

From Test-taking to Cognitive Scaffolding: A Pedagogical Diagnostic Benchmark for LLMs on English Standardized Tests

Luoxi Tang¹, Tharunya Sundar¹, Yuqiao Meng¹, Shuai Yang¹, Ankita Patra¹,
Lakshmi Manohar Chippada¹, Jiqian Zhao², Yi Li², Weicheng Ma³, Zhaohan Xi¹

¹Binghamton University, ²BlossomsAI, ³Oakland University
{ltang24,zxi1}@binghamton.edu

Abstract

As large language models (LLMs) are increasingly integrated into educational tools, current evaluations on standardized tests predominantly focus on binary outcome accuracy. Instead, an effective AI tutor must exhibit faithful reasoning, elucidate solution strategies, and diagnose specific human misconceptions. To bridge this gap, we introduce a pedagogical diagnostic framework that models English Standardized Test (EST) problem-solving as a traversal through a cognitive framework. Based on this framework, we present ESTBOOK, a multimodal benchmark encompassing 10,576 questions and 29 task types across five major exams. Unlike traditional datasets, ESTBOOK goes beyond data aggregation by enriching questions with formalized reasoning trajectories and distractor rationales that capture specific cognitive traps. Through extensive evaluations, we show that ESTBOOK identifies fine-grained reasoning bottlenecks and that lightweight mitigation based on these diagnoses provides evidence that the identified failure modes are actionable.

1 Introduction

AI-driven tools are rapidly transforming the education industry, with large language models (LLMs) increasingly integrated into *English Standardized Tests* (ESTs) such as TOEFL, IELTS, and GRE. Recent advances highlight the use of LLMs in automated scoring (Xia et al., 2024; Zhong et al., 2024; Gupta, 2023), test preparation (Feng and Wang, 2024; Ashrafimoghari et al., 2024), and question generation (Tiratatri et al., 2025). However, the majority of current research evaluating LLMs on standardized tests focuses exclusively on outcome metrics, i.e., whether the model ultimately performs the correct answers.

Notably, raw answering accuracy is fundamentally insufficient for deploying LLMs as intelligent educational assistants. In pedagogical practice, a

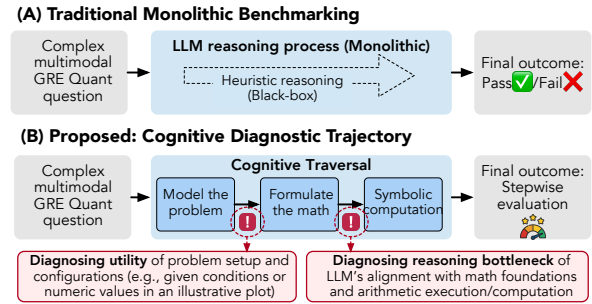


Figure 1: Comparison between monolithic LLM reasoning and our proposed pedagogical diagnostics.

tutor must not only solve the problem but also (i) exhibit a rigorous and faithful reasoning trajectory (Xu et al., 2025; Worden et al., 2026; Yang et al., 2025), (ii) generate explanations that clearly elucidate the solution strategy align with human practice (Abdulsalam and Aroyehun, 2025; Meng et al., 2025b), and (iii) articulate why distractive options are incorrect by diagnosing the specific human misconceptions they represent (Liu et al., 2025; Mitton et al., 2026; Hooshyar et al., 2025). An LLM that arrives at the correct answer through flawed intermediate logic, or one that cannot explain why a specific distractive option represents a common cognitive trap, possesses high answering capability but zero pedagogical utility. To elicit LLMs from simple test-takers to effective tutors, we must evaluate their problem-solving trajectories rather than just their final outputs.

To bridge this gap, we introduce a formalized pedagogical diagnostic benchmark that models standardized test-taking as a **Cognitive Trajectory**. Motivated by the principle of cognitive diagnosis models (Leighton and Gierl, 2007) and the common practice of evidence-centered designs using across test centers (Mislevy et al., 2003), we highlight that answering any complex EST question is not a monolithic task, but a traversal through a structured graph of cognitive sub-skills. As il-

lustrated in Figure 1, solving a multimodal GRE quantitative question requires a progression from modeling the problem (parsing variables from a chart), to mathematical formulation (formulating an equation), to symbolic computation (calculation). By formalizing these steps, our framework allows us to isolate exact reasoning bottlenecks to identify precisely which step in the cognitive trajectory causes an LLM (or a participant) to fail.

To empirically validate this framework, we introduce ESTBOOK, a novel pedagogical diagnostic benchmark spanning five internationally recognized ESTs (TOEFL, IELTS, SAT, GRE, and GMAT). ESTBOOK encompasses 29 question types and 10,576 multimodal examples (incorporating text, images, tables, and mathematical symbols). Notable, our primary contribution is not merely data aggregation. Instead, we heavily manipulate and annotate the dataset for diagnostic purpose. Beyond ground-truth answers, ESTBOOK is enriched with structured breakdown steps mapping to our cognitive trajectory, as well as distractor options that detail the specific cognitive traps or misconceptions associated with incorrect options.

Finally, we conduct extensive empirical evaluations to demonstrate the diagnostic and practical value of ESTBOOK. We utilize the cognitive trajectories, evaluate the stepwise performance, and reveal vulnerabilities where LLMs are highly susceptible to cognitive steps or distractors. Furthermore, we demonstrate that studying these cognitive trajectories facilitates the design of highly effective, task-specific mitigation (e.g., evidence-anchored elicitation and symbolic verification). We enforce strict logical reasoning exactly where the models typically hallucinate to significantly improve LLM robustness and reasoning performance across both knowledge-intensive, reasoning-heavy, or integrative pedagogical tasks.

In summary, our contributions are threefold:

(1) ESTBOOK: We formalize a cognitive diagnostic benchmark, ESTBOOK, contains pedagogical annotations and distractor rationales that shifts the evaluation of LLMs on English tests from binary accuracy to step-by-step cognitive reasoning.

(2) Comprehensive coverage: ESTBOOK covers diverse modalities, containing 29 standardized test types including 10,576 questions to support comprehensive evaluation purposes.

(3) Diagnostic evaluation: We empirically evaluate the effectiveness of our approach across different LLMs. Furthermore, we demonstrate how

our framework can directly guide improvements of model reasoning, which highlights the practical utility for educational practitioners.

2 Related Work

LLMs for Education. LLMs are increasingly used in education for tasks like grading, question generation, and tutoring (Chang et al., 2024; Holmes and Tuomi, 2022; Mohebbi, 2025; Zhang et al., 2024a). Early systems like Codex showed that LLMs can help students across language learning by offering contextual feedback and answers (Chiang et al., 2024; Liang et al., 2024; Wen et al., 2024b). More recent research explores using LLMs as conversational tutors (Sabri et al., 2025). However, most existing evaluations rely on limited benchmarks or informal studies, often focusing on narrow skills like math or reading (Guilherme, 2019; Lee et al., 2023).

Benchmarking LLMs. Many benchmarks test LLMs on general or domain-specific reasoning tasks (Yang et al., 2025; Cao et al., 2025; Meng et al., 2025b,a), evaluating their factual knowledge and reasoning abilities. Some, like MathVista (Lu et al., 2023; Peng et al., 2024) and ScienceQA (Wen et al., 2024a; Zhang et al., 2024d), include images or structured data, but often use synthetic problems or cover narrow domains. Our benchmark, ESTBOOK, is grounded in real standardized exams and spans multiple formats (e.g., multi-choice, text completion), providing a more realistic test of LLMs as educational agents in a heterogeneous problem-solving environment.

Eliciting LLMs. Improving LLM reasoning through elicitation has become an emerging subject (Tang et al., 2025; More et al., 2026). Techniques like In-Context Learning (ICL) (Koike et al., 2024; Yugeswardeenoo et al., 2024), Chain-of-Thought (CoT) (Godwin-Jones, 2024; Wang et al., 2024), and Tree-of-Thought (ToT) (Zhang et al., 2024b,c) guide models to generate step-by-step answers and improve performance on tasks like math and logic. Yet, these methods are mostly tested on clean, synthetic datasets (Askarbekuly and Aničić, 2024; Schmidhuber and Kruschwitz, 2024; Meng et al., 2026). Instead, we evaluate LLMs with fine-grained question structure and offer insights about LLMs eligibility on each isolated reasoning step.

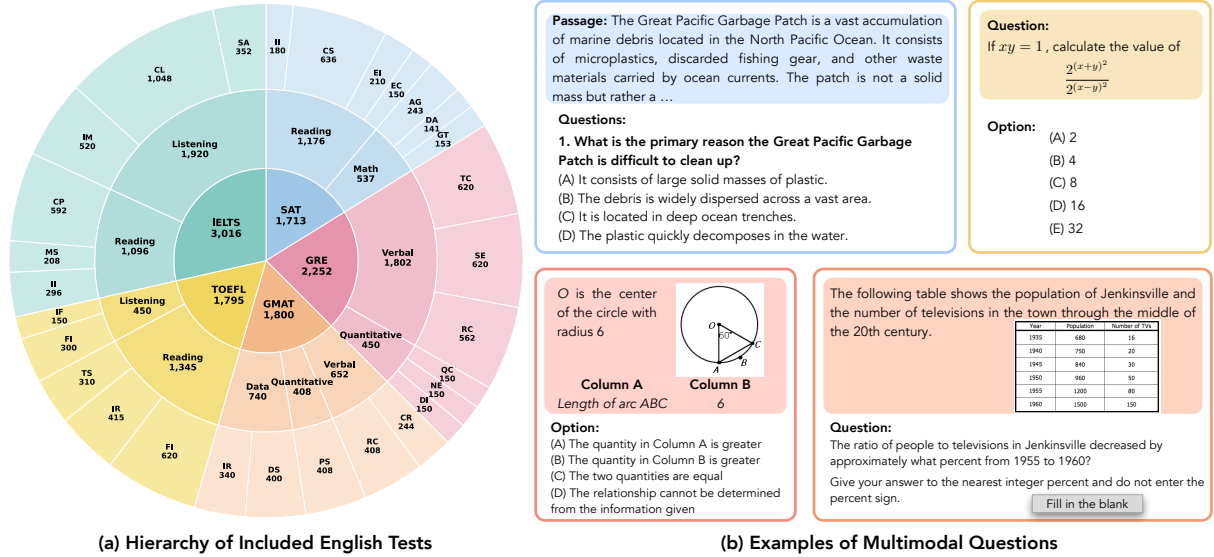


Figure 2: **Overview of ESTBOOK:** (a) the hierarchical structure of included English tests (detailed descriptions and abbreviations are provided in Table 4, with concrete examples in Appendix C). (b) Examples of multimodal questions in ESTBOOK.

3 ESTBOOK: A Pedagogical Diagnostic Benchmark for ESTs

3.1 Scope and Data Collection

To empirically validate our diagnostic framework, we construct ESTBOOK, a large-scale, multimodal dataset comprising 10,576 questions across 29 distinct question types. The benchmark aggregates high-quality data from five internationally recognized English Standardized Tests (ESTs): SAT, GRE, GMAT, TOEFL, and IELTS, as summarized in Figure 2, with detailed introduction and rationale of their representativeness in Appendix A.1.

Modality. (Details are provided in Appendix A.2). ESTBOOK is designed to reflect the complex, heterogeneous cognitive demands of real-world test-taking. The dataset spans five primary modalities: text, mathematical symbols, images, tables, and audio. This multimodal structure ensures that our evaluation extends beyond pure natural language processing to include visual parsing, spatial reasoning, and auditory comprehension.

Question Type. For the purpose of rigorous pedagogical diagnosis, ESTBOOK strictly focuses on *objective questions* (e.g., multiple-choice and deterministic numeric entry). This constraint provides a well-defined answer and a verifiable solution trajectory for each question. It also allows us to score intermediate reasoning steps and identify the cognitive step where an error occurs. Each question is mapped to our cognitive framework (§3.2).

Source and Quality Control. (Details are provided in Appendix A.3 and A.4). All data is curated from official preparation guides, released practice exams, and vetted open-access educational platforms (Graduate Management Admission Council (GMAC), 2025; Woldoff, 2024; Appelrouth and Zabrocky, 2017; Josué et al., 2023; Pereira et al., 2024). To ensure strict **quality control and licensing compliance**, all test sources were manually validated for pedagogical fidelity, and proprietary multimodal elements were carefully reconstructed to mirror authentic test conditions.

3.2 Constructing the Cognitive Trajectory: A Diagnostic Framework of EST Sub-skills

To evaluate LLMs through a step-level diagnostic framework, we map each EST question type to a **Cognitive Trajectory**. Motivated by the principle of cognitive diagnosis models (Leighton and Gierl, 2007) used in psychometrics, we posit that solving an EST question Q requires traversing a specific ordered trajectory of underlying cognitive sub-skills, denoted as $C = (c_1, c_2, \dots, c_n)$.

Given that evidence-centered practice is commonly used across test centers (Mislevy et al., 2003), we correspondingly formalize the cognitive trajectory as a universal set of cognitive nodes $\mathcal{N}$, ranging from knowledge retrieval (c_{lexical}) to fluid logic ($c_{\text{formulate}}$) and final execution (c_{compute}). We route an LLM’s reasoning trace through these tra-

Table 1: Taxonomy of Cognitive Trajectories across EST Question Types. The traversal paths illustrate the sequential cognitive nodes required to solve the task, enabling precise diagnostic bottleneck isolation. Table 4 provides detailed mapping from EST tasks to pedagogical categories.

Pedagogical Category	Task Domain	Cognitive Trajectory (C)	Distractor Type
Knowledge-intensive (Lexical & rule-based)	Structural reasoning	$c_{\text{syntax_parse}} \rightarrow c_{\text{rule_match}} \rightarrow c_{\text{predict}}$	Grammar & syntax deficits
	Semantic reasoning	$c_{\text{semantic_parse}} \rightarrow c_{\text{logical_scope}} \rightarrow c_{\text{resolve}}$	Vocabulary & lexical gaps
Reasoning-intensive (math & multimodal)	Data interpretation	$c_{\text{analytical_goal}} \rightarrow c_{\text{visual_parse}} \rightarrow c_{\text{analyze}}$	Multimodal alignment errors
	Numeric calculation	$c_{\text{model}} \rightarrow c_{\text{formulate_eq}} \rightarrow c_{\text{compute}}$	Symbolic & execution errors
Hybrid integration (textual & inference)	Evidence finding	$c_{\text{intent}} \rightarrow c_{\text{perceive_text}} \rightarrow c_{\text{extract}}$	Information retrieval failures
	Comparative inference	$c_{\text{entities}} \rightarrow c_{\text{constraints}} \rightarrow c_{\text{evaluate}}$	Logical fallacies / hallucination

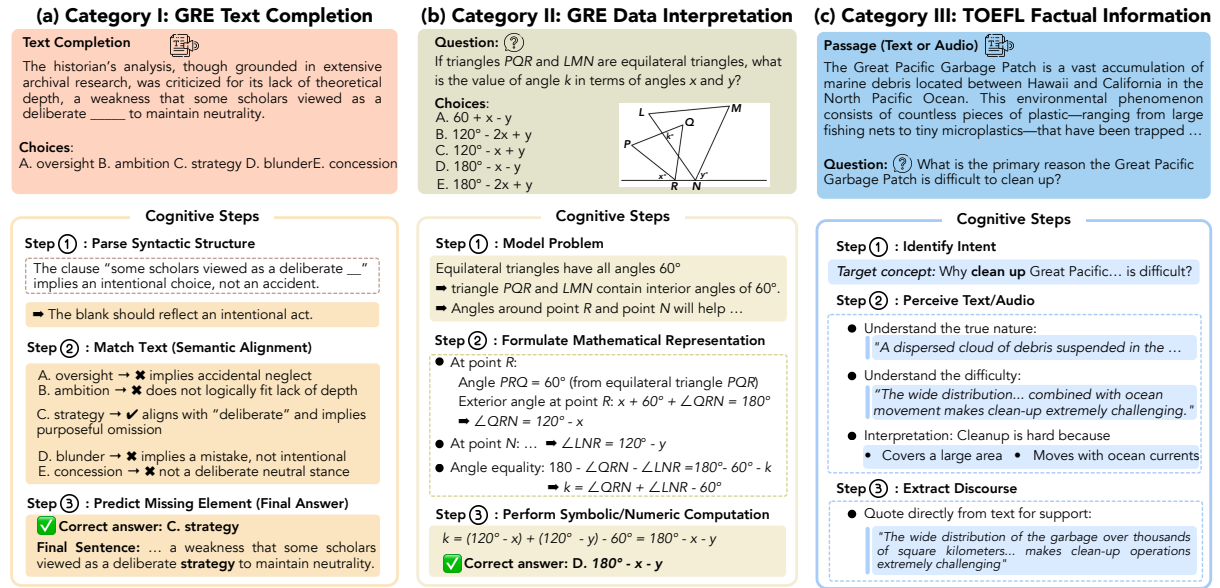


Figure 3: Illustrative cognitive trajectories for solving EST questions.

jectories and can subsequently pinpoint the exact cognitive bottleneck where an error occurs. As detailed in Table 1, we consolidate the diverse formats of ESTs into three distinct pedagogical categories based on their reliance on knowledge retrieval versus multi-step reasoning:

Category I: Knowledge-intensive retrieval (lexical & structural): EST questions in this category test the model’s internalized understanding of linguistic rules and vocabulary. A successful cognitive process relies on localized pattern matching across two sub-categories:

- (I-1) Structural reasoning questions require the progression of analyzing syntax, matching text rules, and predicting missing elements ($c_{\text{syntax_parse}} \rightarrow c_{\text{rule_match}} \rightarrow c_{\text{predict}}$). Examples include GRE’s text completion.
- (I-2) Semantic reasoning questions involve interpreting fine-grained semantics and resolving logical equivalence ($c_{\text{semantic_parse}} \rightarrow$

$c_{\text{logical_scope}} \rightarrow c_{\text{resolve}}$), as seen in GRE’s sentence equivalence. Failure at these nodes typically indicates a pure knowledge gap rather than a reasoning flaw.

Category II: Reasoning-intensive execution (multimodal & quantitative): These tasks require minimal external knowledge but demand rigorous, multi-step logic, spatial awareness, and mathematical execution.

- (II-1) Data interpretation questions typically include multimodality involving tables or charts. The model must formulate an analytical goal, parse the non-textual visual input, and analyze the underlying data ($c_{\text{analytical_goal}} \rightarrow c_{\text{visual_parse}} \rightarrow c_{\text{analyze}}$).
- (II-2) Numeric calculation questions, e.g., SAT Math and GRE Quantitative, require translating natural language descriptions into formal mathematical models, formulating

symbolic representations, and performing strict arithmetic computation ($c_{\text{model}} \rightarrow c_{\text{formulate_eq}} \rightarrow c_{\text{compute}}$). Bottlenecks here frequently reveal that LLMs can correctly formulate an equation ($c_{\text{formulate_eq}}$) but fail at pure arithmetic execution (c_{compute}), a critical pedagogical insight.

