

Can LLMs Estimate Cognitive Complexity of Reading Comprehension Items?

Seonjeong Hwang¹, Hyounhun Kim^{1,2}, Gary Geunbae Lee^{1,2}

¹Graduate School of Artificial Intelligence, POSTECH, Republic of Korea

²Department of Computer Science and Engineering, POSTECH, Republic of Korea
{seonjeongh, h.kim, gblee}@postech.ac.kr

Abstract

Estimating the cognitive complexity of reading comprehension (RC) items is crucial for assessing item difficulty before it is administered to learners. Unlike syntactic and semantic features, such as passage length or semantic similarity between options, cognitive features that arise during answer reasoning are not readily extractable using existing NLP tools and have traditionally relied on human annotation. In this study, we examine whether large language models (LLMs) can estimate the cognitive complexity of RC items by focusing on two dimensions—Evidence Scope and Transformation Level—that indicate the degree of cognitive burden involved in reasoning about the answer. Our experimental results demonstrate that LLMs can approximate the cognitive complexity of items, indicating their potential as tools for prior difficulty analysis. Further analysis reveals a gap between LLMs’ reasoning ability and their metacognitive awareness: even when they produce correct answers, they sometimes fail to correctly identify the features underlying their own reasoning process.

1 Introduction

Estimating the difficulty of reading comprehension (RC) items¹ is essential for delivering appropriate learning materials and constructing balanced test forms. Traditionally, difficulty has been derived from student responses using frameworks such as classical test theory (CTT) or item response theory (IRT) (Lord, 1980; Hambleton and Jones, 1993). However, these approaches are only applicable after test administration and thus cannot support prior difficulty prediction during item development. Expert judgment remains the common alternative, but it is costly, time-consuming, and subject to

¹An RC item typically consists of a passage, a question stem, and sometimes answer options (in the case of the multiple-choice format). In this paper, we use the term “item” interchangeably with “question.”

Passage

- (1) Mia loved baking and often spent her weekends experimenting with new recipes.
- (2) Her specialty was a chocolate cake that her family couldn't get enough of.
- (3) Every Sunday afternoon, the smell of cocoa and vanilla would fill the house.
- (4) Her younger brother, Sam, always asked for a second slice.
- (5) One weekend, Mia decided to enter the local baking competition.
- (6) She worked hard all week preparing the perfect version of her cake.
- (7) The judges praised the moist texture and rich flavor, and Mia won first place.
- (8) After the win, she was invited to share her recipe in the local newspaper.
- (9) Although she was nervous, she accepted the invitation and was excited to see her name in print.
- (10) Mia started thinking about attending a culinary school in the future.

Statement	Evidence Scope & Transformation Level
Mia's specialty was a chocolate cake.	Single-sentence evidence (s1) Word Matching
Mia's house would be filled with the smell of cocoa and vanilla every Sunday afternoon.	Single-sentence evidence (s3) Transformed Word Matching
Despite feeling nervous, Mia was thrilled to have her name published.	Single-sentence evidence (s9) Paraphrasing
Mia received an invitation to publish her recipe in the newspaper following her victory.	Single-sentence evidence (s8) Transformed Paraphrasing
Mia's cake was better than the other contestants' cakes.	Single-sentence evidence (s7) Inference
Mia loved baking and won first place.	Multi-sentence evidence (s1&s7) Word Matching
Mia regularly tried new recipes on weekends and chose to join a baking contest.	Multi-sentence evidence (s1&s5) Paraphrasing
Mia's success in the contest encouraged her to consider becoming a professional baker.	Multi-sentence evidence (s7&s10) Inference
Sam wanted to learn baking from her.	Insufficient evidence

Figure 1: Examples of RC items that require determining the factuality of a statement. Each item is annotated along two cognitively grounded dimensions (*Evidence Scope* and *Transformation Level*) with corresponding supporting sentences highlighted from the passage.

rater variability (AlKhuzayy et al., 2024). Another line of work has attempted to simulate student responses using deep learning models of varying capability (Lalor et al., 2019; Uto et al., 2023; Park et al., 2024), thereby enabling difficulty estimation without real student data. However, this approach incurs substantial computational costs for training or running inference over many models, and the resulting difficulty estimates lack interpretability in terms of what makes an item difficult.

