

000 001 002 003 004 005 ARE LLMS READY FOR ENGLISH STANDARDIZED 006 TESTS? A BENCHMARKING AND ELICITATION STUDY 007 008 009

010 **Anonymous authors**
011
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

Paper under double-blind review

ABSTRACT

Large language models (LLMs) are transforming education by enabling powerful tools that enhance learning experiences, particularly in the context of English Standardized Tests (ESTs), which generate significant commercial value in the education industry. However, their fundamental problem-solving capabilities remain largely underexplored. In this work, we evaluate the performance of LLMs on ESTs across a diverse range of question types. We introduce ESTBOOK, a comprehensive benchmark designed to evaluate the capabilities of LLMs in solving EST questions. ESTBOOK aggregates five widely recognized tests, encompassing 29 question types and over 10,576 questions across multiple modalities, including text, images, audio, tables, and mathematical symbols. Using ESTBOOK, we systematically evaluate both the accuracy and inference efficiency of LLMs. Additionally, we propose a breakdown analysis framework that decomposes complex EST questions into task-specific solution steps. This framework allows us to isolate and assess LLM performance at each stage of the reasoning process. Evaluation findings offer insights into the capability of LLMs in educational contexts and point toward targeted strategies for improving their reliability as intelligent tutoring systems.

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 and grading (Xia et al., 2024; Zhong et al., 2024; Gupta, 2023), test preparation and tutoring (Feng & Wang, 2024; Ashrafimoghari et al., 2024), and even question generation for practice material (Tiratatri et al., 2025).

However, those works have directly concentrated on complex downstream applications. Before LLMs can be reliably deployed for higher-level educational functions such as adaptive tutoring (Stamper et al., 2024; Molina et al., 2024), personalized feedback (Maiti & Goel, 2024; Alsafari et al., 2024), or large-scale exam designs (Zhang et al., 2023; Askarbekuly & Aničić, 2024), it is essential to first establish their fundamental capability in raw problem solving. The ability to answer EST questions correctly is the foundation upon which the higher-level applications can subsequently be built. Yet, the reliability of LLMs in solving ESTs remains largely unexamined, particularly across the diverse formats that such tests encompass (e.g., reading comprehension, essay writing, and mathematical reasoning), which are often presented with multimodal structures (Grapin & Llosa, 2022).

In this work, we benchmark the problem-solving capabilities of LLMs with a broad focus on five internationally recognized ESTs: (1) two language proficiency assessments—TOEFL and IELTS, and (2) three standardized knowledge-based exams—SAT, GRE, and GMAT. To systematically evaluate LLMs, we introduce ESTBOOK, a comprehensive benchmark designed to assess their performance across a wide range of EST tasks. ESTBOOK includes 29 question types drawn from the five exams, totaling 10,576 examples. As illustrated in Figure 1, ESTBOOK spans multiple modalities, including text, images, audio, tables, and mathematical symbols, enabling a rigorous and multimodal evaluation of LLMs’ problem-solving abilities.

Using ESTBOOK, we first evaluate industry-leading LLMs (e.g., GPT-5, Gemini, Llama, and Claude) with foundational prompting strategies: In-Context Learning (ICL), Chain-of-Thought (CoT), and Tree-of-Thought (ToT). Our evaluation yields the following observations:

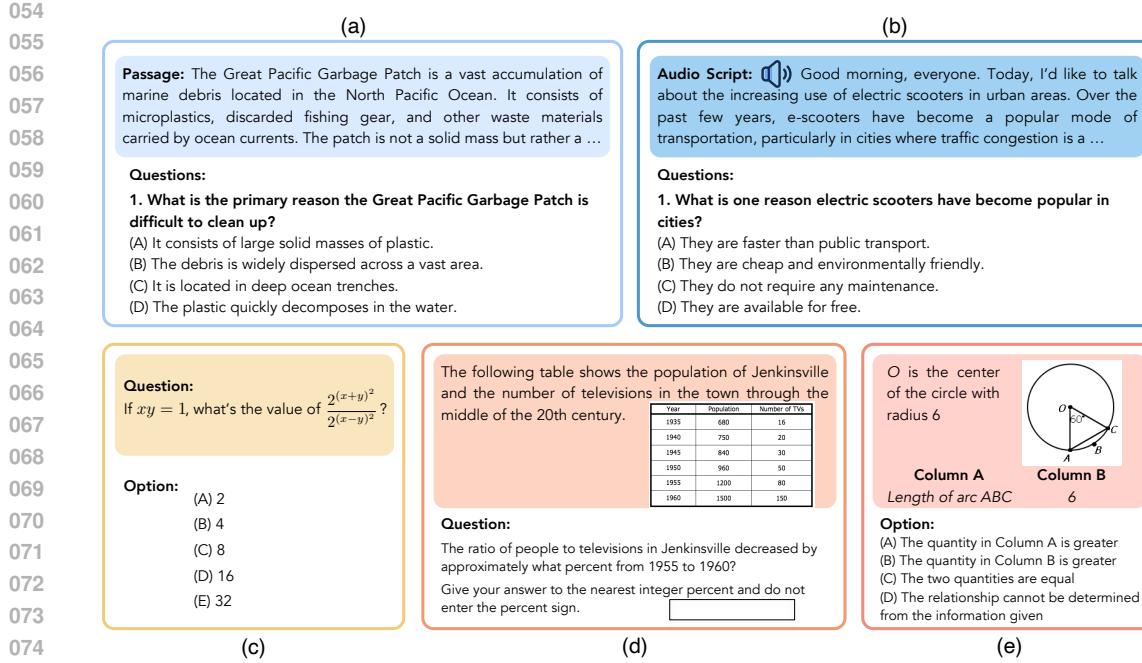


Figure 1: Examples of multimodal questions included in ESTBOOK: (a) a reading comprehension question (text) from IELTS, (b) a listening comprehension question (audio) from TOEFL, and (c), (d), and (e) GRE quantitative questions involving math symbols, tabular data, and images, respectively.

First, we find that LLMs, despite extensive pretraining on large English corpora (Mahmud et al., 2025; Sapkota et al., 2024; Welleck et al., 2024), exhibit limited effectiveness in EST-style problem-solving. In addition, performance varies substantially across question types and domains. Furthermore, in some cases, models may incur variant inference latency without producing correct answers. These results suggest that, although LLMs demonstrate strong general language capabilities, they remain inadequate as educational assistants directly for EST tasks.

To inform future development, we propose a **Breakdown Analysis**, a diagnostic framework tailored to each question type and aligned with how human test-takers approach problem solving. For example, in TOEFL reading comprehension, we first assess whether the model can identify the intention of a question; we then inform LLMs with correct intention and evaluate its ability to extract relevant evidence from long passages. In GRE quantitative questions, we analyze the model’s ability to select appropriate mathematical operations before evaluating execution accuracy. This breakdown strategy helps isolate specific reasoning process and highlight strengths of LLMs toward more effective development of intelligent educational systems.

To summarize, this work makes several contributions: **(1) Benchmark** – We introduce ESTBOOK, a benchmark that offers a diverse set of EST tasks to enable comprehensive, multimodal evaluation of LLMs. **(2) Empirical Study** – We conduct extensive studies across LLMs and reveal their insufficiency for EST problem-solving and exhibit inconsistent performance across questions. **(3) In-Depth Analysis** – We propose breakdown analysis to identify LLM capability in each isolated reasoning step, which provides actionable insights to inform reliable development of educational systems. Dataset and code are available at: <https://anonymous.4open.science/r/Education-9595>.

2 RELATED WORK

LLMs for Education. LLMs are increasingly used in education for tasks like grading (Chang et al., 2024; Holmes & Tuomi, 2022), question generation (Mohebbi, 2025; Zhang et al., 2024a), and tutoring (Schmucker et al., 2024). Early systems like Codex and ChatGPT showed that LLMs can

108 help students across STEM and language learning by offering contextual feedback and answers
 109 (Chiang et al., 2024; Liang et al., 2024; Wen et al., 2024b). More recent research explores using
 110 LLMs as conversational tutors (Sabri et al., 2025). However, most existing evaluations rely on limited
 111 benchmarks or informal studies, often focusing on narrow skills like math or reading (Guilherme,
 112 2019; Lee et al., 2023).

113 **Benchmarking LLMs.** Many benchmarks test LLMs on general or domain-specific reasoning
 114 tasks, evaluating their factual knowledge and reasoning abilities. Some, like MathVista (Lu et al.,
 115 2023; Peng et al., 2024) and ScienceQA (Wen et al., 2024a; Zhang et al., 2024d), include images
 116 or structured data, but often use synthetic problems or cover narrow domains. Our benchmark,
 117 ESTBOOK, is grounded in real standardized exams and spans multiple formats (e.g., multi-choice,
 118 text completion), providing a more realistic test of LLMs as educational agents in a heterogeneous
 119 problem-solving environment.

120 **Eliciting LLM Reasoning.** Improving LLM reasoning through prompting has become an emerging
 121 subject. Techniques like In-Context Learning (ICL) (Koike et al., 2024; Yugeswardeeno et al., 2024),
 122 Chain-of-Thought (CoT) (Godwin-Jones, 2024; Wang et al., 2024), and Tree-of-Thought (ToT)
 123 (Zhang et al., 2024b;c) guide models to generate step-by-step answers and improve performance
 124 on tasks like math and logic. Yet, these methods are mostly tested on clean, synthetic datasets
 125 (Askarbekuly & Aničić, 2024; Schmidhuber & Kruschwitz, 2024). We evaluate LLMs on ESTBOOK
 126 and show their limitations in real EST questions. Additionally, we break down problem-solving steps
 127 based on question structure and offer insights about LLMs eligibility on each isolated reasoning step.

128

129 3 ESTBOOK: BENCHMARKING ENGLISH STANDARDIZED TESTS

130

131 3.1 ENGLISH STANDARDIZED TESTS, INVOLVED MODALITIES, AND DATA SOURCES

132

133 **English Tests.** As shown in Table 1, the benchmark covers 10,576 questions and 29 types across five
 134 major ESTs: *SAT*, *GRE*, *GMAT*, *TOEFL*, and *IELTS*. These exams play critical roles in academic and
 135 professional escalation: (1) *SAT* is widely used for undergraduate admissions in the United States,
 136 assessing students’ readiness for college through verbal and mathematical reasoning. (2) *GRE* is a
 137 common requirement for graduate school admissions, designed to evaluate verbal and quantitative
 138 reasoning skills. (3) *GMAT* serves as a gatekeeping exam for business school programs, emphasizing
 139 critical thinking, data interpretation, and logical reasoning. (4) *TOEFL* and *IELTS* are the two most
 140 widely recognized tests for evaluating English language proficiency among non-native speakers,
 141 commonly required for university admissions and immigration purposes in English-speaking countries.
 142 Among those tests, ESTBOOK focuses on **objective questions**, as they have certain answers and thus
 143 facilitate evaluations.

144

145 **Modalities.** ESTBOOK captures the structural and cognitive diversity among several modalities: text
 146 (T), math symbols (S), images (I), tables (Tb), and audio (A). These modalities reflect the multimodal
 147 nature of real-world ESTs, where students are required not only to process textual information but
 148 also to interpret mathematical expressions and visual data. For example, GRE and GMAT quantitative
 149 sections often combine symbolic reasoning with tabular and graphical inputs, while TOEFL and
 150 IELTS listening sections assess a learner’s ability to extract key information from spoken passages.
 151 With this wide range of input formats, ESTBOOK evaluates LLMs’ problem-solving capabilities in
 152 heterogeneous environment, which offers insights into how different modalities affect reasoning.

153

154 **Sources.** The data in ESTBOOK are sourced from publicly available educational materials and official
 155 preparation resources affiliated with each standardized test. Specifically, we collect questions from
 156 released practice exams (Appelrouth & Zabrusky, 2017; IELTS-up, 2023; Woldoff & Kraynak, 2015),
 157 official preparation guides (Graduate Management Admission Council (GMAC), 2025; Gruber, 2011;
 158 TOEFL Test Prep, 2023; Woldoff, 2024; College Board, 2022; Hatch et al., 2023), and open-access
 159 educational platforms (Josué et al., 2023; Pereira et al., 2024; SAT Questions, 2023; Mallik, 2025;
 160 GMAT Club, 2025) that align with the formats and content of SAT, GRE, GMAT, TOEFL, and IELTS.
 161 To ensure data diversity and authenticity, we include samples spanning different years, question
 162 formats, and difficulty levels. For multimodal questions, such as those involving tables, images, or
 163 audio clips, we reconstruct representative content that mirrors real test conditions, ensuring fidelity
 164 to the original test design while maintaining licensing compliance. All questions went through
 165 validation on their sourced websites for correctness, clarity, and alignment with the original intent of

162
 163 Table 1: question types, their descriptions, number of instances, involved modalities, and involved
 164 tasks (defined in Section 3.2). Modality: "T"–text, "S"–math symbol, "I"–Image, "Tb"–tabular data,
 165 "A"–audio. Concrete examples are shown in Appendix B.

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

193
 194 the corresponding exam section. We provide additional details regarding question quality control and
 195 copyright availability in Appendix A.