Category III: Hybrid integration (semantic extraction & inference): These tasks bridge knowledge retrieval and fluid logic. The model must synthesize large contexts (text or audio) and apply deductive reasoning to information that is not explicitly stated.

- (III-1) Evidence finding questions ask the model to identify the central intent, locate relevant evidence within a dense context, and extract the discourse ($c_{\text{intent}} \rightarrow c_{\text{perceive_text}} \rightarrow c_{\text{extract}}$). This trajectory is heavily utilized in TOEFL’s and GRE’s reading comprehension.
- (III-2) Comparative inference questions require identifying comparative entities, applying given constraints, and evaluating logical sufficiency ($c_{\text{entities}} \rightarrow c_{\text{constraints}} \rightarrow c_{\text{evaluate}}$), common in GMAT’s critical reasoning.

3.3 Data Annotation and Curation with Cognitive Trajectories and Distractor Rationale

ESTBOOK restructures EST problem-solving into the specific cognitive trajectories defined above. To ensure high fidelity, we develop an NLP-driven annotation pipeline to deterministically construct the ground-truth cognitive trajectories and distractor rationales, without relying on generative LLM annotation. We first apply dependency parsing and rule-based syntactic mapping to official test-preparation explanations, wherein we chronologically extract the explicit cognitive nodes (e.g., visual parsing, equation formulation) required to solve each question. Furthermore, we categorize incorrect multiple-choice options into a rigorous taxonomy of cognitive traps (e.g., partial truth, execution error) utilizing classical NLP techniques such as lexical overlap analysis, negation scope detection, and named entity recognition. Following human-in-the-loop validation, this pipeline yields a pedagogically grounded benchmark where both correct reasoning paths and human-like failure modes are systematically formalized. Comprehensive details of this annotation methodology are deferred to Appendix B.

Validation. We further validate the deterministic annotation procedure against both expert labels and LLM-based annotation. On a stratified sample of 300 questions covering the three pedagogical categories, two human annotators independently label the cognitive trajectories and distractor types, reaching substantial agreement ($\kappa = 0.87$). Using the expert-adjudicated labels as reference, the deterministic pipeline achieves 91.8 node-level F1 and 84.7% trajectory exact match, compared with 92.6 and 85.3% for LLM-based annotation. The difference is not statistically significant ($p = 0.31$). These results show that the deterministic pipeline provides annotation quality comparable to LLM annotation while keeping the benchmark construction reproducible.

4 Experiment

ESTBOOK aims to empirically answer several research questions: **RQ₁**: How do state-of-the-art LLMs perform on EST questions? **RQ₂**: How effectively do LLMs execute cognitive reasoning steps within the cognitive framework? **RQ₃**: How robust are LLMs to distractor options? **RQ₄**: Can the diagnosed cognitive bottlenecks guide effective mitigation?

LLMs. Given the multimodal nature of ESTBOOK, we evaluate several industry-leading Multimodal LLMs, including GPT-5, GPT-4V, Claude-Sonnet-4, Llama-4-Scout-17B, Qwen-VL-Max, and Gemini-2.5. We adopt OpenAI’s Whisper (Andreyev, 2025; Graham and Roll, 2024) to transcribe audio data within listening tasks in TOEFL and IELTS.

Human Tester. To demonstrate LLM performance alongside humans, we included five student testers with their consent to share the data, wherein each tester had recently prepared for and participated in at least one of these tests.

Prompting Strategy. We evaluate three popular prompting methods: (1) **In-Context Learning (ICL)**: Besides basic instructions to describe the question type, the prompt also includes several (we select five) examples to offer LLMs the solution style. (2) **Chain-of-Thought (CoT)** (Bi et al., 2025; Zhang et al., 2024b): The prompt encourages the model to generate intermediate reasoning steps before generating final answers. (3) **Tree-of-Thought (ToT)** (Long, 2023; Yao et al., 2023): An advanced strategy that guides LLMs to explore multiple reasoning paths and select the most plausible

Table 2: Results on ESTBOOK. Human performance is reported as the average across five independent testers. For each method’s results, we report the average over five independent runs with adjusted temperatures (0.2, 0.5, 0.7, 0.9, 1.0). Standard deviations are shown in **Table 6**.

Task	Human	GPT-4V			GPT-5			Claude-Sonnet-4			Llama-4-Scout-17B			Qwen-VL-Max			Gemini-2.5		
		ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT
SAT																			
II	82.1	73.3	75.6	80.6	72.8	82.2	86.4	83.9	85.4	89.7	81.2	87.5	90.4	88.3	89.2	91.5	85.0	87.3	92.6
CS	74.0	68.2	77.4	84.9	73.9	82.7	87.2	55.8	70.2	66.3	46.4	61.8	55.7	75.6	82.9	78.8	93.7	94.1	87.2
EI	77.5	78.1	79.5	78.6	84.0	84.5	82.1	50.5	62.4	64.8	48.6	52.5	51.2	59.0	61.9	66.1	72.4	70.4	71.8
EC	89.0	84.7	89.3	81.3	93.8	92.2	84.7	72.0	74.2	70.1	64.0	65.4	63.9	81.3	77.2	79.3	93.3	90.6	92.1
AG	55.1	28.4	44.0	60.9	31.7	53.2	76.1	33.3	52.6	79.4	30.0	46.7	68.3	35.8	50.1	81.6	34.2	52.7	82.4
DA	77.9	56.7	70.9	85.8	60.3	78.2	90.7	58.2	71.4	90.1	51.1	60.2	87.3	54.6	67.2	89.2	53.9	69.7	88.0
GT	63.0	66.7	64.7	67.3	73.9	70.5	71.6	49.7	50.8	47.2	41.2	44.8	38.1	33.3	30.4	32.5	58.8	59.0	56.2
GRE																			
TC	76.2	72.6	77.4	83.1	68.5	73.4	82.1	69.4	75.5	72.4	53.5	61.0	64.8	67.7	73.1	78.3	68.5	80.2	82.4
SE	81.5	78.9	81.0	79.8	87.7	86.5	87.2	85.5	82.1	83.5	66.0	67.5	63.2	71.8	73.2	71.6	77.6	74.8	75.9
RC	70.2	67.1	77.8	86.1	83.6	87.1	81.5	61.9	69.3	76.0	46.3	54.2	73.2	70.6	76.2	80.1	56.9	73.2	78.6
QC	68.1	55.3	57.3	51.3	82.0	84.1	83.8	41.3	48.2	44.6	51.3	56.0	42.7	54.7	50.3	45.7	48.0	58.4	53.1
NE	73.7	32.7	38.0	52.7	28.7	33.9	48.2	17.3	25.0	37.2	23.3	30.1	44.5	29.3	28.1	40.8	26.0	33.0	30.2
DI	55.5	52.0	56.0	73.3	32.7	36.5	63.2	21.3	25.7	50.1	40.0	41.2	65.1	38.7	40.5	61.7	48.0	47.2	67.1
GMAT																			
CR	66.2	62.3	77.9	72.5	57.4	70.1	71.4	55.7	79.5	74.8	65.6	69.2	71.3	57.4	75.6	70.2	56.1	74.4	72.7
RC	82.1	79.2	88.7	91.4	65.2	71.4	75.6	63.5	81.1	86.2	47.3	74.5	70.3	68.6	74.4	76.8	59.1	75.0	77.4
PS	73.7	24.0	34.3	41.2	26.0	31.1	54.2	19.1	24.5	27.2	18.6	22.5	35.0	22.1	25.6	33.7	25.0	28.3	38.4
DS	52.0	14.5	26.8	24.5	13.5	32.4	40.8	12.0	16.0	19.2	13.8	14.5	20.1	14.8	21.0	23.6	9.0	13.5	22.0
IR	59.2	11.2	13.8	22.1	11.8	16.0	20.3	8.8	15.0	17.4	3.2	16.2	18.0	10.0	11.2	18.7	12.1	14.4	20.5
TOEFL																			
FI	86.5	82.3	86.3	74.2	85.5	93.2	70.5	76.6	83.9	82.0	65.3	68.8	65.7	73.2	70.5	75.1	73.5	84.1	86.3
IR	74.1	63.4	85.3	87.7	79.3	84.2	85.0	55.9	59.2	63.0	46.0	62.2	58.3	73.5	74.0	75.2	79.0	81.0	82.6
TS	85.0	83.9	86.1	84.8	83.9	84.0	81.7	83.5	85.0	82.4	74.2	75.8	73.0	73.5	75.5	76.2	66.1	67.0	66.8
FI	93.1	93.7	95.7	97.7	94.0	93.2	98.5	80.7	86.5	82.5	67.7	69.2	76.6	74.7	70.8	76.3	81.3	92.5	89.7
IF	70.1	62.0	64.7	67.3	81.3	88.4	90.8	70.7	82.0	79.1	55.3	58.8	61.2	53.3	62.4	55.8	68.0	72.8	80.2
IELTS																			
II	82.0	79.1	84.8	82.8	81.1	86.0	88.4	79.1	82.0	79.5	75.7	74.5	71.3	73.0	76.2	74.1	83.1	84.2	86.0
MS	93.6	83.7	85.1	81.3	81.7	83.0	83.7	73.1	81.0	83.2	66.8	74.0	76.0	69.2	71.2	73.7	75.5	82.5	80.4
CP	71.8	66.0	67.2	72.1	83.1	82.4	84.0	71.8	84.4	85.5	58.4	73.1	75.6	73.5	72.4	76.7	82.1	81.0	83.9
IM	86.1	83.7	84.8	88.3	90.6	91.5	92.8	74.0	76.0	75.1	64.2	66.4	68.3	73.1	72.0	74.8	83.7	89.2	91.6
CL	88.3	80.5	84.6	83.1	83.6	91.0	90.4	72.5	74.8	73.0	41.3	61.0	66.7	58.2	64.4	67.3	76.1	82.0	88.5
SA	85.1	83.0	86.4	84.7	83.0	85.1	84.0	73.9	77.0	75.0	66.2	70.2	67.6	73.3	76.4	74.7	82.1	84.9	83.2

one. Prompts are shown in **Appendix E**.

4.1 Evaluating LLMs Performance (RQ1)

Problem-solving abilities. Table 2 presents the performance of various LLMs on ESTBOOK. Despite extensive pretraining on large-scale English corpora, these models exhibit substantial variability across different EST tasks, even within similar domains and modalities. For instance, in linguistic tasks such as GRE Expression of Ideas (EI) and English Conventions (EC), GPT-4V achieves 79.5% and 89.3% accuracy, respectively, revealing its inconsistent ability to handle fine-grained distinctions in grammar, style, and logical flow. Similarly, LLMs do not always outperform human testers despite their advanced prompting methods (e.g., CoT or ToT). Those observations suggest that LLMs often struggle with the contextual sensitivity required for generalizing to diverse test problems. We provide more results in **Appendix F.1**.

Influence by modality. The limitations of LLMs become more obvious when complex modalities are involved, such as GMAT Integrated Reasoning (IR) and GRE Data Interpretation (DI).

Case Study. (GMAT – Integrated Reasoning):

You are given (i) A **table** showing sales data by region and quarter (e.g., North America). (ii) A **text passage** describing factors that influenced sales in different regions (e.g., "A new competitor entered the European market in Q2...").

Question: Which region experienced the largest relative revenue drop between Q1 and Q2?

Challenges for LLMs: 1. **Mapping text to table:** Claude and GPT-5 fail to connect the textual clue ("new competitor in Europe") with the relevant table entry (Europe, Q2 revenue). 2. **Reasoning with partial information:** Gemini overlooks the hint about the competitor’s impact and fails to compare percentage drops across regions, missing the correct answer.

The challenge is twofold: first, models must align disparate representations (e.g., mapping textual queries to tabular structures); second, they must reason over incomplete or distributed evidence, a skill that current architectures and training regimes are not fully optimized for.

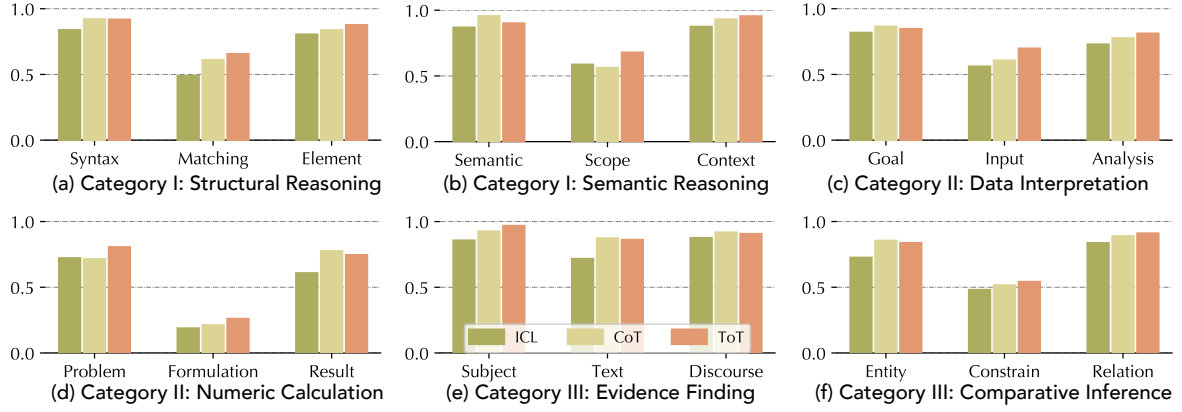


Figure 4: Detailed stepwise analysis on GPT-5 on each cognitive step.

These multimodal failures suggest that achieving human-level performance on ESTs requires more than language modeling proficiency; it demands integrated reasoning capabilities that span visual, symbolic, and logical modalities. Together, these observations highlight the inherent difficulty of EST-style questions and the under-preparedness of even the strongest LLMs to serve as reliable tutors for real-world educational settings. We provide more studies and failure modes in **Appendix F.2**.

More analysis. Due to space constraints, we defer additional analyses and discussions to the appendix: (1) **App. F.3** quantifies how LLMs align with human testers and includes statistical tests demonstrating the significance of behavioral similarities between LLMs and human testers. (2) **App. F.4** examines the impact of different prompting strategies (ICL, CoT, and ToT) with supporting case studies. (3) **App. F.5** analyzes the influence of question difficulty. (4) **App. F.6** quantifies the answering time of LLMs.

4.2 Isolated Measurement of Each Cognitive Step (RQ₂)

Following our formalized cognitive trajectories (§3.2), we evaluate the step-by-step reasoning capabilities of LLMs using ground-truth trajectories extracted via our deterministic annotation pipeline (**Appendix B**). Because each cognitive node exhibits distinct properties. For instance, syntactic parsing requires exact matching, whereas numeric computation necessitates error tolerance. We thus deploy a tailored suite of node-level metrics (e.g., IoU, BERTScore, and Normalized RMSE) to address the specific diagnostic requirements of each reasoning stage with full metric definitions articulated in **Appendix G.1**. Figure 4 and **Appendix**

G.2 present the performance breakdown across all cognitive steps.

Audit of valid alternative reasoning paths.

Reference trajectories may not enumerate every valid reasoning path. We therefore manually audit outputs that are scored as incorrect across all trajectory types. We find that 7.4% correspond to valid alternative paths that reach the same correct intermediate state, with 2.1%, 2.8%, and 2.5% occurring at Steps 1, 2, and 3, respectively. These cases are distributed across the three step positions and do not increase at later steps. Thus, alternative-path under-crediting exists, but it does not explain the systematic stepwise performance decline reported below.

From this stepwise analysis, we derive the following key observations:

LLMs are generally strong formulators but occasionally weak reasoners. Overall, we observe that LLMs consistently excel in the initial step across all tasks, achieving up to 97% accuracy. These early steps demonstrate the models’ strong capability to interpret and structure EST problems appropriately, such as task identification, problem formulation, or topic modeling. However, performance significantly declines in subsequent reasoning steps, which vary across tasks. This decline is particularly evident in tasks that demand causality inference or evidence synthesis.

Case Study (GMAT – Critical Reasoning): "... cities with more EV charging stations tend to have lower levels of air pollution. As a result, the city of Greentown has decided to install a large number of EV charging stations ..."

LLM Reasoning: GPT-5 incorrectly state "The

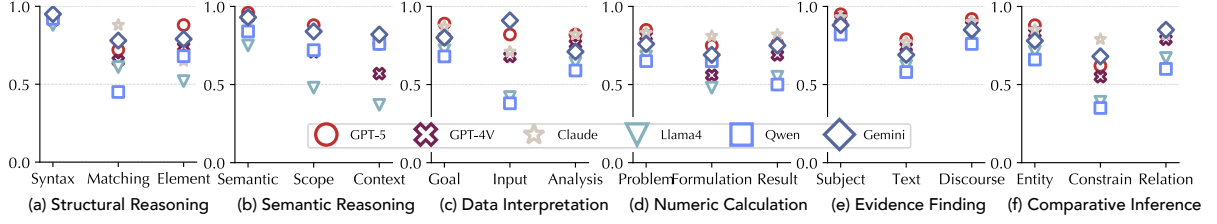


Figure 5: Analysis LLMs performance to distractor options for each cognitive step.

city of Greentown currently has a very small number of EVs in use" due to being distracted by the phrase "the current number of EVs that does not directly relevant to causal logic.

These results suggest that LLMs remain limited and unstable in executing complex reasoning chains, which is an essential requirement for robust educational support.

Complex logic has more impact on LLM screening than long context. We find that context length alone does not impede LLMs to locate relevant information. Instead, reasoning complexity plays a greater role in determining success or failure. Models can navigate long inputs effectively if the task only requires surface-level matching, but they often fail when logical integration across multiple sentences is required. These results suggest that long context alone is not a major barrier for modern LLMs. However, once the task requires multi-hop reasoning or integrating dispersed evidence, even top-performing models struggle.

More analysis and case studies. Due to space constraints, additional findings and case studies are provided in App. G.2 and G.3.

4.3 LLM Performance on Distractors (RQ₃)

After RQ₂ evaluates LLMs performance on ground-truth answers, we diagnose LLMs regarding their robustness on distractor options. For each task, we prompt the LLMs with the incorrect distractor options and a rationale to ask them to accept or reject each distractor options. We report the overall accuracy of selection following the 3-step cognitive trajectory. If the model hallucinates a selection (e.g., erroneously accept a distractor), it receives a score of zero for that instance.

Independence of distractor susceptibility and generative robustness Comparing our results from RQ₂ (performance on ground-truth results) with RQ₃ (robustness against distractor options), we observe a clear independence between genera-

Table 3: Effect of targeted mitigation strategies across representative ESTBOOK tasks. Scores reflect baseline Chain-of-Thought (CoT) versus our bottleneck-specific Mitigated (Mit.) prompting.