In parallel, researchers have explored a complementary direction: leveraging item features to understand the relationship between item characteristics and difficulty in a more interpretable way (Pandiarova et al., 2019; Choi and Moon, 2020;

Benedetto et al., 2021; Byrd and Srivastava, 2022). Syntactic, semantic, and psycholinguistic variables (e.g., sentence length, semantic similarity between options, and word familiarity) have been predominantly leveraged for this analysis, yet these text-based features provide limited insight into the reasoning processes that largely govern difficulty.

Educational psychology research has shown that cognitive factors involved in the answer decision process are more strongly associated with difficulty (Embretson and Wetzel, 1987). Examples include the amount of text that must be referenced to determine the correct answer and the degree of transformation between passage evidence and the answer (Bormuth et al., 1970; Anderson, 1972). However, these cognitive features cannot be automatically extracted with existing NLP tools, and prior studies have relied exclusively on human raters (Hutzler et al., 2014; Lai et al., 2017). This raises a key challenge: *How can we estimate the cognitive complexity of RC items in a scalable way?* We believe that large language models (LLMs), with their powerful reasoning and instruction-following capabilities, may offer a promising approach to this problem.

Recent studies have attempted to leverage LLMs for estimating question difficulty. However, many of these efforts have focused on tasks solvable solely through a model’s internal knowledge—such as mathematics and coding (Rogoz and Ionescu, 2024; Park et al., 2024; Ko et al., 2024; Xu et al., 2024)—or have directly prompted LLMs to predict the difficulty of RC items (Raina and Gales, 2024). Whether LLMs can meaningfully analyze the cognitive complexity involved in solving RC items, however, remains largely unexplored.

Motivated by this gap, we investigate LLMs’ capability to measure the complexity of two cognitive variables: **Evidence Scope** and **Transformation Level**. Evidence Scope reflects the amount of text required to verify an answer—categorized as single-sentence, multi-sentence, or insufficient—while Transformation Level captures the degree of lexical and structural transformation between an option and its supporting evidence in the passage, ranging from word matching to inference. To support empirical evaluation, we constructed **RECO**, a benchmark reading comprehension dataset with cognitive complexity annotations, comprising 776 RC items annotated along these two cognitive dimensions (see Figure 1).

In our experiments, we evaluate eight LLMs,

spanning both proprietary and open-source variants. The results show that LLMs can approximate cognitive complexity, with the best-performing models achieving Macro F1 scores of 74.8 (Evidence Scope) and 82.0 (3-level Transformation Level) in the cognitive complexity classification tasks. Notably, open-source models such as Qwen2.5 (32B) and Mistral-Small (24B) performed comparably to—or even surpassed—GPT-4o. However, we also found that LLMs often struggle to explicitly recognize key features in their reasoning traces—for example, identifying phrase reordering or the evidence sentences they referenced—highlighting a gap between reasoning ability and metacognitive awareness.

Our contributions can be summarized as follows:

- We construct an expert-annotated dataset of RC items along two cognitively grounded dimensions, which are important factors for analyzing item difficulty.²
- We conduct an evaluation of eight instruction-tuned LLMs, demonstrating their potential utility in estimating the cognitive complexity of RC items.
- We probe LLMs on fine-grained cognitive features and observe that, even when they successfully solve items, they do not fully recognize the cognitive processes underlying their problem solving.

2 Related Work

2.1 Difficulty Factors and Taxonomies

Research in educational psychology has long examined the factors influencing RC item difficulty (Bormuth et al., 1970; Anderson, 1972; Freedle and Kostin, 1991; Park, 2004; Rafatbakhsh and Ahmadi, 2023). Researchers have sought to identify correlations between item-based attributes—such as surface-level linguistic features (e.g., sentence complexity, vocabulary difficulty) and cognitive burden factors (e.g., plausibility of distractors, option–text mapping)—and item difficulty, often measured using CTT or IRT (Hsu et al., 2018; Pandarova et al., 2019; Choi and Moon, 2020; Zhou and Tao, 2020; Benedetto et al., 2021). More recently, tools such as Coh-Metrix (McNamara et al., 2014), NLTK (Loper and Bird, 2002), and embedding models like Word2Vec (Mikolov et al., 2013) or BERT (Devlin et al., 2019) enable the

²The dataset, prompt templates, and evaluation codes are available at <https://github.com/SeonjeongHwang/ReCo>.

automatic extraction of syntactic and semantic features. These features have been used as inputs to supervised models such as linear regression for difficulty prediction.

