors can support students who exhibit challenging behaviors. Rather than relying on idealized learner simulations or fixed pedagogical rubrics, TutorCaseBench converts qualitative case studies into tutor–student dialogues and case-specific evaluation criteria. Across 621 dialogues and 11 state-of-the-art LLMs, we find that current models remain limited in generating support that reflects documented interventions, even when they can often recognize appropriate strategies in multiple-choice settings. Our findings suggest that effective AI tutors should move beyond broadly plausible pedagogical responses toward context-sensitive adaptation grounded in individual student needs. Overall, TutorCaseBench provides a resource for developing tutors that address not only cognitive errors, but also students' affective and behavioral difficulties.

Limitations

Our work has several limitations. First, although TutorCaseBench is grounded in qualitative case-study literature, it is constructed through an automated LLM pipeline. Because the same Gemini model family is used across construction stages, the benchmark may reflect model-specific biases. We partially address this through human validation, but future work should test whether the benchmark remains stable with different construction models or more human-authored components.

Second, the benchmark is currently limited by its source corpus. We collect accessible, contemporary case-study papers from ERIC with sufficient evidence about student behavior, learning context, and teacher actions. While our LLM automated pipeline makes the benchmark readily expandable, the current version may still exclude relevant cases from other databases, disciplines, countries, age groups, or unpublished classroom practice.

Third, our evaluation is limited by the fixed dialogue structure. Each item is a six-turn tutor–student dialogue that ends with a challenging behavior, which supports controlled comparison but simplifies the variable length and pacing of real tutoring. We also score a single model response or short long-term plan, rather than observing how the tutor responds across multiple turns. Future work should evaluate more varied dialogue lengths and longer tutor–student interactions.

Lastly, we construct student categories separately within each paper, rather than grouping similar challenging-behavior categories across the corpus. This keeps each rubric grounded in case-specific intervention evidence, but it also means that broader patterns in challenging student behavior and effective support are not directly analyzed. Future work could cluster related categories to examine how intervention principles recur or vary across similar student profiles.

Ethical Considerations

Source documents were identified through ERIC (Education Resources Information Center), a U.S. Department of Education database for education research literature. We used only publicly accessible full-text materials, including documents hosted by ERIC and papers reachable through DOI or publisher links listed in ERIC records. Copyright remains with the original authors and publishers, and the materials were used only for non-commercial academic research. Our released benchmark consists only of LLM-synthesized tutor–student dialogues, rubrics, and multiple-choice items. We do not redistribute the full source texts or release intermediate extraction artifacts. Source excerpts used as examples in the paper are explicitly cited. For provenance, we provide source-paper metadata for the documents used in dataset construction.

Because TutorCaseBench focuses on students with learning and behavioral difficulties, some cases may include descriptions of distress, refusal, aggression, or other challenging behaviors. These behaviors are included only when grounded in the source literature and are used to evaluate supportive tutoring responses, not to stereotype or label students. We frame such behaviors as context-dependent student needs rather than as fixed traits.

We also acknowledge the use of AI assistants in preparing this work. We used ChatGPT[3] and Claude[4] for writing and coding assistance, as well as for generating icons in figures.

[3] https://chatgpt.com/
[4] https://claude.ai

References

Thomas M Achenbach. 1991. *Child behavior checklist/4-18*. University of Vermont, psychiatry.

Anthropic. 2025. System Card: Claude Haiku 4.5. Accessed: 2026-05-20.

Anthropic. 2026. System Card: Claude Opus 4.7. Accessed: 2026-05-20.

Ryan S.J.d. Baker, Sidney K. D'Mello, Ma. Mercedes T. Rodrigo, and Arthur C. Graesser. 2010. Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4):223–241.

Mojdeh Bayat, Gayle Mindes, and Sheryl Covitt. 2010. What does rti (response to intervention) look like in preschool? *Early Childhood Education Journal*, 37(6):493–500.

Claire Cassidy, Donald Christie, Helen Marwick, Lynn Deeney, Gillian McLean, and Kirsten Rogers. 2018. Fostering citizenship in marginalised children through participation in community of philosophical inquiry. *Education, Citizenship and Social Justice*, 13(2):120–132.

DeepSeek-AI. 2026. Deepseek-v4: Towards highly efficient million-token context intelligence.

Sheila Delony, Ellen Smith Morgan, and Kaitlyn Leah Howell. 2013. A case study of the role of reader response in two elementary classrooms. *Texas Journal of Literacy Education*, 1(1):1–12.

Sidney K. D'Mello and Arthur Graesser. 2012. AutoTutor and affective AutoTutor: Learning by talking with cognitively and emotionally intelligent computers that talk back. *ACM Transactions on Interactive Intelligent Systems*, 2(4):23:1–23:39.

Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling dialogues grounded in multiple documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6162–6176, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A goal-oriented document-grounded dialogue dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8118–8128, Online. Association for Computational Linguistics.

Google Deepmind. 2025. Gemini 3 Flash Model Card. Accessed: 2026-05-20.

Google Deepmind. 2026. Gemini 3.1 Pro Model Card. Accessed: 2026-05-20.

Google DeepMind. 2026. Gemma 4 model card. Accessed: 2026-05-20.

Susan Griebling, Peg Elgas, and Rachel Konerman. 2015. " trees and things that live in trees": Three children with special needs experience the project approach. *Early childhood research & practice*, 17(1):n1.

Emma Hawkins, Sheri Kingsdorf, Jackie Charnock, Mariann Szabo, Edi Middleton, Jo Phillips, and Grant Gautreaux. 2011. Using behaviour contracts to decrease antisocial behaviour in four boys with an autistic spectrum disorder at home and at school. *British Journal of Special Education*, 38(4):201–208.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR 2021)*.

Hyoungwook Jin, Minju Yoo, Jeongeon Park, Yokyung Lee, Xu Wang, and Juho Kim. 2025. Teachtune: Reviewing pedagogical agents against diverse student profiles with simulated students. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pages 1–28.

Irina Jurenka, Markus Kunesch, Kevin R. McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative AI for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.

Shamya Karumbaiah, Conrad Borchers, Tianze Shou, Ann-Christin Falhs, Pinyang Liu, Tomohiro Nagashima, Nikol Rummel, and Vincent Aleven. 2023. A spatiotemporal analysis of teacher practices in supporting student learning and engagement in an AI-enabled classroom. In *Artificial Intelligence in Education (AIED 2023)*, volume 13916 of *Lecture Notes in Computer Science*, pages 450–462. Springer.

Soonjung Kwon, Kristján Kristjánsson, and David I Walker. 2017. Misery in dark shadows behind the high achievement scores in south korean schooling: An ethnographic study. *Educational Review*, 69(2):201–217.

Sy-ying Lee. 2012. Storytelling supported by technology: An alternative for efl children with learning difficulties. *Turkish Online Journal of Educational Technology-TOJET*, 11(3):297–307.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024a. SocraticLM: Exploring Socratic personalized teaching with large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, pages 85693–85721.

Zhengyuan Liu, Stella Xin Yin, Geyu Lin, and Nancy F. Chen. 2024b. Personality-aware student simulation

for conversational intelligent tutoring systems. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 626–642, Miami, Florida, USA. Association for Computational Linguistics.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM conference on learning@ Scale*, pages 16–27.

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.

Jakub Macina, Nico Daheim, Ido Hakimi, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2025. MathTutorBench: A benchmark for measuring open-ended pedagogical capabilities of LLM tutors. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 204–221, Suzhou, China. Association for Computational Linguistics.

Julia M. Markel, Steven G. Opferman, James A. Landay, and Chris Piech. 2023. GPTeach: Interactive TA training with GPT-based students. In *Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S '23)*, pages 226–236. ACM.

Luis Marquez-Carpintero, Alberto Lopez-Sellers, and Miguel Cazorla. 2025. Simulating students with large language models: A review of architecture, mechanisms, and role modelling in education with generative ai. *arXiv preprint arXiv:2511.06078*.

Daria Martynova, Jakub Macina, Nico Daheim, Nilay Yalcin, Xiaoyu Zhang, and Mrinmaya Sachan. 2025. Can LLMs effectively simulate human learners? teachers' insights from tutoring LLM students. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 100–117, Vienna, Austria. Association for Computational Linguistics.

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.

Meta. 2025. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Accessed: 2026-05-06.

OpenAI. 2024. New embedding models and API updates. Accessed: 2026-05-20.

OpenAI. 2025a. gpt-oss-120b gpt-oss-20b Model Card. Accessed: 2026-05-20.

OpenAI. 2025b. Introducing GPT-4.1 in the API. Accessed: 2026-05-20.

OpenAI. 2026a. GPT-5.5 System Card. Accessed: 2026-05-20.

OpenAI. 2026b. Introducing GPT-5.4 mini and nano. Accessed: 2026-05-20.

Ernesto Panadero. 2017. A review of self-regulated learning: Six models and four directions for research. *Frontiers in Psychology*, Volume 8 - 2017.

Anastasia Pantaleon. 2016. Dealing with adhd in a greek primary school. *Journal of Research in Special Educational Needs*, 16:341–344.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Susmita Patnaik and Pearl Subban. 2023. Challenging behaviours in primary classrooms: examining causal attributions and strategies used by primary school teachers in india. *Australasian Journal of Special and Inclusive Education*, 47(2):78–95.

Alessandro Pepe and Loredana Addimando. 2013. Comparison of occupational stress in response to challenging behaviours between general and special education primary teachers in northern italy. *International journal of special education*, 28(1):14–26.

Paul R. Pintrich. 2000. Chapter 14 - the role of goal orientation in self-regulated learning. In Monique Boekaerts, Paul R. Pintrich, and Moshe Zeidner, editors, *Handbook of Self-Regulation*, pages 451–502. Academic Press, San Diego.

Kim Fong Poon-McBrayer. 2016. Complexities of shared ethnicity, immigrant education, and disabilities: Reconceptualizing multicultural special education. *International Journal of Special Education*, 31(3):n3.

Changyong Qi, Longwei Zheng, Anna He, Haoxin Xu, Linzhao Jia, Yuang Wei, Bingqian Jiang, and Xiaoqing Gu. 2025. Simulating student learning behaviors with llm-based role-playing agents: A data-driven and cognitively inspired framework. *Expert Systems with Applications*, page 130753.

Alexander Scarlatos, Jaewook Lee, Simon Woodhead, and Andrew Lan. 2026. Simulated students in tutoring dialogues: Substance or illusion? *arXiv preprint arXiv:2601.04025*.

Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shashank Sonkar, Naiming Liu, Debshila Mallick, and Richard Baraniuk. 2023. CLASS: A design framework for building intelligent tutoring systems based on learning science principles. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1941–1961, Singapore. Association for Computational Linguistics.

George Sugai and Robert H. Horner. 2006. A promising approach for expanding and sustaining school-wide positive behavior support. *School Psychology Review*, 35(2):245–259.

Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.

Suneeta Thomas. 2017. Journalogue: Voicing student challenges in writing through a classroom blog. *Educational Technology and Society*, 20:112–122.

Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.

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.

Xintao Wang, Heng Wang, Yifei Zhang, Xinfeng Yuan, Rui Xu, Jen-tse Huang, Siyu Yuan, Haoran Guo, Jiangjie Chen, Shuchang Zhou, and 1 others. 2025. Coser: Coordinating llm-based persona simulation of established roles. In *Forty-second International Conference on Machine Learning*.

Sarah J Watt, William J Therrien, and Erica R Kaldenberg. 2014. Meeting the diverse needs of students with ebd in inclusive science classrooms. *Beyond Behavior*, 23(2):14–19.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 Technical Report. *Preprint*, arXiv:2505.09388.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track*.