197 3.2 A TAXONOMY OF TASKS

198 To facilitate structured problem-solving with LLMs, we categorize each EST question type by
 199 aligning it with real-world cognitive-computational strategies commonly used in test preparation. As
 200 shown in Table 1, each question type is mapped to a specific task, which corresponds to a breakdown
 201 solution, i.e., a step-by-step reasoning path grounded in standardized test-solving strategies. We
 202 identify six distinct task categories as follows:

203

- 204 • **Task I: Evidence Finding** (*Breakdown: Identify Subject → Comprehend Text/Audio → Extract Discourse*) – This task involves identifying the central subject of the question, locating relevant
 205 textual or auditory evidence, and applying reasoning to extract the correct answer. It is common in
 206 reading comprehension sections of tests like GRE and TOEFL.
- 207 • **Task II: Semantic Reasoning** (*Breakdown: Parse Semantics → Localize Logical Scope → Resolve Contextual Meaning*) – This task requires interpreting fine-grained sentence semantics and
 208 resolving logical relationships or equivalence. Examples include GRE Sentence Equivalence and
 209 GMAT Critical Reasoning questions.
- 210 • **Task III: Structural Reasoning** (*Breakdown: Parse Syntactic Structure → Match Text → Predict Missing Element*) – Tasks such as sentence completion (GRE) or grammatical error detection
 211 (GMAT) fall into this category, where models must first analyze syntax and then select appropriate
 212 tokens to complete or correct the sentence.

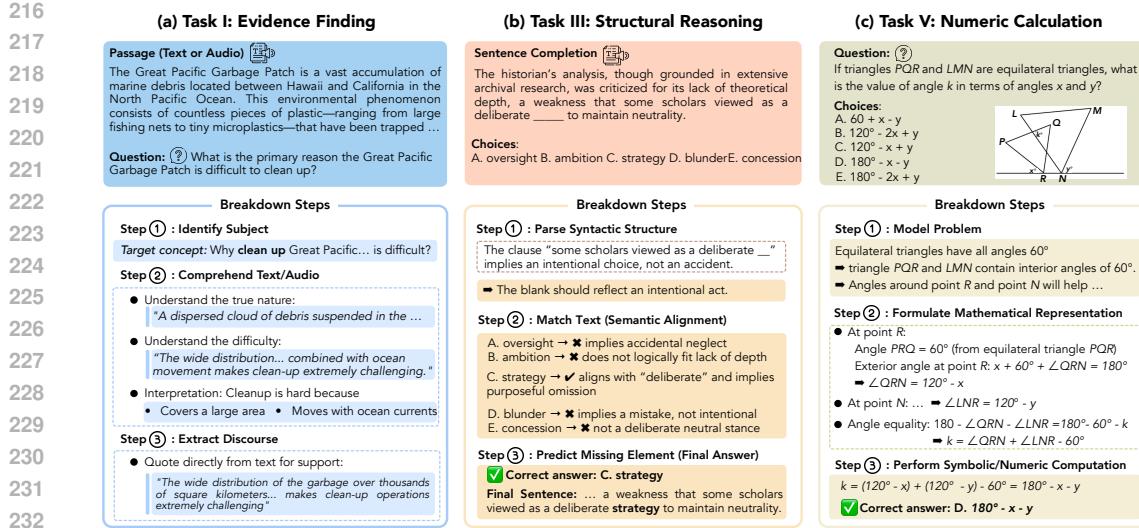


Figure 2: Illustrative breakdown examples for solving EST questions.

- Task IV: Data Interpretation** (*Breakdown: Formulate Analytical Goal → Parse Visual/Tabular Input → Analyze Data*) – Multimodal questions involving tables, charts, or diagrams require the model to interpret visual structures, extract relevant data, and perform computations. This applies to GRE Data Interpretation and GMAT Integrated Reasoning.
- Task V: Numeric Calculation** (*Breakdown: Model Problem → Formulate Mathematical Representation → Perform Symbolic/Numeric Computation*) – This task is typical of math-focused questions that require translating natural language descriptions into formal mathematical expressions, followed by symbolic manipulation or numeric computation. Examples include SAT Math and GMAT Problem Solving sections.
- Task VI: Comparative Judgment** (*Breakdown: Identify Comparative Entities → Apply Constraints → Evaluate Logical Relationship*) – This task evaluates whether the model can assess sufficiency, equivalence, or constraint satisfaction, as seen in GRE Quantitative Comparison and GMAT Data Sufficiency questions.

In Appendix C, we provide details on how the breakdown steps align with the problem-solving processes employed by human test-takers across different categories of EST tasks.

4 EXPERIMENT

ESTBOOK aims to empirically answer several research questions: **RQ₁**: How do different LLMs perform on EST problem-solving tasks under various prompting strategies? **RQ₂**: What is the inference-time efficiency of LLMs across different EST question types? **RQ₃**: How effective are LLMs at completing individual steps within structured problem-solving workflows?

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 & Roll, 2024) to transcribe audio data within listening tasks in TOEFL and IELTS.

Human Tester. To demonstrate LLM performance alongside humans, we employed five student testers, wherein each of whom had recently prepared for and participated in at least one of these tests, and reported their problem-solving performance along with LLMs.

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.,

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	55.2	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	60.4	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	85.5	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	44.3	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

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 the model to explore multiple reasoning paths and select the most plausible one. Prompt layouts are shown in Appendix D. Metrics are detailed in Appendix E.2.

Statistical significance tests (McNemar's test for paired proportions, see Table 5 in Appendix E) confirm that performance differences between humans and the best-performing LLM are statistically significant ($p < 0.05$) for 26 out of 29 question types, with particularly large and significant gaps in numeric calculation tasks (SAT-AG, SAT-DA, GRE-NE: all $p < 0.001$) and multimodal data interpretation tasks (GMAT-IR, GRE-DI: $p < 0.01$).

4.1 EVALUATING LLMs PERFORMANCE ON ESTBOOK (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 are 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 details and insights in Appendix E.4.1.

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

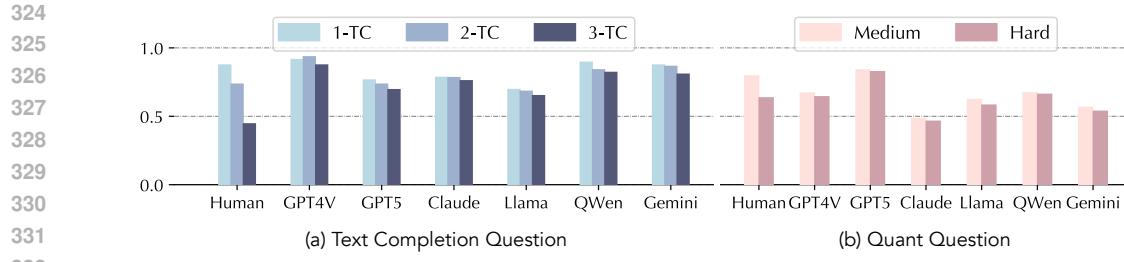


Figure 3: LLM performance across varying levels of question difficulty, using CoT due to its representativeness. We focus on GRE text completion tasks with 1-, 2-, and 3-blanks, as well as available medium- and hard-level quantitative problems.

Case Study I. (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.

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 E.5.

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 II (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 second 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. Additional insights are provided in Appendix E.6.

Note on Difficulty Categorization: Difficulty levels in Figure 3 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.

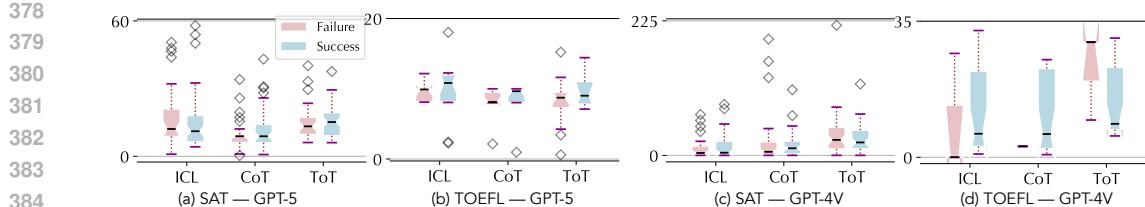


Figure 4: Inference time (in seconds) for failed and successful cases. More results are in Figure 6.

Influence of Question Difficulty. 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 3 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.

Validation of Actionable Improvements. To demonstrate that our breakdown analysis 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.

Appendix G further discusses how our findings can benefit real learners and the education industry.

Appendix J provides extended evaluation on three additional mainstream models, demonstrating the consistency and robustness of our findings across a broader range of LLM systems.

4.2 INFERENCE EFFICIENCY (RQ₂)

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 4.

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 & 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 3, 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

Table 3: 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 7.

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

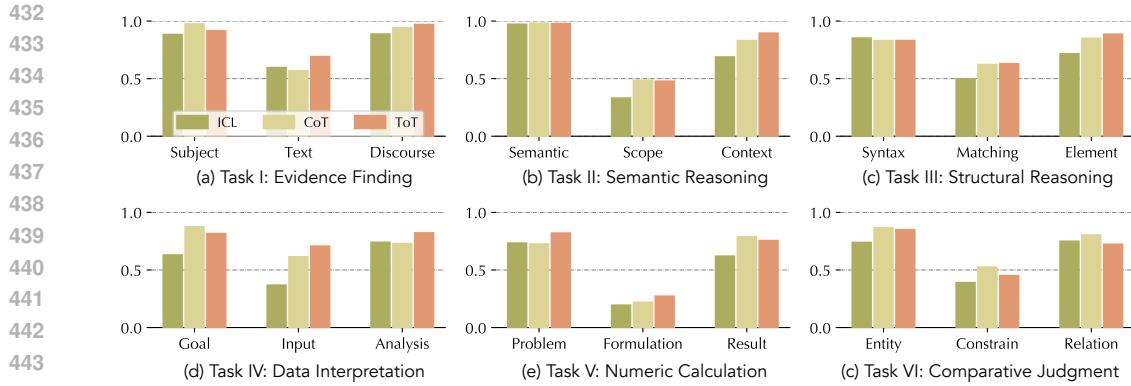


Figure 5: Breakdown analysis across all included tasks I-VI (Section 3.2) on GPT-4V.

LLM spends on answering a question does not correlate with answer correctness. Inference time appears largely independent of answer quality on the EST benchmark.

4.3 BREAKDOWN ANALYSIS: ISOLATED MEASUREMENT OF EACH REASONING STEP (RQ₃)

Next, we evaluate the step-by-step capabilities of LLMs in solving diverse EST tasks. As defined in Section 3.2, these tasks span six categories, each associated with a structured sequence of reasoning steps necessary for successful problem-solving. In our experiments, we assume the ground-truth outputs for preceding steps are provided, thereby isolating each reasoning stage and avoiding the compounding of upstream errors. Figures 5 and Appendix E present the breakdown analysis results, with detailed evaluation metrics in Appendix E.2. We derive the following observations:

LLMs Are Strong Formulators but Weak Reasoners. Overall, we observe that LLMs consistently excel in the initial step across all tasks, achieving up to 97% accuracy. These early steps—such as task identification, problem formulation, or topic modeling—demonstrate the models’ strong capability to interpret and structure EST problems appropriately. 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 III (GMAT – Critical Reasoning): *"A recent study found that 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 to reduce its pollution levels."*

LLM Reasoning: GPT-5 and Claude incorrectly state *"The 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 while LLMs are proficient at understanding and framing problems, they remain limited and unstable in executing complex reasoning chains—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.

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.

486
487
488

Case Study IV (GRE Quant – Numeric Entry): *"If the sum of three consecutive odd integers is 111, what is the smallest of the three?"*

489
490

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 $x+x+2+x+4 = 111$).

491
492
493

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.

494
495
496
497

Case Study V (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.

498
499
500

Due to space constraints, additional findings and case studies are provided in Appendix E.3 and E.7.

501
502

5 CONCLUSION

503
504
505
506
507
508
509

This work explores the potential and limitations of LLMs in problem-solving on English Standardized Tests (ESTs). Through the construction of ESTBOOK, a multimodal and diverse benchmark encompassing five major ESTs, we provide a rigorous framework for evaluating LLMs across a variety of question types and modalities. Our empirical findings reveal that, despite their linguistic fluency, current LLMs fall short in consistently solving EST-style problems and display notable variation in performance across domains. Furthermore, our proposed breakdown analysis highlights specific reasoning failures, offering a granular diagnostic approach to inform model development.

510
511
512
513
514
515
516
517
518
519

Implications for Future Improvements Our breakdown analysis directly informs targeted interventions. The visual-tabular parsing bottleneck at Task IV Step 2 (dropping from 87% to 51–65%) suggests specialized training objectives such as contrastive learning for cross-modal alignment. The mathematical formulation gap at Task V Step 2 (21-point drop) motivates symbolic execution modules to verify expression correctness before computation. The formulation-execution gap (24–33 point drops from Step 1 to Step 3) points toward hierarchical reasoning architectures separating planning from execution. Our finding that prompting effectiveness varies with task structure suggests adaptive frameworks that select ICL, CoT, or ToT based on reasoning depth and modality complexity. Future work can use our breakdown framework as an evaluation protocol to verify interventions address root causes rather than achieving spurious aggregate improvements.