Several taxonomies have been proposed to systematically classify the complexity of RC items. Bloom’s taxonomy, for instance, organizes learning objectives by levels of cognitive demand (Bloom et al., 1956). Lai et al. (2017) employed a five-level taxonomy to characterize reasoning across RC items, which is a simplified combination of the two dimensions adopted in our study. While previous taxonomies have been used to categorize different types of RC items—such as main idea, author’s intent, fill-in-the-blank, and detail information questions—our study focuses on distinguishing variations in cognitive complexity, even within items of the same type.

2.2 LLM-based Difficulty Estimation

Recently, various approaches have been proposed to leverage LLMs for predicting item difficulty. Some studies directly prompted LLMs to estimate difficulty (Xu et al., 2024), while others inferred difficulty from model-generated outcomes such as answering accuracy or confidence scores (Rogoz and Ionescu, 2024; Park et al., 2024; Lu and Wang, 2024; Jain et al., 2025). However, much of this prior work has focused on domains such as mathematics, medicine, or coding, where models rely solely on their internal knowledge to solve the problems. This differs from RC, where the model must reference the information provided in the passage while applying its own reasoning ability.

Several studies have explored LLMs for predicting RC item difficulty. Raina and Gales (2024) found that comparative prompting—asking an LLM to compare the relative difficulty of two items—aligned better with human judgments than absolute prompting, where the model assigns a difficulty score to a single item. Dutulescu et al. (2024) predicted item difficulty using indicators derived from LLMs’ question answering (QA) loss. Kapoor et al. (2025) showed that combining item text features, LLM embeddings, and contextual information (e.g., grade level, year) improved prediction performance, underscoring the importance of item-feature-based analyses. Instead of directly predicting item difficulty, this paper investigates whether LLMs can estimate two cognitively grounded features that influence difficulty.

2.3 Datasets

While QA datasets such as SQuAD (Rajpurkar et al., 2016) and BoolQ (Clark et al., 2019) are widely used, they lack per-item difficulty annotations, limiting their usefulness for difficulty analysis. RACE++ (Lai et al., 2017; Liang et al., 2019) contains RC items spanning middle school through college levels and has been used with grade level as a proxy for difficulty (Raina and Gales, 2024; Liusie et al., 2023), but this approach does not capture fine-grained variation within a single learner group. Multi-hop QA datasets such as HotpotQA (Yang et al., 2018) assess complexity through multi-hop reasoning across documents, yet this setup differs from standard RC formats in educational assessment, which typically involve a single reading passage.

Huang et al. (2017) predicted item difficulty using student error rates on English reading problems in China, but the full dataset was not released. Mullooly et al. (2023) released the CMCQRD dataset, which contains 289 RC items labeled with CEFR levels and IRT-based difficulty scores derived from pretesting, providing holistic estimates of item difficulty. Dutulescu et al. (2024) annotated Fairy-taleQA (Xu et al., 2022) along two dimensions: explicit vs. implicit and local vs. summary. This dataset is the closest to our ours, but the two dimensions considered in this work capture more fine-grained cognitive features involved in the answer decision process of RC items.

3 Data Construction

To construct our dataset, we used True/False/Not Given (TFNG) items, where the task is to assess the factuality of a statement given a passage. Each item comprises a reading passage and a declarative statement, as illustrated in Figure 1. This format is particularly suitable for our study, as it spans a wide range of cognitive complexity—from direct span matching to multi-sentence inference—and is commonly featured in the RC sections of standardized proficiency exams.

3.1 Dimensions of Cognitive Complexity

Evidence Scope. Items that can be solved by referencing a single sentence in the passage are generally easier than those requiring integration of information scattered across multiple sentences (Bor-muth et al., 1970; Park, 2004). In this study, Evidence Scope refers to the span of text required

to determine the truth value of a statement, and is categorized into three levels:

- **Single-sentence evidence:** All necessary information to evaluate the statement is contained within a single sentence in the passage.
- **Multi-sentence evidence:** The required information is distributed across multiple sentences (i.e., inter-sentence comprehension).
- **Insufficient evidence:** The passage lacks adequate information to definitively confirm or refute the statement. In such cases, learners are required to examine the entire passage before concluding that the passage provides no supporting evidence.