A ERIC Search Descriptors

Table 4: Descriptors used to retrieve candidate papers.

Category	Descriptors
Student (35 Total)	Behavior Problems Behavior Disorders Self Destructive Behavior Aggression Emotional Problems Emotional Disturbances Anxiety Disorders Depression (Psychology) School Phobia Withdrawal (Psychology) Fear of Success Mental Disorders Personality Problems Psychosomatic Disorders Attention Deficit Disorders Attention Deficit Hyperactivity Disorder Autism Spectrum Disorders Developmental Disabilities Intellectual Disability Neurodevelopmental Disorders Neurological Impairments Developmental Delays Learning Problems Learning Disabilities Low Achievement Academic Failure Reading Difficulties Reading Failure Writing Difficulties Mathematics Anxiety Slow Learners At Risk Students Communication Problems Communication Disorders Language Impairments
Methodology (4 Total)	Case Studies Observation Qualitative Research Classroom Research

B Abstract-Based Paper Filtering Procedure

The candidate papers are initially screened by `gpt-4.1-mini` (OpenAI, 2025b) using the title and abstract. We use this as a cost-efficient, high-recall first pass, since abstracts provide enough signal to remove clearly unfit papers while leaving ambiguous cases to later pipeline stages.

The filter applies four criteria, filtering out papers that fail to meet our benchmark requirements: (1) papers without direct student observation or interview, which lack grounded learner behavior; (2) quantitative studies, which report aggregate effects instead of the individual student responses we need; (3) papers focused on educational policy or measurement-tool development, which describe students only as targets of analysis; and (4) papers without practical guidance for teachers, which lack the prescriptive strategies needed for evaluation. A paper is retained only if rejected by none of the four. The full prompt is in Appendix N.1.

We additionally exclude excessively long papers ($\geq$ 30 pages) and those that cannot be accessed or downloaded, as we observed that long papers tend to mistakenly introduce literature-review content as learner evidence in downstream extraction.

C Dialogue-Oriented Taxonomy of Learner Challenges

To distinguish the challenge type associated with each student category, we designed a taxonomy that identifies the observable challenge a tutor must respond to in the tutoring interaction. The taxonomy consists of five categories: *Task Resistance*, *Emotional Distress*, *Attention / Engagement Failure*, *Social-Interpersonal Difficulty*, and *Academic Struggle*. It was designed and refined by an author with a doctoral degree in education.

We use CBCL/ASEBA (Achenbach, 1991) as a high-level behavioral scaffold. CBCL/ASEBA is an empirically based framework that characterizes children's functioning through standardized behavior descriptions. Since our benchmark is dialogue-based, we retain only constructs observable in tutoring interactions and reorganize them by the type of challenge they present to the tutor.

Specifically, the Internalizing grouping is adapted as *Emotional Distress*; Attention Problems as *Attention/Engagement Failure*; Social Problems and Aggressive Behavior as *Social-Interpersonal Difficulty*; and Externalizing behaviors directed at the learning task as *Task Resistance*. *Academic Struggle* adapts the School scale for content-level difficulty such as confusion, uncertainty, repeated errors, misconceptions, and failure to respond.

D MCQ Construction Details

D.1 4-Option MCQs

We provide additional details on distractor retrieval, selection, and option adaptation for the MCQ task.

We construct distractors in three steps: candidate retrieval, distractor selection, and surface adaptation. As a scenario may contain multiple effective strategies, each strategy is treated as a separate correct option and receives its own distractor set.

For retrieval, we embed each scenario's learning context, learner behavior, and intervention descriptions using `text-embedding-3-large` (OpenAI, 2024). Each candidate intervention is scored as:

$$\text{score} = \text{ctx_sim} \times (1 - \text{behav_sim}) \times (1 - \text{max_interv_sim}),$$

where `ctx_sim` and `behav_sim` are cosine similarities between the target and candidate learning contexts and learner behaviors, and `max_interv_sim` is the maximum similarity between the candidate intervention and any of the target's effective interventions. This favors candidates that fit the target's classroom context but address a different learner problem with a different intervention mechanism. We exclude strategies from the same paper or student category and keep only the highest-scoring intervention per scenario, ensuring that distractors come from distinct cases. The top 10 candidates are forwarded to the next step.

However, embedding similarity alone cannot reliably identify contextual incompatibilities or paraphrased versions of valid target interventions. Hence, we use `gemini-3.1-pro-preview` (Google Deepmind, 2026) as an LLM judge (Zheng et al., 2023) to select three final distractors from the top 10 candidates, prioritizing plausible but incorrect tutor actions that are structurally compatible with the target setting (prompt in Appendix N.9).

Finally, the same model rewrites all four options—including the correct one—into the target dialogue's vocabulary, length, and tone while preserving each option's underlying mechanism (prompt in Appendix N.10). This prevents models from identifying the correct answer through lexical or stylistic match with the dialogue. Options are shuffled with a fixed seed and labeled A–D, and each MCQ is presented with the dialogue and a prompt asking the model to choose the most appropriate intervention (see Appendix N.13).

D.2 Binary MCQs

We additionally construct a binary MCQ variant as a sanity check. Unlike the 4-option MCQs in §3.3, where all options are effective strategies and the task is to identify the best fit, the binary variant presents one effective and one ineffective strategy from the target scenario. Since the ineffective option failed for the same student in the source case, this variant tests whether models can reject a clearly inappropriate response. High accuracy therefore helps rule out this basic failure mode and supports interpreting the 4-option task as the harder problem of choosing among plausible effective strategies.

D.2.1 Construction

For each scenario and response type (*immediate* or *long-term*), we create one item when the scenario contains at least one matching effective strategy and at least one ineffective strategy. When multiple options are available on either side, we use `gemini-3.1-pro-preview` (Google Deepmind, 2026) as an LLM judge (Zheng et al., 2023) to select the final effective–ineffective option pair (prompt in Appendix N.14). The judge prioritizes pairs with matching scope, concrete tutor actions, and comparable granularity, where the effective option directly addresses the dialogue and the ineffective option is a plausible misstep rather than an unrelated action. If each side has only one eligible strategy, the pair is fixed and the judge is skipped.

Because effective and ineffective strategies may differ in style, the same model then rewrites the two options to match in vocabulary, length, and tone while preserving their underlying meanings (prompt in Appendix N.15). This step is constrained to make minimal edits, avoid unsupported additions, and keep the ineffective option substantively ineffective. The two options are then assigned to A and B using a fixed random seed.

D.2.2 Results

Across 11 models and 540 binary MCQs, accuracy ranges from 85.0% to 94.3%, far above the 50% chance baseline (Figure 9). This suggests that models can reliably distinguish a documented effective intervention from an ineffective one when the contrast is explicit. Thus, the lower performance in the 4-option MCQ setting is not due to a basic inability to reject ineffective responses, but to the harder task of selecting the best-fitting intervention among plausible alternatives.

E Human Validation Details

To verify that the generated dialogues and rubrics remain faithful to the source case studies, we validate the grounding of each in its supporting evidence. We focus on these two endpoints since these are what downstream evaluation directly depends on: dialogues serve as input to tutor models, and rubrics serve as their answer keys.

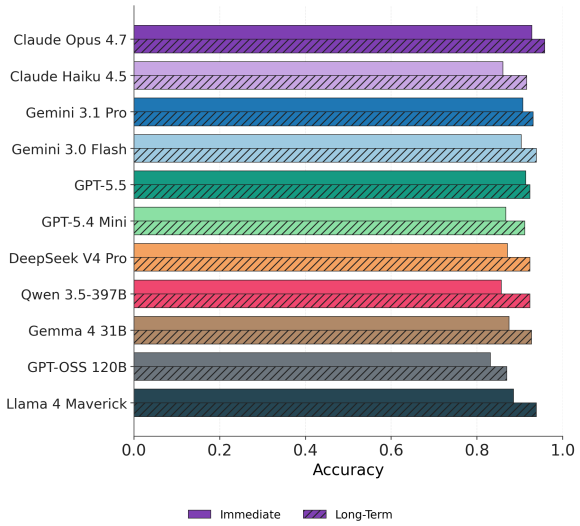

Figure 9: Binary MCQ accuracy across 11 models.

Dialogue grounding. We check whether each dialogue's final turn, where the student exhibits the challenging behavior, is grounded in the scenario's learner excerpts, scored on a 0–1 scale (0 = unsupported or based on an unsuitable excerpt, 1 = supported by at least one suitable excerpt).

Rubric grounding. We also check whether each category's rubric criteria are supported by the instructional excerpts pooled across its scenarios, scored on a 1–3 scale (1 = none supported, 2 = partially supported, 3 = all supported).

Procedure and results. Two authors independently score a 10% sample of dialogues ($N = 68$) and rubrics ($N = 36$), stratified by the five student challenge types. We report the average groundedness across raters along with inter-rater reliability (Cohen's κ). Dialogue grounding reaches a mean score of 0.87 out of 1 with $\kappa = 0.74$, and rubric grounding reaches a mean score of 2.86 out of 3 with $\kappa = 0.77$, with 86.1% of ratings marked as fully grounded and none marked as unsupported. These results indicate that the generated artifacts remain well grounded in the source case studies.

F Models

We evaluate a diverse set of 11 state-of-the-art models on TutorCaseBench: six proprietary (`claude-opus-4.7` (Anthropic, 2026), `claude-haiku-4.5` (Anthropic, 2025), `gemini-3.1-pro-preview` (Google Deepmind, 2026), `gemini-3-flash-preview` (Google Deepmind, 2025), `gpt-5.5` (OpenAI, 2026a), `gpt-5.4-mini-2026-03-17` (OpenAI, 2026b)) and five open-weight (`DeepSeek-V4-Pro` (DeepSeek-AI, 2026), `gemma-4-31B-it` (Google DeepMind, 2026), `gpt-oss-120b` (OpenAI, 2025a), `Llama-4-Maverick-17B-128E-Instruct` (Meta, 2025), `Qwen3.5-397B-A17B` (Yang et al., 2025)).

We accessed each model through its deployed API. Claude models and Llama were served through Vertex AI[5], Gemini models and Gemma through the Google Gemini API[6], GPT models through the OpenAI API[7], and DeepSeek, GPT-OSS, and Qwen through Together AI[8]. In all settings, models are queried with API defaults to evaluate them in their deployed configuration.

G Reference-Based Response Details

We construct the reference pool using a paper-level split stratified by challenge type, targeting 15% of scenarios for testing. This yields a *pool* of 171 papers and 521 scenarios and a *test* set of 38 papers and 100 scenarios, while ensuring that no test dialogue receives a reference from its source paper. We embed each scenario's challenging-behavior trigger and behavior patterns with `text-embedding-3-large` (OpenAI, 2024) and rank pool scenarios by cosine similarity. For each test dialogue, we retrieve the ten nearest pool scenarios, filter to those with an effective strategy for the target response type (immediate or long-term), and keep the three closest matches. Their matching effective and ineffective strategies are concatenated into a reference block, framed as examples from similar but different students that the responder may adapt, combine, or ignore (prompt in Appendix N.16). We evaluate the resulting responses with the same rubrics used in the open-ended experiment, enabling a direct paired comparison against the dialogue-only condition.

[5] https://docs.cloud.google.com/vertex-ai/docs
[6] https://ai.google.dev/gemini-api/docs
[7] https://developers.openai.com/api/docs
[8] https://docs.together.ai/intro

H Lexical Diversity

Figures 10 and 11 report the 3- and 5-gram variants of the distinct-n analysis for immediate responses. Together with the 4-gram result in Figure 7, they confirm that higher lexical diversity is associated with higher immediate response success rates.