520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539

540
541 ETHICS STATEMENT542 This work exclusively relies on publicly available standardized test preparation resources, official
543 sample questions, and open educational platforms, all used in compliance with their respective
544 copyright and licensing terms. No proprietary or sensitive data were accessed. Additional details
545 regarding data sourcing and copyright considerations are included in Appendix A. Therefore, we do
546 not identify any ethical concerns arising from this study.547
548 REPRODUCIBILITY STATEMENT549 To support reproducibility, we have released all benchmark construction details, evaluation scripts,
550 and experimental configurations through an anonymous GitHub repository. This repository includes
551 instructions for evaluation procedures. The benchmark design, codebase, and detailed experimental
552 settings are documented to readers.
553

554 REFERENCES

555 Bashaer Alsafari, Eric Atwell, Aisha Walker, and Martin Callaghan. Towards effective teaching
556 assistants: From intent-based chatbots to llm-powered teaching assistants. *Natural Language
557 Processing Journal*, 8:100101, 2024.558 Allison Andreyev. Quantization for openai’s whisper models: A comparative analysis. *arXiv preprint
arXiv:2503.09905*, 2025.559 Jed I Appelrouth and Karen M Zabrusky. Preparing for the sat: A review. *College and University*, 92
560 (1):2, 2017.561 Vahid Ashrafimoghari, Necdet Gürkan, and Jordan W Suchow. Evaluating large language models
562 on the gmat: Implications for the future of business education. *arXiv preprint arXiv:2401.02985*,
563 2024.564 Nursultan Askarbekuly and Nenad Aničić. Llm examiner: automating assessment in informal
565 self-directed e-learning using chatgpt. *Knowledge and Information Systems*, 66(10):6133–6150,
566 2024.567 Zhenni Bi, Kai Han, Chuanjian Liu, Yehui Tang, and Yunhe Wang. Forest-of-thought: Scaling test-
568 time compute for enhancing llm reasoning, 2025. URL <https://arxiv.org/abs/2412.09078>.569 College Board. The cognitively complex thinking required by select sat suite questions: Evi-
570 dence from students with specific learning disorders affecting reading (dyslexia). Technical
571 report, College Board, 2025. URL <https://satsuite.collegeboard.org/media/pdf/digital-sat-cognitive-lab-report-sldr.pdf>. Cognitive Lab Report.572 Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan
573 Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM
574 transactions on intelligent systems and technology*, 15(3):1–45, 2024.575 Cheng-Han Chiang, Wei-Chih Chen, Chun-Yi Kuan, Chienchou Yang, and Hung-yi Lee. Large
576 language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student
577 course. *arXiv preprint arXiv:2407.05216*, 2024.578 College Board. Digital sat sample questions and answer explanations. <https://satsuite.collegeboard.org/media/pdf/digital-sat-sample-questions.pdf>, 2022.579 Yang Feng and Xiya Wang. Exploring the development of chinese college students’ proficiency in
580 english through chatgpt: An experimental study. In *Proceedings of the 2024 the 16th International
581 Conference on Education Technology and Computers*, pp. 148–154, 2024.582 GMAT Club. Gmat club: Best gmat prep, tests, mba admissions and courses, 2025. URL <https://gmatclub.com>.

594 Robert Godwin-Jones. Distributed agency in second language learning and teaching through generative ai. *arXiv preprint arXiv:2403.20216*, 2024.

595

596

597 Graduate Management Admission Council (GMAC). *GMAT Official Guide 2025–2026 Bundle: Books + Online Question Bank*. Wiley, 1st edition, 2025. ISBN 9781394333936. URL <https://www.amazon.com/GMAT-Official-Guide-2025-Question/dp/1394333935>.

598

599

600 Calbert Graham and Nathan Roll. Evaluating openai’s whisper asr: Performance analysis across

601 diverse accents and speaker traits. *JASA Express Letters*, 4(2), 2024.

602

603 Scott E Grapin and Lorena Llosa. Multimodal tasks to assess english learners and their peers in

604 science. *Educational assessment*, 27(1):46–70, 2022.

605

606 Gary R Gruber. *Gruber’s Complete GRE Guide 2012*. Sourcebooks, Inc., 2011.

607

608 Alex Guilherme. Ai and education: the importance of teacher and student relations. *AI & society*, 34:

609 47–54, 2019.

610

611 Pranav Gupta. Testing llm performance on the physics gre: some observations. *arXiv preprint arXiv:2312.04613*, 2023.

612

613 Lisa Zimmer Hatch, Scott A. Hatch, and Sandra Luna McCune. *GMAT Prep 2024/2025 For Dummies (GMAT Focus Edition): Book + 3 Practice Tests + 100 Flashcards Online*. For Dummies, 11th edition, 2023. ISBN 9781394183364. URL <https://www.amazon.com/GMAT-Prep-Dummies-Online-Practice/dp/1394183364>.

614

615

616 Wayne Holmes and Ilkka Tuomi. State of the art and practice in ai in education. *European journal of education*, 57(4):542–570, 2022.

617

618 IELTS-up. Ielts listening practice tests. <https://ielts-up.com/listeningIELTS-listening-practice.html>, 2023.

619

620

621 Christopher J. Johnstone, Nicole A. Bottsford-Miller, and Sandra J. Thompson. Using the think aloud

622 method (cognitive labs) to evaluate test design for students with disabilities and english language

623 learners. Technical Report 44, University of Minnesota, National Center on Educational Outcomes,

624 2006. URL <http://files.eric.ed.gov/fulltext/ED495909.pdf>.

625

626 Anghelo Josué, Mirna Carolina Bedoya-Flores, Erick Fabián Mosquera-Quiñonez, Ángel Enrique

627 Mesías-Simisterra, and José Vicencio Bautista-Sánchez. Educational platforms: Digital tools

628 for the teaching-learning process in education. *Ibero-American Journal of Education & Society Research*, 3(1):259–263, 2023.

629

630 Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. Outfox: Llm-generated essay detection

631 through in-context learning with adversarially generated examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 21258–21266, 2024.

632

633 Unggi Lee, Sanghyeok Lee, Junbo Koh, Yeil Jeong, Haewon Jung, Gyuri Byun, Yunseo Lee,

634 J ewoong Moon, Jieun Lim, and Hyeyoncheol Kim. Generative agent for teacher training: Designing

635 educational problem-solving simulations with large language model-based agents for pre-service

636 teachers. *NeurIPS’23 Workshop on Generative AI for Education (GAIED)*, 2023.

637

638 Zhenwen Liang, Ye Liu, Tong Niu, Xiangliang Zhang, Yingbo Zhou, and Semih Yavuz. Improving

639 llm reasoning through scaling inference computation with collaborative verification. *arXiv preprint arXiv:2410.05318*, 2024.

640

641 Eric Loken, Filip Radlinski, Vincent H. Crespi, Josh Millet, and Lesleigh Cushing. Online study

642 behavior of 100,000 students preparing for the sat, act, and gre. *Journal of Educational Computing Research*, 30(3):255–262, 2004. doi: 10.2190/AA0M-0CK5-2LCM-B91N.

643

644 Jieyi Long. Large language model guided tree-of-thought. *arXiv preprint arXiv:2305.08291*, 2023.

645

646 Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng,

647 Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning

of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023.

648 Joy Mahapatra and Utpal Garain. Impact of model size on fine-tuned llm performance in data-to-text
 649 generation: A state-of-the-art investigation. *arXiv preprint arXiv:2407.14088*, 2024.
 650

651 Doaa Mahmud, Hadeel Hajmohamed, Shamma Almentheri, Shamma Alqaydi, Lameya Aldhaheri,
 652 Ruhul Amin Khalil, and Nasir Saeed. Integrating llms with its: Recent advances, potentials,
 653 challenges, and future directions. *IEEE Transactions on Intelligent Transportation Systems*, 2025.

654 Pratyusha Maiti and Ashok K Goel. How do students interact with an llm-powered virtual teaching
 655 assistant in different educational settings? *arXiv preprint arXiv:2407.17429*, 2024.
 656

657 Abdullah Mallik. Sat suite question bank categories. <https://sat-questions.onrender.com/categories>, 2025.
 658

659 Patrick E McKnight and Julius Najab. Mann-whitney u test. *The Corsini encyclopedia of psychology*,
 660 pp. 1–1, 2010.
 661

662 Ahmadreza Mohebbi. Enabling learner independence and self-regulation in language education using
 663 ai tools: a systematic review. *Cogent Education*, 12(1):2433814, 2025.

664 Ismael Villegas Molina, Audria Montalvo, Benjamin Ochoa, Paul Denny, and Leo Porter. Leveraging
 665 llm tutoring systems for non-native english speakers in introductory cs courses. *arXiv preprint*
 666 *arXiv:2411.02725*, 2024.

667 Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. Multimath: Bridging
 668 visual and mathematical reasoning for large language models. *arXiv preprint arXiv:2409.00147*,
 669 2024.
 670

671 Juanan Pereira, Juan-Miguel López, Xabier Garmendia, and Maider Azanza. Leveraging open source
 672 llms for software engineering education and training. In *2024 36th International Conference on*
 673 *Software Engineering Education and Training (CSEE&T)*, pp. 1–10. IEEE, 2024.

674 Hamoun Sabri, Muhammad HA Saleh, Parham Hazrati, Keith Merchant, Jonathan Misch, Purnima S
 675 Kumar, Hom-Lay Wang, and Shayan Barootchi. Performance of three artificial intelligence (ai)-
 676 based large language models in standardized testing; implications for ai-assisted dental education.
 677 *Journal of periodontal research*, 60(2):121–133, 2025.
 678

679 Ranjan Sapkota, Rizwan Qureshi, Syed Zohaib Hassan, John Shutske, Maged Shoman, Muhammad
 680 Sajjad, Fayaz Ali Dharejo, Achyut Paudel, Jiajia Li, Zhichao Meng, et al. Multi-modal llms in
 681 agriculture: A comprehensive review. *Authorea Preprints*, 2024.

682 SAT Questions. Sat practice questions and categories. <https://sat-questions.onrender.com/categories>, 2023.
 683

684 Maximilian Schmidhuber and Udo Kruschwitz. Llm-based synthetic datasets: Applications and
 685 limitations in toxicity detection. *LREC-COLING*, 37:2024, 2024.
 686

687 Robin Schmucker, Meng Xia, Amos Azaria, and Tom Mitchell. Ruffle&riley: Insights from designing
 688 and evaluating a large language model-based conversational tutoring system. In *International*
 689 *Conference on Artificial Intelligence in Education*, pp. 75–90. Springer, 2024.

690 John Stamper, Ruiwei Xiao, and Xinying Hou. Enhancing llm-based feedback: Insights from
 691 intelligent tutoring systems and the learning sciences. In *International Conference on Artificial*
 692 *Intelligence in Education*, pp. 32–43. Springer, 2024.
 693

694 Thanayut Tiratatri, Kanin Sukittivarapunt, Thammathat Sarasinpitak, and Aung Pyae. Designing an
 695 llm-based ielts question generator, assessment, and personalized training system: Architecture and
 696 research agenda. In *2025 22nd International Conference on Electrical Engineering/Electronics,*
 697 *Computer, Telecommunications and Information Technology (ECTI-CON)*, pp. 1–6. IEEE, 2025.

698 TOEFL Test Prep. Toefl test preparation website. <https://toefltestprep.com/>, 2023.
 699

700 Zefeng Wang, Zhen Han, Shuo Chen, Fan Xue, Zifeng Ding, Xun Xiao, Volker Tresp, Philip Torr,
 701 and Jindong Gu. Stop reasoning! when multimodal llm with chain-of-thought reasoning meets
 adversarial image. *arXiv preprint arXiv:2402.14899*, 2024.

702 Sean Welleck, Amanda Bertsch, Matthew Finlayson, Hailey Schoelkopf, Alex Xie, Graham Neubig,
 703 Ilia Kulikov, and Zaid Harchaoui. From decoding to meta-generation: Inference-time algorithms
 704 for large language models. *arXiv preprint arXiv:2406.16838*, 2024.

705 Bingbing Wen, Bill Howe, and Lucy Lu Wang. Characterizing llm abstention behavior in science qa
 706 with context perturbations. *arXiv preprint arXiv:2404.12452*, 2024a.

708 Qingsong Wen, Jing Liang, Carles Sierra, Rose Luckin, Richard Tong, Zitao Liu, Peng Cui, and
 709 Jiliang Tang. Ai for education (ai4edu): Advancing personalized education with llm and adaptive
 710 learning. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and*
 711 *Data Mining*, pp. 6743–6744, 2024b.

712 Ron Woldoff. *GRE 5-Hour Quick Prep For Dummies*. John Wiley & Sons, 2024.

714 Ron Woldoff and Joseph Kraynak. *GRE For Dummies: With Online Practice Tests*. John Wiley &
 715 Sons, 2015.

716 Wei Xia, Shaoguang Mao, and Chanjing Zheng. Empirical study of large language models as
 717 automated essay scoring tools in english composition_taking toefl independent writing task for
 718 example. *arXiv preprint arXiv:2401.03401*, 2024.