A special case arises when the supporting evidence includes anaphoric expressions. While [Bormuth et al. \(1970\)](#) treated such items as a separate category, they found little difference in difficulty compared to single-sentence evidence. Accordingly, we apply the following rules: 1) If the anaphora clearly refers to a frequently mentioned and easily identifiable entity in the prior sentences, the item is classified as single-sentence evidence. 2) However, if resolving the anaphora requires referring back to a prior sentence, we label it as multi-sentence evidence. This approach reflects that many reading passages in RC assessments employ anaphoric references, and accounting for such subtleties is essential for accurate difficulty prediction.

Transformation Level. When the degree of transformation between a statement and its supporting evidence is higher, identifying the corresponding passage text and assessing the statement’s truth value imposes greater cognitive demands ([Bormuth et al., 1970](#); [Anderson, 1972](#)). We adopt a *5-level* taxonomy inspired by previous work, which captures the type of transformation required to derive a statement from the evidence:

- **Word Matching:** The content words in the statement appear verbatim in the evidence, and the phrase order is preserved.
- **Transformed Word Matching:** The content words are still present in the evidence but have been rearranged.
- **Paraphrasing:** The statement rephrases the content words without changing the order of the words.
- **Transformed Paraphrasing:** The content words are rephrased and the phrase order is altered, combining lexical and structural transformation.

- **Inference:** The statement cannot be directly derived from any surface form in the passage, even through paraphrasing or reordering; instead, it requires inference.

In contrast to the *single-sentence evidence* cases, phrase reordering is either trivial or pervasive in *multi-sentence* cases; therefore, we label these items using a simplified *3-level* taxonomy: *word matching*, *paraphrasing*, and *inference*. In addition, anaphora resolution, identifying the antecedent of an anaphor within a text, is not considered paraphrasing unless additional lexical transformation is involved.

3.2 Data Annotation

RACE++ ([Lai et al., 2017](#); [Liang et al., 2019](#)) is a RC dataset consisting of English RC items sourced from exams administered to Chinese middle school, high school, and college students. To construct RECO, we use Multiple-choice True/False (MTF) items from the RACE++, each consisting of a reading passage and four options. We collected only items requiring holistic passage comprehension, excluding those targeting specific entities or local paragraph-level details. Items from the middle and high school levels were drawn from the test split, while college-level items were taken from the validation and test splits, due to limited data in its test set. Each MTF item was divided into four True/False/Not Given (TFNG) items, yielding triplets of (passage, statement, factuality label [*True* or *Not True*]).

Three experts with prior experience tutoring students for standardized English exams or authoring RC items independently identified the evidence sentences in the passage and labeled each item along the two cognitive dimensions. For statements identified as *False* within the *Not True* cases, they produced minimally revised *True* statements to enable annotation of transformation level. We retained only items where at least two annotators agreed on the same label; in cases of partial agreement, discrepancies were resolved through author adjudication. The resulting annotated dataset, RECO, is released for non-commercial research purposes under the RACE license. Further details on the annotation process and inter-annotator agreement are provided in [Appendix A](#).

3.3 Data Statistics

[Table 1](#) presents the statistics for the ReCo demonstration and test sets, and [Table 2](#) summarizes the

Split	# Passage	# Statement	
		<i>True</i>	<i>Not True</i>
Test set	151	236 (409)	262
		498 (671)	
Demonstration set	83	129 (222)	149
		278 (371)	

Table 1: Number of passages and statements (*True / Not True / Total*) in the test and demonstration sets of ReCo. Parentheses indicate *True* statements including annotator-revised versions of *False* statement.

Evidence Scope			
Single-sentence Evidence			388
Multi-sentence Evidence			243
Insufficient Evidence			145
Transformation Level			
<i>5-level</i>		<i>3-level</i>	
Word Matching	73	Word Matching	123
Transformed Word Matching	36		
Paraphrasing	55	Paraphrasing	189
Transformed Paraphrasing	78		
Inference	146	Inference	319

Table 2: Distribution of examples in the RECO dataset across Evidence Scope and Transformation Level. For the Transformation Level dimension, items labeled as insufficient evidence are excluded, and the 5-level scheme applies only to single-sentence comprehension items.

distribution of items across labels. For Evidence Scope, 50% items are labeled as single-sentence evidence, while insufficient evidence items account for the lowest proportion. For Transformation Level, items requiring inference are the most frequent, whereas transformed word matching items are the least common. In the 3-level scheme, which includes multi-sentence comprehension items, inference items remain dominant. These statistics reveal an imbalanced label distribution in our dataset, which is expected because the items are drawn from real exams, whose distributions may vary depending on the target proficiency level. Further analysis of our dataset can be found in Appendix B.