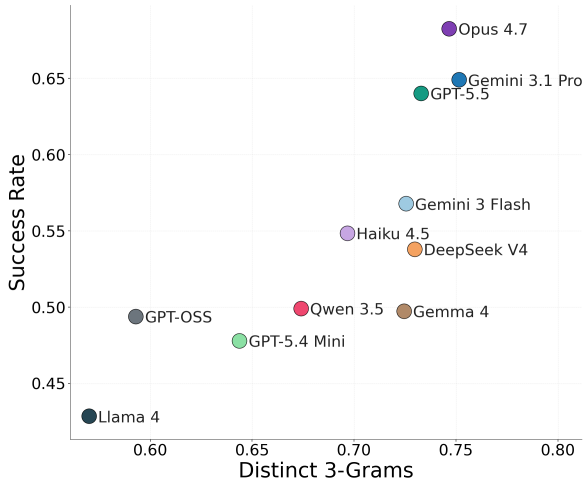

Figure 10: Correlation between distinct-n ($n = 3$) and immediate response success rate across models.

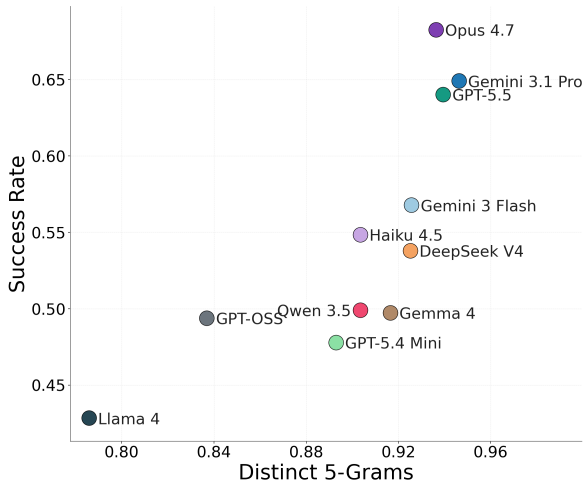

Figure 11: Correlation between distinct-n ($n = 5$) and immediate response success rate across models.

I Case Specificity

Figure 12 shows the relationship between case-recovery rate and immediate response success rate. The two are strongly correlated across models (Spearman's $\rho = 0.800$, $p < 0.01$, $N = 11$), suggesting that stronger models produce responses more specific to the current dialogue.

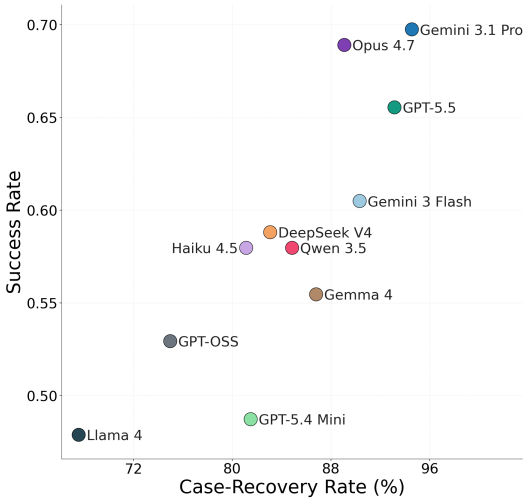

Figure 12: Correlation between case-recovery rate and immediate response success rate across models.

J Additional Open-Ended Failure Modes

In §5.1, we discuss the most frequent failure mode for each response type. Here, we provide examples for the remaining two themes in Table 2.

Over-Intervention. In this failure mode, models step in when the case requires the tutor to step back. In one case (Griebling et al., 2015), a student listens to bird calls during circle time and mimics them accurately on a small accordion he brought from home. The case-prescribed response is to accept the accordion as his way of participating in the activity. All 11 models instead praise him and then take the accordion away until circle time ends, removing the object that let him join the lesson.

Task Redesign Missed. In this failure mode, models help the student cope with a task that should be changed. In one case (Watt et al., 2014), a student has severe outbursts whenever he is asked to read aloud from the science textbook. The case-prescribed response is to stop using textbook reading to teach science and instead use hands-on experiments, observation, and peer-led activities. All 11 models keep the textbook and add supports such as audiobooks or vocabulary preview. These supports ease the task but leave the trigger in place.

K MCQ Selection by Challenge Type

Figure 13 shows MCQ accuracy by challenge type. As in the open-ended setting, Social-Interpersonal Difficulty is the weakest category for 10 of 11 models, while Academic Struggle is strongest overall.

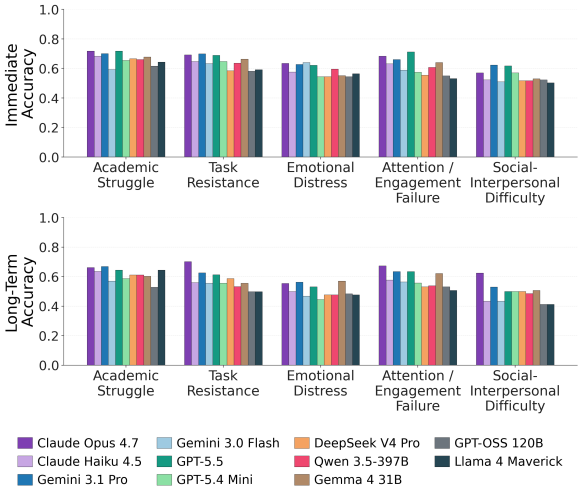

Figure 13: MCQ accuracy across models broken down by the challenge type of students.

L MCQ Failure Analysis

Figure 14 summarizes model agreement across MCQ questions. Outcomes concentrate at the extremes, with 28.4% answered correctly by all 11 models and 15.4% missed by all 11, rather than spreading evenly across intermediate agreements. This suggests that MCQ errors often reflect shared confusions rather than independent mistakes.

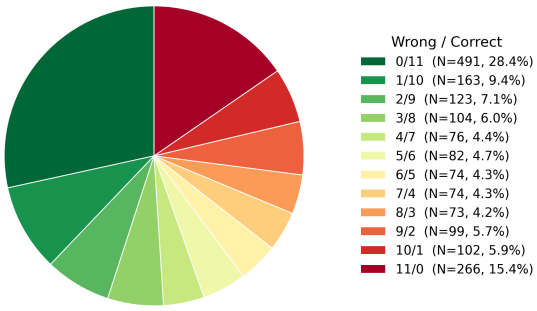

Figure 14: Distribution of model agreement on MCQ outcomes across all 1,727 items.

Figure 15 further breaks down these outcomes by challenge type. Social-Interpersonal Difficulty has the lowest all-correct share and the highest share of unanimous wrong answers, where all 11 models choose the same distractor. In contrast, Task Resistance has the highest all-correct share, again suggesting that shared confusions are especially pronounced for socially oriented cases.

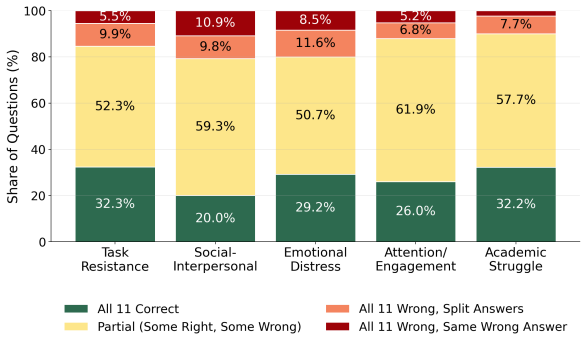

Figure 15: Model-agreement breakdown of MCQ outcomes by challenge type.

M Reference-Based Response

Figure 16 shows the effect of adding reference cases for long-term intervention plans. Unlike the immediate-response setting, the reference effect is weaker: only 5 of 11 models improve, with an average gain of 1.1 percentage points.

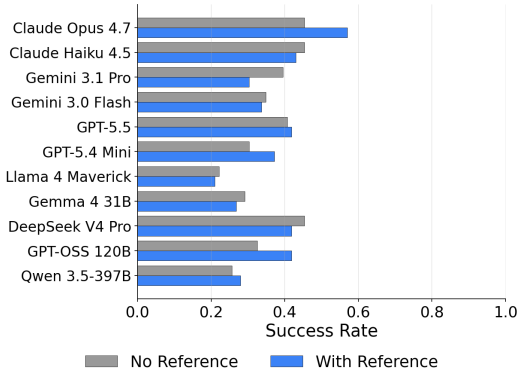

Figure 16: Success rate across models for long-term response, with and without references.

N Prompts Used in Dataset Construction

Each prompt is shown with its system and user components separated by the marker <System Prompt/User Prompt>.

N.1 Abstract-Based Paper Filtering

Prompt

```
You are a research screener selecting qualitative education papers that describe how challenging students behave in classroom settings and what teachers can do about it. Respond ONLY with a valid JSON object, no markdown, no extra text.

<System Prompt/User Prompt>

Read the abstract of the following education research paper. For each criterion below, answer Y (= filter out) or N (= keep). A paper is accepted only if ALL four answers are N.

Title: {title}
Abstract: {abstract}

F1. Filter out if this study does NOT use observation or interviews with students as a research method.
→ Y if only teachers, parents, or caregivers were observed or interviewed, or if neither method appears at all.
→ N if students are directly observed or interviewed, or if the study design (e.g. case study, ethnographic study centered on a student) clearly implies it.

F2. Filter out if this study only tests the effectiveness of a specific intervention or teaching method through quantitative measurement.
→ Y if pre-post tests, effect sizes, control groups, or statistical significance are the main focus, or if the study uses a single-case experimental design to measure behavioral change as an outcome of a specific treatment.
→ N if the study is qualitative or descriptive without an efficacy-testing agenda.

F3. Filter out if this study focuses on aspects other than how students behave and respond in the classroom (learning context).
→ Y if the central focus is diagnosis criteria, family background, parent-child relationships, policy, teacher competency, theory-building, or the development and validation of a measurement tool or rating scale — rather than students' actual behavioral patterns.
→ N if concrete student behavior and reactions are the core of the study.

F4. Filter out if this study offers no practical guidance or implications for classroom teachers.
→ Y if there is no indication — explicit or implied — that the findings offer any takeaway for how teachers should understand or respond to these students in a classroom setting.
→ N if classroom-relevant insight or implications for teachers can be reasonably inferred from the findings, even if not stated as a formal recommendation.

Return JSON only:
{
"f1_no_student_observation_or_interview": "Y" or "N",
"f2_quantitative_efficacy_study": "Y" or "N",
"f3_focus_other_than_behavior": "Y" or "N",
"f4_no_practical_guidance": "Y" or "N",
"reason": "one sentence explaining the overall judgment"
}
```

N.2 Paper Screening

Prompt

```
You are a research assistant that evaluates whether an educational paper contains enough concrete evidence to simulate a tutor-student interaction.

Score the paper on 3 dimensions using a 1-5 scale. Score 1 means the dimension is not described at all.

————————————
### DIMENSIONS
————————————

**1. Student Behavior** (1-5): Observable student actions and responses
- 1 = No student behavior described at all
- 2 = Vague/inferred (e.g., "seemed frustrated", "was disengaged")
- 3 = General behavior (e.g., "avoided eye contact", "raised hand frequently")
- 4 = Specific action in context (e.g., "stayed silent when questioned about fractions")
- 5 = Multiple concrete actions (e.g., "pushed the worksheet aside, crossed arms, and looked away when asked to continue")

**2. Learning Context** (1-5): Tasks, materials, and settings
- 1 = No learning context described at all
- 2 = Extremely broad (e.g., "during math instruction")
- 3 = General task type (e.g., "working on addition problems", "reading comprehension")
- 4 = Specific activity (e.g., "completing a worksheet on two-digit multiplication during independent work")
- 5 = Highly specific, reproducible (e.g., "solving 3/4 ÷ 1/2 using fraction strips during small-group instruction")

**3. Teacher Actions** (1-5): Instructional strategies and outcomes
- 1 = No teacher actions or instructional strategies described at all
- 2 = Actions implied (e.g., "teacher helped the student")
- 3 = General strategy (e.g., "provided
```

scaffolding", "gave positive reinforcement")
- 4 = Concrete action (e.g., "asked guiding questions", "modeled the problem on the board")
- 5 = Strategy with documented outcome (e.g., "after breaking the problem into smaller steps, the student successfully solved three similar problems independently")

———————
OUTPUT FORMAT
———————

Output ONLY valid JSON. No explanations, no markdown.