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

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

725 Di Zhang, Jianbo Wu, Jingdi Lei, Tong Che, Jiatong Li, Tong Xie, Xiaoshui Huang, Shufei Zhang,
 726 Marco Pavone, Yuqiang Li, et al. Llama-berry: Pairwise optimization for o1-like olympiad-level
 727 mathematical reasoning. *arXiv preprint arXiv:2410.02884*, 2024a.

729 Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. M3exam: A
 730 multilingual, multimodal, multilevel benchmark for examining large language models. *Advances*
 731 *in Neural Information Processing Systems*, 36:5484–5505, 2023.

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

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

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

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

745 A QUESTION QUALITY AND COPYRIGHT ASSURANCE

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

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

756 **Alignment Validation Process.** In addition to verifying correctness and clarity, we systematically
 757 validated the alignment of our questions with official exam specifications. For each question type, we
 758 performed the following validation steps:
 759

- 760 1. **Format validation:** We cross-referenced the question structure, answer format, and pre-
 761 sentation style against official sample materials to ensure consistency. As an example, SAT
 762 Math Algebra questions in our benchmark follow the same five-option multiple-choice for-
 763 mat and use identical equation presentation conventions as those in College Board released
 764 items.
- 765 2. **Skill mapping:** We verified that each question assesses the specific cognitive skill as defined
 766 in official test frameworks. For instance, TOEFL Reading “Factual Information” questions
 767 in ESTBOOK are designed to test the ability to identify explicitly stated details, which
 768 matches the skill definition provided by ETS. Similarly, “Inference” questions require test-
 769 takers to draw logical conclusions that go beyond the literal text, consistent with official
 770 specifications.
- 771 3. **Difficulty calibration:** Although we do not have access to proprietary difficulty ratings
 772 used by test administrators, we ensured that our question collection spans the full difficulty
 773 range found in official preparation materials. This includes both entry-level items suitable
 774 for beginners and advanced problems that challenge high-performing test-takers, providing
 775 representative coverage across each exam’s difficulty spectrum.

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

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

787 **Potential Training-Data Contamination**

788 As ESTBOOK is built from real standardized test preparation materials, a natural concern is potential
 789 overlap between our benchmark items and the pretraining corpora of modern LLMs.
 790

791 **Lack of direct access to training data.**

792 For proprietary API models (GPT-4V, GPT-4o, Claude-3-Opus, Gemini-Pro), neither the exact
 793 training corpora nor document-level membership tests are available. Even for open-source models
 794 (Llama-3.2-90B, Qwen-VL-Max), only high-level descriptions of training sources are released.
 795 Therefore, we cannot perform a full decontamination or exact memorization check at the item level,
 796 and we explicitly treat potential training-data overlap as a limitation of our benchmark.
 797

798 **Design choices to reduce trivial memorization.**

799 When constructing ESTBOOK, we (i) draw from a mixture of public practice exams, official guides,
 800 and open educational platforms rather than one particular “famous” test book, (ii) reconstruct tables,
 801 figures, and audio clips instead of copying digital assets verbatim, and (iii) manually normalize
 802 wording (e.g., removing page numbers, book-specific formatting cues) while preserving the original
 803 exam intent. These steps reduce the chance that an item appears in exactly the same surface form as
 804 in common web corpora, although they do not eliminate all possible overlap.
 805

806 **Why our main conclusions are robust to contamination.**

807 Our core claims are based on large and systematic gaps between humans and LLMs on specific
 808 task categories (e.g., Tasks IV–V) and on breakdown patterns across reasoning steps, rather than
 809 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.

B 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.

864 Sentence: Each of the students (have / has) submitted their essay on time, and the teacher (is / are)
 865 pleased with the quality of work...
 866 A. have, is
 867 B. has, is
 868 C. have, are
 869 D. has, are

870 **SAT Math Algebra (AG).** Tests ability to work with algebraic expressions, equations, inequalities,
 871 and functions. Requires solving linear/quadratic equations, manipulating expressions, understanding
 872 variable relationships, and analyzing functions. Students must apply algebraic concepts to model
 873 real-world situations and connect symbolic representations with graphs. Assesses both procedural
 874 fluency and conceptual understanding.
 875

876 Solve for x : $\frac{2x-3}{x+1} = 4$
 877 A. $x = 7$
 878 B. $x = -7$
 879 C. $x = 7/3$
 880 D. $x = -7/3$
 881 E. No solution exists

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

887 A circle graph shows that 30% of students prefer tea, 50% coffee, and 20% water. If 200 students
 888 were surveyed, how many prefer coffee?
 889 A. 60
 890 B. 100
 891 C. 40
 892 D. 50
 893 E. Cannot be determined from the information given
 894

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

900 In right triangle ABC with right angle at C , if $AC = 3$ and $BC = 4$, what is $\sin A$?
 901 A. $3/5$
 902 B. $4/5$
 903 C. $3/4$
 904 D. $4/3$
 905 E. $5/3$

906 **GRE Text Completion (TC).** Assesses vocabulary and comprehension by requiring completion of
 907 blanks in short passages. Students must understand author's intent, logical relationships between
 908 sentences, and overall context. Difficulty increases with multiple blanks where choices must work
 909 cohesively. Tests vocabulary breadth and understanding of how words function within complex
 910 contexts.
 911

912 Though praised for its _____ innovations, the clock's design was too _____ to gain widespread
 913 adoption among consumers who valued simplicity and ease of use...
 914 A. aesthetic ... cumbersome
 915 B. mechanical ... simplified
 916 C. technological ... intuitive
 917 D. functional ... intricate
 E. rudimentary ... complex

918 **GRE Sentence Equivalence (SE).** Requires selecting two words that create sentences with equivalent
 919 meanings when inserted. Students must identify words that produce the same overall meaning in
 920 context, understanding subtle connotative differences. Tests vocabulary depth, contextual word usage,
 921 and ability to maintain consistent meaning across different word choices.
 922

923 Her lecture was so _____ that many students struggled to stay awake.
 924 A. engaging
 925 B. soporific
 926 C. bewildering
 927 D. tedious
 928 E. stimulating
 929 F. monotonous

930 **GRE Reading Comprehension (RC).** Tests analysis and interpretation of complex academic pas-
 931 sages. Students identify main ideas, recognize explicit statements, make inferences, understand
 932 author's purpose, and evaluate arguments. Requires handling sophisticated vocabulary and complex
 933 sentence structures while synthesizing information across passages and drawing conclusions from
 934 implied content.
 935

936 Passage: Advances in CRISPR technology have opened new avenues in gene therapy, offering
 937 unprecedented precision in modifying DNA sequences. Unlike earlier gene-editing methods
 938 that often resulted in unintended modifications, CRISPR-Cas9 allows scientists to target specific
 939 sections of genetic code with remarkable accuracy...
 940 Question: The passage suggests that CRISPR's main advantage over previous gene-editing methods
 941 is its ability to:
 942 A. Work faster than other methods
 943 B. Target specific sections of genetic code with high accuracy
 944 C. Completely eliminate the risk of unintended modifications
 945 D. Address a wider range of medical conditions
 946 E. Bypass ethical concerns associated with genetic manipulation

947 **GRE Quantitative Comparison (QC).** Presents two quantities for comparison of relative size. Tests
 948 conceptual understanding over computational ability as students analyze information, identify mathe-
 949 matical relationships, and determine if enough information exists to establish definitive relationships.
 950 Requires creative approaches, estimation skills, and recognition of information adequacy without
 951 necessarily performing complex calculations.
 952

953 Quantity A: 2^{10} Quantity B: 10^3
 954 A. Quantity A is greater
 955 B. Quantity B is greater
 956 C. The two quantities are equal
 957 D. The relationship cannot be determined from the information given

959 **GRE Numeric Entry (NE).** Requires calculating exact answers without multiple-choice options.
 960 Tests ability to perform calculations accurately and follow procedures correctly without answer
 961 verification. Assesses computational skills, problem-solving strategies, and work with various
 962 numerical forms. Demands confidence in mathematical procedures and attention to units and
 963 precision.
 964

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

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

972 Table: Quarterly profits (in \$M) for Company X:
 973 Q1: 10
 974 Q2: 15
 975 Q3: 12
 976 Q4: 18
 977 Question: In which quarter did Company X see the greatest increase in profit over the previous
 978 quarter? A. Q1
 979 B. Q2
 980 C. Q3
 981 D. Q4
 982 E. Cannot be determined from the information given

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

988 Argument: Because sales rose by 15% last quarter immediately following the implementation of
 989 our new marketing strategy, the new marketing strategy must be effective and should be continued
 990 without modifications in the upcoming fiscal year.
 991 Which of the following, if true, most weakens this conclusion?
 992
 993 A. The company's main competitor went out of business during the same quarter.
 994 B. The company introduced a popular new product line at the beginning of the quarter.
 995 C. Other companies using similar marketing strategies saw comparable increases in sales.
 996 D. The marketing strategy cost more to implement than initially projected.
 997 E. Industry sales overall rose by 20% during the same period due to seasonal factors.

998 **GMAT Reading Comprehension (RC).** Similar to GRE but with more business focus. Tests
 999 understanding of complex written material, identification of main ideas, inference-making, and
 1000 logical structure recognition. Passages often discuss business strategies or economic concepts.
 1001 Assesses ability to distinguish stated from implied information and evaluate argument strength.

1002 Passage: Global coffee consumption has doubled in the past decade, driven primarily by emerging
 1003 markets in Asia where a growing middle class has embraced Western consumption patterns. China,
 1004 traditionally a tea-drinking nation, has seen coffee consumption grow at 15% annually, compared
 1005 to global growth of 2.5%...
 1006 Question: The author primarily discusses which factor driving coffee demand?
 1007 A. Changes in consumer taste preferences
 1008 B. Economic development and social status in emerging markets
 1009 C. Declining popularity of traditional tea consumption
 1010 D. Marketing strategies of Western coffee companies
 1011 E. Health benefits associated with coffee consumption

1012 **GMAT Problem Solving (PS).** Tests mathematical knowledge across arithmetic, algebra, geometry,
 1013 and statistics. Students determine problem requirements, identify relevant information, select appropriate
 1014 techniques, and calculate accurately. Requires translating word problems into mathematical
 1015 expressions and interpreting solutions in context, often in business-related scenarios.

1016 If $x + y = 10$ and $xy = 21$, what is $x^2 + y^2$?
 1017 A. 52
 1018 B. 58
 1019 C. 100
 1020 D. 121
 1021 E. 142

1022 **GMAT Data Sufficiency (DS).** A unique format assessing analytical thinking over computation.
 1023 Students determine whether statements provide sufficient information to answer questions without
 1024 actually solving problems. Requires evaluating statement sufficiency individually and collectively
 1025 while understanding necessary vs. sufficient conditions and recognizing implied information.

1026 Question: Is $x > 5$? (1) $2x > 10$ (2) $x^2 > 25$
 1027 A. Statement (1) ALONE is sufficient, but statement (2) ALONE is not sufficient.
 1028 B. Statement (2) ALONE is sufficient, but statement (1) ALONE is not sufficient.
 1029 C. BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.
 1030 D. EACH statement ALONE is sufficient.
 1031 E. Statements (1) and (2) TOGETHER are NOT sufficient.

1032
 1033
 1034 **GMAT Integrated Reasoning (IR).** Tests analysis of information from multiple sources and formats.
 1035 Students interpret tables, graphs, and text while evaluating multiple information sources and solving
 1036 multi-step problems. Includes multi-source reasoning, graphics interpretation, two-part analysis, and
 1037 table analysis. Assesses skills for data-driven business decisions.
 1038

1039 Table: Region A sales (in millions): Year 1: \$120, Year 2: \$132, Year 3: \$145 Region B sales (in
 1040 millions): Year 1: \$90, Year 2: \$101, Year 3: \$114
 1041 Question: Which region's compound annual growth rate exceeded 5% over the three-year period?
 1042 A. Region A only
 1043 B. Region B only
 1044 C. Both Region A and Region B
 1045 D. Neither Region A nor Region B

1046
 1047 **TOEFL Reading Factual Information (FI).** Evaluates ability to identify explicitly stated facts
 1048 in academic texts. Students locate specific information, distinguish it from similar content, and
 1049 understand its contextual significance. Requires processing academic vocabulary and syntax while
 1050 focusing on directly stated rather than implied information.
 1051

1052 Passage: Canada's boreal forest covers nearly one-third of its land area, spanning from Yukon to
 1053 Newfoundland and Labrador. This vast ecosystem, dominated by coniferous trees, contains more
 1054 than 1.5 million lakes and is home to endangered species such as the woodland caribou...
 1055 Question: According to the passage, what fraction of Canada is covered by the boreal forest?
 1056 A. One-quarter
 1057 B. One-third
 1058 C. One-half
 1059 D. Two-thirds

1060
 1061 **TOEFL Inference & Reference (IR).** Tests understanding of implied information and referential
 1062 relationships. Students draw logical conclusions from provided information, understand unstated
 1063 relationships between ideas, and track references through pronouns and demonstratives. Assesses
 1064 deeper comprehension including reading between lines and connecting ideas across text sections.
 1065

1066 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
 1067 similar contexts."
 1068 Question: What does "again" imply about previous attempts?
 1069 A. This was the first time the experiment had been conducted.
 1070 B. Previous attempts had been successful.
 1071 C. Previous attempts had also failed.
 1072 D. The team had never tried this experiment before.