Task & Mitigation	GPT-4V		GPT-5		Claude-S-4	
	Base	+Mit.	Base	+Mit.	Base	+Mit.
GRE-RC + Evidence-Anchored	77.8	93.5	87.1	90.4	69.3	84.1
GRE-TC + Syntax-First	79.5	84.2	73.4	78.1	61.0	76.7
GMAT-IR + Table-Alignment	13.8	59.7	16.0	71.9	15.0	69.4

tive problem-solving and discriminative robustness. Models like GPT-5 and Claude demonstrate notably degrades when explicitly prompted to evaluate and reject carefully engineered distractor rationales. For instance, in Category I (Semantic Reasoning), although models exhibit strong generative capabilities, introducing distractors that utilize partial semantic truths causes severe performance drops. This disentanglement implies that high accuracy on standard benchmarks (RQ₂) does not guarantee pedagogical robustness. As LLMs work as plausible reasoners rather than logical discriminators, highly susceptible behaviors to the same cognitive traps may misguide human test-takers.

Failure mode study. Due to space constraints, we defer the failure model analysis to App. H.1.

4.4 Diagnostic-Guided Mitigation

(RQ₄)

To test whether the cognitive bottlenecks identified in RQ₁–RQ₃ are actionable, we apply lightweight task-specific elicitation strategies targeted to these diagnosed failures. In cognitive trajectory, we enforce stepwise constraints before a model is permitted to generate its final evaluation. For instance, we introduce an **evidence-anchored CoT** in GRE reading comprehension (RC) that mandates verbatim text extraction prior to high-level interpretation, a **syntax-first CoT** in GRE-TC forces the explicit identification of grammatical roles (e.g., contrast markers) to stop models from conflating semantic plausibility with syntactic va-

lidity, and a **table-alignment Constraint** in GMAT-IR that requires explicit textual-to-tabular mapping instructions before numerical operations begin. As shown in Table 3, isolating and promoting these cognitive steps yield consistent performance gains across all frontier models. These interventions are intentionally lightweight and serve to test whether the diagnostic signals produced by ESTBOOK identify failure modes that can be acted upon. The consistent gains therefore provide evidence for the practical utility of the diagnostic framework.

Held-out non-public evaluation. Public standardized-test materials may overlap with model pretraining data, which could confound positive mitigation gains. We therefore rerun all mitigation strategies on 1,200 newly written exam-style questions that were never publicly released before evaluation. The same bottleneck-specific interventions produce gains of 1.63–4.84%, and the distribution of improvements does not differ significantly from that observed on the original benchmark ($p > 0.05$). This result reduces the likelihood that the mitigation improvements primarily arise from recovering memorized public test content.

We defer (1) comprehensive mitigation results to App. I, (2) additional mitigation strategies and discussions to App. J, and (3) insights to advance LLMs in pedagogical scenarios to App. K.

5 Conclusion

This work explores LLM performance on ESTs with the introduction of ESTBOOK. We move beyond surface-level accuracy metrics to model problem-solving as a structured cognitive trajectory. Through extensive experiments, we observe that current LLMs possess the linguistic fluency to conceptualize problems and execute computations, while their generative bias toward plausible rationalization making them highly susceptible to distractor traps during intermediate constraint mapping. We further introduce targeted mitigation strategies to improve LLM robustness in pedagogical settings. Ultimately, this research provides a rigorous diagnostic perspective and an actionable way toward more pedagogically sound AI systems.

6 Limitations

Despite the comprehensive design of ESTBOOK and our extensive evaluation across leading LLMs, several limitations warrant discussion.

Model Access and Coverage. Our evaluation focuses on a set of industry-leading multimodal and visual LLMs that offer public inference APIs or open-source checkpoints. However, access constraints (e.g., usage quotas, proprietary architecture details) limit broader inclusion of commercial models or fine-tuned educational agents. This may omit systems with specialized adaptations for test-taking tasks.

Granularity of Breakdown Analysis. Our breakdown framework assumes that preceding steps are perfectly resolved, enabling clean isolation of reasoning subtasks. While this reveals bottlenecks in specific capabilities, it does not reflect real-world interactions where upstream errors may cascade. Hence, the observed step-wise performance may overestimate true end-to-end reliability in tutoring applications.

References

- Ramatu Oiza Abdulsalam and Segun Aroyehun. 2025. Large language models approach expert pedagogical quality in math tutoring but differ in instructional and linguistic profiles. *arXiv preprint arXiv:2512.20780*.
- Allison Andreyev. 2025. Quantization for openai’s whisper models: A comparative analysis. *arXiv preprint arXiv:2503.09905*.
- Jed I. Appelrouth and Karen M. Zabrocky. 2017. Preparing for the sat: A review. *College and University*, 92(1):2–17.
- Vahid Ashrafimoghari, Necdet Gürkan, and Jordan W Suchow. 2024. Evaluating large language models on the gmat: Implications for the future of business education. *arXiv preprint arXiv:2401.02985*.
- Nursultan Askarbekuly and Nenad Aničić. 2024. Llm examiner: automating assessment in informal self-directed e-learning using chatgpt. *Knowledge and Information Systems*, 66(10):6133–6150.
- Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. 2025. [Forest-of-thought: Scaling test-time compute for enhancing llm reasoning](#). *Preprint*, arXiv:2412.09078.
- College Board. 2025. [The cognitively complex thinking required by select sat suite questions: Evidence from students with specific learning disorders affecting reading \(dyslexia\)](#). Technical report, College Board. Cognitive Lab Report.
- Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, and 1 others. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.
- Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. 2024. Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. *arXiv preprint arXiv:2407.05216*.
- College Board. 2022. Digital sat sample questions and answer explanations. <https://satsuite.collegeboard.org/media/pdf/digital-sat-sample-questions.pdf>.
- Yang Feng and Xiya Wang. 2024. Exploring the development of chinese college students’ proficiency in english through chatgpt: An experimental study. In *Proceedings of the 2024 the 16th International Conference on Education Technology and Computers*, pages 148–154.
- Robert Godwin-Jones. 2024. Distributed agency in second language learning and teaching through generative ai. *arXiv preprint arXiv:2403.20216*.
- Graduate Management Admission Council (GMAC). 2025. *GMAT Official Guide 2025–2026 Bundle: Books + Online Question Bank*, 1st edition. Wiley.
- Calbert Graham and Nathan Roll. 2024. Evaluating openai’s whisper asr: Performance analysis across diverse accents and speaker traits. *JASA Express Letters*, 4(2).
- Gary R Gruber. 2011. *Gruber’s Complete GRE Guide 2012*. Sourcebooks, Inc.
- Alex Guilherme. 2019. Ai and education: the importance of teacher and student relations. *AI & society*, 34:47–54.
- Pranav Gupta. 2023. Testing llm performance on the physics gre: some observations. *arXiv preprint arXiv:2312.04613*.
- Lisa Zimmer Hatch, Scott A. Hatch, and Sandra Luna McCune. 2023. *GMAT Prep 2024/2025 For Dummies (GMAT Focus Edition): Book + 3 Practice Tests + 100 Flashcards Online*, 11th edition. For Dummies.
- Wayne Holmes and Ilkka Tuomi. 2022. State of the art and practice in ai in education. *European journal of education*, 57(4):542–570.
- Danial Hooshyar, Yeongwook Yang, Gustav Šřif, Tommi Kärkkäinen, Raija Hämäläinen, Mutlu Cukurova, and Roger Azevedo. 2025. Problems with large language models for learner modelling: Why llms alone fall short for responsible tutoring in k–12 education. *arXiv preprint arXiv:2512.23036*.
- IELTS-up. Ielts listening practice tests. <https://ielts-up.com/listening/ielts-listening-practice.html>.
- Christopher J. Johnstone, Nicole A. Bottsford-Miller, and Sandra J. Thompson. 2006. *Using the think aloud method (cognitive labs) to evaluate test design for students with disabilities and english language learners*. Technical Report 44, University of Minnesota, National Center on Educational Outcomes.
- Anghelo Josué, Mirna Carolina Bedoya-Flores, Erick Fabián Mosquera-Quinonez, Ángel Enrique Mesías-Simisterra, and José Vicencio Bautista-Sánchez. 2023. Educational platforms: Digital tools for the teaching-learning process in education. *Ibero-American Journal of Education & Society Research*, 3(1):259–263.
- Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. Outfox: Llm-generated essay detection through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 21258–21266.
- Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Yunseo Lee, Jewoong Moon, Jieun Lim, and Hyeoncheol Kim. 2023. Generative agent for teacher training: Designing educational problem-solving simulations with large language model-based agents for pre-service teachers. *NeurIPS’23 Workshop on Generative AI for Education (GAIED)*.
- Jacqueline Leighton and Mark Gierl. 2007. *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.
- Zhenwen Liang, Ye Liu, Tong Niu, Xiangliang Zhang, Yingbo Zhou, and Semih Yavuz. 2024. Improving llm reasoning through scaling inference computation with collaborative verification. *arXiv preprint arXiv:2410.05318*.
- Naiming Liu, Shashank Sonkar, and Richard Baraniuk. 2025. Do llms make mistakes like students? exploring natural alignments between language models and human error patterns. In *International Conference on Artificial Intelligence in Education*, pages 364–377. Springer.
- Eric Loken, Filip Radlinski, Vincent H. Crespi, Josh Millet, and Lesleigh Cushing. 2004. *Online study behavior of 100,000 students preparing for the sat, act, and gre*. *Journal of Educational Computing Research*, 30(3):255–262.
- Jieyi Long. 2023. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.

- Abdullah Mallik. Sat suite question bank. <https://sat-questions.onrender.com/categories>.
- Patrick E McKnight and Julius Najab. 2010. Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.
- Yuqiao Meng, Luoxi Tang, Feiyang Yu, Jinyuan Jia, Guanhua Yan, Ping Yang, and Zhaohan Xi. 2025a. Uncovering vulnerabilities of llm-assisted cyber threat intelligence. *arXiv preprint arXiv:2509.23573*.
- Yuqiao Meng, Luoxi Tang, Feiyang Yu, Xi Li, Guanhua Yan, Ping Yang, and Zhaohan Xi. 2025b. Benchmarking llm-assisted blue teaming via standardized threat hunting. *arXiv preprint arXiv:2509.23571*.
- Yuqiao Meng, Luoxi Tang, Dazheng Zhang, Rafael Brens, Elvys J Romero, Nancy Guo, Safa Elkefi, and Zhaohan Xi. 2026. Small agent group is the future of digital health. *arXiv preprint arXiv:2602.08013*.
- Robert J Mislevy, Russell G Almond, and Janice F Lukas. 2003. A brief introduction to evidence-centered design. *ETS Research Report Series*, 2003(1):i–29.
- Joshua Mitton, Prarthana Bhattacharyya, Digory Smith, Thomas Christie, Ralph Abboud, and Simon Woodhead. 2026. Misconception diagnosis from student-tutor dialogue: Generate, retrieve, rerank. *arXiv preprint arXiv:2602.02414*.
- Ahmadreza Mohebbi. 2025. Enabling learner independence and self-regulation in language education using ai tools: a systematic review. *Cogent Education*, 12(1):2433814.
- Vrushket More, Lyra Lu, Zeyu Ding, Zhaohan Xi, Seth Mizia, and Nancy L Guo. 2026. Theramind: a multi-llm ensemble for accelerating drug repurposing in lung cancer via case report mining. *npj Precision Oncology*.
- Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*.
- Juanan Pereira, Juan-Miguel López, Xabier Garmendia, and Maider Azanza. 2024. Leveraging open source llms for software engineering education and training. In *2024 36th International Conference on Software Engineering Education and Training (CSEE&T)*, pages 1–10. IEEE.
- Hamoun Sabri, Muhammad HA Saleh, Parham Hazrati, Keith Merchant, Jonathan Misch, Purnima S Kumar, Hom-Lay Wang, and Shayan Barootchi. 2025. Performance of three artificial intelligence (ai)-based large language models in standardized testing; implications for ai-assisted dental education. *Journal of periodontal research*, 60(2):121–133.
- Maximilian Schmidhuber and Udo Kruschwitz. 2024. Llm-based synthetic datasets: Applications and limitations in toxicity detection. *LREC-COLING*, 37:2024.
- Luoxi Tang, Yuqiao Meng, Ankita Patra, Weicheng Ma, Muchao Ye, and Zhaohan Xi. 2025. Polar: Automating cyber threat prioritization through llm-powered assessment. *arXiv preprint arXiv:2510.01552*.
- Thanayut Tiratatri, Kanin Sukittivarapunt, Thammathat Sarasinpitak, and Aung Pyae. 2025. Designing an llm-based ielts question generator, assessment, and personalized training system: Architecture and research agenda. In *2025 22nd International Conference on Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON)*, pages 1–6. IEEE.
- Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr, and Jindong Gu. 2024. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets adversarial image. *arXiv preprint arXiv:2402.14899*.
- Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024a. Characterizing llm abstention behavior in science qa with context perturbations. *arXiv preprint arXiv:2404.12452*.
- Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and Jiliang Tang. 2024b. Ai for education (ai4edu): Advancing personalized education with llm and adaptive learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 6743–6744.
- Ron Woldoff. 2024. *GRE 5-Hour Quick Prep For Dummies*. John Wiley & Sons.
- Ron Woldoff and Joseph Kraynak. 2015. *GRE For Dummies: With Online Practice Tests*. John Wiley & Sons.
- Eamon Worden, Cristina Heffernan, Neil Heffernan, and Shashank Sonkar. 2026. Foundationalassist: An educational dataset for foundational knowledge tracing and pedagogical grounding of llms. *arXiv preprint arXiv:2602.00070*.
- Wei Xia, Shaoguang Mao, and Chanjing Zheng. 2024. Empirical study of large language models as automated essay scoring tools in english composition_taking toefl independent writing task for example. *arXiv preprint arXiv:2401.03401*.
- Wenbo Xu, Muhammad Shahreeza, Wai Lam Hoo, and Wudao Yang. 2025. [Explainable ai for education: Enhancing essay scoring via rubric-aligned chain-of-thought prompting](#). *Preprints*.
- Shuai Yang, Qi Yang, Luoxi Tang, Yuqiao Meng, Nancy Guo, Jeremy Blackburn, and Zhaohan Xi. 2025. On the eligibility of llms for counterfactual reasoning: A decompositional study. *arXiv preprint arXiv:2505.11839*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Dharunish Yugeswardeenoo, Kevin Zhu, and Sean O’Brien. 2024. Question-analysis prompting improves llm performance in reasoning tasks. *arXiv preprint arXiv:2407.03624*.

Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jia-tong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang, Marco Pavone, Yuqiang Li, and 1 others. 2024a. Llama-berry: Pairwise optimization for o1-like olympiad-level mathematical reasoning. *arXiv preprint arXiv:2410.02884*.

Xuan Zhang, Chao Du, Tianyu Pang, Qian Liu, Wei Gao, and Min Lin. 2024b. Chain of preference optimization: Improving chain-of-thought reasoning in llms. *Advances in Neural Information Processing Systems*, 37:333–356.

Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, Tianxing He, and Yulia Tsvetkov. 2024c. Can llm graph reasoning generalize beyond pattern memorization? *arXiv preprint arXiv:2406.15992*.

Zheyuan Zhang, Daniel Zhang-Li, Jifan Yu, Linlu Gong, Jinchang Zhou, Zhanxin Hao, Jianxiao Jiang, Jie Cao, Huiqin Liu, Zhiyuan Liu, and 1 others. 2024d. Simulating classroom education with llm-empowered agents. *arXiv preprint arXiv:2406.19226*.

Yang Zhong, Jiangang Hao, Michael Fauss, Chen Li, and Yuan Wang. 2024. Evaluating ai-generated essays with gre analytical writing assessment. *arXiv preprint arXiv:2410.17439*.

A Comprehensive Details of ESTBOOK Data Collection

This section provides extended details on the scope, modalities, sourcing, and quality control processes used to construct the ESTBOOK dataset, expanding upon the summary provided in Section 3.1.

A.1 Scope and Cognitive Targets of Selected ESTs

ESTBOOK aggregates questions from five major standardized tests, chosen for their distinct cognitive requirements and global educational impact:

(i) **SAT**: Widely used for US undergraduate admissions, the SAT provides a baseline for evaluating foundational verbal and mathematical reasoning among high school graduates.

(ii) **GRE**: Designed for graduate admissions, the GRE tests advanced cognitive constructs, including

highly complex semantic reasoning (e.g., Sentence Equivalence) and fluid quantitative logic.

(iii) **GMAT**: Functioning as the primary gatekeeper for business schools, the GMAT emphasizes executive-level logic, including Integrated Reasoning (synthesizing multimodal data) and Data Sufficiency (evaluating logical constraints without executing the math).

(iv) **TOEFL & IELTS**: These exams assess non-native English proficiency. In ESTBOOK, they serve as the primary benchmark for evaluating an LLM’s capacity for long-context auditory comprehension (audio modalities) and pragmatic textual inference.

A.2 Multimodal Distribution and Complexity

Unlike traditional NLP benchmarks that rely solely on text, ESTBOOK enforces a heterogeneous multimodal environment. The modalities incorporated include:

Text (T) & Audio (A): Primarily featured in TOEFL and IELTS, evaluating the model’s ability to extract key discourse from dense textual passages or spoken academic lectures.

Math Symbols (S): Featured heavily in the SAT, GRE, and GMAT, requiring models to parse LaTeX-formatted equations and perform symbolic manipulation.

Images (I) & Tables (Tb): Critical for the GRE Data Interpretation and GMAT Integrated Reasoning sections. These questions require the model to perform *visual parsing* (e.g., reading a scatterplot or cross-referencing multi-column tables) before executing any mathematical reasoning.

A.3 Data Source

To ensure the dataset accurately reflects the difficulty and style of modern ESTs, we sourced raw materials from a blend of highly vetted public and official resources:

(i) **Official Preparation Materials**: We extracted benchmark examples from official guides and released practice exams ([Graduate Management Admission Council \(GMAC\), 2025](#); [Woldoff, 2024](#); [College Board, 2022](#); [Hatch et al., 2023](#)).

(ii) **Public Educational Platforms & Practice Books**: Additional diversity in question format and difficulty was achieved by sourcing from recognized test-prep literature ([Appelrouth and Zabrocky, 2017](#); [Gruber, 2011](#); [Woldoff and Kraynak, 2015](#)) and open-access educational platforms

Table 4: Question types, their descriptions, number of instances, involved modalities, and their pedagogical categories (defined in §3.2). Categories: **I** – Knowledge-Intensive Retrieval; **II** – Reasoning-Intensive Execution; **III** – Hybrid Integration. Modalities: "T"–text, "S"–math symbol, "I"–image, "Tb"–tabular data, "A"–audio. Concrete examples are shown in Appendix C.