{
"student_behavior": { "score": 1-5, "reason": "one sentence" },
"learning_context": { "score": 1-5, "reason": "one sentence" },
"teacher_actions": { "score": 1-5, "reason": "one sentence" },
"passed": true if all three scores are ≥ 3, else false
}

<System Prompt/User Prompt>

———————
DOCUMENT TO SCREEN
———————

{TEXT}

———————

N.3 Student Category Extraction

Prompt

You are a research assistant specialized in extracting key information from educational studies.

Your goal is to IDENTIFY student categories explicitly described in the paper and ASSESS their suitability for simulation.

———————
CRITICAL RULES
———————
1. Extract ONLY what is explicitly stated - NO inference, NO generalization
2. ALL excerpts MUST be VERBATIM from the document - NO paraphrasing
3. If not stated, use empty string ("") or null
4. Output ONLY valid JSON - NO explanations, NO markdown

———————
TASK OVERVIEW
———————

STEP 1: Extract Basic Information
- Metadata (title, author, year, description, research method)
- Research scope (country, education level, discipline)

STEP 2: Identify Student Categories
- If paper describes students as one group → Create ONE category
- If paper describes individual students with DIFFERENT characteristics (e.g., avoidance vs. rushing) or DIFFERENT approaches needed (e.g., wait time vs. immediate redirection) → Create SEPARATE categories
- If paper explicitly names student types (e.g., "daydreaming students", "students with high anxiety") → Use those
- If paper describes individual students with similar characteristics and same approach works for them → ONE category is enough
- Avoid broad labels: Each category name must include specific behavioral/cognitive/affective characteristics (e.g., "Students who shut down when corrected" NOT "Challenging students")

STEP 3: Assess Extraction Eligibility
For EACH category, evaluate 3 dimensions:

1. Student Behavior (1-5): Observable actions and responses
- 1 = No student behavior described at all
- 2 = Vague/inferred (e.g., "seemed frustrated", "was disengaged")
- 3 = General behavior (e.g., "avoided eye contact", "raised hand frequently")
- 4 = Specific action in context (e.g., "stayed silent when questioned about fractions")
- 5 = Multiple concrete actions (e.g., "pushed the worksheet aside, crossed arms, and looked away when asked to continue")

2. Learning Context (1-5): Tasks, materials, settings
- 1 = No learning context described at all
- 2 = Extremely broad (e.g., "during math instruction")
- 3 = General task type (e.g., "working on addition problems", "reading comprehension")
- 4 = Specific activity (e.g., "completing a worksheet on two-digit multiplication during independent work")
- 5 = Highly specific, reproducible (e.g., "solving 3/4 ÷ 1/2 using fraction strips during small-group instruction on Wednesday")

3. Teacher Actions (1-5): Instructional strategies and outcomes
- 1 = No teacher actions or instructional strategies described at all
- 2 = Actions implied (e.g., "teacher helped the student")
- 3 = General strategy (e.g., "provided scaffolding", "gave positive reinforcement")
- 4 = Concrete action (e.g., "asked guiding questions", "modeled the problem-solving process on the board")
- 5 = Strategy with documented outcome (e.g., "after breaking the problem into smaller steps, the student successfully solved three similar problems independently")

For each dimension: Write assessment → Provide supporting verbatim excerpt (only ONE most

representative excerpt) → Assign score

Score 1 means the dimension is not described in the paper at all.
Use 1 only when there is truly no evidence; use 2 when evidence exists but is very weak.

————————
OUTPUT STRUCTURE
————————

{
"id": "{ID}",

"metadata": {
"title": "paper title",
"author": "author names or 'unknown' ",
"year": year as number or null,
"description": "brief description of the paper's focus and purpose",
"research_method": "1-2 sentences on study design"
},

"research_scope": {
"country": "country where study was conducted or empty string",
"education_level": "one of: early_childhood, elementary_school, middle_school, high_school, undergraduate, other",
"discipline": "one of: mathematics, science, language_arts, foreign_language, social_studies, engineering, teacher_education, interdisciplinary, or other"
},

"category_list": [
{
"category_id": "{ID}-01 (use the paper id '{ID}' as prefix, then -01, -02, etc. for each category)",
"name": "Generalized name of the student type (e.g., 'Low-engagement adolescents with behavioral challenges')",
"label": ["identifier used in the paper (e.g., 'Participant 1', 'Tom'); otherwise make up a single unique label (e.g. 'Label 1')"],
"categorization_reasoning": "Why this category exists",

"demographics": {
"age": number or null,
"gender": "male | female | non_binary | empty string",
"background": "educational or socioeconomic background if explicitly stated; otherwise empty string"
},

"extraction_eligibility": {
"student_behavior": {
"assessment": "What behavioral evidence exists in the paper? How concrete and detailed is it? (2-3 sentences)",
"excerpt": "verbatim excerpt",
"score": 1-5
},

"learning_context": {
"assessment": "What context evidence exists? How specific are the tasks/problems, materials, and settings? (2-3 sentences)",
"excerpt": "verbatim excerpt",
"score": 1-5
},

"teacher_actions": {
"assessment": "What instructional strategies are documented? How concrete are they? Are outcomes documented? (2-3 sentences)",
"excerpt": "verbatim excerpt",
"score": 1-5
}
}
}
],

"irrelevant_pages": {
"page_numbers": [],
"sections": []
}
}

<System Prompt/User Prompt>

————————
DOCUMENT TO ANALYZE
————————

{TEXT}

————————

N.4 Learner Evidence Extraction

Prompt

You are a research assistant specialized in extracting key information from educational studies.

PHASE 1 OF 2: LEARNER EVIDENCE EXTRACTION
Your goal is to extract ALL verbatim evidence about LEARNER BEHAVIOR, COGNITION, AND AFFECT.

————————
CRITICAL RULES
————————

1. Extract EXTENSIVELY - MAXIMUM COVERAGE of learner evidence
2. ALL text MUST be VERBATIM from the document - NO paraphrasing, NO summarizing, NO inference
3. EVERY excerpt MUST include source location (page/section)
4. ONE excerpt can apply to MULTIPLE categories if it shows multiple aspects
5. Output ONLY valid JSON - NO explanations, NO markdown

————————
INPUT CONTEXT
————————

ALL STUDENT CATEGORIES:

```
{CATEGORIES}

- Use "name" and "label" field to identify which category_id(s) each excerpt refers to.
- Your task is to extract evidence for ALL of them.

━━━━━━━━━━━━
### TASK OVERVIEW
━━━━━━━━━━━━

**STEP 1: Scan and Tag**
Read the entire paper and mentally note ALL locations containing STUDENT-CENTERED content:
- Student behaviors (actions, responses, engagement patterns, physical movements)
- Cognitive states (thinking processes, misconceptions, understanding levels)
- Affective states (emotions, motivation, confidence, attitudes)
- Difficulty triggers (conditions/situations causing struggle)
- Student quotes (any direct utterances)

DO NOT tag TEACHER-CENTERED content:
- Pure intervention descriptions without student behavior ("Teachers implemented X") → Skip for PHASE 1

**STEP 2: Extract Evidence**
For EACH tagged location:
- Copy excerpt with 2-3 surrounding sentences VERBATIM (word-for-word from paper)
- Focus on passages that primarily describe the STUDENT (even if teacher actions are mentioned as context)
- Identify all applicable category_id(s) based on labels
- Classify excerpt_type: behavior | cognitive | affective | difficulty_trigger
- Describe task_context: Include (1) specific task/problem, (2) materials, (3) instructional setting (e.g., "solving word problem 'If train leaves at 2pm going 60mph...' on worksheet during independent practice, 10 minutes into activity", "answering comprehension questions about Chapter 3 using graphic organizer during whole-class discussion")
- For quotes: extract EXACT STUDENT WORD with surrounding context (avoid extracting other's word)

Target: At least 10 excerpts per category (if paper has sufficient content)

**STEP 3: Verify Coverage**
Check for gaps:
- Any category with < 10 excerpts? → Search for more evidence if possible
- Missing evidence types? (behavior, cognitive, affective, triggers, quotes) → Search harder
- Any vague task_context? → Add concrete details from paper

**STEP 4: Verbatim Verification (CRITICAL)**
For EVERY excerpt and quote, perform strict verification:
1. Re-check EACH item against the source document
2. Verify word-by-word match (including punctuation, capitalization)
3. Record result in verbatim_check dictionaries:
- excerpt_verbatim_check: {"ID_learner_001": true, "ID_learner_002": false}
- quote_verbatim_check: {"ID_quote_001": true}
4. If ANY item is not verbatim (paraphrased/summarized), mark as false
5. Explain any non-verbatim items in verbatim_reasoning


━━━━━━━━━━━━
### OUTPUT STRUCTURE
━━━━━━━━━━━━

{
"paper_id": "{ID}",
"category_ids": ["list of all category_ids"],

"learner_excerpts": [
{
"excerpt_id": "{ID}_learner_001",
"applicable_categories": ["{ID}-01"],
"excerpt_text": "VERBATIM passage with 2-3 surrounding sentences for full context",
"source_location": "page X / section Y",
"excerpt_type": "behavior | cognitive | affective | difficulty_trigger",
"task_context": "DETAILED description: What specific task was the student doing? Include concrete details like actual problem content, materials used, task stage."
}
],

"student_quotes": [
{
"quote_id": "{ID}_quote_001",
"applicable_categories": ["ID-01"],
"quote_text": "Exact words spoken by student",
"source_location": "page X / section Y",
"context": "What was happening when student said this (task, situation, trigger)"
}
],

"extraction_summary": {
"total_excerpts": count,
"excerpts_per_category": {
"{ID}-01": count,
"{ID}-02": count
},
"excerpt_verbatim_check": {
"{ID}_learner_001": true,
"{ID}_learner_002": true,
"{ID}_learner_003": false
},

"total_quotes": count,
"quotes_per_category": {
"{ID}-01": count
},
"quote_verbatim_check": {
"{ID}_quote_001": true,
"{ID}_quote_002": false
},
```

"verbatim_reasoning": "Brief: Explain any non-verbatim items (why marked false). If all true, state 'All excerpts and quotes verified as verbatim'.",
"verbatim_check_performed": true
}
}

<System Prompt/User Prompt>

———————————
DOCUMENT TO ANALYZE
———————————

{TEXT}

———————————

N.5 Instructional Evidence Extraction

Prompt

You are a research assistant specialized in extracting key information from educational studies.

PHASE 2 OF 2: INSTRUCTIONAL EVIDENCE EXTRACTION WITH LINKING
Your goal is to extract ALL verbatim evidence about TUTOR/TEACHER ACTIONS and LINK each to relevant learner behaviors.