4 Experimental Setup

We formulate the measurement of cognitive complexity as a classification task. In the **Evidence Scope (ES) classification**, the model receives an instruction, a passage, a statement, and its factuality label (*True* or *Not True*), and predicts one of three evidence types: *single*, *multi*, or *insufficient*. In the **Transformation Level (TL) classification**,

applied to items with *True* statements, the model estimates the degree of transformation using the task definition, passage, and statement. We report performance using both the 5-level taxonomy—*word matching (WM)*, *transformed word matching (TWM)*, *paraphrasing (P)*, *transformed paraphrasing (TP)*, and *inference (I)*—and a simplified 3-level version (*WM*, *P*, *I*), which omits distinctions based on phrase reordering. In the 3-level setting, predictions of TWM and TP are mapped to WM and P, respectively, and ground-truth labels for single-sentence evidence items are converted accordingly. We report model performance using Macro F1; Micro F1 scores are provided in Appendix D.

We evaluate eight *instruction-tuned* LLMs, including open-source models Gemma2-9B/27B (Team, 2024a), Mistral-7B/24B (Jiang et al., 2023), and Qwen2.5-7B/32B (Team, 2024b), as well as proprietary models GPT-4o and GPT-4o-mini (Hurst et al., 2024). Two prompting strategies are considered: Standard Prompting (SP), where the model receives a task definition and input and returns a label directly; and Chain-of-Thought Prompting (CoT), which encourages step-by-step reasoning before prediction (Wei et al., 2022). Greedy decoding is used as the default inference method across all models and prompting strategies. In the CoT setting, we additionally apply self-consistency decoding (Wang et al., 2022), generating 10 samples with top- $k=20$, top- $p=0.8$, and temperature=0.7 and using priority answer.

Each strategy is evaluated under three prompting conditions—zero-shot, one-shot, and few-shot (Brown et al., 2020)—with the few-shot demonstrations covering items from all labels. Exemplars are sampled from the RECO demonstration split and fixed across models to ensure consistency. To filter out overly trivial items that might inflate model performance, we exclude those that GPT-4o correctly classifies with a zero-shot CoT prompt. Details on model versions and the experimental environment are provided in Appendix C.

5 Results

Table 3 presents the performance of LLMs and human experts on the ES and TL classification tasks. Human performance is computed using annotators’ initial labels, before applying inter-annotator agreement filtering and adjudication. The table also reports LLMs’ performance on the RC task, which