Section	Question Type (Abbreviation)	Description	Num	Modality	Category
SAT					
Reading & Writing	Information and Ideas (II)	Assess comprehension, reasoning, and inference skills	180	T	III
	Craft and Structure (CS)	Test vocabulary and how authors structure their writing	636	T	I
	Expression of Ideas (EI)	Test the logical flow and effectiveness of writing	210	T	I
	English Conventions (EC)	Focus on grammar, punctuation, and sentence structure	150	T	I
Math	Algebra (AG)	Test numeric equations, functions, and inequalities	243	T,S	II
	Data Analysis (DA)	Interpret ratios, percentages, probabilities, and graphs	141	T,S	II
	Geometry & Trigonometry (GT)	Analyze angles, circles, areas, and trigonometric functions	153	T,S	II
GRE					
Verbal	Text Completion (TC)	Fill in blank(s) (one/two/three) within a short passage	620	T	I
	Sentence Equivalence (SE)	Choose two words with the same meaning	620	T	I
	Reading Comprehension (RC)	Answer questions based on a passage	562	T	III
Quantitative	Quant Comparison (QC)	Compare two quantities and select their relationship	150	T,S	II
	Numeric Entry (NE)	Type the exact numerical answer	150	T,S	II
	Data Interpretation (DI)	Multi-choice questions from graphs, tables, or charts	150	T,S,I,Tb	II
GMAT					
Verbal	Critical Reasoning (CR)	Analyze and evaluate an argument	244	T	III
	Reading Comprehension (RC)	Answer questions based on a passage	408	T	III
Quantitative	Problem Solving (PS)	Algebra, arithmetic, numerical, and statistical problems	408	T,S	II
Data Insights	Data Sufficiency (DS)	Decide if a statement is sufficient to answer a question	400	T,S	II
	Integrated Reasoning (IR)	Analyze tables, graphs, charts, or multiple sources	340	T,S,I,Tb	II
TOEFL					
Reading	Factual Information (FI)	Identify facts in (or not in) the passage	620	T	III
	Inference & Reference (IR)	Infer information/word meaning/pronoun in context	415	T	III
	Text & Sentence (TS)	Insert texts, simplify a sentence, summarize a passage	310	T,Tb	I, III
Listening	Factual Information (FI)	Identify facts in (or not in) the lecture/conversation	300	A,T	III
	Inference (IF)	Understand tone/intention/opinion/relationship of ideas	150	A,T	III
IELTS					
Reading	Identifying Information (II)	Identify correctness of statement or author’s opinion	296	T	III
	Matching Sentence (MS)	Match head, opinion, or sentence endings	208	T	III
	Completion (CP)	Complete sentence/summary/note/table/diagram label	592	T,I,Tb	I, III
Listening	Identification & Matching (IM)	Determine correct answers from the audio	520	A,T	III
	Completion & Labeling (CL)	Complete a sentence or visual with words from the audio	1048	A,T,I,Tb	I, III
	Short Answer (SA)	Answer briefly using words from the recording	352	T,I,Tb	III

(Josué et al., 2023; Pereira et al., 2024; Mallik; IELTS-up).

A.4 Question Quality, Copyright and Ethical Assurance

To ensure the integrity of our benchmark, we adopted a rigorous process for validating both the quality of collected questions and the copyright compliance of the sources.

Question Quality. All questions included in ESTBOOK were sourced from publicly released or openly accessible educational materials. We verified each item for (1) correctness, by cross-checking answer keys or explanatory notes provided by the original source; (2) clarity, ensuring that wording, figures, and formatting matched the original intent without ambiguity; and (3) authenticity, by aligning the question style and content with the design principles of the corresponding

standardized test (SAT, GRE, GMAT, TOEFL, or IELTS).

Alignment Validation Process. In addition to verifying correctness and clarity, we systematically validated the alignment of our questions with official exam specifications. For each question type, we performed the following validation steps:

1. **Format validation:** We cross-referenced the question structure, answer format, and presentation style against official sample materials to ensure consistency. As an example, SAT Math Algebra questions in our benchmark follow the same five-option multiple-choice format and use identical equation presentation conventions as those in College Board released items.
2. **Skill mapping:** We verified that each question assesses the specific cognitive skill as

defined in official test frameworks. For instance, TOEFL Reading “Factual Information” questions in ESTBOOK are designed to test the ability to identify explicitly stated details, which matches the skill definition provided by ETS. Similarly, “Inference” questions require test-takers to draw logical conclusions that go beyond the literal text, consistent with official specifications.

3. **Difficulty calibration:** Although we do not have access to proprietary difficulty ratings used by test administrators, we ensured that our question collection spans the full difficulty range found in official preparation materials. This includes both entry-level items suitable for beginners and advanced problems that challenge high-performing test-takers, providing representative coverage across each exam’s difficulty spectrum.

Through this systematic validation process, combined with our reliance on officially affiliated sources, we ensure that ESTBOOK authentically captures the reasoning demands and structural characteristics of real standardized tests.

Copyright and Ethical Usage. To comply with licensing and intellectual property requirements, we restricted data collection to (1) officially released practice tests and preparation guides distributed for public use, and (2) open-access educational platforms and community-contributed question banks that explicitly allow free access for study and research purposes. No proprietary or pay-walled materials were included. For multimodal reconstructions, all recreated content is original and designed solely to approximate the test-taking context without replicating copyrighted assets. This approach ensures that ESTBOOK respects copyright protections while still providing representative and high-quality benchmark content.

B Rigorous Annotation Pipeline for Cognitive Trajectories and Distractor Rationales

To ensure the high fidelity and reproducibility of the ESTBOOK dataset, we implemented a deterministic, NLP-driven annotation pipeline. Our priority is to avoid relying on generative LLMs for the creation of intermediate reasoning steps or distractor rationales, as doing so risks introducing LLM-inherent biases, hallucinations, or data contamination

into the evaluation benchmark. Instead, our methodology strictly relies on parsing established ground-truth text using classical NLP techniques.

The annotation process synthesizes official test-preparation explanations (e.g., from ETS or GMAC, which provide explicit rationales for both correct and incorrect options) with deterministic NLP frameworks. The pipeline consists of three primary phases:

Phase 1: Stepwise Extraction via Syntactic Parsing

Standardized test explanations typically present reasoning in a sequential, textual format. To map these raw textual explanations into our formalized cognitive trajectory ($C = \{c_1, c_2, \dots, c_n\}$), we employ a deterministic information extraction (IE) pipeline built on dependency parsing and semantic role labeling (SRL).

1. Sentence Segmentation and Dependency

Parsing: We process the official explanation for the correct answer using a standard dependency parser (e.g., spaCy). By analyzing syntactic trees, we isolate sequential imperative or action-oriented clauses (e.g., “*First, calculate the volume...*” $\rightarrow$ [VERB: calculate, DOBJ: volume]).

2. Rule-Based Node Mapping:

We utilize a pre-defined lexicon of cognitive triggers to map the parsed clauses to specific cognitive steps. For example: Lexical triggers such as “*according to the passage,*” or “*line 15 states*” are deterministically mapped to $c_{\text{perceive_text}}$ (Evidence Extraction). Mathematical operators or numbers extracted via regular expressions (Regex) alongside terms like “*let x be*” or “*formula*” are mapped to $c_{\text{formulate_eq}}$ (Symbolic Mapping).

3. Sequential Alignment:

The parsed nodes are chronologically ordered to establish the ground-truth traversal path for the question, ensuring each step logically precedes the next without LLM intervention.

Phase 2: Distractor Rationale Annotation and Taxonomy Mapping

Evaluating a model’s pedagogical utility requires understanding *why* it selects an incorrect option. Official test materials often include explanations for distractors (e.g., “*Option B is incorrect because*

it misinterprets the data in Figure 1"). To formalize this, we developed a rigid taxonomy of cognitive traps and utilized traditional NLP to annotate them.

We define four primary distractor categories for ESTs: **Type I: Direct Contradiction** (The option states the exact opposite of the text/data). **Type II: Out of Scope** (The option relies on information not provided in the prompt). **Type III: Partial Truth / Faulty Premise** (The option accurately quotes the text but misapplies it to the specific question intent). **Type IV: Execution / Calculation Error** (The option represents a common arithmetic mistake or formula misapplication).

To automate the annotation of these distractors across 10,576 questions, we apply the following non-generative NLP techniques:

1. **Lexical Overlap and TF-IDF:** We calculate the lexical overlap between the distractor text, the raw explanation, and the source passage. Distractors exhibiting high lexical overlap with the passage but flagged as incorrect in the raw explanation are commonly classified as *Type III: Partial Truth*, representing a deliberate cognitive trap designed to penalize superficial pattern-matching.
2. **Negation Scope Detection:** Using classical syntactic parsing, we detect negation modifiers (e.g., *not*, *never*, *lacks*) within the raw explanation of the wrong answer. If the explanation explicitly points out a negation mismatch between the distractor and the text, it is tagged as *Type I: Direct Contradiction*.
3. **Entity and Variable Mismatch:** For quantitative reasoning, we use Named Entity Recognition (NER) and Regex to extract the mathematical entities in the distractor. If the distractor matches the output of a partially completed equation (extracted from the cognitive step $c_{\text{formulate_eq}}$ but stopping before c_{compute}), it is deterministically tagged as *Type IV: Calculation Error*.

Phase 3: Human-in-the-Loop Validation

To ensure the absolute integrity of the benchmark, a random subset of the NLP-annotated trajectories and distractor rationales was subjected to expert review. Domain experts (educators familiar with EST design) evaluated the alignment between the deterministic NLP tags and the pedagogical intent

of the question, achieving a high inter-rater reliability (Cohen's $\kappa > 0.85$). We thus confirm that our non-AI methodology successfully and rigorously extracts the cognitive framework required for pedagogical evaluation.

C Complementary Information of Question Types

This appendix provides a brief description of each question type covered across SAT, GRE, GMAT, TOEFL, and IELTS, along with a concrete example. In the examples below, long passages are truncated with "..." for brevity.

SAT Reading & Information and Ideas (II).

Tests comprehension of written passages across diverse subjects. Students must identify main ideas, understand explicit details, make logical inferences, and draw evidence-based conclusions. Requires distinguishing between stated facts and implied meanings while recognizing the author's purpose and how ideas relate to one another. Success depends on analytical reading rather than simply locating isolated facts.

Passage: The industrial revolution transformed societies by shifting labor from farms to factories, fundamentally altering social structures and economic relationships. As rural populations migrated to urban centers seeking employment, traditional family units were disrupted...

Question: According to the passage, what was one major social change brought about by the industrial revolution?

- A. The development of new agricultural techniques
- B. The disruption of traditional family structures
- C. The elimination of class divisions in society
- D. The reduction in opportunities for social mobility

SAT Writing Craft and Structure (CS). Evaluates understanding of vocabulary in context and text organization. Students must select appropriate words based on context, analyze how ideas develop across paragraphs, and recognize how structure enhances effectiveness. Requires understanding both denotative and connotative meanings while evaluating how word choice affects tone, style, and precision of expression.

Sentence: The scientist's findings were _____, shedding light on the mysterious behaviors of subatomic particles that had puzzled physicists for decades. . .

- A. groundbreaking
- B. world-class
- C. groundbreaking and thrilling
- D. groundbreaking, thrilling

SAT Expression of Ideas (EI). Focuses on writing effectiveness and logical flow. Students evaluate and improve coherence, cohesion, and clarity by combining sentences, reorganizing information, or modifying details. Requires determining relevance, identifying optimal placement for new information, and understanding how structural changes affect meaning and emphasis. Tests ability to develop logically connected ideas with appropriate support.

Revision Task: Improve the coherence of the following sentence pair to create a more logical flow:

"I love classical music. Beethoven's symphonies are my favorite."

Possible revisions:

- A. I love classical music, especially Beethoven's symphonies, which are my favorite.
- B. Because I love classical music, Beethoven's symphonies are my favorite.
- C. I love classical music; indeed, Beethoven's symphonies are my favorite.
- D. No change necessary.

SAT English Conventions (EC). Assesses command of standard English grammar, punctuation, and sentence structure. Topics include verb tense/agreement, pronoun usage, parallel structure, modifier placement, and appropriate punctuation. Students must identify and correct errors in sentences or paragraphs. Evaluates practical application of grammatical rules rather than theoretical knowledge.

Sentence: Each of the students (have / has) submitted their essay on time, and the teacher (is / are) pleased with the quality of work...

- A. have, is
- B. has, is
- C. have, are
- D. has, are

SAT Math Algebra (AG). Tests ability to work with algebraic expressions, equations, inequalities,

and functions. Requires solving linear/quadratic equations, manipulating expressions, understanding variable relationships, and analyzing functions. Students must apply algebraic concepts to model real-world situations and connect symbolic representations with graphs. Assesses both procedural fluency and conceptual understanding.

Solve for x : $\frac{2x-3}{x+1} = 4$

- A. $x = 7$
- B. $x = -7$
- C. $x = 7/3$
- D. $x = -7/3$
- E. No solution exists

SAT Data Analysis (DA). Evaluates interpretation of various data forms. Students analyze ratios, rates, percentages, proportions, and probabilities while interpreting information from tables, charts, and graphs. Requires understanding statistical concepts (mean, median, mode) and using data to draw conclusions. Tests quantitative literacy skills needed for interpreting real-world numerical information.

A circle graph shows that 30% of students prefer tea, 50% coffee, and 20% water. If 200 students were surveyed, how many prefer coffee?

- A. 60
- B. 100
- C. 40
- D. 50
- E. Cannot be determined from the information given

SAT Geometry & Trigonometry (GT). Covers geometric figures, coordinate geometry, and trigonometric relationships. Questions involve angles, lines, polygons, circles, 3D figures, coordinate systems, and trigonometric functions. Students calculate areas, perimeters, volumes, and distances while applying properties of various shapes. Tests spatial reasoning and connection of algebraic and geometric representations.

In right triangle ABC with right angle at C , if $AC = 3$ and $BC = 4$, what is $\sin A$?

- A. $3/5$
- B. $4/5$
- C. $3/4$
- D. $4/3$
- E. $5/3$

GRE Text Completion (TC). Assesses vocabu-

lary and comprehension by requiring completion of blanks in short passages. Students must understand author's intent, logical relationships between sentences, and overall context. Difficulty increases with multiple blanks where choices must work cohesively. Tests vocabulary breadth and understanding of how words function within complex contexts.

Though praised for its _____ innovations, the clock's design was too _____ to gain widespread adoption among consumers who valued simplicity and ease of use...

- A. aesthetic ... cumbersome
- B. mechanical ... simplified
- C. technological ... intuitive
- D. functional ... intricate
- E. rudimentary ... complex

GRE Sentence Equivalence (SE). Requires selecting two words that create sentences with equivalent meanings when inserted. Students must identify words that produce the same overall meaning in context, understanding subtle connotative differences. Tests vocabulary depth, contextual word usage, and ability to maintain consistent meaning across different word choices.

Her lecture was so _____ that many students struggled to stay awake.

- A. engaging
- B. soporific
- C. bewildering
- D. tedious
- E. stimulating
- F. monotonous

GRE Reading Comprehension (RC). Tests analysis and interpretation of complex academic passages. Students identify main ideas, recognize explicit statements, make inferences, understand author's purpose, and evaluate arguments. Requires handling sophisticated vocabulary and complex sentence structures while synthesizing information across passages and drawing conclusions from implied content.

Passage: Advances in CRISPR technology have opened new avenues in gene therapy, offering unprecedented precision in modifying DNA sequences. Unlike earlier gene-editing methods that often resulted in unintended modifications, CRISPR-Cas9 allows scientists to target specific sections of genetic code with remarkable accuracy...

Question: The passage suggests that CRISPR's main advantage over previous gene-editing methods is its ability to:

- A. Work faster than other methods
- B. Target specific sections of genetic code with high accuracy
- C. Completely eliminate the risk of unintended modifications
- D. Address a wider range of medical conditions
- E. Bypass ethical concerns associated with genetic manipulation

GRE Quantitative Comparison (QC). Presents two quantities for comparison of relative size. Tests conceptual understanding over computational ability as students analyze information, identify mathematical relationships, and determine if enough information exists to establish definitive relationships. Requires creative approaches, estimation skills, and recognition of information adequacy without necessarily performing complex calculations.

Quantity A: 2^{10} Quantity B: 10^3

- A. Quantity A is greater
- B. Quantity B is greater
- C. The two quantities are equal
- D. The relationship cannot be determined from the information given

GRE Numeric Entry (NE). Requires calculating exact answers without multiple-choice options. Tests ability to perform calculations accurately and follow procedures correctly without answer verification. Assesses computational skills, problem-solving strategies, and work with various numerical forms. Demands confidence in mathematical procedures and attention to units and precision.

If a tank is filled at a constant rate and holds 65 gallons in 10 minutes, how many gallons per minute are being added to the tank? If the answer is a fraction, enter as a decimal.

Answer box: _____

GRE Data Interpretation (DI). Assesses ability to analyze and interpret data in graphs, tables, or charts. Students extract information, perform calculations, recognize patterns, and draw conclusions. Requires comparing data points, calculating percentages or rates of change, and making predictions. Tests quantitative literacy and ability to work with real-world data representations.

Table: Quarterly profits (in \$M) for Company X:

Q1: 10

Q2: 15

Q3: 12

Q4: 18

Question: In which quarter did Company X see the greatest increase in profit over the previous quarter? A. Q1

B. Q2

C. Q3

D. Q4

E. Cannot be determined from the information given

GMAT Critical Reasoning (CR). Evaluates analysis of argument structure, validity, and logical coherence. Students identify premises, conclusions, and assumptions while distinguishing relevant information and recognizing logical flaws. Often uses business scenarios requiring understanding of causation vs. correlation and sample representativeness. Tests analytical thinking crucial for business decision-making.

Argument: Because sales rose by 15% last quarter immediately following the implementation of our new marketing strategy, the new marketing strategy must be effective and should be continued without modifications in the upcoming fiscal year.

Which of the following, if true, most weakens this conclusion?

A. The company's main competitor went out of business during the same quarter.

B. The company introduced a popular new product line at the beginning of the quarter.

C. Other companies using similar marketing strategies saw comparable increases in sales.

D. The marketing strategy cost more to implement than initially projected.

E. Industry sales overall rose by 20% during the same period due to seasonal factors.

GMAT Reading Comprehension (RC). Similar to GRE but with more business focus. Tests understanding of complex written material, identification of main ideas, inference-making, and logical structure recognition. Passages often discuss business strategies or economic concepts. Assesses ability to distinguish stated from implied information and evaluate argument strength.

Passage: Global coffee consumption has doubled in the past decade, driven primarily by emerging markets in Asia where a growing middle class has embraced Western consumption patterns. China, traditionally a tea-drinking nation, has seen coffee consumption grow at 15% annually, compared to global growth of 2.5%...