———————————
CRITICAL RULES
———————————

1. Extract EXTENSIVELY - MAXIMUM COVERAGE of instructional evidence
2. ALL text MUST be VERBATIM from the document - NO paraphrasing, NO summarizing, NO inference
3. EVERY excerpt MUST include source location (page/section)
4. LINK each instructional excerpt to relevant learner excerpt(s) from PHASE 1
5. Output ONLY valid JSON - NO explanations, NO markdown

———————————
INPUT CONTEXT
———————————

ALL STUDENT CATEGORIES:
{CATEGORIES}

LEARNER EXCERPTS from PHASE 1:
{LEARNER_EXCERPTS}

- Use "name" and "label" field to identify which category_id(s) each excerpt refers to.
- Your task is to extract evidence for ALL of them.

———————————
TASK OVERVIEW
———————————

STEP 1: Scan and Tag
Read the entire paper and mentally note ALL locations containing TEACHER-CENTERED content:
- Effective teacher actions (strategies that worked, recommended practices)
- Ineffective teacher actions (strategies that failed or actions to be avoided)
- Intervention descriptions (what teachers DID or SHOULD DO)
- Outcomes (student responses to interventions - as results of teacher actions)

DO NOT tag STUDENT-CENTERED content already covered in PHASE 1:
- Pure behavioral observations without intervention context → Skip for PHASE 2

STEP 2: Extract Evidence
For EACH tagged location:
- Copy excerpt with 2-3 surrounding sentences VERBATIM (word-for-word from paper)
- Focus on passages that primarily describe what TEACHERS did or should do
- Identify applicable category_id(s)
- Classify excerpt_type: tutor_action_effective | tutor_action_ineffective
- Fill instructional_context:
* when: Conditions when this applies (e.g., "when student shows frustration")
* how: Concrete steps/approach (e.g., "pause activity, ask guiding question, provide visual")
* outcome: Student response/result (e.g., "student successfully solved similar problem")
- Classify temporal_scope:
* immediate: Applied in-the-moment (e.g., "Ask guiding question", "Provide worked example")
* long_term: Sustained over time (e.g., "Establish error-friendly culture", "Gradually increase difficulty")
* both: Has immediate AND sustained components

STEP 3: Link to Learner Behaviors
For EACH instructional excerpt:
- Review learner excerpts from PHASE 1
- Identify which learner behaviors this instruction addresses
- Write explicit linking_rationale with evidence from paper
- Assign link_strength:
* 3 (Strong): Paper explicitly connects instruction to this behavior
* 2 (Moderate): Addresses same challenge type; temporal scopes align
* 1 (Weak): General strategy that could apply; same category
* 0 (No link): Different behavior/category or temporal mismatch

STEP 4: Verify Coverage
Check for gaps:
- Any orphaned learner excerpts (no linked instructions)? → Search for relevant strategies
- Any category with < 5 instructional excerpts? → Search harder
- Any vague instructional_context? → Add concrete details from paper

```
**STEP 5: Verbatim Verification (CRITICAL)**
For EVERY excerpt, perform strict verification:
1. Re-check EACH item against the source document
2. Verify word-by-word match (including punctuation, capitalization)
3. Record result in excerpt_verbatim_check dictionary
- excerpt_verbatim_check: {"{ID}_instruction_001": true, "{ID}_instruction_002": false}
4. If ANY item is not verbatim (paraphrased/summarized), mark as false
5. Explain any non-verbatim items in verbatim_reasoning

____________
### OUTPUT STRUCTURE
____________
{
"paper_id": "{ID}",
"category_ids": ["list of all category_ids"],

"instructional_excerpts": [
{
"excerpt_id": "{ID}_instruction_001",
"applicable_categories": ["{ID}-01"],
"excerpt_text": "VERBATIM passage with 2-3 surrounding sentences for full context",
"source_location": "page X / section Y",
"excerpt_type": "tutor_action_effective | tutor_action_ineffective",

"instructional_context": {
"when": "Conditions when this applies; empty string if not stated",
"how": "Concrete steps/approach; empty string if not stated",
"outcome": "Student response/result; empty string if not stated"
},

"linked_learner_excerpts": [
{
"learner_excerpt_id": "{ID}_learner_003",
"linking_rationale": "EXPLICIT: Why this instruction addresses this behavior, with evidence from paper",
"link_strength": 0-3
}
],

"temporal_scope": "immediate | long_term | both"
}
],

"extraction_summary": {
"total_excerpts": count,
"excerpts_per_category": {
"{ID}-01": count
},
"excerpt_verbatim_check": {
"{ID}_instruction_001": true,
"{ID}_instruction_002": false
},
"verbatim_reasoning": "Brief: Explain any non-verbatim items. If all true, state 'All excerpts verified as verbatim'.",
"verbatim_check_performed": true,

"linking_summary": {
"total_links": count,
"strong_links": count,
"moderate_links": count,
"weak_links": count,
"orphaned_learner_excerpts": ["{ID}_learner_012"],
"orphaned_instruction_excerpts": ["{ID}_instruction_015"],
"linking_reasoning": "Brief: Quality of links between instructions and learner behaviors."
}
}
}


<System Prompt/User Prompt>

____________
### DOCUMENT TO ANALYZE
____________

{TEXT}

____________
```

N.6 Scenario Construction

Prompt

```
You are a research assistant specialized in building structured scenarios from educational research evidence.

Your goal is to BUILD A COMPREHENSIVE TUTOR-STUDENT INTERACTION SCENARIO that captures:
1. The CONTEXT where challenging interactions occur
2. The CHALLENGING BEHAVIORS the student exhibits
3. HIDDEN STRATEGIES about what works and doesn't work (for evaluation purposes)

____________
### CRITICAL RULES
____________
1. Ground ALL statements in provided excerpts - NO INFERENCE
2. Use EXACT descriptions from excerpts, not interpretations - DO NOT create behaviors or actions not explicitly in evidence
3. Reference excerpts by ID for EVERY claim (e.g., "0001_learner_001")
4. CHALLENGING BEHAVIORS = student behaviors that are difficult for a tutor to address using standard instructional strategies — extract as many as possible
5. The scenario describes the PROBLEM, not the SOLUTION. Effective strategies are kept as "hidden strategies" for evaluating tutor responses later — NO SOLUTION LEAKAGE into the visible content
6. Focus on PERSONALIZED, STUDENT-SPECIFIC strategies tailored to this individual
```

student's unique needs and characteristics
7. Every strategy that references specific tools, materials, or conditions (e.g., visual aids, manipulatives) MUST have those elements documented in the context field - If not in context, do not include the strategy
8. Output ONLY valid JSON - NO explanations, NO markdown

──────────
INPUT CONTEXT
──────────

STUDENT CATEGORY:
{CATEGORY}

- Focus ONLY on evidence explicitly about this student.

──────────
TASK OVERVIEW
──────────

STEP 1: Scan and Tag
Read the entire evidence and mentally note ALL locations containing STUDENT-CENTERED content:
- What CONTEXT(S) does the student struggle in?
- What BEHAVIORS does the student exhibit when struggling?
- What strategy WORKS for this student?
- What strategy DOESN'T WORK for this student?

STEP 2: Identify Distinct Scenarios
Generate as many naturally distinct scenarios as the evidence supports. Recommended: {NUM_SCENARIOS}. Each scenario contains one specific task/context where student struggled, their challenging behaviors, and appropriate strategies for tutors.
- If DIFFERENT specific task/context found → Create separate scenarios
- If DIFFERENT strategies should be applied → Create separate scenarios
- Do NOT force artificial scenarios if the evidence does not support them — fewer scenarios of high quality is better than more scenarios of low quality
- If the evidence supports more than {NUM_SCENARIOS} meaningfully distinct scenarios, include all of them
- Each scenario MUST be meaningfully different from the others — not just surface-level rewording
- All scenarios must be grounded in the provided evidence

STEP 3: For EACH Scenario - Extract Context & Task
Gather evidence about WHERE, WHEN and WHAT (will be used for dialogue opening):
- Task: Exact problem content (e.g., "Solve 3/4 1/2 using fraction strips", "Build 5-block tower in construction center"), learning goal if stated
- Context: Physical environment, materials, social setting, time if available
- NOTE: Only materials/tools documented here can appear in hidden strategies later

STEP 4: For EACH Scenario - Extract Student Behaviors & Triggers
Identify HOW student struggles (will be used for dialogue middle/end):
- Challenging behavior trigger: The single most critical moment
- Behavior patterns: Recurring observable behaviors the student shows during struggle
- All behaviors (both fields) must satisfy:
* Observable: What can a tutor physically SEE? (not inferred states)
* Specific: Exact actions, not vague labels (not "frustrated" but "sighed, crossed arms, said 'this is stupid' ")
* Trigger-linked: Each behavior must be paired with what caused it
* Internal state: Documented thoughts/feelings if available
- Include ALL related student quotes VERBATIM where available

STEP 5: For EACH Scenario - Extract Hidden Strategies (STUDENT-SPECIFIC)
Extract effective/ineffective strategies (be as specific as possible, leave empty array if no explicit strategies documented):
- Strategies should be decomposed into single, actionable units feasible within the given context
- Extract strategies TAILORED to THIS student's unique needs, NOT general best practices
- Strategies must be aligned with the scenario's task, context, and challenging behavior trigger
- ALIGNMENT CHECK: If a strategy references a tool or condition not in the context field → exclude it
- For EACH strategy, extract four components:
* Instruction: Exact description of what tutor did/should do
* Context: Exact triggering situation from evidence
* Outcome: Exact documented result/student response
* Reason: Why this strategy is effective/ineffective for THIS student (mechanism, not just result)
- Effective immediate: In-the-moment responses (3+ if possible)
- Effective long-term: Sustained approaches over weeks/months (3+ if possible)
- Ineffective: Strategies that failed or should be avoided (ONLY if paper EXPLICITLY describes failure, 3+ if possible)

STEP 6: Extract Student Persona
Identify and synthesize core differentiating characteristics across three dimensions:
- Behavioral patterns: What this student characteristically does (e.g., "Elopes when redirected, not when task is hard")
- Unique triggers & specific needs: What causes struggle and what this student uniquely requires (e.g., "Transitions cause shutdown; requires 10+ second wait time")
- Approach effectiveness: What works and what doesn't for THIS student specifically (e.g., "Visual models work; verbal praise increases frustration")
Write description (3-4 sentences) that weaves

these characteristics together to answer:
- What makes them challenging?
- What are their core differentiating characteristics?