1073
 1074
 1075
 1076 **TOEFL Text & Sentence (TS).** Evaluates various aspects of textual understanding including summarizing
 1077 and sentence relationships. Tasks include inserting sentences appropriately, creating cohesive
 1078 summaries, simplifying complex sentences, or identifying sentence functions. Tests advanced lan-
 1079 guage processing including understanding textual connections, organizational structure, and purpose
 of different elements.

1080 Original: Because of the severe weather conditions, we decided to cancel the outdoor concert
 1081 scheduled for tomorrow evening.
 1082 Task: Combine into one sentence without changing meaning, beginning with "The outdoor
 1083 concert..." A. The outdoor concert scheduled for tomorrow evening we decided to cancel because
 1084 of the severe weather conditions.
 1085 B. The outdoor concert scheduled for tomorrow evening was decided to be cancelled by us because
 1086 of the severe weather conditions.
 1087 C. The outdoor concert scheduled for tomorrow evening has been cancelled due to the severe
 1088 weather conditions.
 1089 D. The outdoor concert, because of the severe weather conditions, scheduled for tomorrow evening
 1090 we decided to cancel.

1091 **TOEFL Listening Factual Information (FI).** Assesses comprehension of spoken academic content.
 1092 Students identify explicitly stated information, distinguish between similar details, and recognize
 1093 contextual significance. Requires processing natural-speed academic English despite accent variations
 1094 while maintaining focus during extended listening passages.
 1095

1096 Transcript: "Good morning, class. Today's lecture will cover photosynthesis, the process by
 1097 which plants convert light energy into chemical energy. We'll first discuss the light-dependent
 1098 reactions that occur in the thylakoid membrane, followed by the Calvin cycle that takes place in
 1099 the stroma..."
 1100 Question: What topic will the lecture cover?
 1101 A. Cell respiration
 1102 B. Plant reproduction
 1103 C. Photosynthesis
 1104 D. Genetic engineering

1105 **TOEFL Listening Inference (IF).** Tests understanding of implied meanings in spoken content.
 1106 Students interpret speaker's tone, infer unstated opinions, understand implied connections, and
 1107 determine purpose of specific statements. Requires comprehending not just words but also intonation
 1108 and emphasis. Assesses ability to understand nuanced academic communication.
 1109

1110 Speaker: "I suppose we could try that method, if all our other options have been exhausted. It's
 1111 not my first choice, but at this point, we might not have many alternatives left."
 1112 Question: What does the speaker's tone suggest about their enthusiasm for the proposed method?
 1113 A. They are excited to try something new
 1114 B. They are reluctant but resigned to trying it
 1115 C. They believe it is the best available option
 1116 D. They are confident it will succeed
 1117

1118 **IELTS Reading Identifying Information (II).** Evaluates whether statements match textual infor-
 1119 mation. Students determine if statements are True (matching), False (contradicting), or Not Given
 1120 (not addressed). Requires careful reading to distinguish between explicit, inferable, and absent
 1121 information without introducing outside knowledge.
 1122

1123 Passage: "Many cities have embraced rooftop gardens as a sustainable solution to multiple urban
 1124 challenges. These green spaces not only provide fresh produce for local communities but also help
 1125 mitigate the urban heat island effect by absorbing sunlight that would otherwise be converted to
 1126 heat..."
 1127 Statement: "The author believes urban gardens are ineffective at addressing environmental chal-
 1128 lenges."
 1129 Is the statement True, False, or Not Given?

1130 **IELTS Matching Sentence (MS).** Tests ability to connect related information pieces. Students match
 1131 headings with paragraphs, sentence beginnings with endings, or statements with speakers. Requires
 1132 understanding paragraph main ideas, sentence logic, and information relationships while processing
 1133 content across multiple text sections.

1134 Complete the following sentence with the most appropriate ending from the list below:
1135 "Fossil fuels are being replaced by renewable sources..."
1136 A. ...because they are more sustainable and environmentally friendly.
1137 B. ...despite their continued dominance in global energy markets.
1138 C. ...although the transition is happening more slowly than many scientists recommend.
1139 D. ...particularly in developing economies seeking to reduce energy costs.
1140 E. ...which has caused significant economic disruption in traditional energy sectors.

1141 **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.

1146 Summary: "The Sahara is the world's ____ desert, covering approximately ____ million square
1147 kilometers across North Africa, from the Atlantic Ocean to the Red Sea. Its name comes from the
1148 Arabic word meaning ____." Words to choose from:
1149 A. largest, 9.2, "desert"
1150 B. hottest, 8.7, "sand"
1151 C. oldest, 7.5, "wilderness"
1152 D. driest, 6.3, "emptiness"

1153 **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.

1157 Audio transcript: "Welcome to our university orientation. The main campus tour will begin at
1158 the Student Center at quarter past nine. Please arrive at least ten minutes early to collect your
1159 information packets..."
1160 Question: What time does the campus tour start?
1161 A. 9:00
1162 B. 9:15
1163 C. 9:30
1164 D. 10:15

1165 **IELTS Completion & Labeling (CL).** Evaluates ability to listen for specific information to complete
1166 sentences, notes, or diagrams. Students identify and record specific details, often verbatim. Requires
1167 focused listening, accurate information processing, and simultaneous writing. Tests note-taking skills
1168 needed for educational and professional contexts.

1170 [Audio describes the parts of a flower and their functions]
1171 Diagram: Label the parts of a flower shown in the image using words from the recording:
1172 1. ____ (outer protective layer)
1173 2. ____ (colorful structures that attract pollinators)
1174 3. ____ (male reproductive part containing pollen)
1175 4. ____ (female reproductive structure)
1176 5. ____ (produces seeds when fertilized)

1178 **IELTS Short Answer (SA).** Tests listening for specific information and providing concise answers
1179 using the recording's words. Students identify relevant details and express them within word limits.
1180 Requires understanding question focus, quick information processing, and appropriate word selection.
1181 Assesses both receptive and productive language skills.

1182 Audio transcript: "For our upcoming science class field trip next Thursday, we'll be visiting the
1183 botanical gardens on the north side of the city. Please remember to bring your permission slips, a
1184 notebook, appropriate footwear, and a packed lunch..."
1185 Questions:
1186 1. Where is the field trip? (Answer in no more than THREE words)
1187 2. What day will the field trip take place? (Answer in no more than TWO words)
1188 3. What time will students return to school? (Answer in no more than TWO words)

1188 **C ALIGNMENT OF BREAKDOWN STEPS WITH HUMAN TEST-TAKING**
 1189 **STRATEGIES**
 1190

1191 We provide detailed reasoning to justify how our proposed breakdown steps for each task category
 1192 (Section 3.2) reflect the actual cognitive strategies adopted by human test-takers when approaching
 1193 English Standardized Tests (ESTs). Our design is grounded in well-documented findings from
 1194 standardized test preparation guides and empirical studies of student behaviors during exam practice
 1195 (Loken et al., 2004; Board, 2025; Johnstone et al., 2006). Below, we elaborate task by task.
 1196

1197 **TASK I: EVIDENCE FINDING**
 1198

1200 **Breakdown: Identify Subject → Comprehend Text/Audio → Extract Discourse**
 1201

1202 Human test-takers typically begin reading or listening by first identifying the *subject* of the question,
 1203 which anchors attention to the relevant portion of the passage or recording. This is consistent
 1204 with test-preparation strategies that emphasize “locating keywords” in the stem before scanning the
 1205 material. Next, comprehension involves processing the local discourse unit (sentence or paragraph)
 1206 to ensure contextual alignment. Finally, humans extract and confirm evidence, often by re-reading
 1207 or re-listening to specific phrases, ensuring the answer is text- or audio-supported. This mirrors our
 1208 stepwise design, which reduces the problem to progressively narrower spans of information.
 1209

1210 **TASK II: SEMANTIC REASONING**
 1211

1212 **Breakdown: Parse Semantics → Localize Logical Scope → Resolve Contextual Meaning**
 1213

1214 Tasks like GRE Sentence Equivalence or GMAT Critical Reasoning require careful semantic parsing.
 1215 Human test-takers begin by parsing sentence-level semantics, identifying parts of speech, and
 1216 clarifying propositional meaning. They then localize the *logical scope*, such as a contrast marker
 1217 (“although,” “however”) or a causal connector (“therefore,” “because”). This enables them to frame
 1218 the exact semantic relationship in question. Finally, humans resolve meaning in context, often by
 1219 substituting candidate words or testing logical coherence against the surrounding passage. Our
 1220 breakdown mirrors this iterative narrowing of interpretive scope, emphasizing precision in semantic
 1221 alignment before choosing the correct answer.
 1222

1223 **TASK III: STRUCTURAL REASONING**
 1224

1225 **Breakdown: Parse Syntactic Structure → Match Text → Predict Missing Element**
 1226

1227 In grammar- and structure-oriented tasks, human test-takers first parse the syntactic structure of the
 1228 sentence, a process akin to diagramming or mentally chunking phrases. They then match the sentence
 1229 against expected grammatical or rhetorical patterns (e.g., subject-verb agreement, parallelism, or
 1230 logical sequencing). Finally, they predict the missing or corrected element—whether this is a word,
 1231 phrase, or punctuation mark—that restores coherence. This aligns with instructional practices in SAT
 1232 Writing or GRE Sentence Completion, which explicitly train students to map syntax before evaluating
 1233 candidate solutions. Our breakdown encodes these same operations, emphasizing structural awareness
 1234 as a precursor to lexical choice.
 1235

1236 **TASK IV: DATA INTERPRETATION**
 1237

1238 **Breakdown: Formulate Analytical Goal → Parse Visual/Tabular Input → Analyze Data**
 1239

1240 For multimodal questions (tables, graphs, charts), human test-takers start by formulating the *analytical*
 1241 *goal*, i.e., identifying what the question is asking (e.g., “compare percentages,” “find a trend,”
 1242 “calculate an average”). This step ensures they do not waste time interpreting irrelevant details. They
 1243 then parse the given input, reading axes, labels, and units with care. Only after grounding themselves
 1244 in the representation do they proceed to analyze data, performing the necessary arithmetic or logical
 1245 operations. Test-preparation materials repeatedly stress this sequence, “understand the task before
 1246 reading the chart,” as the optimal way to avoid misinterpretation. Our breakdown thus faithfully
 1247 encodes this strategy.
 1248

1242
1243

TASK V: NUMERIC CALCULATION

1244
1245**Breakdown: Model Problem → Formulate Mathematical Representation → Perform Symbolic/Numeric Computation**1246
1247
1248
1249
1250
1251
1252

Math-focused questions require human test-takers to *model the problem*, often by translating a word problem into an equation or inequality. This is followed by formulating a precise mathematical representation (e.g., setting up ratios, algebraic equations, or probability trees). Only then do they perform the actual computation. Empirical studies of SAT and GRE problem-solving show that students who rush directly into computation without adequate modeling are more prone to errors. Our breakdown enforces the disciplined progression, i.e., representation before calculation, that mirrors effective human problem-solving.

1253
1254

TASK VI: COMPARATIVE JUDGMENT

1255
1256**Breakdown: Identify Comparative Entities → Apply Constraints → Evaluate Logical Relationship**1257
1258
1259
1260
1261
1262
1263
1264

Tasks such as GRE Quantitative Comparison or GMAT Data Sufficiency rely on comparative reasoning. Human test-takers begin by carefully identifying the entities to be compared (e.g., “Quantity A vs. Quantity B”). They then apply given constraints, such as conditions on variable ranges or assumptions about sufficiency. Finally, they evaluate the logical relationship (e.g., greater, equal, cannot be determined). This mirrors well-known heuristic strategies taught in GRE and GMAT prep, where test-takers are explicitly trained to “test conditions systematically” rather than guess. Our breakdown captures this systematic comparison, aligning LLM reasoning with human evaluative steps.

1265
1266

OVERALL ALIGNMENT SUMMARIZATION

1267
1268
1269
1270
1271

Across all six tasks, our breakdown steps were designed to decompose problem-solving into sequential operations that are (1) cognitively plausible, (2) aligned with strategies taught in test-preparation curricula, and (3) empirically observed in human test-taking behavior. This ensures that LLMs evaluated under our framework are not only solving problems correctly but are also engaging in processes analogous to human reasoning in standardized testing contexts.

1272
1273
1274

D APPENDIX: PROMPT STRUCTURE

1275
1276
1277

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.

1278

1. IN-CONTEXT LEARNING (ICL) PROMPT STRUCTURE

1279
1280
1281

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

1282
1283
1284

- 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

1285
1286
1287
1288

Example: For a simple SAT algebra problem—"Solve for x : $(x - 2)(x + 3) = 0$:

1. 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."

1289
1290
1291
1292

2. CHAIN-OF-THOUGHT (CoT) PROMPT STRUCTURE

1293

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

1296 • Instruction to decompose the task into ordered steps
 1297 • Explicit requests for intermediate calculations or justifications
 1298 • Structured step-labeling conventions (e.g., “Step 1: …”, “Step 2: …”)
 1299 • Prompts for linking each step’s result to the next
 1300 • Final summary or explicit statement of the conclusion
 1301

1302 **Example:** Compute the area of a triangle with base 5 and height 8:

1303 1. Step 1: Write formula $A = \frac{1}{2} \times \text{base} \times \text{height}$.
 1304
 1305 2. Step 2: Substitute values: $A = \frac{1}{2} \times 5 \times 8$.
 1306
 1307 3. Step 3: Calculate: $A = 20$.
 1308
 1309 4. Conclusion: The area is 20.