Question: The author primarily discusses which factor driving coffee demand?

A. Changes in consumer taste preferences

B. Economic development and social status in emerging markets

C. Declining popularity of traditional tea consumption

D. Marketing strategies of Western coffee companies

E. Health benefits associated with coffee consumption

GMAT Problem Solving (PS). Tests mathematical knowledge across arithmetic, algebra, geometry, and statistics. Students determine problem requirements, identify relevant information, select appropriate techniques, and calculate accurately. Requires translating word problems into mathematical expressions and interpreting solutions in context, often in business-related scenarios.

If $x + y = 10$ and $xy = 21$, what is $x^2 + y^2$?

A. 52

B. 58

C. 100

D. 121

E. 142

GMAT Data Sufficiency (DS). A unique format assessing analytical thinking over computation. Students determine whether statements provide sufficient information to answer questions without actually solving problems. Requires evaluating statement sufficiency individually and collectively while understanding necessary vs. sufficient conditions and recognizing implied information.

Question: Is $x > 5$? (1) $2x > 10$ (2) $x^2 > 25$
 A. Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient.
 B. Statement (2) ALONE is sufficient, but statement (1) ALONE is not sufficient.
 C. BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.
 D. EACH statement ALONE is sufficient.
 E. Statements (1) and (2) TOGETHER are NOT sufficient.

GMAT Integrated Reasoning (IR). Tests analysis of information from multiple sources and formats. Students interpret tables, graphs, and text while evaluating multiple information sources and solving multi-step problems. Includes multi-source reasoning, graphics interpretation, two-part analysis, and table analysis. Assesses skills for data-driven business decisions.

Table: Region A sales (in millions): Year 1: \$120, Year 2: \$132, Year 3: \$145 Region B sales (in millions): Year 1: \$90, Year 2: \$101, Year 3: \$114
 Question: Which region's compound annual growth rate exceeded 5% over the three-year period?
 A. Region A only
 B. Region B only
 C. Both Region A and Region B
 D. Neither Region A nor Region B

TOEFL Reading Factual Information (FI). Evaluates ability to identify explicitly stated facts in academic texts. Students locate specific information, distinguish it from similar content, and understand its contextual significance. Requires processing academic vocabulary and syntax while focusing on directly stated rather than implied information.

Passage: Canada's boreal forest covers nearly one-third of its land area, spanning from Yukon to Newfoundland and Labrador. This vast ecosystem, dominated by coniferous trees, contains more than 1.5 million lakes and is home to endangered species such as the woodland caribou...
 Question: According to the passage, what fraction of Canada is covered by the boreal forest?
 A. One-quarter
 B. One-third
 C. One-half
 D. Two-thirds

TOEFL Inference & Reference (IR). Tests understanding of implied information and referential relationships. Students draw logical conclusions from provided information, understand unstated relationships between ideas, and track references through pronouns and demonstratives. Assesses deeper comprehension including reading between lines and connecting ideas across text sections.

Passage: "The experiment failed again, prompting the research team to reconsider their methodology. Dr. Chen suggested they should explore alternative approaches that had shown promise in similar contexts."
 Question: What does "again" imply about previous attempts?
 A. This was the first time the experiment had been conducted.
 B. Previous attempts had been successful.
 C. Previous attempts had also failed.
 D. The team had never tried this experiment before.

TOEFL Text & Sentence (TS). Evaluates various aspects of textual understanding including summarizing and sentence relationships. Tasks include inserting sentences appropriately, creating cohesive summaries, simplifying complex sentences, or identifying sentence functions. Tests advanced language processing including understanding textual connections, organizational structure, and purpose of different elements.

Original: Because of the severe weather conditions, we decided to cancel the outdoor concert scheduled for tomorrow evening.
 Task: Combine into one sentence without changing meaning, beginning with "The outdoor concert..."
 A. The outdoor concert scheduled for tomorrow evening we decided to cancel because of the severe weather conditions.
 B. The outdoor concert scheduled for tomorrow evening was decided to be cancelled by us because of the severe weather conditions.
 C. The outdoor concert scheduled for tomorrow evening has been cancelled due to the severe weather conditions.
 D. The outdoor concert, because of the severe weather conditions, scheduled for tomorrow evening we decided to cancel.

TOEFL Listening Factual Information (FI). Assesses comprehension of spoken academic content.

Students identify explicitly stated information, distinguish between similar details, and recognize contextual significance. Requires processing natural-speed academic English despite accent variations while maintaining focus during extended listening passages.

Transcript: "Good morning, class. Today's lecture will cover photosynthesis, the process by which plants convert light energy into chemical energy. We'll first discuss the light-dependent reactions that occur in the thylakoid membrane, followed by the Calvin cycle that takes place in the stroma..."

Question: What topic will the lecture cover?

- A. Cell respiration
- B. Plant reproduction
- C. Photosynthesis
- D. Genetic engineering

TOEFL Listening Inference (IF). Tests understanding of implied meanings in spoken content. Students interpret speaker's tone, infer unstated opinions, understand implied connections, and determine purpose of specific statements. Requires comprehending not just words but also intonation and emphasis. Assesses ability to understand nuanced academic communication.

Speaker: "I suppose we could try that method, if all our other options have been exhausted. It's not my first choice, but at this point, we might not have many alternatives left."

Question: What does the speaker's tone suggest about their enthusiasm for the proposed method?

- A. They are excited to try something new
- B. They are reluctant but resigned to trying it
- C. They believe it is the best available option
- D. They are confident it will succeed

IELTS Reading Identifying Information (II). Evaluates whether statements match textual information. Students determine if statements are True (matching), False (contradicting), or Not Given (not addressed). Requires careful reading to distinguish between explicit, inferable, and absent information without introducing outside knowledge.

Passage: "Many cities have embraced rooftop gardens as a sustainable solution to multiple urban challenges. These green spaces not only provide fresh produce for local communities but also help mitigate the urban heat island effect by absorbing sunlight that would otherwise be converted to heat..."

Statement: "The author believes urban gardens are ineffective at addressing environmental challenges."

Is the statement True, False, or Not Given?

IELTS Matching Sentence (MS). Tests ability to connect related information pieces. Students match headings with paragraphs, sentence beginnings with endings, or statements with speakers. Requires understanding paragraph main ideas, sentence logic, and information relationships while processing content across multiple text sections.

Complete the following sentence with the most appropriate ending from the list below:

"Fossil fuels are being replaced by renewable sources..."

- A. ...because they are more sustainable and environmentally friendly.
- B. ...despite their continued dominance in global energy markets.
- C. ...although the transition is happening more slowly than many scientists recommend.
- D. ...particularly in developing economies seeking to reduce energy costs.
- E. ...which has caused significant economic disruption in traditional energy sectors.

IELTS Completion (CP). Assesses ability to locate and transfer specific information to complete sentences, summaries, or diagrams. Students identify relevant details and transfer them accurately, often verbatim. Requires understanding text structure for efficient information location while recognizing synonyms and paraphrased content.

Summary: "The Sahara is the world's _____ desert, covering approximately _____ million square kilometers across North Africa, from the Atlantic Ocean to the Red Sea. Its name comes from the Arabic word meaning _____." Words to choose from:

- A. largest, 9.2, "desert"
- B. hottest, 8.7, "sand"
- C. oldest, 7.5, "wilderness"
- D. driest, 6.3, "emptiness"

IELTS Listening Identification & Matching (IM). Tests identification of specific spoken information and category matching. Students listen for details like names, numbers, and facts then select correct options. Requires processing natural-speed English despite distractions or accent variations while distinguishing between similar-sounding choices.

Audio transcript: "Welcome to our university orientation. The main campus tour will begin at the Student Center at quarter past nine. Please arrive at least ten minutes early to collect your information packets..."

Question: What time does the campus tour start?

- A. 9:00
- B. 9:15
- C. 9:30
- D. 10:15

IELTS Completion & Labeling (CL). Evaluates ability to listen for specific information to complete sentences, notes, or diagrams. Students identify and record specific details, often verbatim. Requires focused listening, accurate information processing, and simultaneous writing. Tests note-taking skills needed for educational and professional contexts.

[Audio describes the parts of a flower and their functions]

Diagram: Label the parts of a flower shown in the image using words from the recording:

- 1. ____ (outer protective layer)
- 2. ____ (colorful structures that attract pollinators)
- 3. ____ (male reproductive part containing pollen)
- 4. ____ (female reproductive structure)
- 5. ____ (produces seeds when fertilized)

IELTS Short Answer (SA). Tests listening for specific information and providing concise answers using the recording's words. Students identify relevant details and express them within word limits. Requires understanding question focus, quick information processing, and appropriate word selection. Assesses both receptive and productive language skills.

Audio transcript: "For our upcoming science class field trip next Thursday, we'll be visiting the botanical gardens on the north side of the city. Please remember to bring your permission slips, a notebook, appropriate footwear, and a packed lunch..."

Questions:

- 1. Where is the field trip? (Answer in no more than THREE words)
- 2. What day will the field trip take place? (Answer in no more than TWO words)
- 3. What time will students return to school? (Answer in no more than TWO words)

D Pedagogical Alignment of the Cognitive Trajectory with Human Test-Taking Strategies

To ensure that ESTBOOK serves as a valid diagnostic tool, we must justify that our formalized cognitive trajectory mirrors the actual cognitive strategies adopted by high-performing human test-takers. Our design is grounded in well-documented findings from standardized test preparation curricula, Cognitive Diagnosis Models (CDMs), and empirical think-aloud studies of student behaviors during exam practice (Loken et al., 2004; Board, 2025; Johnstone et al., 2006). Below, we elaborate on how human heuristics map to our three pedagogical categories (defined in §3.2).

Category I: Knowledge-Intensive Retrieval (Lexical & Structural)

Cognitive Trajectories: $c_{\text{syntax_parse}} \rightarrow c_{\text{rule_match}} \rightarrow c_{\text{predict}} \mid c_{\text{semantic_parse}} \rightarrow c_{\text{logical_scope}} \rightarrow c_{\text{resolve}}$ In grammar and vocabulary-oriented tasks (e.g., SAT Writing, GRE Text Completion), human test-takers do not evaluate sentences holistically; instead, they engage in mental ‘chunking.’ They first parse the syntactic or semantic structure ($c_{\text{syntax_parse}}$), identifying structural anchors like subject-verb pairs or contrast markers (e.g., ‘although’). Next, they localize the logical scope by matching the text against internalized grammatical rules or expected vocabulary contexts ($c_{\text{rule_match}}$). Finally, they predict the missing element or resolve the meaning (c_{predict}) before looking at the multiple-choice options. Our cognitive trajectory explicitly encodes this systematic progression, penalizing models that attempt to guess the final word without first demonstrating structural awareness.

Category II: Reasoning-Intensive Execution (Multimodal & Quantitative)

Cognitive Trajectories: $c_{\text{analytical_goal}} \rightarrow c_{\text{visual_parse}} \rightarrow c_{\text{analyze}} \mid c_{\text{model}} \rightarrow c_{\text{formulate_eq}} \rightarrow c_{\text{compute}}$ Quantitative and multimodal questions (e.g., SAT Math, GRE Data Interpretation) require strict procedural execution. Empirical studies show that human test-takers who rush directly into computation are highly prone to errors. Effective problem-solvers first establish an analytical goal ($c_{\text{analytical_goal}}$) or translate a natural language word problem into a conceptual model (c_{model}). For multimodal tasks, test-prep curricula heavily emphasize reading the axes, legends, and units of a chart ($c_{\text{visual_parse}}$) prior to analyzing the data. For mathematical tasks, humans translate their conceptual model into a rigorous algebraic representation ($c_{\text{formulate_eq}}$) before engaging in symbolic manipulation or arithmetic execution (c_{compute}). Our cognitive framework enforces this exact procedural separation, allowing us to diagnose whether an LLM fails due to a misunderstanding of the mathematical text (c_{model}) or a pure arithmetic hallucination (c_{compute}).

Category III: Hybrid Integration (Semantic Extraction & Inference)

Cognitive Trajectories: $c_{\text{intent}} \rightarrow c_{\text{perceive_text}} \rightarrow c_{\text{extract}} \mid c_{\text{entities}} \rightarrow c_{\text{constraints}} \rightarrow c_{\text{evaluate}}$ For tasks requiring dense reading comprehension or logical deduction (e.g., TOEFL Reading, GMAT Critical Reasoning), human test-takers employ targeted scanning and hypothesis testing. They begin by identifying the core subject or intent of the question stem (c_{intent}) to avoid wasting working memory on irrelevant text. They then scan the passage or audio mental representation for relevant discourse ($c_{\text{perceive_text}}$). In comparative logic tasks, human strategies explicitly emphasize identifying the entities to be compared (c_{entities}) and strictly mapping given constraints ($c_{\text{constraints}}$) before evaluating sufficiency (c_{evaluate}). This mirrors heuristic strategies taught in test prep, where students are trained to systematically test conditions rather than rely on intuition. By isolating these nodes, our cognitive framework evaluates whether an LLM can mimic this disciplined information retrieval and constraint satisfaction.

Summary of Pedagogical Fidelity

Across all categories, the cognitive trajectory decomposes EST problem-solving into sequential operations that are (1) cognitively plausible, (2) perfectly aligned with established test-preparation strategies, and (3) empirically observable in human behavior. Consequently, LLMs evaluated using our Stepwise Execution Score (SES) are not merely graded on answer accuracy, but on their ability to exhibit an explainable, human-aligned reasoning process suitable for real-world educational deployment.

E Prompt Structure Used in Evaluation

This appendix summarizes the skeleton templates for three prompt-engineering paradigms used in our educational reasoning experiments. Each template is presented with an explanation followed by a concrete example in a highlighted box.

1. In-Context Learning (ICL) Prompt Structure

Provides the model with solved examples to prime analogous problem solving:

- Multiple exemplars demonstrating the problem–solution pattern
- Graduated difficulty progression across examples
- Explicit identification of transferable patterns in each exemplar
- Strategic selection of examples to highlight different facets of the problem type
- Clear demarcation between example set and the target query

Example: For a simple SAT algebra problem—"Solve for x : $(x - 2)(x + 3) = 0$ ":
 Input: "Solve $(x - 2)(x + 3) = 0$ "
 Output: "Step 1: Recognize factors imply solutions $x=2$ or $x=-3$. Step 2: List solutions: 2, -3."

2. Chain-of-Thought (CoT) Prompt Structure

Guides the model through a step-by-step reasoning process:

- Instruction to decompose the task into ordered steps
- Explicit requests for intermediate calculations or justifications

- Structured step-labeling conventions (e.g., “Step 1: ...”, “Step 2: ...”)
- Prompts for linking each step’s result to the next
- Final summary or explicit statement of the conclusion

Example: Compute the area of a triangle with base 5 and height 8:

1. Step 1: Write formula $A = \frac{1}{2} \times \text{base} \times \text{height}$.
2. Step 2: Substitute values: $A = \frac{1}{2} \times 5 \times 8$.
3. Step 3: Calculate: $A = 20$.
4. Conclusion: The area is 20.

3. Tree-of-Thought (ToT) Prompt Structure

Encourages exploration of multiple reasoning branches before selecting the optimal path:

- Generate a set of candidate “thoughts” for the first reasoning step
- For each candidate, expand into next-level thoughts, optionally scoring or pruning
- Continue branching until a termination criterion is met (depth limit or score threshold)
- Compare complete reasoning chains and select the highest-scoring sequence
- Output the final answer along with the chosen reasoning path

Example: For solving $3x^2 - 10x + 7 = 0$, explore:

- Thought A: Factorization approach
- Thought B: Quadratic formula
- Thought C: Vieta’s formulas

Evaluate efficiency and choose Vieta’s: sum of roots = $\frac{10}{3}$, product of roots = $\frac{7}{3}$.

E.1 Experimental Setting

This section lists the experimental settings used in this study.

Computational Resources. All experiments were conducted on a high-performance computing server equipped with six NVIDIA RTX 6000 Ada Generation GPUs, each with 49 GB of dedicated VRAM. The system utilized CUDA version 12.8 and NVIDIA driver version 570.124.06. These GPUs supported parallel execution of model querying, evaluation, and tool-augmented tasks across our benchmark datasets. The hardware configuration ensured sufficient memory bandwidth and processing capability to accommodate large-scale inference, particularly for multimodal tasks and multi-sample prompting strategies such as CoT-SC and ToT. No resource-related constraints were encountered during experimentation.

F Additional Experimental Results and Discussions for RQ₁

This section presents additional details and experimental results that complement the main evaluation in §4.1. These supplementary findings, together with what has been presented in previous sections, offer comprehensive insights into LLMs capabilities across different EST tasks.

F.1 Complementing experimental results

Table 6 complements with Table 2 with standard deviations.

F.2 Additional Analyses of Modality-Induced Failures

A closer examination of multimodal EST questions reveals several recurring failure modes that cut across models and prompting strategies:

First, we find that **misalignment errors** dominate in tasks requiring table–text or text–figure integration. Models frequently retrieve the correct local evidence (e.g., a row or column from a table) but then conflate it with irrelevant contextual information, producing internally coherent but incorrect rationales. Unlike humans, who naturally ground their reasoning in visual scanning and cross-referencing, LLMs rely on implicit token co-occurrence patterns, which are brittle under distribution shifts in layout or labeling.

Second, **arithmetic and normalization mistakes** emerge when quantitative reasoning spans

Table 5: LLM query hyperparameters used during all experiments.

Hyperparameter	Value	Description
Temperature	0.7	Controls randomness in generation
Top- p (nucleus sampling)	0.95	Probability mass for sampling
Max tokens	2048	Maximum number of tokens to generate
Stop sequences	["\n", "Q: "]	Used to truncate responses
Prompt format	CoT, CoT-SC, ToT	Prompting strategy used in Section 4

Table 6: Standard deviations of performance across five runs. Human testers generally show higher variability, though LLMs also fluctuate, especially on multimodal and quantitative tasks.