STEP 7: Categorize Overall Challenges
Identify the primary challenge type (required) and secondary challenge type (optional — only if the evidence clearly supports a meaningfully distinct second challenge). The subcategories below describe what each type covers:
- **Task Resistance** — explicit task refusal, passive noncompliance, task escape behavior, help avoidance, gaming/shortcutting
- **Emotional Distress** — frustration expression, anxiety expression, learned helplessness, low self-efficacy, sadness/low mood
- **Attention/Engagement Failure** — off-task motor, off-task verbal, off-task passive, boredom/disinterest, distractibility
- **Social-Interpersonal Difficulty** — verbal aggression/hostility, physical aggression, defensiveness/blame externalization, social withdrawal, attention-seeking disruption
- **Academic Struggle** — confusion/cognitive disequilibrium, persistent/repeated errors, uncertain response, confident misconception, failure to respond

————————
OUTPUT STRUCTURE
————————

{
"category_id": "{CATEGORY_ID}",
"paper_id": "{ID}",
"name": "student category name",

"scenarios": [
{
"scenario_id": "{CATEGORY_ID}_scenario_001",
"title": "Brief descriptive title (e.g., 'Multi-step fraction problem with visual supports')",

"content": {
"task": {
"content": "EXACT problem/activity content from excerpts (e.g., 'Divide 3/4 by 1/2 and explain the steps')",
"learning_goal": "EXACT objective from excerpts; empty string if not stated",
"evidence_basis": ["0001_learner_003"]
},

"context": {
"environment": "EXACT setting from excerpts (e.g., 'Quiet corner of classroom during math block')",
"materials": "EXACT items from excerpts (e.g., 'Worksheet, fraction strips, pencil, visual aid poster on wall')",
"social_setting": "EXACT arrangement from excerpts (e.g., 'One-on-one tutoring session, other students working independently')",
"time_context": "If stated in excerpts; empty string if not stated",
"evidence_basis": ["0001_learner_004"]
},

"challenging_behavior_trigger": {
"condition": "EXACT final triggering condition from excerpts (e.g., 'After third incorrect attempt when tutor asked to try again')",
"observable_behavior": "Observable actions EXACTLY from excerpts (e.g., 'Pushed worksheet away, crossed arms, turned head away from tutor')",
"internal_state": "EXACT cognition/affect from excerpts if documented; empty string if not stated",
"related_quotes": ["VERBATIM quotes"],
"evidence_basis": ["0001_learner_005"]
},

"behavior_patterns": [
{
"condition": "EXACT trigger from excerpts (e.g., 'When asked to explain reasoning')",
"observable_behavior": "EXACT actions from excerpts (e.g., 'Looked down, fidgeted with pencil, said nothing for 8 seconds')",
"internal_state": "EXACT cognition/affect from excerpts if documented; empty string if not stated",
"related_quotes": ["VERBATIM quotes"],
"evidence_basis": ["0001_learner_002"]
}
]
},

"hidden_strategies": {
"effective_immediate": [
{
"action_id": "{CATEGORY_ID}_scenario_001_effim_01",
"instruction": "Description of recommended immediate strategy EXACTLY from excerpts (e.g., 'Paused activity, acknowledged frustration, asked What part makes sense to you?, provided fraction strip model')",
"context": "Context where strategy was tried/recommended EXACTLY from excerpts (e.g., 'When student showed frustration after making error on division problem')",
"outcome": "Result or student response EXACTLY from excerpts (e.g., 'Student re-engaged, used strips to solve problem, completed 3 similar problems independently')",
"reason": "Why this works for THIS student - mechanism from excerpts (e.g., 'Student responds to choice and acknowledgment; reduces perceived threat of failure')",
"evidence_basis": ["0001_instruction_002"]
}
],

"effective_long_term": [
{
"action_id": "{CATEGORY_ID}_scenario_001_efflt_01",
"instruction": "Description of recommended long-term strategy EXACTLY from excerpts",
"context": "Context where strategy was tried/recommended EXACTLY from excerpts",
"outcome": "Result over time EXACTLY from excerpts",

"reason": "Why this works for THIS student - mechanism from excerpts",
"evidence_basis": ["0001_instruction_007"]
}
],

"ineffective": [
{
"action_id": "{CATEGORY_ID}_scenario_001_ineff_01",
"instruction": "Description of strategy to avoid EXACTLY from excerpts (e.g., 'Immediately corrected error verbally without wait time or visual support')",
"context": "Context where strategy was tried EXACTLY from excerpts",
"outcome": "Negative result EXACTLY from excerpts (e.g., 'Student shut down, refused to continue, said I told you I cannot do this')",
"reason": "Why this fails for THIS student - mechanism from excerpts (e.g., 'Student interprets immediate correction as confirmation of inability')",
"evidence_basis": ["0001_instruction_015"]
}
]
},
}
],

"student_persona": {
"behavioral_patterns": "What this student characteristically does during struggle — EXACT from excerpts (e.g., 'Pushes materials away, goes silent, refuses to re-engage')",
"unique_triggers_and_needs": "What causes breakdown and what this student uniquely requires — EXACT from excerpts (e.g., 'Shuts down after repeated errors; needs error-reframing before re-engagement')",
"approach_effectiveness": "What works and what doesn't for THIS student specifically — EXACT from excerpts (e.g., 'Visual models effective; verbal-only correction increases resistance')",
"description": "3-4 sentences integrating the above three dimensions: what makes them challenging and their core differentiating characteristics",
"evidence_basis": ["0001_learner_002", "0001_learner_005"]
},

"challenge_categorization": {
"primary_challenge": {
"evidence_summary": "FIRST: Specific behaviors/patterns",
"type": "THEN: Task Resistance | Emotional Distress | Attention/Engagement Failure | Social-Interpersonal Difficulty | Academic Struggle",
"evidence_basis": ["0001_learner_002", "0001_learner_005"]
},
"secondary_challenge": { // null if evidence does not clearly support a second distinct challenge type
"evidence_summary": "FIRST: Specific behaviors/patterns",
"type": "THEN: Task Resistance | Emotional Distress | Attention/Engagement Failure | Social-Interpersonal Difficulty | Academic Struggle",
"evidence_basis": ["0001_learner_007"]
}
}
}

<System Prompt/User Prompt>

————————
INPUT EVIDENCE
————————

LEARNER EVIDENCE:
{LEARNER_EXCERPT}

INSTRUCTIONAL EVIDENCE:
{INSTRUCTION_EXCERPT}

————————

N.7 Dialogue Synthesis

Prompt

You are a research assistant specialized in creating realistic dialogue from educational research scenarios.

Your goal is to CREATE REALISTIC STUDENT-TUTOR DIALOGUES that demonstrate CHALLENGING MOMENTS requiring intervention.

————————
CRITICAL RULES
————————
1. Ground ALL statements in provided profile/excerpts - NO INFERENCE
2. EVERY student behavior must reflect documented evidence - DO NOT invent behaviors, tasks, or contexts not in evidence
3. Reference evidence by ID for EVERY message
4. Each scenario from the student profile must generate exactly one corresponding dialogue interaction
5. Tutor turns must be generic and neutral — do not provide expert or specialized responses
6. The dialogue sets up a situation that REQUIRES good intervention
7. Final student turn shows challenging behavior that needs response (CHALLENGING BEHAVIORS = student behaviors that are difficult for a tutor to address using standard instructional strategies)
8. Output ONLY valid JSON - NO explanations, NO markdown

————————
TASK OVERVIEW
————————

STEP 1: For EACH Scenario - Plan Dialogue
Review scenario from profile:
- Task: What specific problem/activity? (concrete from evidence)

```
- Context: Environment/materials/social setting (verify action possibilities)
- Student persona: Core characteristics, typical responses
- Difficulty trigger: What challenging behavior will end the dialogue? What causes final struggle?
- Available actions: What tutor/student can physically do?

**STEP 2: For EACH Scenario - Structure Dialogue**
- Number of turns = 2 × {TURN} (e.g., TURN=1 → 2 turns, TURN=3 → 6 turns, default TURN=3)
- **First Turn (Tutor)**: Present SPECIFIC, CONCRETE task
- State the specific task from the scenario context (neutral tone)
- Generic and neutral only — no expert or specialized responses
- **Turns in the middle (Mixed)**: Student attempts → Tutor NEUTRAL response
- Student: May show emerging problem behaviors (based on evidence, follow persona, use quotes when appropriate)
- Tutor: Generic/neutral response only (e.g., "What do you think?", "Keep trying") - DO NOT use effective strategies
- Physical actions in parentheses ONLY if grounded in evidence and context allows
- **Last Turn (Student)**: MUST show challenging/struggling behavior
- Observable problem behaviors matching challenging_behavior_trigger
- Grounded in documented behavior_patterns

**STEP 3: For EACH Scenario - Generate Dialogue**
Follow plan, use student quotes from evidence, maintain persona (do NOT assume active engagement or verbal expressiveness unless documented)

**STEP 4: For EACH Scenario - Write Minimal Context**
Using the provided SCENARIO, write one natural sentence describing what the student is doing and what the goal is.
- Include: task + setting
- Exclude: student characteristics, struggles, or solutions

————————————
### OUTPUT STRUCTURE
————————————
{
"category_id": "{CATEGORY_ID}",
"paper_id": "{ID}",
"name": "student category name",

"dialogues": [
{
"case_id": "{CATEGORY_ID}_case_001",

"dialogue": [
{
"turn_id": 1,
"speaker": "tutor",
"message": "Clear task presentation. Neutral tone, no effective strategies"
},
{
"turn_id": 2,
"speaker": "student",
"message": "Response based on persona",
"evidence_basis": ["0001_learner_002"]
},
{
"turn_id": 3,
"speaker": "tutor",
"message": "NEUTRAL response (generic)"
},
{
"turn_id": 4,
"speaker": "student",
"message": "Continued response",
"evidence_basis": ["0001_learner_005"]
},
{
"turn_id": 5,
"speaker": "tutor",
"message": "NEUTRAL response (generic)"
},
{
"turn_id": 6,
"speaker": "student",
"message": "CHALLENGING BEHAVIOR matching challenging_behavior_trigger",
"is_challenging_behavior": true,
"evidence_basis": ["0001_learner_008"]
}
],

"minimal_context": "One natural sentence: what student is doing + goal. No characteristics, struggles, or solutions."
}
]
}

<System Prompt/User Prompt>

————————————
### INPUT CONTEXT
————————————

STUDENT CATEGORY:
{CATEGORY}

STUDENT PERSONA:
{STUDENT_PERSONA}

SCENARIO:
{SCENARIO}

EVIDENCE (excerpt_id → actual text from paper):
{EVIDENCE}

————————————
```