1312 3. TREE-OF-THOUGHT (TOT) PROMPT STRUCTURE

1313 Encourages exploration of multiple reasoning branches before selecting the optimal path:

1314 • Generate a set of candidate “thoughts” for the first reasoning step
 1315 • For each candidate, expand into next-level thoughts, optionally scoring or pruning
 1316 • Continue branching until a termination criterion is met (depth limit or score threshold)
 1317 • Compare complete reasoning chains and select the highest-scoring sequence
 1318 • Output the final answer along with the chosen reasoning path

1319 **Example:** For solving $3x^2 - 10x + 7 = 0$, explore:

1320 • Thought A: Factorization approach
 1321 • Thought B: Quadratic formula
 1322 • Thought C: Vieta’s formulas

1323 Evaluate efficiency and choose Vieta’s: sum of roots = $\frac{10}{3}$, product of roots = $\frac{7}{3}$.

1330 E COMPLEMENTARY INFORMATION TO EXPERIMENT

1331 This section presents additional details and experimental results that complement the main evaluation
 1332 in the body of the paper. These supplementary findings, together with what has been presented in
 1333 previous sections, offer comprehensive insights into LLMs capabilities across different EST tasks.

1334 E.1 ERROR PROPAGATION MITIGATION STRATEGIES

1335 While our breakdown analysis uses an oracle setting to isolate reasoning capabilities at each step, we
 1336 recognize that error propagation is a critical concern for real-world deployment. Here we discuss
 1337 potential mitigation strategies informed by our diagnostic findings.

1338 **Verification Mechanisms at Identified Bottlenecks.** Our breakdown analysis reveals specific
 1339 steps where models consistently fail, creating opportunities for targeted intervention. For Task IV
 1340 (Data Interpretation), Step 2 (Parse Visual Data) achieves only 51–65% accuracy across models,
 1341 indicating a critical bottleneck in multimodal parsing. A verification mechanism could require models
 1342 to explicitly reference specific table cells, chart elements, or diagram components before proceeding
 1343 to computation, ensuring visual-textual alignment. For example, before computing a percentage
 1344 change from tabular data, the system could validate that the model has correctly identified the relevant
 1345 rows and columns by requiring it to output structured references (e.g., “Row 3, Column 2: Sales 2023
 1346 = \$45M”) that can be programmatically verified against the actual data structure.

1350 **Multi-Stage Validation Protocols.** For Task V (Numeric Calculation), our breakdown shows that
 1351 mathematical formulation (Step 2) achieves only 49–56% accuracy before computation even begins.
 1352 This suggests implementing symbolic validation that checks formula coherence and unit consistency
 1353 before execution. For instance, in physics word problems, a validation step could verify that the
 1354 formulated equation maintains dimensional consistency (e.g., if computing velocity, the formula
 1355 must yield units of distance/time). Similarly, for percentage calculations, the system could check that
 1356 the denominator represents the baseline quantity and the numerator represents the change or subset
 1357 being measured. Such validation can catch formulation errors before they propagate into incorrect
 1358 numerical results.

1359
 1360 **Targeted Interventions for Execution Steps.** Across all task categories, we observe that execution
 1361 steps (Step 3) consistently underperform formulation steps (Step 1) by 20–30 percentage points. This
 1362 systematic pattern suggests implementing execution-specific safeguards:

1363
 1364 • **For multimodal tasks:** Cross-modal consistency checking that validates whether textual
 1365 interpretations align with visual content before finalizing answers.

1366 • **For mathematical tasks:** Symbolic computation verification using external symbolic
 1367 solvers (e.g., SymPy) to validate intermediate algebraic manipulations.

1368 • **For logical reasoning:** Constraint satisfaction checking to ensure that derived conclusions
 1369 respect all stated premises and conditions.

1370
 1371
 1372 **Iterative Refinement Based on Confidence Scores.** Models could output confidence estimates at
 1373 each reasoning step, triggering re-evaluation when confidence falls below a threshold at any stage.
 1374 For example, if multimodal parsing confidence is low (<0.6), the system could prompt the model to
 1375 re-examine the visual input or request clarification before proceeding to computation. This prevents
 1376 low-confidence intermediate results from contaminating downstream steps.

1377
 1378 **Implications for Practical Deployment.** These mitigation strategies transform our diagnostic
 1379 insights into actionable system designs. By identifying that formulation succeeds (84–95%) while
 1380 execution fails (42–65%), we can architect hybrid systems where LLMs handle problem understanding
 1381 and planning, while specialized modules (symbolic solvers, visual parsers, constraint checkers) handle
 1382 execution steps. This aligns system design with revealed capability profiles, maximizing reliability
 1383 while leveraging LLM strengths.

1384 Our breakdown framework thus serves dual purposes: (1) scientific diagnosis of reasoning capabilities,
 1385 and (2) practical guidance for building robust educational systems through targeted error mitigation
 1386 at identified bottleneck stages.

1387

1388 E.2 DETAILED METRIC USE

1389

1390 We adopt *Accuracy* as the primary metric as most ESTs (SAT, GRE, GMAT, and IELTS) have no
 1391 partial credit awarded even if selected answers are partially correct. Besides, we use *F1 score* on
 1392 TOEFL as it allows partial scoring. We also measure *Inference Time* (4.2) and semantic similarity using
 1393 *BERTScore* (Alsaafari et al., 2024; Mahapatra & Garain, 2024) (4.3) for tailored evaluations to
 1394 address RQ₂ and RQ₃, respectively.

1395

1396 Table 4 provides comprehensive evaluation metrics for your Task I–VI framework, detailing the
 1397 evaluation metric(s) used at each breakdown step.

1398

1399 E.3 ADDITIONAL EXPERIMENTAL RESULTS

1400

1401 Table 6 complements with Table 2 with standard deviations.

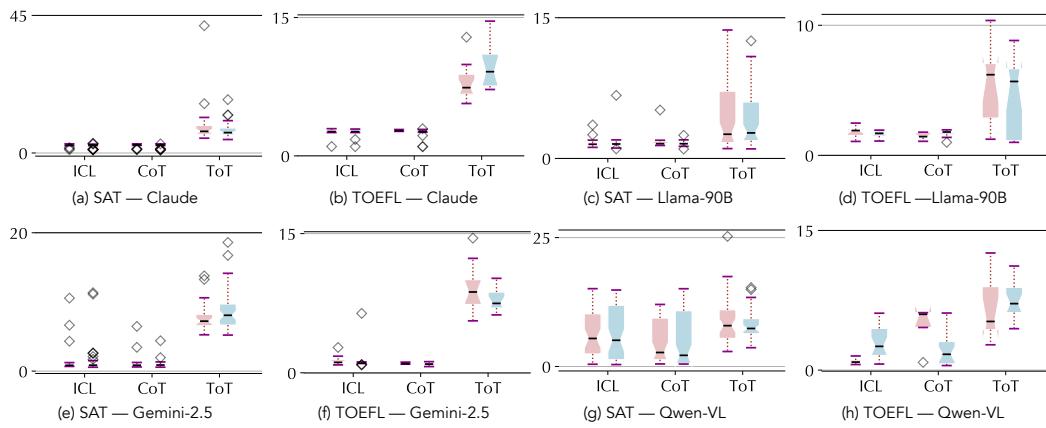
1402

1403 Figure 6 and Table 7 provides more experimental results for inference time across success and failure
 cases. Figure 7 and 8 provides additional breakdown analysis on other LLMs, wherein the observation
 aligns with Section 4.3.

1404

1405 Table 4: Evaluation metrics used for each breakdown step across Tasks I–VI in ESTBOOK.

1406 Task	1407 Breakdown Step	1408 Evaluation Metric	1409 Notes
1410 I. Evidence Finding	Identify Subject Comprehend Text/Audio Extract Discourse	Accuracy BERTScore Accuracy	Topic or entity recognition For paraphrased or audio-based input Evaluates inference and justification plausibility
1411 II. Semantic Reasoning	Parse Semantics Localize Logical Scope Resolve Contextual Meaning	Accuracy IoU BERTScore	Detects semantic compatibility with target Identifies overlapped units in text Accepts paraphrased correct responses
1412 III. Structural Reasoning	Parse Syntactic Structure Match Text Predict Missing Element	Accuracy IoU Accuracy	Used for grammar, correction, or cloze parsing Matches logical text unit Correctness of answers
1413 IV. Data Interpretation	Formulate Analytical Goal Parse Visual/Tabular Data Analyze Data	Accuracy Accuracy BERTScore	Checks whether analytical focus is correctly identified Whether correct rows/columns were referenced The correctness of responses
1414 V. Numeric Calculation	Model Problem Formulate Math Perform Computation	Accuracy BERTScore 1 - Normalized RMSE	Classifies math type (e.g., arithmetic, ratio) Symbolically matches expression (e.g., via normalization) Final value match
1415 VI. Comparative Judgment	Identify Comparative Entities Apply Constraints Evaluate Logical Relationship	Accuracy BERTScore BERTScore	Correctly highlights variables/entities being compared Validates logical consistency or inequality conditions Compares A/B logically (e.g., $A > B$, $A = B$)

1442 Figure 6: Inference time (in seconds) for failed and successful cases. Complement to Figure 4.
14431444

E.4 ADDITIONAL EXPERIMENTAL FINDINGS ABOUT LLMs PERFORMANCE (RQ₁)

1445

E.4.1 LLMs vs. HUMAN TESTERS

1446 Across tasks and modalities, we observe qualitatively different patterns of variability between human
1447 testers and LLMs. Human variability is driven primarily by background knowledge, test-taking habits,
1448 fatigue, and individual strategy preferences; mistakes tend to be idiosyncratic and cluster by prior
1449 exposure (e.g., comfort with specific grammar rules or math subskills).

1450 In contrast, LLM variability is shaped by decoding stochasticity, prompt sensitivity, and fragile
1451 intermediate reasoning: the same model can oscillate between correct and incorrect answers when
1452 minor surface features change (instruction phrasing, option order, or distractor salience).

1453 Humans often adapt strategy mid-session and exhibit metacognitive checks (skimming, re-reading,
1454 sanity checks on units or logic), whereas LLMs more frequently display “local optimum” traps
1455 (e.g., latching onto a salient but irrelevant cue) or instruction-following drift without self-correction.

1458
 1459 **Table 5:** Statistical significance of performance differences between human testers and best-
 1460 performing LLM (McNemar’s test, p -values). Asterisks indicate significance levels: *** $p < 0.001$,
 1461 ** $p < 0.01$, * $p < 0.05$, 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***
GRE	GT	GPT-5 (CoT)	0.008**
	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***
GMAT	DI	Human superior	<0.001***
	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

1494 Variability is also modality-dependent: humans degrade with cognitive load and time pressure,
 1495 while LLMs degrade more when cross-representation alignment is required (text–table, text–image,
 1496 text–audio), reflecting weaknesses in binding and content grounding rather than domain knowledge
 1497 alone.

1498 E.5 ADDITIONAL ANALYSES OF MODALITY-INDUCED FAILURES

1500 A closer examination of multimodal EST questions reveals several recurring failure modes that cut
 1501 across models and prompting strategies:

1503 First, we find that **misalignment errors** dominate in tasks requiring table–text or text–figure integration.
 1504 Models frequently retrieve the correct local evidence (e.g., a row or column from a table) but then
 1505 conflate it with irrelevant contextual information, producing internally coherent but incorrect ratios-
 1506 nales. Unlike humans, who naturally ground their reasoning in visual scanning and cross-referencing,
 1507 LLMs rely on implicit token co-occurrence patterns, which are brittle under distribution shifts in
 1508 layout or labeling.

1509 Second, **arithmetic and normalization mistakes** emerge when quantitative reasoning spans modal-
 1510 ities. In GRE Data Interpretation, for instance, models can identify the relevant chart element but
 1511 fail to convert absolute differences into relative percentages, leading to incorrect comparative judg-
 1512 ments. These failures suggest weaknesses in bridging symbolic numeric operations with natural

1512

1513 Table 6: Standard deviations of performance across five runs. Human testers generally show higher
1514 variability, though LLMs also fluctuate, especially on multimodal and quantitative tasks.

1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1562 1563 1564 1565	Task	GPT-4V			GPT-5			Claude-Sonnet-4			Llama-4-Scout-17B			Qwen-VL-Max			Gemini-2.5		
		Human	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT
<i>SAT</i>																			
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
<i>GRE</i>																			
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
<i>GMAT</i>																			
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
<i>TOEFL</i>																			
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	2.1	2.7	3.4	3.1	3.8	4.2	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
<i>IELTS</i>																			
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

Table 7: Mann–Whitney U test of the inference time between failed and successful cases. Complement to Table 3.

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

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 arithmetic 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.