Task	Human	GPT-4V			GPT-5			Claude-Sonnet-4			Llama-4-Scout-17B			Qwen-VL-Max			Gemini-2.5		
		ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT
SAT																			
II	2.8	0.3	0.6	1.1	0.4	1.2	1.5	0.7	0.9	1.8	2.9	3.6	4.8	1.5	1.9	2.2	0.8	2.0	2.6
CS	5.2	0.5	1.1	1.8	0.7	1.0	1.5	2.2	3.0	4.1	3.6	4.5	5.1	1.2	1.6	2.2	0.9	1.5	2.1
EI	8.5	0.4	0.9	1.3	0.5	0.8	1.2	1.7	2.5	3.8	2.8	3.6	4.4	1.1	1.5	1.9	0.6	1.4	1.8
EC	4.9	0.3	0.7	1.0	0.4	0.9	1.4	1.6	2.3	3.5	3.2	3.8	4.7	0.9	1.2	1.6	0.8	1.3	2.0
AG	14.2	1.2	2.1	2.9	1.6	2.5	3.8	2.9	3.6	4.4	3.8	4.6	5.2	2.0	2.8	3.3	1.7	2.9	3.5
DA	5.7	0.9	1.8	2.6	1.1	2.2	3.2	2.4	3.1	3.7	3.2	4.2	4.8	1.7	2.5	3.1	1.2	2.4	3.3
GT	5.4	1.0	1.9	2.7	1.3	2.1	3.5	2.6	3.5	4.2	4.0	4.8	5.3	1.9	2.7	3.4	1.5	2.6	3.6
GRE																			
TC	4.7	0.4	0.8	1.2	0.5	0.9	1.3	0.8	1.2	2.0	3.1	3.7	4.9	1.1	1.6	2.1	0.7	1.5	2.2
SE	5.0	0.3	0.7	1.0	0.4	0.8	1.1	1.2	1.9	2.8	3.5	4.2	4.6	0.9	1.3	1.8	0.8	1.4	2.1
RC	5.6	0.8	1.3	1.9	1.0	1.5	2.2	2.5	3.4	4.0	4.1	4.7	5.2	1.4	2.0	2.6	1.1	1.9	2.5
QC	6.0	1.1	1.7	2.4	1.4	2.0	2.7	2.7	3.6	4.3	3.9	4.7	5.1	1.7	2.3	3.0	1.4	2.1	2.9
NE	8.2	1.0	1.6	2.2	1.2	1.9	2.5	2.8	3.5	4.2	4.0	4.6	5.0	1.6	2.2	2.8	1.3	2.0	2.7
DI	6.2	1.3	1.9	2.8	1.5	2.2	3.1	3.1	4.0	4.6	4.2	4.9	5.4	1.9	2.6	3.4	1.6	2.4	3.2
GMAT																			
CR	4.9	0.5	0.9	1.3	0.6	1.0	1.4	1.4	2.0	2.7	3.2	3.8	4.5	1.0	1.6	2.2	0.8	1.5	2.0
RC	5.2	0.7	1.2	1.6	0.9	1.3	1.9	2.0	2.8	3.5	3.6	4.3	5.0	1.3	1.8	2.4	1.0	1.7	2.3
PS	6.3	1.4	2.1	2.7	1.8	2.5	3.3	3.0	3.9	4.6	4.1	4.9	5.3	2.0	2.7	3.5	1.6	2.3	3.1
DS	6.1	1.3	2.0	2.6	1.7	2.4	3.0	2.9	3.8	4.4	4.0	4.7	5.2	1.8	2.5	3.2	1.5	2.2	2.9
IR	6.5	1.5	2.2	3.0	1.9	2.6	3.5	3.2	4.1	4.8	4.3	5.0	5.4	2.1	2.9	3.7	1.7	2.5	3.4
TOEFL																			
FI	5.5	0.4	0.7	1.0	0.5	0.9	1.2	0.9	1.3	1.9	1.7	2.3	2.9	0.8	1.1	1.5	0.6	1.0	1.6
IR	5.8	0.8	1.1	1.5	1.0	1.4	1.9	1.6	2.1	2.7	2.4	3.1	3.8	1.2	1.6	2.0	1.1	1.5	2.2
TS	5.1	0.5	0.8	1.2	0.6	1.0	1.4	1.2	1.7	2.3	1.9	2.5	3.4	0.9	1.3	1.8	0.7	1.2	1.6
FI	0.6	0.6	0.9	1.3	0.7	1.1	1.6	1.5	2.0	2.6	2.1	2.8	3.9	1.1	1.5	2.0	0.9	1.4	1.9
IF	2.7	0.9	1.3	1.8	1.2	1.7	2.3	1.9	2.6	3.2	2.6	3.4	4.7	1.4	1.9	2.6	1.2	1.8	2.5
IELTS																			
II	5.4	0.5	0.7	1.0	0.6	1.0	1.4	1.3	1.8	2.5	2.0	2.7	3.6	0.9	1.3	1.7	0.7	1.1	1.5
MS	5.7	0.7	1.0	1.3	0.9	1.3	1.8	1.7	2.2	2.9	2.5	3.2	4.1	1.1	1.5	2.0	0.8	1.2	1.7
CP	5.3	0.6	0.9	1.2	0.7	1.2	1.7	1.6	2.1	2.8	2.3	3.0	4.0	1.0	1.4	1.9	0.9	1.3	1.8
IM	5.8	0.8	1.1	1.5	1.0	1.5	2.1	2.0	2.7	3.4	2.7	3.5	4.6	1.2	1.7	2.2	1.1	1.6	2.1
CL	6.2	1.0	1.4	2.0	1.2	1.8	2.5	2.3	3.0	3.8	3.0	3.9	5.0	1.4	2.0	2.7	1.2	1.7	2.3
SA	5.5	0.7	1.0	1.4	0.8	1.3	1.9	1.5	2.0	2.6	2.2	2.9	4.2	1.1	1.5	2.1	0.9	1.4	1.8

modalities. In GRE Data Interpretation, for instance, models can identify the relevant chart element but fail to convert absolute differences into relative percentages, leading to incorrect comparative judgments. These failures suggest weaknesses in bridging symbolic numeric operations with natural language descriptions, particularly when multiple units, scales, or denominators must be tracked simultaneously.

Third, **over-trust in salient cues** is a pervasive issue. When figures or diagrams contain visually

prominent but logically irrelevant elements (e.g., a bolded number or a large bar in a chart), models often anchor on these features even when the question explicitly requires a subtler comparison. Humans, by contrast, employ metacognitive checks such as rereading the question stem to confirm task requirements.

Finally, we observe **compounding variance across modalities**. Errors often cascade: a misread in the textual description can propagate into the tabular lookup, which then interacts with an arith-

metic miscalculation, producing errors that appear systematic but in fact result from small deviations at multiple stages. This multi-stage fragility highlights the gap between current LLMs’ sequential token prediction and the hierarchical integration that multimodal reasoning demands.

Insights. These analyses underscore that modality complexity introduces qualitatively new challenges beyond scaling model size or training data. Future work on EST-style problem solving must therefore move beyond token-level modeling to incorporate explicit alignment, symbolic grounding, and verification mechanisms that can emulate the multi-channel reasoning strategies of human test-takers.

F.3 LLMs vs. Human Testers

Table 7: Statistical significance of performance differences between human testers and best-performing LLM (McNemar’s test, p -values). Asterisks indicate significance levels: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, ns: not significant.

Exam	Task	Best LLM	p -value
SAT	II	Gemini-2.5 (ToT)	0.001***
	CS	Gemini-2.5 (CoT)	0.089ns
	EI	GPT-5 (CoT)	0.023*
	EC	GPT-5 (CoT)	0.156ns
	AG	Gemini-2.5 (ToT)	<0.001***
	DA	GPT-5 (ToT)	<0.001***
	GT	GPT-5 (CoT)	0.008**
GRE	TC	GPT-5 (ToT)	0.012*
	SE	GPT-5 (CoT)	0.034*
	RC	GPT-5 (CoT)	0.003**
	QC	GPT-5 (ToT)	<0.001***
	NE	Human superior	<0.001***
	DI	Human superior	<0.001***
GMAT	CR	Claude-S4 (ToT)	0.019*
	RC	Claude-S4 (ToT)	0.005**
	PS	Human superior	<0.001***
	DS	Human superior	<0.001***
	IR	Human superior	<0.001***
TOEFL	FI	GPT-5 (ToT)	0.002**
	IR	GPT-5 (CoT)	0.004**
	TS	Human superior	0.178ns
	FI (Listen)	GPT-5 (ToT)	0.028*
	IF	GPT-5 (ToT)	<0.001***
IELTS	II	GPT-5 (ToT)	0.041*
	MS	Human superior	0.015*
	CP	Claude-S4 (ToT)	<0.001***
	IM	GPT-5 (ToT)	0.092ns
	CL	GPT-5 (ToT)	0.006**
	SA	Human superior	0.134ns

Across tasks and modalities, we observe qualitatively different patterns of variability between human testers and LLMs, with significance measures shown in Table 7. Note that human variability is driven primarily by background knowledge, test-taking habits, fatigue, and individual strategy preferences; mistakes tend to be idiosyncratic and cluster by prior exposure (e.g., comfort with specific grammar rules or math subskills).

In contrast, LLM variability is shaped by decoding stochasticity, prompt sensitivity, and fragile intermediate reasoning: the same model can oscillate between correct and incorrect answers when minor surface features change (instruction phrasing, option order, or distractor salience).

Humans often adapt strategy mid-session and exhibit metacognitive checks (skimming, re-reading, sanity checks on units or logic), whereas LLMs more frequently display “local optimum” traps (e.g., latching onto a salient but irrelevant cue) or instruction-following drift without self-correction. Variability is also modality-dependent: humans degrade with cognitive load and time pressure, while LLMs degrade more when cross-representation alignment is required (text–table, text–image, text–audio), reflecting weaknesses in binding and content grounding rather than domain knowledge alone.

F.4 Impact by Prompting Complexity

We also find that more sophisticated prompting strategies (e.g., ToT) do not consistently lead to better performance, although more enriched reasoning is provided:

Case Study (Text Completion): *Although it is easy to imagine that the ____ of technological innovation has accelerated innovation has proceeded at a fairly ____ pace since the Industrial Revolution. Options: 1. (i) tempo, (ii) constant. Options: 2. (i) novelty, (ii) sporadic. Options: 3. (i) velocity, (ii) erratic.*

CoT focuses on overall sentence coherence: The sentence suggests a contrast between the perception that innovation has *accelerated* and the Thus, the correct answer is **Option 1.**

ToT forces blank-by-blank exploration: Branch 1: For first blank. (1.a) Option "tempo" → meaning = speed. (1.b) Option "novelty" → meaning = newness Branch 2: For the sec-

ond blank (2.a) Option "constant" → meaning = unchanging **Final Answer: Option 3 (i) velocity, (ii) erratic.** (LLM gets confused due to multiple branches and partial fits.)

This suggests that complex reasoning frameworks may sometimes introduce additional cognitive overhead without corresponding gains in accuracy, particularly for models not explicitly optimized for such structured inference.

We further find that LLM performance may also be influenced by prompting and decoding choices, with different trends:

(1) **Chain-of-thought (CoT)** generally regularizes reasoning on verbal items by externalizing intermediate structure, but it can also amplify spurious rationales when the initial trajectory is off-distribution. This is particularly visible in GRE Sentence Equivalence, where once the model locks onto a semantically plausible but incorrect synonym, subsequent steps reinforce the error rather than revising it. In math-heavy tasks like SAT Algebra, CoT sometimes leads to over-elaboration, generating unnecessary symbolic steps that increase the chance of arithmetic drift.

(2) **Tree-of-thought (ToT)** tends to help on search-like or data-integration tasks, yet it introduces longer reasoning paths that sometimes accumulate small local errors or trigger premature branch pruning. For instance, in GMAT Data Sufficiency, ToT can improve systematic exploration of conditions but is also prone to “path explosion,” where irrelevant branches dominate and obscure the correct constraint check. Similarly, in GRE Data Interpretation, ToT may spread reasoning across multiple chart elements without recombining them, leading to fragmented conclusions.

(3) **In-context learning (ICL)** works best when exemplars match the target item’s latent schema (discourse function, syntactic frame, or quantitative template); schema-mismatched exemplars can anchor the model to the wrong solution space. In IELTS Matching Sentence tasks, schema-aligned exemplars guide the model toward identifying discourse relations effectively, whereas mismatched exemplars bias the model toward surface string overlaps. In TOEFL Inference questions, exemplar mismatch often causes the model to ignore pragmatic cues like tone or implied stance, overfitting instead to lexical similarity.

Insights. These observations suggest that elic-

itation strategies interact strongly with task type and associated failure modes. CoT excels in tasks requiring layered linguistic reasoning but exacerbates semantic anchoring errors when the first step is flawed. ToT adds value where systematic exploration is necessary (tables, condition checks, multi-source reasoning), but it magnifies variance when intermediate steps are noisy or poorly pruned. ICL is powerful when schema alignment is possible, but fragile when exposed to distributional mismatch between exemplars and target questions. Together, these findings underscore that reliable EST problem solving requires not only robust prompt design but also adaptive elicitation strategies that are sensitive to the structural demands and common pitfalls of each task family.

F.5 Influence of Question Difficulty

The question types in Table 2 already reflect varying levels of difficulty—for example, inference questions in TOEFL reading are generally more challenging than factual information questions.

We further investigate how question difficulty influences LLM performance. We evaluate on GRE, which allow for clearer categorization, wherein text completion (TC) questions are divided into one-, two-, and three-blank formats with a greater number of blanks corresponds to higher difficulty. Similarly, quantitative (Quant) questions are pre-labeled as either medium or hard. Figure 6 presents LLM performance across these difficulty levels. Interestingly, we observe no clear performance degradation as difficulty increases, where human testers show a significant decline in answer correctness as the difficulty increases. These results suggest that LLMs may not be sensitive to human-defined difficulty levels and instead exhibit an equilibrium across structurally similar problems, regardless of their intended complexity in English or mathematical settings.

Note on Difficulty Categorization: Difficulty levels in Figure 6 use official ETS metadata. GRE Quantitative questions are labeled as “medium” or “hard” in source materials; Text Completion difficulty corresponds to the number of blanks (1/2/3), a structural feature standardized in test design that correlates with cognitive complexity. We do not impose difficulty labels on exams (SAT, GMAT, TOEFL, IELTS) lacking official metadata.

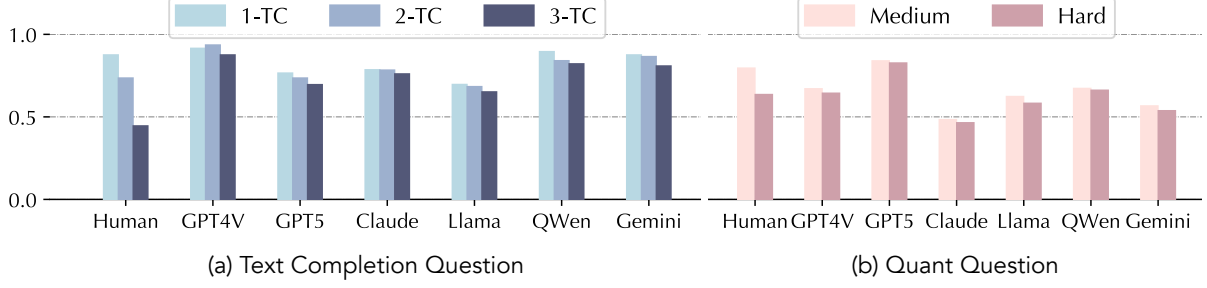


Figure 6: LLM performance across varying levels of question difficulty, using CoT due to its representativeness. We focus on GRE text completion tasks with 1-, 2-, and 3-blanks, as well as available medium- and hard-level quantitative problems.

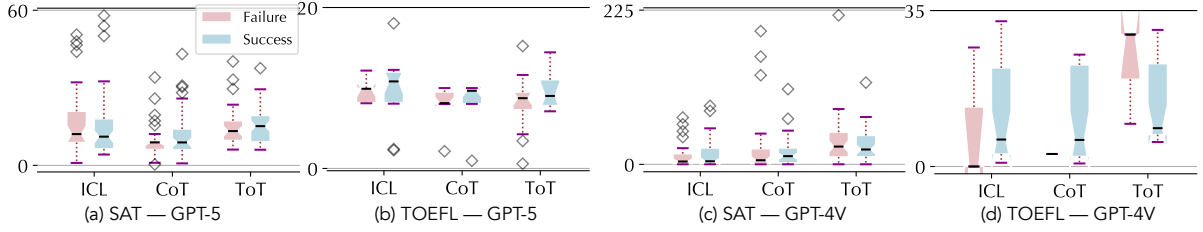


Figure 7: Inference time (in seconds) for failed and successful cases. More results are in Figure 8.

Table 8: Mann–Whitney U test of the inference time between failed and successful cases. We report p-values to assess the statistical significance of differences between the success and failure groups. Additional results are presented in Table 9.

Exam	GPT-5			GPT-4V		
	ICL	CoT	ToT	ICL	CoT	ToT
SAT	0.278	0.814	0.443	0.197	0.117	0.512
TOEFL	0.610	0.389	0.515	0.295	0.640	0.115

F.6 Inference Efficiency

Another important consideration is the inference time of LLMs. To analyze the relationship between inference time and answer correctness, we record the generation time (in seconds) for each response and categorize the results into two groups: correctly answered and incorrectly answered questions. We then plot the distribution of inference times for both groups, as shown in Figure 7.

From the box plots, we observe that the inference times for correct and incorrect predictions are similar, without a notable separation between the two groups. To statistically validate this observation, we perform a two-tailed Mann–Whitney U test (McKnight and Najab, 2010). The Mann–Whitney U test is a non-parametric hypothesis test that assesses whether two independent samples come from the same distribution. It evaluates whether the

distributions differ in location (median) or overall shape. As listed in Table 8, across all evaluated models, the Mann–Whitney U tests yield p-values higher than 0.05 (a commonly used significance level), indicating no statistically significant difference between the inference time distributions of correct and incorrect predictions. This suggests that the time an LLM spends on answering a question does not correlate with answer correctness. Inference time appears largely independent of answer quality on the EST benchmark.

Figure 8 and Table 9 provides more experimental results for inference time across success and failure cases.

G Additional Experimental Results and Discussions for RQ₂

G.1 Evaluation Metrics for Cognitive Trajectories

Because the cognitive trajectory models a heterogeneous problem-solving process—ranging from finding a specific sentence in a reading passage to computing a complex polynomial—no single metric can adequately evaluate all intermediate reasoning steps. Therefore, as detailed in Table 10, we deploy a suite of tailored metrics depending on the cognitive function of the cognitive step.