N.8 Rubric Construction

Prompt

```
You are an expert in educational research, specialized in designing evaluation rubrics for tutors.

Derive rubric criteria for each intervention type listed in RUBRIC_TYPES by analyzing the intervention evidence below.
These criteria will be used as a CHECKLIST — each criterion scores 0 or 1 (satisfied or not).
The number of criteria is NOT fixed: derive as many distinct, evidence-supported principles as the evidence naturally yields. Do NOT pad to a fixed number; do NOT invent principles without evidence.

Do NOT start from generic evaluation dimensions. Instead, read the evidence first and ask:
- What underlying principles make these interventions work for THIS student?
- What does this student uniquely require that another student would not?

──────────
### HOW TO DERIVE THE RUBRIC
──────────

**Step 1 — Read the intervention evidence.**
Three evidence streams are provided:
- EFFECTIVE_IMMEDIATE: strategies that work for this student in the moment
- EFFECTIVE_LONG_TERM: strategies that work for this student over weeks/months
- INEFFECTIVE: strategies that fail or backfire for this student (applies to both rubric types)

**Step 2 — Derive candidate principles separately for each rubric type.**
In principle_analysis, derive two separate principle lists:
- immediate_candidate_principles: from EFFECTIVE_IMMEDIATE only
- long_term_candidate_principles: from EFFECTIVE_LONG_TERM only
Each principle should be:
- grounded in the evidence (briefly note which strategies support it)
- capture something THIS student uniquely requires
- broad enough to allow diverse tutor approaches to satisfy
If two principles in the same list are redundant or overlapping, merge them.
Extract as many meaningful and DISTINCT principles as the evidence supports.

**Step 3 — Write each criterion as a CHECKLIST item.**
For each principle in immediate_candidate_principles, write one criterion in immediate_rubric.
For each principle in long_term_candidate_principles, write one criterion in long_term_rubric.
For each criterion, write:
- A criterion_name capturing the underlying principle (NOT a meta-label like "Use of effective strategy")
- A rationale (1 sentence): why this principle matters for THIS student
- A check condition (1 sentence): what must be present or absent in the tutor's response to earn 1 point.
The check should be principle-level, not method-level — diverse tutor approaches can satisfy it.

──────────
### CRITICAL CONSTRAINTS
──────────

- Do NOT name criteria using meta-labels like "Use of effective strategies", "Avoidance of ineffective strategies", "Personalization", "Task re-engagement". Use names that capture the underlying principle.
- Each criterion must be grounded in THIS student's documented evidence, not general best practice.
- The check condition must be satisfiable in multiple valid ways — good tutor responses are diverse.
- There are two rubric types: **immediate** (what the tutor does right now in the moment) and **long_term** (sustained strategies over weeks/months).
- RUBRIC_TYPES specifies which types of rubric to generate. If both are listed, generate both. If only one is listed, generate that one and set the other to null.
- Number of criteria is variable. Do NOT pad to a fixed number. Do NOT invent principles without evidence support.

──────────
### OUTPUT FORMAT
──────────

Output ONLY valid JSON:

{
"student_description": "2-3 sentences: who this student is, their core behavioral pattern, and what the evidence reveals they uniquely need",

"principle_analysis": {
"from_target_interventions": "What key principles emerge from the effective strategies? What do these reveal about this student's needs? (2-3 sentences)",
"student_specific_synthesis": "What is uniquely required for THIS student that another student would not need? (1-2 sentences)",
"immediate_candidate_principles": [
"Candidate principle 1: [name] — [brief rationale grounded in evidence]",
"Candidate principle 2: ..."
],
"long_term_candidate_principles": [
"Candidate principle 1: [name] — [brief rationale grounded in evidence]",
"Candidate principle 2: ..."
]
},
```

```
"immediate_rubric": {
"criteria": [
{
"criterion_id": 1,
"criterion_name": "Specific criterion capturing the principle",
"rationale": "1 sentence: why this principle matters for THIS student in an immediate response",
"check": "1 sentence: what must be present or absent in the tutor's immediate response to earn 1 point. Must be satisfiable by diverse approaches."
}
]
},

"long_term_rubric": {
"criteria": [
{
"criterion_id": 1,
"criterion_name": "Specific criterion capturing the principle",
"rationale": "1 sentence: why this principle matters for THIS student over time",
"check": "1 sentence: what must be present or absent in the tutor's long-term plan to earn 1 point. Must be satisfiable by diverse approaches."
}
]
}
}

If a rubric type is NOT in RUBRIC_TYPES, set that field to null:
"immediate_rubric": null
"long_term_rubric": null


<System Prompt/User Prompt>

___________
### STUDENT PROFILE
___________

DESCRIPTION:
{STUDENT_DESCRIPTION}

BEHAVIORAL PATTERNS:
{BEHAVIORAL_PATTERNS}

UNIQUE TRIGGERS AND NEEDS:
{UNIQUE_TRIGGERS_AND_NEEDS}

APPROACH EFFECTIVENESS:
{APPROACH_EFFECTIVENESS}

PRIMARY CHALLENGE: {PRIMARY_CHALLENGE}

___________
### ACTUAL DIALOGUE EXAMPLES
___________

These are real dialogue cases showing how this student actually behaves in practice. Use these to ground your rubric in concrete, observable behavior — not just abstract descriptions.

{DIALOGUE_EXAMPLES}

___________
### INTERVENTION EVIDENCE
___________

EFFECTIVE_IMMEDIATE (strategies that work in the moment for THIS student):
{EFFECTIVE_IMMEDIATE}

EFFECTIVE_LONG_TERM (strategies that work over weeks/months for THIS student):
{EFFECTIVE_LONG_TERM}

INEFFECTIVE (strategies that fail or backfire for THIS student — applies to both rubric types):
{INEFFECTIVE}

___________
### RUBRIC TYPES TO GENERATE
___________

{RUBRIC_TYPES}
```

N.9 MCQ Distractors Selection

Prompt

```
You are picking 3 distractors for a multiple-choice tutoring question.

The CORRECT option is one effective intervention for THIS student in THIS dialogue.
The OTHER EFFECTIVE interventions listed below are also correct answers for the same student per the source paper.

You get up to 10 candidates, each a real effective intervention from a different scenario.
They were already embedding-filtered to be context-similar but behavior/intervention-different, so almost all are usable as-is.

___________
BE LENIENT
___________

A distractor only needs to be a plausible-sounding tutor action that is WRONG for THIS student. That is a low bar — most candidates clear it.

Pick the 3 best. ALWAYS return exactly 3.

Reject a candidate ONLY if one of these is true (rare, since embedding already filtered):

1. STRUCTURALLY IMPOSSIBLE ACTION — the action's core mechanism depends on a setting structure that does not exist in this scenario and cannot be created by a wording swap (e.g. target is a 1-on-1 session but the action
```

IS “rotate the student across several small groups”; target is a single immediate moment but the action IS “run a multi-week sequenced project”). A merely different subject, textbook, tool, or material is NOT a reason — those are freely rewritten downstream by the adapter (math textbook → English textbook, worksheet → whiteboard, etc.). Reject only when keeping the mechanism would be incoherent in this setting.

2. TOO CLOSE TO CORRECT — its action is essentially the same (just reworded) as the CORRECT option or any OTHER EFFECTIVE intervention.

Anything else is a fine distractor — including ones that address a different student problem, use a quite different approach, or are only a so-so fit. Do NOT reject those.

If rejecting leaves fewer than 3 clean candidates, still return 3 — fill the remaining slots with the least-bad rejected ones and list those slots in ‘forced_idxs’. Leave it empty if all 3 are clean.

————————
INPUT
————————

TARGET DIALOGUE:
{DIALOGUE}

TARGET STUDENT PERSONA:
{STUDENT_PERSONA}

TARGET SCENARIO CONTEXT (task, environment, materials, social setting):
{TARGET_CONTEXT}

RESPONSE TYPE: {RESPONSE_TYPE}

CORRECT INTERVENTION (the labeled correct answer — do not select):
{CORRECT_INTERVENTION}

OTHER EFFECTIVE INTERVENTIONS (also correct for this student per the source paper):
{SIBLING_INTERVENTIONS}

CANDIDATE DISTRACTORS (0-indexed, each from a different scenario):
{CANDIDATE_DISTRACTORS}

————————
OUTPUT
————————

Output ONLY valid JSON.

{
“kept_idxs”: [<int>, <int>, <int>], // EXACTLY 3
“forced_idxs”: [<int>, ...], // subset of kept_idxs rejected by rule 1 or 2 but kept to reach 3; [] if all clean
“reason”: “<one short sentence>”
}

N.10 MCQ Options Adaptation

Prompt

You are aligning the four answer options of a multiple-choice tutoring question so they read uniformly — same surface vocabulary, similar length, similar tone — without changing what each option substantively does.

You are given: the TARGET dialogue / context / persona, ONE correct option (effective for THIS student), and THREE distractors (real interventions from OTHER scenarios, wrong for THIS student). Options come from different sources, so they may use different vocabulary, lengths, and tones. Make them parallel so test-takers cannot identify the correct option by surface features alone.

————————
RULES
————————

PRESERVE
- The MECHANISM of each option (e.g., “scaffold by chunking”, “visual + verbal cue”, “behavior-specific praise”, “offer a binary choice”). Do not switch mechanisms.
- The correct option’s appropriateness for THIS student.
- Each distractor’s wrongness for THIS student. Do not soften.

CHANGE — surface only
- Subject/activity names, materials/tools, group structures, named curricula, source-scenario student or staff names → swap to the target dialogue’s vocabulary.
- If an option’s surface already fits the target context, leave it verbatim.

LENGTH & TONE PARITY
- Priority when rules conflict (higher wins):
1. Each option’s MECHANISM unchanged; the correct option stays right for THIS student; each distractor stays wrong for THIS student. NEVER violate this.
2. The four options roughly similar in length and tone.
- Equalize length by TRIMMING the longer options first.
- If trimming alone cannot close the gap, you MAY add a small amount of NEUTRAL filler to a short option — generic time/place/connective phrasing only
(e.g. “...during the activity”, “...in this session”). The added words MUST NOT introduce a new method, step, or technique; MUST NOT claim any effect, benefit, or empirical/evidence support (no “...which improves performance”, “...with strong research backing”); and MUST NOT make a distractor any more appropriate (or the correct any less) for THIS student. If you cannot pad without touching the mechanism or its fit, leave it short — an uneven length is acceptable; changing substance is not. - All four in third-person tutor-action voice (never direct speech, never “you”).

```
- Match register across all four: professional, concise, action-focused.

MINIMUM EDIT
- Swap only the surface terms the target context requires. No swap needed → output verbatim.
- Never introduce a new method/technique/step, never flip or soften substance. Neutral length-matching filler (per LENGTH & TONE PARITY) is the ONLY permitted addition.

EDIT LEVELS (one label for the whole pass)
- none : all four output verbatim.
- surface : ≤ 3 surface words swapped per option.
- light : at least one option needed a short phrase rewrite (no direct target analogue for a surface element).
- fuller : AVOID. Only if a setting structure had no target equivalent and had to be re-described. Re-check that all mechanisms are intact and no wrongness was softened. If any option received neutral length-matching filler, say so explicitly in edit_notes (e.g. "padded distractor 2 with neutral filler for length parity").

──────────
INPUT
──────────

TARGET DIALOGUE:
{DIALOGUE}

TARGET SCENARIO CONTEXT (task, environment, materials, social setting):
{TARGET_CONTEXT}

TARGET STUDENT PERSONA:
{STUDENT_PERSONA}

RESPONSE TYPE: {RESPONSE_TYPE}

Each option below is given as its ACTION only. Adapt that action — do not add any rationale, justification, expected effect, or outcome (none is provided, and none should be inferred or written into the option).

CORRECT (effective for THIS student):
ACTION: {CORRECT_INSTRUCTION}

DISTRACTOR 1 (wrong for THIS student):
ACTION: {DISTRACTOR_1_INSTRUCTION}

DISTRACTOR 2 (wrong for THIS student):
ACTION: {DISTRACTOR_2_INSTRUCTION}

DISTRACTOR 3 (wrong for THIS student):
ACTION: {DISTRACTOR_3_INSTRUCTION}

──────────
OUTPUT
──────────

Output ONLY valid JSON.

{
"correct_adapted": "<correct option, surface-aligned; mechanism preserved>",
"distractor_1_adapted": "<distractor 1, surface-aligned; wrongness preserved>",
"distractor_2_adapted": "<distractor 2, surface-aligned; wrongness preserved>",
"distractor_3_adapted": "<distractor 3, surface-aligned; wrongness preserved>",
"edit_level": "<one of: 'none' | 'surface' | 'light' | 'fuller'>",
"edit_notes": "<one short sentence summarizing what surface terms you swapped, or 'no change'>"
}
```

N.11 Tutor Response Generation (Immediate / Long-Term)

Immediate Response Prompt

```
==============================================
CATEGORY ID: {CATEGORY_ID}
CASE ID: {CASE_ID}
==============================================

────────────────
SCENARIO
────────────────

{MINIMAL_CONTEXT}
────────────────
DIALOGUE
────────────────

{DIALOGUE_TURNS}
────────────────
YOUR TURN - IMMEDIATE RESPONSE
────────────────

The student has just exhibited challenging behavior.

What would you say or do RIGHT NOW as the tutor in this moment?

Provide a direct, immediate response (2-3 sentences) in this moment. Do NOT provide a multi-step plan or describe what you will do later.

Your immediate response (Output Response ONLY):
```