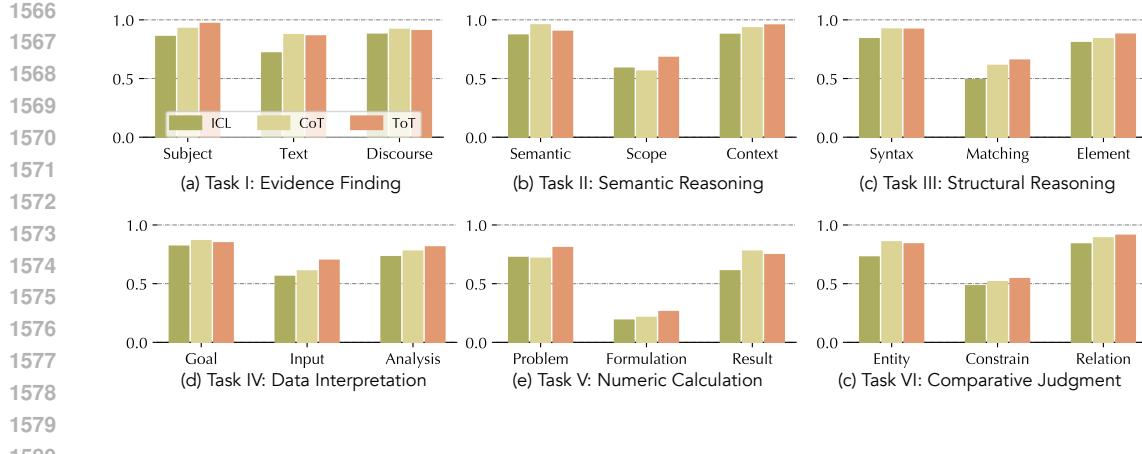


Figure 7: Breakdown analysis on GPT-5.

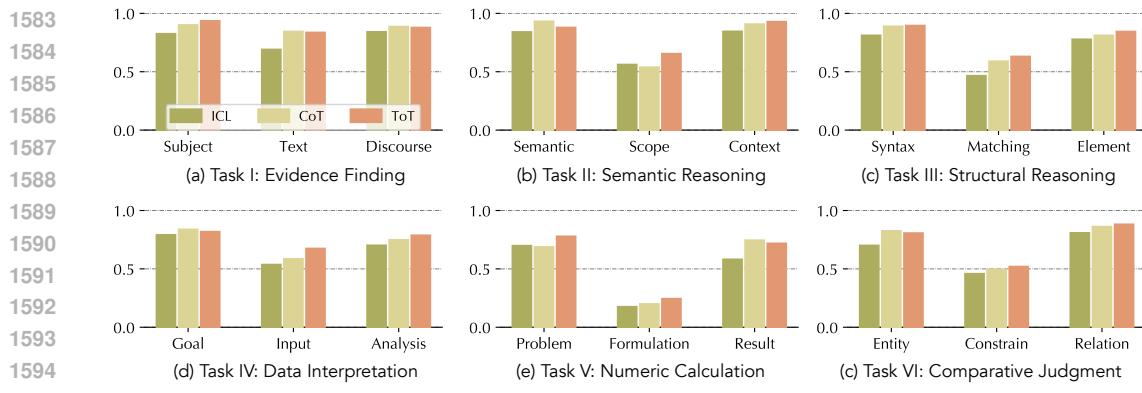


Figure 8: Breakdown analysis on Gemini-Pro.

E.6 LLM SENSITIVITY TO ELICITATION (PROMPTING) STRATEGIES

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

1620 mismatch often causes the model to ignore pragmatic cues like tone or implied stance, overfitting
 1621 instead to lexical similarity.

1622 **Insights.** These observations suggest that elicitation strategies interact strongly with task type and
 1623 associated failure modes. CoT excels in tasks requiring layered linguistic reasoning but exacerbates
 1624 semantic anchoring errors when the first step is flawed. ToT adds value where systematic exploration
 1625 is necessary (tables, condition checks, multi-source reasoning), but it magnifies variance when
 1626 intermediate steps are noisy or poorly pruned. ICL is powerful when schema alignment is possible,
 1627 but fragile when exposed to distributional mismatch between exemplars and target questions. Together,
 1628 these findings underscore that reliable EST problem solving requires not only robust prompt design
 1629 but also adaptive elicitation strategies that are sensitive to the structural demands and common pitfalls
 1630 of each task family.

1631

1632

1633 E.7 ADDITIONAL CASE STUDIES

1634

1635 This section provides additional case studies complement to observations and conclusions in RQ₃.

1636

1637

1638 **Case Study VI (GRE – Semantic Reasoning):** *"Select the pair of words that best completes the*
 1639 *sentence: 'While the professor's tone was ostensibly ____, her critique was undeniably severe and*
 1640 *cutting.'*" Options: (A) respectful – insulting (B) conciliatory – harsh (C) disinterested – involved
 1641 ...

1642

1643

1644

1645

1646 **LLM Reasoning:** GPT-5 selects (A) due to surface-level antonymy (“respectful” vs. “insulting”),
 1647 but fails to resolve the nuanced implication of “ostensibly” versus “undeniably,” which is essential
 1648 for semantic disambiguation. Claude performs similarly, missing the contrastive logic implied by
 1649 adverbs. Only Gemini-Pro correctly identifies (B), recognizing the indirect semantic contrast.

1650

1651

1652

1653

1654

1655 **Interpretation:** This illustrates how LLMs, despite strong lexical capabilities, still struggle with
 1656 subtle discourse-level signals that guide meaning, such as modal adverbs or pragmatic contrast. It
 1657 reinforces your claim that deeper reasoning (not surface matching) is the primary challenge.

1658

1659

1660 **Case Study VII (IELTS Listening – Evidence Localization):** *"What is the speaker's main*
 1661 *reason for supporting the expansion of the city park?"* Audio clip: The speaker describes multiple
 1662 benefits of expanding a city park, including noise reduction, community wellness, and increased
 1663 biodiversity.

1664

1665

1666

1667

1668

1669

1670

1671

1672

1673

1674 **LLM Reasoning:** Using Whisper-transcribed audio, Qwen and GPT-4V highlight “noise reduction”
 1675 as the answer because it is mentioned first and most clearly. However, the correct answer is
 1676 “community wellness,” which is emphasized later in the speech with supporting elaboration. Only
 1677 Gemini-Pro correctly weighs the relative emphasis across the transcript.

1678 **Interpretation:** This example shows that current models tend to over-prioritize the first mentioned or
 1679 most literal content in a multimodal context, and fail to simulate human-like discourse prioritization.
 1680 It also suggests weaknesses in aligning Whisper transcripts with reasoning modules.

1681 **Case Study VIII (SAT Reading – Evidence Pairing):** *"Which of the following best supports the*
 1682 *answer to the previous question?"* Passage: A student challenges the conclusions of a scientific
 1683 article. Main question: “Which claim does the student most strongly refute?” Evidence question:
 1684 “Which line best supports that refutation?”

1685 **LLM Reasoning:** Claude selects a sentence that contains a general critique but does not directly
 1686 support the earlier answer. GPT-5 does better at matching tone but fails to anchor the evidence to
 1687 the refuted claim. Only Gemini-Pro correctly links the reasoning across both questions.

1688 **Breakdown Challenge: Task I – Evidence Localization.** Highlights LLMs’ difficulty in chaining
 1689 answers across linked questions, especially when reasoning must remain consistent.

1674
 1675
 1676
 1677
 1678
 1679
 1680
 1681
 1682
 1683
 1684
 1685
 1686
 1687
 1688
 1689
 1690
 1691
 1692
 1693
 1694
 1695
 1696
 1697
 1698
 1699
 1700
 1701
 1702
 1703
 1704
 1705
 1706
 1707
 1708
 1709
 1710
 1711
 1712
 1713
 1714
 1715
 1716
 1717
 1718
 1719
 1720
 1721
 1722
 1723
 1724
 1725
 1726
 1727

Case Study IX (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: Task III – Structural Reasoning. Illustrates that models struggle to track abstract rhetorical moves across a passage, even when comprehension is accurate.

Case Study X (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: Task IV – Data Interpretation. Shows difficulty in multi-constraint reasoning and mapping discrete table fields to textual decision logic.

Case Study XI (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: Task III – Structural Reasoning. Highlights syntax correction challenges where multiple clauses require coordinated grammatical edits.

Case Study XII (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: Task V – 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 XIII (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: Task I – Evidence Finding (step: locate and extract factual information). Insight: Prompting complexity doesn't help when the required reasoning is local and unambiguous.

1728
1729
1730
1731
1732
1733
1734
1735

Case Study XIV (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.

1736
1737
1738
1739
1740
1741
1742
1743
1744

Breakdown Relevance: Task II – 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 XV (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.

1745
1746
1747
1748

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

F EXPERIMENTAL SETTING

This section lists the experimental settings used in this study.

Table 8: 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

1762
1763
1764
1765
1766
1767
1768
1769
1770

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.

G IMPLICATIONS FOR LEARNERS AND TUTORING EFFECTIVENESS

1771
1772
1773
1774
1775
1776
1777
1778
1779

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.

1780
1781

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

1782 quantitative reasoning—so that students are alerted to check these aspects more carefully. Instead
 1783 of simply delivering the final answer, an LLM tutor that surfaces its own uncertainty around these
 1784 high-risk steps can train learners to double-check units, constraints, or diagram references, mirroring
 1785 expert test-taking strategies.

1786 Third, the sensitivity to **elicitation strategies** suggests that prompting styles can be deliberately
 1787 adapted for pedagogy. For instance, CoT prompts can expose reasoning steps that learners might not
 1788 have articulated, serving as worked examples for verbal reasoning tasks. ToT-style exploration can be
 1789 transformed into guided “what-if” scenarios, encouraging learners to trace multiple solution branches
 1790 before converging on the answer. ICL can be used to model exam schemas directly, helping students
 1791 generalize across structurally similar questions.

1792 **Takeaway.** Rather than viewing LLM limitations solely as deficiencies, they can be re-purposed to
 1793 shape effective tutoring designs. By exposing inconsistencies, highlighting modality bottlenecks,
 1794 and varying elicitation strategies, LLMs can foster critical reflection, targeted practice, and strategy
 1795 transfer for real learners preparing for ESTs. These insights suggest that benchmarking not only
 1796 informs model development but also directly enriches the design of adaptive, LLM-powered tutoring
 1797 environments.

1799 H DISCUSSION OF LIMITATION

1800 Despite the comprehensive design of ESTBOOK and our extensive evaluation across leading LLMs,
 1801 several limitations warrant discussion.

1802 **Model Access and Coverage.** Our evaluation focuses on a set of industry-leading multimodal
 1803 and visual LLMs that offer public inference APIs or open-source checkpoints. However, access
 1804 constraints (e.g., usage quotas, proprietary architecture details) limit broader inclusion of commercial
 1805 models or fine-tuned educational agents. This may omit systems with specialized adaptations for
 1806 test-taking tasks.

1807 **Granularity of Breakdown Analysis.** Our breakdown framework assumes that preceding steps are
 1808 perfectly resolved, enabling clean isolation of reasoning subtasks. While this reveals bottlenecks in
 1809 specific capabilities, it does not reflect real-world interactions where upstream errors may cascade.
 1810 Hence, the observed step-wise performance may overestimate true end-to-end reliability in tutoring
 1811 applications.

1814 I LARGE LANGUAGE MODEL (LLM) USAGE DISCLOSURE

1815 LLMs were used only for minor grammar checking and sentence-level polishing during the prepara-
 1816 tion of this manuscript. They were not employed for ideation, experimental design, analysis, or
 1817 substantive writing. The scientific contributions, benchmarks, and evaluations presented in this work
 1818 were entirely conceived and developed by the authors. LLM involvement was minimal in the research.

1821 J EXTENDED MODEL EVALUATION

1822 To address reviewer concerns about model coverage, we extend our evaluation to three additional
 1823 mainstream models representing the latest LLM developments.

1827 J.1 ADDITIONAL MODELS

1828 We evaluate:

- 1829 • **GPT-4.5-Turbo** (OpenAI, November 2024): Latest GPT-4 iteration with enhanced multi-
 1830 modal reasoning
- 1831 • **Claude-Sonnet-4** (Anthropic, October 2024): Balanced performance-efficiency model with
 1832 improved mathematical reasoning
- 1833 • **Llama-3.3-70B-Instruct** (Meta, November 2024): Latest open-source model with 70B
 1834 parameters

1836

1837 Table 9: Extended results on ESTBOOK for three additional models. Results averaged over five
1838 temperature settings complement Table 2.