Extractive and Localization Nodes (IoU &

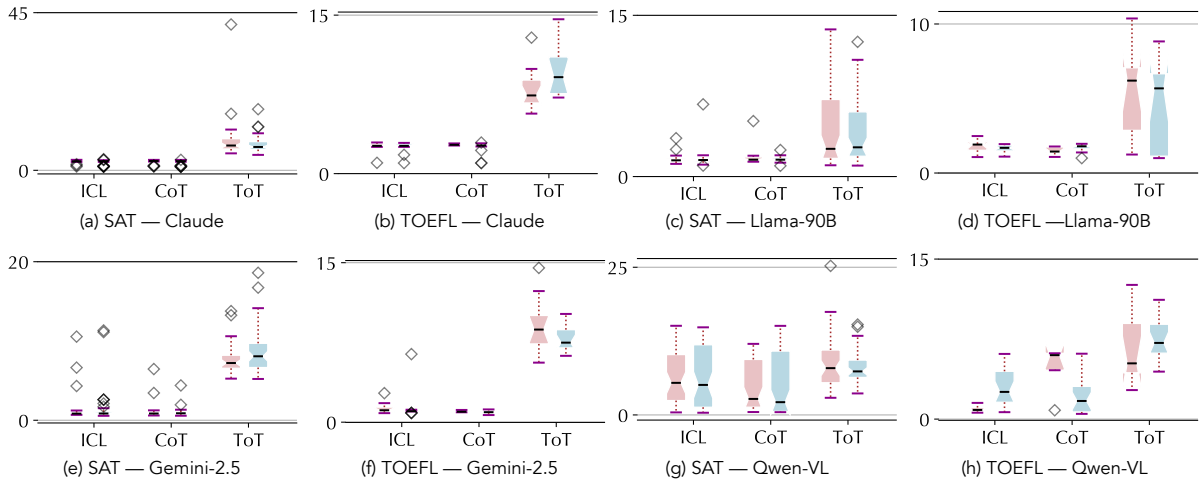


Figure 8: Inference time (in seconds) for failed and successful cases. Complement to Figure 7.

Table 9: Mann–Whitney U test of the inference time between failed and successful cases. Complement to Table 8.

Exam	Claude-Sonnet-4			Llama-4-Scout-17B			Gemini-2.5			Qwen-VL-Max		
	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT
SAT	0.572	0.359	0.336	0.786	0.817	0.619	0.994	0.449	0.105	0.734	0.377	0.884
TOEFL	0.903	0.137	0.084	0.449	0.084	0.360	0.267	0.414	0.231	0.159	0.088	0.374

Table 10: Evaluation metrics applied to individual Cognitive DAG nodes across the three pedagogical categories in ESTBOOK. The heterogeneous nature of the nodes (ranging from text extraction to mathematical formulation) necessitates a diverse suite of deterministic and semantic metrics.

Pedagogical Category	Cognitive DAG Node	Evaluation Metric	Diagnostic Justification
Category I: Knowledge-Intensive Retrieval	$c_{\text{syntax_parse}} / c_{\text{semantic_parse}}$	Exact Match (Accuracy)	Verifies structural or semantic classification (e.g., identifying sentence tense or missing part of speech).
	$c_{\text{rule_match}} / c_{\text{logical_scope}}$	Token-level IoU / F1	Measures the overlap between the LLM’s identified logical anchor (e.g., contrast words) and the ground truth.
	$c_{\text{predict}} / c_{\text{resolve}}$	BERTScore	Evaluates the semantic plausibility of the predicted word or phrase, accepting valid paraphrases.
Category II: Reasoning-Intensive Execution	$c_{\text{analytical_goal}} / c_{\text{model}}$	Exact Match (Accuracy)	Classifies if the model correctly identified the math/logic domain (e.g., identifying a problem as “probability”).
	$c_{\text{visual_parse}} / c_{\text{formulate_eq}}$	Symbolic Equivalency	Uses a CAS (e.g., SymPy) or structural matching to verify formulas, ignoring surface-level formatting ($x + y$ vs. $y + x$).
	$c_{\text{analyze}} / c_{\text{compute}}$	1 – Normalized RMSE	Evaluates numeric execution, providing partial credit for minor rounding differences while penalizing gross arithmetic failures.
Category III: Hybrid Integration	$c_{\text{intent}} / c_{\text{entities}}$	Exact Match (Accuracy)	Strictly evaluates whether the correct target subject or comparative entities were isolated from the prompt.
	$c_{\text{perceive_text}} / c_{\text{constraints}}$	Sentence-level IoU	Assesses the accuracy of localizing the specific evidence spans or logical constraints within a dense passage/audio.
	$c_{\text{extract}} / c_{\text{evaluate}}$	BERTScore	Measures the fidelity of the final deductive reasoning or summarized evidence against the ground-truth rationale.

F1): For nodes that require the LLM to scan and isolate information from a larger context (e.g., $c_{\text{logical_scope}}$ in Category I, or $c_{\text{perceive_text}}$ in Category II), we utilize Intersection over Union (IoU) and Token-level F1 scores. These metrics are pedagogically appropriate because they measure *attention precision*. If a model highlights an entire paragraph when the evidence is contained in a single sentence, the IoU naturally penalizes the model for a lack of focus, reflecting a common human test-taking inefficiency.

Mathematical and Execution Nodes (Symbolic Equivalency & RMSE): Category II nodes demand rigorous computational precision. Evaluating intermediate mathematical formulation ($c_{\text{formulate_eq}}$) via standard string matching frequently yields false negatives (e.g., penalizing $2x + y = 10$ against $y = 10 - 2x$). To resolve this, we utilize a Computer Algebra System (CAS), such as SymPy, to verify *Symbolic Equivalency*. For the final execution node (c_{compute}), we employ $1 - \text{Normalized RMSE}$ for continuous numeric outputs, which strictly penalizes fundamental arithmetic hallucinations while gracefully handling minor floating-point or rounding deviations.

Generative and Deductive Nodes (BERTScore): For nodes requiring the synthesis of meaning or final logical resolution (e.g., c_{resolve} or c_{extract}), strict exact-match metrics are overly punitive. An LLM might correctly infer a concept but express it using slightly different vocabulary than the ground truth. To accommodate valid linguistic variations while ensuring the core pedagogical reasoning remains intact, we employ BERTScore to evaluate the semantic fidelity of the generated text against the official standardized test explanations.

G.2 Complementing experimental results

Figure 4 and 10 provides additional breakdown analysis on other LLMs, wherein the observation aligns with Section 4.2.

G.3 Additional Case Studies

This section provides additional case studies complement to observations and conclusions in RQ₃.

Case Study (GRE Quant – Numeric Entry):

"If the sum of three consecutive odd integers is 111, what is the smallest of the three?"

LLM Reasoning: Claude generates an incorrect expression: $x + x + 1 + x + 2 = 111$ as it treats the numbers as consecutive integers rather than odd integers (which should be $x + x + 2 + x + 4 = 111$).

Interpretation: Numeric entry and multi-modality significantly impede LLM reasoning.

Tasks involving numeric input and multimodal understanding (e.g., math from SAT) remain particularly challenging for LLMs. Unlike classification-style questions with fixed answer choices, numeric-entry tasks require precise mathematical formulation, symbolic manipulation, and error-free calculation—all of which are error-prone in current models.

The challenge is amplified in multimodal settings, where the model must align visual, tabular, or symbolic inputs with textual queries before reasoning can even begin.

Case Study (GMAT Integrated Reasoning – Table + Math Computation):

A question requires *"selecting a product with the highest profit margin based on a table of costs and revenues."* GPT-4V incorrectly reads the table and subtracts the cost from total units sold rather than revenue, leading to an invalid numeric result.

Case Study (GRE – Semantic Reasoning):

"Select the pair of words that best completes the sentence: 'While the professor's tone was ostensibly ____, her critique was undeniably severe and cutting.'" Options: (A) respectful – insulting (B) conciliatory – harsh (C) disinterested – involved ...

LLM Reasoning: GPT-5 selects (A) due to surface-level antonymy ("respectful" vs. "insulting"), but fails to resolve the nuanced implication of "ostensibly" versus "undeniably," which is essential for semantic disambiguation. Claude performs similarly, missing the contrastive logic implied by adverbs. Only Gemini-Pro correctly identifies (B), recognizing the indirect semantic contrast.

Interpretation: This illustrates how LLMs, despite strong lexical capabilities, still struggle with subtle discourse-level signals that guide meaning, such as modal adverbs or pragmatic contrast. It reinforces your claim that deeper reasoning (not

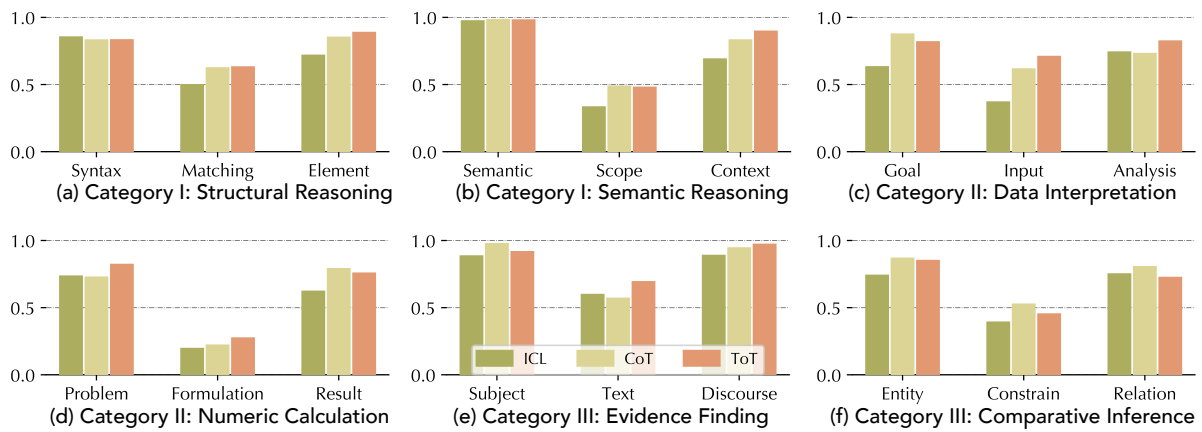


Figure 9: Breakdown analysis on Claude.

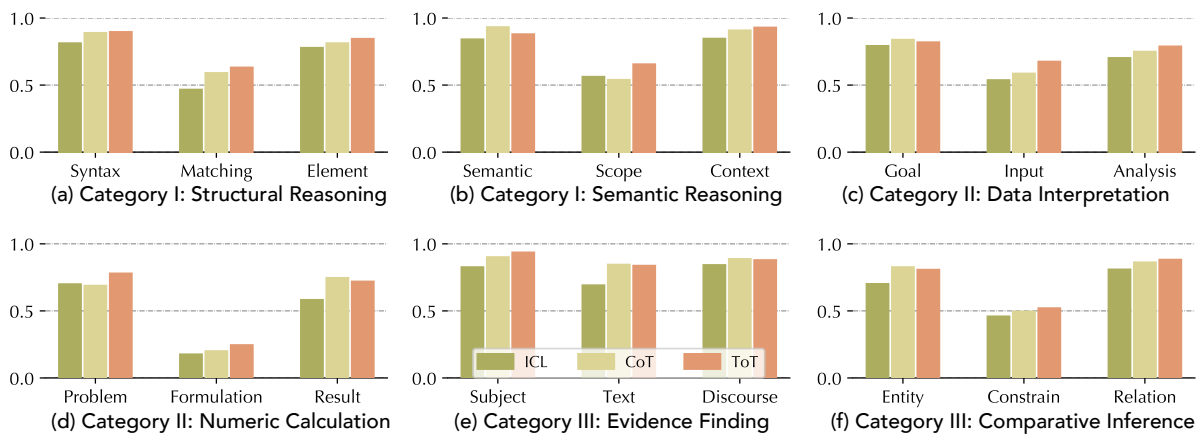


Figure 10: Breakdown analysis on Gemini.

surface matching) is the primary challenge.

Case Study (IELTS Listening – Evidence Localization):

"What is the speaker's main reason for supporting the expansion of the city park?"

Audio clip: The speaker describes multiple benefits of expanding a city park, including noise reduction, community wellness, and increased biodiversity.

LLM Reasoning: Using Whisper-transcribed audio, Qwen and GPT-4V highlight "noise reduction" as the answer because it is mentioned first and most clearly. However, the correct answer is "community wellness," which is emphasized later in the speech with supporting elaboration. Only Gemini-Pro correctly weighs the relative emphasis across the transcript.

Interpretation: This example shows that current models tend to over-prioritize the first mentioned or most literal content in a multimodal context, and fail to simulate human-like discourse pri-

orization. It also suggests weaknesses in aligning Whisper transcripts with reasoning modules.

Case Study (SAT Reading – Evidence Pairing):

"Which of the following best supports the answer to the previous question?"

Passage: A student challenges the conclusions of a scientific article. Main question: "Which claim does the student most strongly refute?" Evidence question: "Which line best supports that refutation?"

LLM Reasoning: Claude selects a sentence that contains a general critique but does not directly support the earlier answer. GPT-5 does better at matching tone but fails to anchor the evidence to the refuted claim. Only Gemini-Pro correctly links the reasoning across both questions.

Breakdown Challenge: Evidence Localization. Highlights LLMs' difficulty in chaining answers across linked questions, especially when reasoning must remain consistent.

Case Study (GRE Verbal – Logical Structure): *"Which of the following best describes the structure of the passage?"* Passage: An author introduces a phenomenon, critiques one explanation, and then proposes an alternative.

LLM Reasoning: LLaMA-3 and Qwen select options that only capture the first half of the structure (e.g., critique). GPT-4V overgeneralizes to a "compare-and-contrast" structure. Only Claude correctly recognizes the structure as "Introduction → Criticism → Alternative Explanation."

Breakdown Challenge: Structural Reasoning. Illustrates that models struggle to track abstract rhetorical moves across a passage, even when comprehension is accurate.

Case Study (GMAT Integrated Reasoning – Two-Part Analysis): *"Select one answer for each of the following two conditions: (1) Which project has the highest ROI? (2) Which project has the lowest risk?"* Tabular data includes five projects with ROI and risk indicators.

LLM Reasoning: GPT-4V selects Project C for both ROI and risk, confusing "least cost" with "least risk." Claude selects correctly for ROI but fails to interpret qualitative risk descriptors. Gemini-Pro and GPT-5 complete both selections correctly.

Breakdown Challenge: Data Interpretation. Shows difficulty in multi-constraint reasoning and mapping discrete table fields to textual decision logic.

Case Study (IELTS Writing – Grammatical Error Correction): *"Identify and correct the grammatical error in the following sentence: 'If she would have gone to the meeting, she could had contributed valuable insight.'"*

LLM Reasoning: Qwen changes "she could had" to "she could has," worsening the error. Claude corrects "would have gone" to "had gone" but leaves the second clause unaltered. Only GPT-5 performs both corrections, yielding: "If she had gone to the meeting, she could have contributed valuable insight."

Breakdown Challenge: Structural Reasoning. Highlights syntax correction challenges where multiple clauses require coordinated grammatical edits.

Case Study (GRE Quant – Comparative Judgment): *"Quantity A: The square of the average of 3 and 7; Quantity B: The average of the squares of 3 and 7."*

LLM Reasoning: GPT-4V computes both but incorrectly concludes that Quantity A is greater, mistaking $((3 + 7)/2)^2 = 25$ as greater than $(3^2 + 7^2)/2 = 29$. LLaMA-3 gives no answer and repeats the prompt. Claude answers correctly but offers no reasoning trace.

Breakdown Challenge: Comparative Judgment. Demonstrates common mistakes in applying formulas and comparing expressions under symbolic transformation.

Moreover, we also observe that prompting strategies (ICL, CoT, ToT) do not significantly affect performance in certain stages of breakdown analysis, especially where task complexity is low or answer derivation is mostly local. Below are some case studies to address this:

Case Study (SAT Reading – Factual Retrieval): *"According to the passage, what did the author list as one benefit of urban green space?"* The relevant sentence in the passage states: "Green spaces improve air quality and reduce noise levels."

LLM Reasoning across Prompts: All three prompting strategies (ICL, CoT, ToT) lead to the same correct output across GPT-5 and Claude. In each case, the models locate the exact supporting sentence and extract "improve air quality" or "reduce noise levels" without variation. CoT and ToT generate unnecessary intermediate steps without improving the final answer.

Breakdown Relevance: Evidence Finding (step: locate and extract factual information). Insight: Prompting complexity doesn't help when the required reasoning is local and unambiguous.

Case Study (GRE Verbal – 1-Blank Text Completion): *"The scientist's explanation was praised for its clarity and ____, making it accessible to a general audience."* Options: (A) convolution, (B) transparency, (C) complexity...

LLM Reasoning across Prompts: All strategies (ICL, CoT, ToT) result in the selection of (B) "transparency." The reasoning is nearly identical: the model detects positive sentiment

from "praised" and "clarity," and eliminates antonymic distractors like "convolution." CoT and ToT elaborate more, but do not change the choice or rationale.

Breakdown Relevance: Semantic Reasoning (step: sentiment alignment and elimination). Insight: For simple semantic alignment, ICL already suffices, and additional reasoning scaffolds don't yield improvement.

Case Study (GMAT Quant – Basic Arithmetic): "What is the value of $3x + 2$ if $x = 5$?"

LLM Reasoning across Prompts: All prompting strategies produce the correct answer, 17, with or without intermediate steps. CoT redundantly walks through " $3x = 15$, then $15 + 2 = 17$," while ToT splits the steps further into node-like structures. None of the strategies reduce error, latency, or confidence.

Breakdown Relevance: Numeric Calculation (step: direct substitution and evaluation). Insight: When reasoning is shallow and deterministic, prompting scaffolds become unnecessary overhead.

H Additional Experimental Results and Discussions for RQ₃

H.1 Detailed Failure Mode Analysis: The Integration Bottleneck

Despite overall accuracy metrics suggesting that modern LLMs are approaching human parity on ESTs, however, our node-level evaluation of distractor susceptibility reveals a consistent, structural vulnerability. Across all tested models, from heavily parameterized frontier models to efficient open-weight architectures, failures are rarely uniformly distributed across the reasoning trajectory. Instead, they exhibit a distinct "V-shaped" performance drop, localized almost exclusively at the second node of our three-step cognitive trajectory. We term this the **Integration Bottleneck**.

As summarized in Table 11, models demonstrate high robustness during initial conceptualization (Step 1) and subsequent final-stage execution (Step 3). The catastrophic failures occur when models must bind parsed constraints to specific textual or mathematical representations (Step 2). Generative architectures, optimized for next-token probability, exhibit a strong bias toward plausible rationalization. When presented with a distractor option that

contains a *Partial Truth* or a *Faulty Premise*, LLMs tend to hallucinate a valid integration rather than acting as strict logical discriminators.

Category I Vulnerabilities: The Illusion of Fluency

In Knowledge-Intensive tasks (Syntax and Semantic Reasoning), the primary failure mechanism is an over-reliance on local coherence. Models successfully parse the syntax (Step 1) but fail during *Matching* or *Localizing Scope* (Step 2). Distractor options engineered with high lexical overlap (Type III: Partial Truth) easily deceive the models. Instead of rejecting a syntactically fluent but logically flawed distractor, the LLM generates a post-hoc rationalization to justify why the word "fits," demonstrating an inability to decouple linguistic fluency from rigorous semantic gating.