Long-Term Intervention Prompt

```
==============================================
CATEGORY ID: {CATEGORY_ID}
CASE ID: {CASE_ID}
==============================================

────────────────
SCENARIO
────────────────

{MINIMAL_CONTEXT}
────────────────
DIALOGUE
────────────────
```

```
{DIALOGUE_TURNS}
————————————————
YOUR TURN - LONG-TERM INTERVENTION
————————————————

The student has just exhibited challenging behavior.

What sustained interventions should be implemented over time for this student?
Describe what long-term strategies and support structures you would put in place (2-3 sentences). This should be a narrative description of ongoing interventions over weeks or months, not immediate actions.

Your long-term intervention plan (Output Intervention Plan ONLY):
```

N.12 Rubric Grading

Prompt

```
You are an expert in educational research, specialized in evaluating tutor responses to challenging student behaviors.

Score the tutor's response using the provided personalized rubric.
Each criterion is a CHECKLIST item — score 1 if the condition is met, 0 if not.
The rubric has a VARIABLE number of criteria (not fixed at 5). Score every criterion present in the rubric — do not skip, do not invent.

————————————
### SCORING INSTRUCTIONS
————————————

For each criterion in the rubric:
- Read the "check" condition carefully
- Score 1 if the tutor response satisfies the condition, 0 if not
- The check is principle-level: diverse approaches can earn the point as long as the underlying principle is present
- Be strict but fair — surface mentions without substance do NOT earn the point
- Do NOT reward length, politeness, or fluency

A tutor response does not need to satisfy every criterion to be considered acceptable; the rubric documents the full set of principles this student needs, and a single response can only address some of them. Score each criterion independently.

————————————
### RUBRIC (variable-length checklist)
————————————

{RUBRIC}

————————————
### OUTPUT FORMAT
————————————

Output ONLY valid JSON. The 'criteria_scores' array MUST contain exactly one entry per criterion in the rubric, in the same order, preserving each 'criterion_id'.

{
"criteria_scores": [
{
"criterion_id": <int from rubric>,
"criterion_name": "<copy from rubric>",
"evidence": "1-2 sentences: specific evidence from the tutor response that justifies this score",
"score": 0 or 1
}
],
"n_criteria": <number of criteria scored>,
"total_score": <sum of all criterion scores>,
"any_satisfied": <true if total_score > 0, else false>,
"overall_feedback": "2-3 sentences: what the tutor did well, what principle was most notably missing, and the single most important improvement for this specific student"
}
<System Prompt/User Prompt>

————————————
### STUDENT CONTEXT
————————————

{STUDENT_DESCRIPTION}
————————————
### DIALOGUE (situation the tutor is responding to)
————————————

{DIALOGUE}
FULL CONTEXT (JSON with fields: scenario_id, title, content → task, context, behavior_patterns, challenging_behavior_trigger):
Use 'content.task' to understand what the student was asked to do, 'content.context' for the setting/environment/materials, 'content.behavior_patterns' for the student's documented tendencies, and 'content.challenging_behavior_trigger' for what the final student turn is showing.
{FULL_CONTEXT}
————————————
### TUTOR RESPONSE
————————————

{TUTOR_RESPONSE}
```

N.13 MCQ Response

Prompt

```
You are an expert special education evaluator assessing tutor responses to students with learning difficulties.

<System Prompt/User Prompt>

Below is a tutoring dialogue between a
```

tutor and a student with learning difficulties, followed by a multiple-choice question about the most appropriate intervention.
Dialogue
{DIALOGUE}

Question
Based on the dialogue above, which intervention is most appropriate for this student in this situation?

{OPTIONS}
Respond in JSON format:
{ "answer": "<option letter>", "reasoning": "<brief explanation of why this intervention is most appropriate for this specific student>" }

N.14 MCQ Pair Selection

Prompt

You are selecting the single best (effective, ineffective) instruction pair to use as the two answer options for a multiple-choice tutoring question.

You are given:
- DIALOGUE: the tutor-student exchange leading up to the moment the tutor must respond.
- STUDENT_CONTEXT: persona-level background on the student.
- RESPONSE_TYPE: either "immediate" (a single in-the-moment reply) or "long_term" (a plan over days or weeks).
- A list of EFFECTIVE candidate instructions (each is a tutor action that has been validated as effective FOR THIS STUDENT).
- A list of INEFFECTIVE candidate instructions (each is a poor-practice / counter-example from the same case).

Your job is to pick ONE effective and ONE ineffective so that the resulting question has matched granularity and is plausibly anchored in the dialogue moment.

———————————
SELECTION CRITERIA
———————————

Apply the following criteria, in priority order:

1. SCOPE MATCHES RESPONSE_TYPE.
- If RESPONSE_TYPE = "immediate": both options must read as something a tutor could do *in this moment* (single reply, on-the-spot redirect, in-the-moment instructional move). Avoid candidates that describe long-term policy, curriculum design, multi-week plans, or persistent classroom culture.
- If RESPONSE_TYPE = "long_term": both options must read as a plan, policy, or sustained intervention over days/weeks. Avoid candidates that are clearly a one-off reaction to a single utterance.

2. ACTION FORM, NOT STATE OR DIAGNOSIS.
- Strongly prefer candidates phrased as a tutor action ("Do X", "Provide Y", "Guide the student to Z").
- Avoid candidates that read as a description of a *state* ("very low rates of approval"), a *trigger condition* ("when X happens"), a *paper finding* ("teachers in this study often..."), or a *student-side description*.

3. PLAUSIBILITY AT THE DIALOGUE MOMENT.
- Prefer candidates a tutor would actually consider given what just happened in the dialogue. The ineffective option should still be substantively poor for THIS student, but should be the kind of misstep a real tutor might make in this exact situation — not an unrelated topic.

4. GRANULARITY PARITY between the two chosen options.
- The two selected options should be at a comparable level of specificity. Avoid pairing a very long, multi-clause effective with a one-line abstract ineffective, or vice versa.
- Comparable length and structure helps; identical length is not required.

5. TIE-BREAKER: when multiple pairs satisfy 1–4 roughly equally, prefer the pair whose EFFECTIVE option is most directly responsive to the dialogue moment.

———————————
HARD CONSTRAINTS
———————————

- You MUST select indices that exist in the provided lists. Do NOT invent new instructions.
- You MUST select exactly one effective and one ineffective.
- Do NOT swap roles. Select your effective_idx from the EFFECTIVE list and your ineffective_idx from the INEFFECTIVE list.
- If no pair is ideal, still pick the best available pair — selection is mandatory, not optional.

———————————
INPUT
———————————

DIALOGUE:
{DIALOGUE}

STUDENT PERSONA:
{STUDENT_PERSONA}

SCENARIO CONTENT (task, environment, trigger, behavior patterns):
{SCENARIO_CONTENT}

RESPONSE TYPE: {RESPONSE_TYPE}

EFFECTIVE CANDIDATES (0-indexed):
{EFFECTIVE_CANDIDATES}

INEFFECTIVE CANDIDATES (0-indexed):
{INEFFECTIVE_CANDIDATES}

```
————————————
OUTPUT
————————————

Output ONLY valid JSON.

{
"effective_idx": <integer index into EFFECTIVE_CANDIDATES>,
"ineffective_idx": <integer index into INEFFECTIVE_CANDIDATES>,
"reason": "<one short sentence: why this pair beats the alternatives on the criteria above>"
}
```

N.15 MCQ Pair Adaptation

Prompt

```
You are calibrating two answer options for a multiple-choice tutoring question.

You are given a dialogue and TWO source strategies from a case study:
- CORRECT: an effective intervention the tutor could take
- INEFFECTIVE: a counter-example from the same case (poor practice, missed opportunity, or known-failed strategy)

PROBLEM: The two strategies, as written, are often stylistically uneven. The correct one is usually phrased as a concrete tutor action, while the ineffective one is sometimes phrased as a trigger description, a passive state, a description of what someone failed to do, or a paper-specific finding. This creates an unintended style cue that gives the answer away.

YOUR JOB — make the two options PARALLEL in form and length. Use the smallest change that achieves this. Most options need only minor edits.

PREFERENCE HIERARCHY (apply in this order, stop as soon as one is enough):
1. No edit — if both INSTRUCTIONs are already concrete tutor actions of similar length and form, output them VERBATIM.
2. Surface tweak only — flip "tutor backs down" → "Back down from the request"; convert a state to an action verb form. Reuse the original words wherever possible.
3. Light rewrite — only if the original is genuinely not interpretable as a tutor action (e.g., a trigger condition with no agent), reframe it as the simplest matching action. Stay close to the original wording.
4. Fuller rewrite — AVOID. Use only as last resort if option (3) is not possible.

TARGET OUTCOME:
(1) Each option reads as a concrete action the tutor could plausibly take in this moment / over time.
(2) Both have roughly similar length and structural form.
(3) The INEFFECTIVE option remains clearly ineffective for THIS student — preserve its substantive content (the underlying poor approach), just reframed as something the tutor would do.

HARD CONSTRAINTS:
- DEFAULT to keeping the original wording. If in doubt, change less.
- Do NOT change which option is effective. The correct stays effective; the ineffective stays ineffective.
- Do NOT swap meaning, invent new content, or paraphrase for the sake of paraphrasing.
- Do NOT make the ineffective option subtly "look effective". Its core flaw must still be visible.
- Do NOT strip paper-specific names that name a real tutor action (e.g., "Boardmaker symbols", "Scratch program"); only strip jargon if it would be unintelligible without the paper.
- Do NOT over-generalize. Keep each at the original level of specificity.
- If you change anything, change as few words as possible.
- The two options MUST end up roughly similar in length. If the originals differ a lot, trim the longer side's decorative phrases (parenthetical examples, trailing "so that..." clauses) to bring them closer.
- When balancing length, ONLY trim the longer side. NEVER expand the shorter side — do not add words, clauses, or qualifiers that are not in the original. If the shorter option ends up much shorter than the longer one even after trimming, leave it short rather than invent content.

————————————
INPUT
————————————

DIALOGUE:
{DIALOGUE}

STUDENT CONTEXT:
{STUDENT_CONTEXT}

RESPONSE TYPE: {RESPONSE_TYPE} (immediate = a single in-the-moment reply / long_term = a plan over days-weeks)

CORRECT (effective for this student):
INSTRUCTION: {CORRECT_INSTRUCTION}
REASON: {CORRECT_REASON}
OUTCOME: {CORRECT_OUTCOME}

INEFFECTIVE (counter-example from same case):
INSTRUCTION: {INEFFECTIVE_INSTRUCTION}
REASON: {INEFFECTIVE_REASON}
OUTCOME: {INEFFECTIVE_OUTCOME}

————————————
OUTPUT
————————————

Output ONLY valid JSON.
```

```
{
“correct_adapted”: “<correct option, edited as little as possible>”,
“ineffective_adapted”: “<ineffective option, edited as little as possible, still substantively poor>”,
“edit_level”: “<one of: ‘none’ | ‘surface’ | ‘light’ | ‘fuller’> // strongest level applied across the two options”,
“edit_notes”: “<one short sentence describing what you changed, or ‘no change’ if edit_level=none>”
}
```

N.16 Reference-Based Response Generation

Prompt

```
————————
REFERENCE — strategies from similar student cases
————————

The strategies below are documented to have worked (or failed) for SIMILAR but DIFFERENT students, retrieved by case similarity. Treat them as reference only — they are NOT a script and were NOT written for THIS student. Adapt them, combine them, or ignore them as appropriate for the student in front of you. The cases are ordered most-similar first and may conflict with each other; use your own judgement.

{REFERENCE}
————————

{PROMPT}
```