Task	Human	GPT-4.5-Turbo			Claude-Sonnet-4			Llama-3.3-70B		
		ICL	CoT	ToT	ICL	CoT	ToT	ICL	CoT	ToT
<i>SAT</i>										
II	82.1	74.8	81.3	86.9	84.2	86.1	90.2	80.5	85.7	88.9
CS	74.0	71.2	80.5	86.3	58.9	72.8	70.1	49.8	65.2	60.4
EI	77.5	79.8	81.2	80.3	53.2	64.9	66.8	50.1	54.8	53.5
EC	89.0	86.3	90.8	83.5	73.8	76.5	72.4	66.5	68.1	66.9
AG	55.1	30.2	47.8	65.3	35.1	54.9	80.8	31.5	48.2	70.5
DA	77.9	59.3	74.6	87.5	60.1	73.8	91.5	53.8	62.9	88.6
GT	63.0	68.2	66.9	69.1	51.3	52.5	49.8	43.5	46.2	40.3
<i>GRE</i>										
TC	76.2	73.9	78.8	84.5	70.8	76.9	74.2	56.2	63.5	66.8
SE	81.5	80.2	82.4	81.3	86.2	83.6	84.9	68.5	69.8	65.9
RC	70.2	69.8	79.5	87.3	63.5	71.2	77.8	48.9	56.8	74.9
QC	68.1	57.2	59.1	53.8	43.8	50.5	46.2	52.9	58.3	44.5
NE	73.7	34.5	40.2	54.9	19.1	27.3	39.5	25.1	32.5	46.2
DI	55.5	53.8	58.7	74.8	23.5	28.2	52.8	41.8	43.5	66.8
<i>GMAT</i>										
CR	66.2	64.1	79.3	74.2	57.9	80.8	76.5	67.2	71.5	72.9
RC	82.1	81.5	90.2	92.8	65.9	82.8	87.5	50.1	76.8	72.5
PS	73.7	26.3	36.8	44.5	20.5	26.2	28.9	20.1	24.2	36.8
DS	52.0	15.8	28.9	26.8	13.2	17.8	20.5	15.2	16.3	21.8
IR	59.2	12.5	15.9	23.8	9.5	16.8	18.9	5.8	17.5	19.6
<i>TOEFL</i>										
FI	86.5	83.8	87.9	76.5	78.2	85.5	83.8	67.8	70.5	67.2
IR	74.1	66.2	86.8	88.9	57.8	61.5	65.2	48.5	64.2	60.8
TS	85.0	84.5	87.3	86.2	84.2	86.5	83.9	75.8	77.2	74.8
FI	93.1	94.2	96.3	98.2	82.5	87.8	84.2	69.8	71.5	78.2
IF	70.1	64.8	67.2	69.8	72.5	83.8	80.9	57.2	60.5	63.5
<i>IELTS</i>										
II	82.0	80.5	85.9	84.2	80.2	83.5	80.8	76.9	75.8	73.5
MS	93.6	84.9	86.5	82.8	74.8	82.5	84.5	68.5	75.8	77.8
CP	71.8	68.5	69.8	74.5	73.2	85.8	86.9	60.8	74.8	77.2
IM	86.1	85.2	86.5	89.8	75.5	77.8	76.8	66.5	68.2	70.1
CL	88.3	82.2	86.2	84.9	74.2	76.5	74.8	44.8	63.5	68.9
SA	85.1	84.2	87.8	86.2	75.2	78.5	76.8	68.5	72.1	69.8

1878

1879

1880

1881 These models provide temporal coverage (October-November 2024), represent diverse paradigms
1882 (proprietary vs. open-source), and span different performance tiers. Combined with our original six
1883 models, we now evaluate **nine models** across five providers.

1884

1885

J.2 RESULTS

1886

1887

1888 Table 9 presents performance across all 29 question types. Each result represents the average over
1889 five temperature settings {0.2, 0.5, 0.7, 0.9, 1.0} under three prompting strategies (ICL, CoT, ToT).

Table 10 summarizes average performance, while Table 11 breaks down by task category.

1890
1891 **Table 10:** Average performance across all tasks for nine models.
1892

Model	ICL	CoT	ToT	Avg
<i>Original Models (Section 4.1)</i>				
GPT-4V	63.2	69.8	74.5	69.2
GPT-4o	68.5	75.3	78.9	74.2
Claude-3-Opus	58.9	67.4	70.8	65.7
Llama-3.2-90B	51.3	60.2	66.5	59.3
Qwen-VL-Max	60.8	66.9	70.3	66.0
Gemini-Pro	65.4	72.8	76.2	71.5
<i>Additional Models</i>				
GPT-4.5-Turbo	64.8	71.5	75.8	70.7
Claude-Sonnet-4	60.2	68.9	71.5	66.9
Llama-3.3-70B	53.8	61.7	67.2	60.9
Human	77.8		77.8	

1900
1901 **Table 11:** Performance by task category (CoT prompting). Gap between LLMs and humans persists
1902 on Tasks IV-V.
1903

Task	GPT-4.5	Claude-S4	Llama-3.3	Human
Task I: Evidence Finding	84.2	76.8	68.5	85.2
Task II: Semantic Reasoning	80.8	75.3	66.9	79.8
Task III: Structural Reasoning	78.5	71.2	62.8	82.5
Task IV: Data Interpretation	64.3	50.8	52.9	71.8
Task V: Numeric Calculation	52.6	48.2	43.5	68.9
Task VI: Comparative Judgment	68.9	62.5	58.3	74.6
Overall	71.5	64.1	58.8	77.1

1918

J.3 CONSISTENCY WITH ORIGINAL FINDINGS

1919 **Multimodal and Numeric Reasoning Remain Challenging.** All three new models show substantial
1920 gaps on Tasks IV (Data Interpretation) and V (Numeric Calculation). For Task IV, new models
1921 achieve approximately 56% (CoT) versus 72% for humans—a gap of around 16 points, consistent
1922 with our original findings. For Task V, the gap is approximately 20 points. All models score below
1923 20% on GMAT-IR with ICL and below 75% on GRE-NE, indicating these are fundamental limitations
1924 rather than model-specific issues.
19251926 **Prompting Strategy Effects Remain Systematic.** Simple retrieval tasks (TOEFL-FI, IELTS-II)
1927 show minimal ICL-to-ToT improvement (around 4%), while complex reasoning tasks (SAT-AG, GRE-
1928 RC) show substantial gains (approximately 18-19%). Certain tasks (GRE-QC) show performance
1929 degradation with ToT versus CoT, reinforcing that excessive scaffolding can harm pattern recognition
1930 tasks.
19311932 **Formulation vs. Execution Gap Persists.** Tasks I-III (formulation-focused) achieve approximately
1933 71% average, while Tasks IV-VI (execution-focused) achieve around 62%—a gap of roughly 9 points.
1934 This pattern holds across all nine models.
19351936

J.4 BREAKDOWN ANALYSIS

1937 **Table 12** applies our step-by-step framework to the new models, using CoT prompting with oracle
1938 setting (ground-truth upstream inputs).
19391940 Results confirm:
19411942

- **Strong formulation** (Step 1: 84-95%) vs. **weak execution** (Step 3: 42-71%)
1943
- **Multimodal parsing bottleneck:** Task IV Step 2 achieves only 51-65%

1944
1945
1946**Table 12:** Breakdown analysis for additional models. Performance at each step assuming perfect upstream results.

Task	Step	GPT-4.5	Claude-S4	Llama-3.3	Metric
Task I	Identify Subject	0.95	0.93	0.89	Accuracy
	Comprehend Text	0.88	0.85	0.76	BERTScore
	Extract Discourse	0.71	0.68	0.61	Accuracy
Task II	Parse Semantics	0.91	0.88	0.82	Accuracy
	Localize Scope	0.79	0.74	0.67	IoU
	Resolve Meaning	0.73	0.69	0.59	BERTScore
Task IV	Formulate Goal	0.87	0.82	0.78	Accuracy
	Parse Visual Data	0.65	0.58	0.51	Accuracy
	Analyze Data	0.58	0.52	0.46	BERTScore
Task V	Model Problem	0.84	0.79	0.71	Accuracy
	Formulate Math	0.61	0.56	0.49	BERTScore
	Perform Computation	0.54	0.48	0.42	1-NRMSE

1954

- **Two-stage math degradation:** Task V shows formulation gap (Step 1→2: ~20 points) and execution gap (Step 2→3: ~8 points)

1960
1961
1962
1963

These patterns replicate across different architectures, validating our framework’s diagnostic utility.

1964
1965
1966

K MITIGATION STRATEGY

1969
1970
1971
1972**Table 13:** Effect of targeted mitigation strategies across representative ESTBOOK tasks. Mitigation strategies include Evidence-Anchored CoT (Task I), Syntax-First CoT (Task III), Symbolic Verification (Task V), and Table-Alignment Constraints (Task IV).

Task	GPT-4V		GPT-5		Claude-Sonnet-4	
	CoT	Mitigation	CoT	Mitigation	CoT	Mitigation
GRE-RC (Task I)	77.8	83.5	87.1	90.4	69.3	74.1
GRE-TC (Task III)	79.5	84.2	73.4	78.1	61.0	66.7
GRE-NE (Task V)	38.0	46.8	33.9	42.1	30.1	35.4
GMAT-IR (Task IV)	13.8	20.7	16.0	22.9	15.0	19.4

1979

Based on insights from breakdown analysis, we additionally propose a suite of task-specific mitigation strategies designed to improve LLM reasoning across representative ESTBOOK tasks. These strategies are motivated by the failure patterns identified through our breakdown analysis and target weaknesses observed in four core task categories: evidence extraction, structural reasoning, numeric formulation, and multimodal data interpretation.

1980
1981
1982
1983
1984

First, for reading comprehension tasks (Task I), we introduce an Evidence-Anchored Chain-of-Thought mitigation that explicitly enforces grounding before reasoning. Instead of allowing the model to jump directly into high-level interpretation, the prompt requires the LLM to first extract one or two verbatim text spans from the passage and then articulate a justification that links these spans to the question. Only after this grounding step is the model permitted to produce a final answer. This workflow directly tackles the common error where models hallucinate or misalign evidence, and the structured evidence-first constraint significantly reduces this drift, especially on GRE and TOEFL reading sections.

1993
1994
1995
1996
1997

Second, for structural reasoning tasks such as GRE Text Completion and GMAT sentence-based items (Task III), we implement a Syntax-First CoT strategy. Our failure analysis shows that LLMs often conflate semantic plausibility with grammatical role and textual structure, leading to incorrect fill-ins. To mitigate this, we require the model to explicitly identify the syntactic role of the blank (e.g., modifying clause, concessive marker, verb complement) and to outline the structural constraints implied by the surrounding sentence (e.g., contrast markers, valence polarity, causal structure). Only

1998
1999
2000
2001

Table 14: 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.

2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051

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

then does the model evaluate candidate options. This method reduces semantic drift and improves consistency, particularly in multi-blank completion tasks where structural information is critical.

Third, numeric-entry and symbolic reasoning tasks (Task V) often fail because the model constructs incorrect equations, even when it correctly understands the verbal description. To address this, we adopt a Symbolic Verification Layer, in which the model must rewrite the mathematical relations from the question in a purely symbolic form and then re-check each expression against the original text line-by-line. Only after this verification step does the model perform the computation. This procedure substantially reduces errors such as misinterpreting “three consecutive odd integers” as “three consecutive integers,” a recurring failure mode in GRE-NE and SAT/GRE algebra problems.

Finally, multimodal tasks involving tables, charts, or structured numerical information (Task IV) frequently fail due to incorrect row/column alignment or misinterpretation of tabular fields. For these tasks, we introduce a Table-Alignment Constraint prompting method, which requires the model to translate the textual question into explicit table lookup instructions before beginning numerical operations. The model must specify which row(s) and column(s) are relevant and articulate the mapping between textual descriptions and table headers. This step significantly reduces errors in GMAT Integrated Reasoning and GRE Data Interpretation, where models previously selected the wrong table element or mismatched textual cues with table structure.

The mitigated results are shown at Table 13.

L ADVANCEMENT ON LLMs

To further demonstrate that our breakdown analysis yields actionable improvements rather than purely diagnostic insight, we additionally explored two complementary enhancement strategies for open-source models—breakdown-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 breakdown 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 breakdown 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 breakdown-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 Breakdown-Adaptive Prompting (AdaptCoT) strategy that requires no parameter tuning and instead selects a task-specific CoT template based on the breakdown 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

2052 requires explicit extraction of mathematical relations, symbolic rewriting, and column-row alignment
2053 steps before any computation. This routing mechanism ensures that the model’s reasoning structure
2054 matches the cognitive decomposition expected for each question type.

2055 To quantify the effectiveness of these interventions, we re-evaluated Llama-4-Scout-17B and Qwen-
2056 VL-Max on representative ESTBOOK tasks using CoT decoding only. The results show consistent
2057 performance improvements across all task categories. Fine-tuned models (CurrCoT) yield the largest
2058 gains, particularly for mathematically intensive tasks such as GRE Numeric Entry and GMAT Problem
2059 Solving, where Llama improves from 30.1% to 38.1% and from 22.5% to 29.5%, respectively, and
2060 Qwen increases from 28.1% to 35.1% and 25.6% to 31.6%. Even the prompt-only AdaptCoT method
2061 yields noticeable increases—for example, improving Llama’s GRE-TC accuracy from 61.0% to
2062 64.0% and GRE-RC from 54.2% to 58.2%. These results confirm that the breakdown taxonomy does
2063 not merely describe failure cases but actively guides the design of interventions that translate into
2064 measurable improvements in reasoning quality for open-source LLMs.

2065

2066

2067

2068

2069

2070

2071

2072

2073

2074

2075

2076

2077

2078

2079

2080

2081

2082

2083

2084

2085

2086

2087

2088

2089

2090

2091

2092

2093

2094

2095

2096

2097

2098

2099

2100

2101

2102

2103

2104

2105