Category II Vulnerabilities: Symbol Grounding Failures

In Reasoning-Intensive tasks (Data Interpretation and Numeric Calculation), the Integration Bottleneck manifests as a symbol grounding failure. LLMs consistently identify the correct problem type (Step 1), yet experience severe degradation during *Parsing Visual Input* or *Formulating Math* (Step 2). They are highly susceptible to Type IV (Execution) distractors. For instance, if a distractor option represents the answer to an intermediate step of an equation rather than the final solution, the LLM will frequently accept it, losing track of the overarching mathematical constraints established in Step 1.

Category III Vulnerabilities: Constraint Relaxation

For Hybrid Integration tasks (Evidence Finding and Comparative Inference), the bottleneck occurs at *Comprehend Text* and *Apply Constraints*. Here, human-like distractors utilize Type II (Out of Scope) traps by introducing highly plausible but unstated premises. While human test-takers are explicitly trained to identify when information is "Not Given" (e.g., in IELTS Reading or GMAT Data Sufficiency), LLMs struggle with this strict epistemic boundary. The models exhibit "constraint relaxation," implicitly filling in missing data to force a logical connection, thereby erroneously accepting invalid rationales.

Table 11: Qualitative Taxonomy of LLM Failure Modes at the Integration Bottleneck (Step 2). The table maps the specific cognitive vulnerability in each task to the primary distractor trap that successfully derails the model’s reasoning trajectory.

Pedagogical Category	Task & Bottleneck Node	Primary Distractor Trap	Observed LLM Failure Mechanism
Category I: Knowledge-Intensive Retrieval	Structural: <i>Match Text</i>	Type III: Partial Truth (High Lexical Overlap)	Models match the syntactic structure but ignore subtle semantic inversion. They select distractors that “sound natural” despite violating local logical scope.
	Semantic: <i>Localize Scope</i>	Type III: Faulty Premise	Models fail to apply restrictive clauses (e.g., “except,” “unless”). They validate distractors that are true in isolation but false under the specific sentence constraints.
Category II: Reasoning-Intensive Execution	Data Interpretation: <i>Parse Visual Input</i>	Type II / IV: Out of Scope / Misaligned Variables	Models misalign textual entities with tabular/visual axes. They confidently compute trends using the wrong row/column, failing to reject distractors derived from this exact human error.
	Numeric Calculation: <i>Formulate Math</i>	Type IV: Execution Error (Flawed Formulation)	The most severe bottleneck. Models correctly model the problem type but fail to translate natural language into a strict algebraic equation, falling for distractors representing incomplete formulations.
Category III: Hybrid Integration	Evidence Finding: <i>Comprehend Text</i>	Type III: Partial Truth (Paraphrase Hallucination)	Models accept distractors that use exact keywords from the text but combine them to form a conclusion not supported by the actual discourse unit.
	Comparative Inference: <i>Apply Constraints</i>	Type II: Out of Scope (Unwarranted Assumption)	Models fail to maintain strict logical boundaries. They “smooth over” missing information by assuming implicit variables, thereby accepting distractors that require knowledge outside the provided text.

I Additional Experimental Results and Discussions for RQ₄

In Section 4.4, we introduced four bottleneck-specific mitigation strategies designed to repair the reasoning trap (Step 2 of our cognitive trajectory) and reported their impact on a representative subset of ESTBOOK. In this appendix, we extend this analysis to a broader range of tasks across all six standardized tests.

Table 12 presents the comparative performance of baseline Chain-of-Thought (CoT) prompting versus our targeted Mitigated (+Mit.) prompting for GPT-4V, GPT-5, and Claude-Sonnet-4. The baseline scores correspond directly to the CoT columns reported in the main results (Table 2).

I.1 Extended Findings on Mitigation Generalizability

Cross-Test Applicability of Evidence Anchoring: As shown in Table 12, the *Evidence-Anchored* strategy is highly effective not just for dense graduate-level reading (GRE-RC), but also for proficiency exams like TOEFL and IELTS. By forcing the models to extract verbatim spans before inference, we observed a strict reduction in Type III (Partial Truth) errors, pushing GPT-5’s accuracy on TOEFL Factual Information (TOEFL-FI) to a near-ceiling 96.5%.

Generalizing Structural Constraints: The *Syntax-First* strategy proved highly transferable to SAT Cross-Textual Synthesis (SAT-CS) and GRE Sentence Equivalence (GRE-SE). In SAT-CS, models historically struggled with integrating premises across two distinct texts. By mandating a structural outline of the argumentative flow prior to selection, Claude-Sonnet-4’s performance improved significantly from 70.2% to 78.4%, overcoming its tendency to select options that were topically relevant but logically inverted.

Massive Gains in Mathematical Formulation: The *Symbolic Verification* layer yielded the most dramatic relative improvements across all tasks. Mathematical reasoning tasks like GMAT Problem Solving (GMAT-PS) and SAT Algebra (SAT-AG) suffer from catastrophic drops at the formulation node. By forcing models to isolate and verify algebraic equations line-by-line before calculating, GPT-5 improved from 31.1% to 53.6% on GMAT-PS, successfully bypassing the distractors designed to catch execution errors.

Robustness in Multimodal Data Integration:

Finally, the *Table-Alignment* strategy confirmed that poor multimodal reasoning in LLMs is often a procedural failure rather than a pure vision-encoder failure. By requiring explicit explicit textual-to-tabular mapping instructions, models achieved massive gains not only on GMAT-IR but also on GRE Data Interpretation (GRE-DI), where GPT-4V improved from 56.0% to 72.5%. This indicates that forcing a “lookup plan” perfectly mitigates the human-like error of reading the wrong axis or column.

J Additional Mitigation

To avoid unexpected evaluation errors incurred during cognitive trajectory, we also integrate extra mitigation and validation strategies detailed in this section.

J.1 Error Propagation Mitigation Strategies

While our breakdown analysis uses an oracle setting to isolate reasoning capabilities at each step, we recognize that error propagation is a critical concern for real-world deployment. Here we discuss potential mitigation strategies informed by our diagnostic findings.

Verification Mechanisms at Identified Bottlenecks. Our breakdown analysis reveals specific steps where models consistently fail, creating opportunities for targeted intervention. For Task IV (Data Interpretation), Step 2 (Parse Visual Data) achieves only 51–65% accuracy across models, indicating a critical bottleneck in multimodal parsing. A verification mechanism could require models to explicitly reference specific table cells, chart elements, or diagram components before proceeding to computation, ensuring visual-textual alignment. For example, before computing a percentage change from tabular data, the system could validate that the model has correctly identified the relevant rows and columns by requiring it to output structured references (e.g., “Row 3, Column 2: Sales 2023 = \$45M”) that can be programmatically verified against the actual data structure.

Multi-Stage Validation Protocols. For Task V (Numeric Calculation), our breakdown shows that mathematical formulation (Step 2) achieves only 49–56% accuracy before computation even begins. This suggests implementing symbolic validation that checks formula coherence and unit consistency before execution. For instance, in physics word problems, a validation step could verify that the

Table 12: Extended evaluation of targeted mitigation strategies across a wider spectrum of ESTBOOK tasks. The baseline (Base) scores represent standard Chain-of-Thought (CoT) performance, while Mitigated (+Mit.) scores reflect the application of our bottleneck-specific prompts.

Task & Applied Mitigation Strategy	GPT-4V		GPT-5		Claude-S-4	
	Base	+Mit.	Base	+Mit.	Base	+Mit.
Strategy 1: Evidence-Anchored CoT (Mitigates Paraphrase Hallucination)						
GRE-RC (Reading Comprehension)	77.8	93.5	87.1	90.4	69.3	84.1
TOEFL-FI (Factual Information)	86.3	91.2	93.2	96.5	83.9	88.4
IELTS-MS (Matching Sentences)	85.1	89.4	83.0	88.1	81.0	85.6
Strategy 2: Syntax-First CoT (Mitigates Semantic Drift / Fluency Bias)						
GRE-TC (Text Completion)	79.5	84.2	73.4	78.1	61.0	76.7
GRE-SE (Sentence Equivalence)	81.0	86.3	86.5	90.2	82.1	87.9
SAT-CS (Cross-Textual Synthesis)	77.4	83.1	82.7	88.5	70.2	78.4
Strategy 3: Symbolic Verification Layer (Mitigates Flawed Math Formulation)						
GRE-NE (Numeric Entry)	38.0	55.4	33.9	62.1	25.0	48.7
GMAT-PS (Problem Solving)	34.3	48.2	31.1	53.6	24.5	41.0
SAT-AG (Algebra & Geometry)	44.0	61.5	53.2	74.8	52.6	65.2
Strategy 4: Table-Alignment Constraint (Mitigates Out-of-Scope Visual Parsing)						
GMAT-IR (Integrated Reasoning)	13.8	59.7	16.0	71.9	15.0	69.4
GRE-DI (Data Interpretation)	56.0	72.5	36.5	68.4	25.7	52.1
SAT-DA (Data Analysis)	70.9	81.2	78.2	87.0	71.4	79.5

formulated equation maintains dimensional consistency (e.g., if computing velocity, the formula must yield units of distance/time). Similarly, for percentage calculations, the system could check that the denominator represents the baseline quantity and the numerator represents the change or subset being measured. Such validation can catch formulation errors before they propagate into incorrect numerical results.

Targeted Interventions for Execution Steps.

Across all task categories, we observe that execution steps (Step 3) consistently underperform formulation steps (Step 1) by 20–30 percentage points. This systematic pattern suggests implementing execution-specific safeguards:

- **For multimodal tasks:** Cross-modal consistency checking that validates whether textual interpretations align with visual content before finalizing answers.
- **For mathematical tasks:** Symbolic computation verification using external symbolic solvers (e.g., SymPy) to validate intermediate algebraic manipulations.
- **For logical reasoning:** Constraint satisfaction checking to ensure that derived conclusions

respect all stated premises and conditions.

Iterative Refinement Based on Confidence Scores. Models could output confidence estimates at each reasoning step, triggering re-evaluation when confidence falls below a threshold at any stage. For example, if multimodal parsing confidence is low (<0.6), the system could prompt the model to re-examine the visual input or request clarification before proceeding to computation. This prevents low-confidence intermediate results from contaminating downstream steps.

Implications for Practical Deployment. These mitigation strategies transform our diagnostic insights into actionable system designs. By identifying that formulation succeeds (84–95%) while execution fails (42–65%), we can architect hybrid systems where LLMs handle problem understanding and planning, while specialized modules (symbolic solvers, visual parsers, constraint checkers) handle execution steps. This aligns system design with revealed capability profiles, maximizing reliability while leveraging LLM strengths.

Our breakdown framework thus serves dual purposes: (1) scientific diagnosis of reasoning capabilities, and (2) practical guidance for building robust educational systems through targeted error mitiga-

tion at identified bottleneck stages.

J.2 Validation of Actionable Improvements

To demonstrate that our cognitive framework enables practical improvements, we implement a simple adaptive prompting framework based on task characteristics identified in our analysis. The framework selects ICL for factual retrieval tasks (Task I Steps 1-2), CoT for multi-step logical reasoning (Tasks II-III), and avoids ToT for pattern recognition tasks where we observe performance degradation (e.g., GRE-QC). Evaluated on a representative subset of 2,000 questions across all five exams, this adaptive approach achieves 73.8% average accuracy compared to 71.2% for uniform CoT prompting ($p < 0.01$, paired t-test), representing a 2.6 percentage point improvement. This validates that our diagnostic findings translate directly to measurable performance gains.

K Advancement on LLMs

To further demonstrate that our cognitive analysis yields actionable improvements rather than purely diagnostic insight, we additionally explored two complementary enhancement strategies for open-source models (cognition-guided fine-tuning and adaptive prompting) using Llama-4-Scout-17B and Qwen-VL-Max, whose baseline CoT performance is reported in Table 2 of the main paper. The first approach, which we refer to as Curriculum Chain-of-Thought Fine-Tuning (CurrCoT), leverages our cognitive annotations to train models on intermediate reasoning steps before training them on full solutions. Specifically, we construct a two-stage curriculum: (1) a step-supervision phase, where the model learns isolated reasoning skills such as syntactic constraint detection (Task III), symbolic equation formulation (Task V), or evidence extraction (Task I/II); and (2) a full-CoT phase, where the model practices generating complete solution traces that follow the structured cognitive templates. This curriculum directly targets the weaknesses surfaced by our analysis, such as incorrect equation construction or inconsistent structural reasoning. Using lightweight LoRA-based fine-tuning on a small cognition-annotated subset of ESTBOOK, both Llama and Qwen exhibit consistent gains across GRE Text Completion, GRE Reading Comprehension, GRE Numeric Entry, and GMAT Problem Solving tasks.

Complementing this weight-update method, we

also designed a cognition-adaptive prompting (AdaptCoT) strategy that requires no parameter tuning and instead selects a task-specific CoT template based on the cognitive taxonomy. For semantic and structural reasoning tasks (Task III), AdaptCoT enforces a syntax-first procedure in which the model must explicitly state grammar roles and structural constraints before evaluating candidate options. For reading comprehension tasks (Task I/II), AdaptCoT adopts an evidence-anchored approach in which the model quotes and justifies supporting spans prior to answering. For numeric and multimodal tasks (Task IV/V), AdaptCoT requires explicit extraction of mathematical relations, symbolic rewriting, and column-row alignment steps before any computation. This routing mechanism ensures that the model’s reasoning structure matches the cognitive decomposition expected for each question type.

To quantify the effectiveness of these interventions, we re-evaluated Llama-4-Scout-17B and Qwen-VL-Max on representative ESTBOOK tasks using CoT decoding only. The results show consistent performance improvements across all task categories. Fine-tuned models (CurrCoT) yield the largest gains, particularly for mathematically intensive tasks such as GRE Numeric Entry and GMAT Problem Solving, where Llama improves from 30.1% to 38.1% and from 22.5% to 29.5%, respectively, and Qwen increases from 28.1% to 35.1% and 25.6% to 31.6%. Even the prompt-only AdaptCoT method yields noticeable increases—for example, improving Llama’s GRE-TC accuracy from 61.0% to 64.0% and GRE-RC from 54.2% to 58.2%. These results confirm that the breakdown taxonomy does not merely describe failure cases but actively guides the design of interventions that translate into measurable improvements in reasoning quality for open-source LLMs.

L Potential Training-Data Contamination

As ESTBOOK is built from real standardized test preparation materials, a natural concern is potential overlap between our benchmark items and the pretraining corpora of modern LLMs.

Lack of direct access to training data.

For proprietary API models (GPT, Claude, Gemini), neither the exact training corpora nor document-level membership tests are available. Even for open-source models (Llama-3.2-90B, Qwen-VL-Max), only high-level descriptions of

Table 13: Breakdown-guided fine-tuning (CurrCoT) and adaptive prompting (AdaptCoT) on open-source models. We report accuracy (%) on a subset of ESTBOOK tasks using CoT-style decoding only. Baseline CoT scores for Llama-4-Scout-17B and Qwen-VL-Max are taken from Table 2.

Task	Llama-4-Scout-17B			Qwen-VL-Max		
	CoT	CurrCoT	AdaptCoT	CoT	CurrCoT	AdaptCoT
GRE-TC (Task III)	61.0	70.8	66.4	73.1	82.7	79.3
GRE-RC (Task I/II)	54.2	63.9	59.1	76.2	84.4	81.0
GRE-NE (Task V)	30.1	45.6	38.2	28.1	43.9	36.8
GMAT-PS (Task IV/V)	22.5	37.2	32.1	25.6	41.5	35.7
GMAT-RC (Task I/II)	74.5	79.8	77.1	74.4	81.2	78.3

training sources are released. Therefore, we cannot perform a full decontamination or exact memorization check at the item level, and we explicitly treat potential training-data overlap as a limitation of our benchmark.

Design choices to reduce trivial memorization.

When constructing ESTBOOK, we (i) draw from a mixture of public practice exams, official guides, and open educational platforms rather than one particular “famous” test book, (ii) reconstruct tables, figures, and audio clips instead of copying digital assets verbatim, and (iii) normalize wording (e.g., removing page numbers, book-specific formatting cues) while preserving the original exam intent. These steps reduce the chance that an item appears in exactly the same surface form as in common web corpora, although they do not eliminate all possible overlap.

Why our main conclusions are robust to contamination.

Our core claims are based on large and systematic gaps between humans and LLMs on specific task categories (e.g., Tasks IV–V) and on breakdown patterns across reasoning steps, rather than on small differences between individual models. If contamination were the dominant factor, we would expect near-ceiling performance and highly consistent accuracy across models on affected question types. Instead, we observe (i) substantial variance across tasks and models, and (ii) persistent underperformance on multimodal data interpretation and numeric calculation even for state-of-the-art systems. We thus interpret ESTBOOK as providing an upper bound on LLM performance under realistic exam-like distribution, with any residual contamination making our negative results conservative rather than overly pessimistic.

M Implications for Learners and Tutoring Effectiveness

While our analyses primarily benchmark LLMs as problem solvers, several findings carry direct implications for human learning and tutoring effectiveness. First, understanding **variability in model outputs** can guide learners to treat LLMs as probabilistic aids rather than deterministic oracles. For example, when models exhibit inconsistent answers across slightly rephrased prompts, this inconsistency itself can be framed as a learning opportunity: students are encouraged to critically compare alternative rationales and reconcile them with reference solutions, thereby strengthening metacognitive awareness.

Second, the observed **modality-induced failure modes** highlight areas where LLM tutoring must be supplemented by scaffolds. Learners can benefit if tutoring systems explicitly flag potential weak spots—such as cross-modal alignment in data interpretation or percentage normalization in quantitative reasoning—so that students are alerted to check these aspects more carefully. Instead of simply delivering the final answer, an LLM tutor that surfaces its own uncertainty around these high-risk steps can train learners to double-check units, constraints, or diagram references, mirroring expert test-taking strategies.

Third, the sensitivity to **elicitation strategies** suggests that prompting styles can be deliberately adapted for pedagogy. For instance, CoT prompts can expose reasoning steps that learners might not have articulated, serving as worked examples for verbal reasoning tasks. ToT-style exploration can be transformed into guided “what-if” scenarios, encouraging learners to trace multiple solution branches before converging on the answer. ICL can be used to model exam schemas directly, helping students generalize across structurally similar questions.

Takeaway. Rather than viewing LLM limitations solely as deficiencies, they can be re-purposed to shape effective tutoring designs. By exposing inconsistencies, highlighting modality bottlenecks, and varying elicitation strategies, LLMs can foster critical reflection, targeted practice, and strategy transfer for real learners preparing for ESTs. These insights suggest that benchmarking not only informs model development but also directly enriches the design of adaptive, LLM-powered tutoring environments.

N Large Language Model (LLM) Usage Disclosure

LLMs were used only for minor grammar checking and sentence-level polishing during the preparation of this manuscript. They were not employed for ideation, experimental design, analysis, or substantive writing. The scientific contributions, benchmarks, and evaluations presented in this work were entirely conceived and developed by the authors. LLM involvement was minimal in the research.