Method	#Demo	Gemma2		Mistral		Qwen2.5		GPT-4o	
		9B	27B	7B	24B	7B	32B	mini	-
<i>Reading Comprehension</i>									
CoT	1	78.6	82.4	59.8	74.2	74.8	82.8	80.6	84.4
<i>Evidence Scope Classification [Human: 87.0]</i>									
SP	0	<u>48.6</u>	53.0	43.0	58.4	45.6	57.4	49.6	58.5
	1	47.1	<u>56.6</u>	<u>44.7</u>	<u>60.2</u>	47.7	<u>61.3</u>	<u>52.7</u>	62.0
	6	48.5	51.7	39.4	56.8	<u>48.5</u>	58.9	52.0	<u>65.6</u>
CoT	0	53.6	58.3	15.9	62.9	<u>55.0</u>	70.0	63.4	71.1
	1	<u>61.3</u>	<u>69.7</u>	<u>53.1</u>	65.9	53.0	<u>73.4</u>	<u>66.7</u>	74.8
	6	56.7	65.9	50.0	<u>66.4</u>	53.8	70.5	63.0	68.4
CoT (SC)	0	55.5	63.4	12.7	67.2	55.6	73.1	65.1	71.8
	1	64.2	71.5	58.0	70.9	57.5	73.5	71.0	73.2
	6	58.3	68.1	49.6	65.9	51.7	71.0	66.5	72.3
<i>(3-level) Transformation Level Classification [Human: 84.9]</i>									
SP	0	60.4	50.8	46.9	72.9	<u>55.3</u>	53.7	53.1	<u>62.3</u>
	1	54.2	<u>55.1</u>	56.2	<u>73.0</u>	50.5	<u>56.5</u>	<u>53.9</u>	59.3
	8	54.6	51.9	42.4	70.4	42.3	46.6	45.4	55.3
CoT	0	<u>52.9</u>	51.4	43.3	<u>77.2</u>	62.3	66.1	59.9	73.9
	1	49.9	53.7	44.9	<u>70.2</u>	69.4	<u>68.6</u>	62.3	72.9
	8	50.9	<u>55.0</u>	<u>46.6</u>	71.3	56.8	67.2	65.1	68.4
CoT (SC)	0	58.5	55.9	<u>55.2</u>	82.0	<u>68.7</u>	74.0	58.4	69.9
	1	49.6	51.7	41.1	63.0	61.0	69.3	<u>60.3</u>	68.0
	8	57.2	56.3	43.0	72.9	58.4	65.7	58.6	61.0
<i>(5-level) Transformation Level Classification [Human: 83.0]</i>									
SP	0	32.2	27.4	22.1	45.6	31.5	32.1	28.6	39.8
	1	37.7	30.7	23.2	<u>51.7</u>	<u>33.0</u>	<u>35.0</u>	27.5	<u>40.5</u>
	8	30.2	<u>33.9</u>	<u>26.8</u>	47.5	24.4	28.4	27.6	37.7
CoT	0	32.1	32.9	24.4	<u>55.0</u>	35.1	<u>52.2</u>	43.5	61.3
	1	32.0	32.8	26.3	42.6	<u>42.7</u>	45.4	44.5	60.0
	8	<u>35.7</u>	<u>34.0</u>	<u>27.6</u>	45.4	39.6	43.4	43.3	49.6
CoT (SC)	0	33.8	38.9	26.6	58.7	31.9	56.0	<u>42.1</u>	53.1
	1	35.3	35.9	28.0	38.4	44.1	48.2	40.0	54.6
	8	<u>36.9</u>	32.1	26.4	48.3	35.6	43.4	37.5	43.4

Table 3: Performance of LLMs on the RC task and the ES and TL classification tasks. Greedy decoding is the default inference method, and SC indicates that self-consistency decoding is used. **Bolded** values denote each model’s best score per task; underlined values indicate the best score within each method group across demonstration settings.

requires the model to determine whether a statement is true or not based on the passage.

Before evaluating LLMs’ ability to analyze the cognitive complexity, we first examined their performance on the RC task itself. According to the results, all models except the Mistral family achieved Macro F1 scores above 74.0, with larger models approaching 85.0, indicating strong reading comprehension ability. This result confirms that the RC items in RECO are generally easy for current LLMs, and thus that errors in cognitive complexity prediction are unlikely to stem from failures in basic comprehension or answer reasoning. However, for smaller models such as Mistral-7B, incomplete comprehension may still contribute to some degree

of performance variation.

In the ES classification task, GPT-4o achieves the highest score (74.8) with a one-shot CoT prompt, and Qwen2.5-32B performs comparably (73.5). Yet all models fall short of expert performance (87.0), highlighting the difficulty of modeling human cognitive processes in evidence selection. Within open-source model families, larger models tend to yield comparable performance across variants—particularly under self-consistency decoding—, while smaller models exhibit more divergent results. Mistral-7B, in particular, exhibits unstable behavior, occasionally yielding unexpectedly low scores despite identical prompts. As expected, zero-shot prompting underperforms compared to demonstration-based settings; however, in large models, few-shot prompting occasionally led to performance degradation.

In the TL classification task with the 3-level taxonomy, Mistral-24B (82.0) and Qwen2.5-32B (74.0) outperform GPT-4o (73.9), approaching human performance (84.9). Under the 5-level taxonomy, however, overall model performance decreases, underscoring the challenge of capturing phrase reordering between statements and evidence. Nevertheless, open-source models again achieve performance comparable to GPT-4o. By contrast, Gemma2-27B consistently underperforms, trailing even smaller models, which suggests limitations in handling lexical and syntactic transformations. Prompting effects are less consistent than in ES classification: optimal configurations vary by model, and in some cases zero-shot prompting even outperforms demonstration-based prompting. Self-consistency decoding improves results in several settings, particularly zero-shot CoT, but its gains are less stable than in ES classification.

In summary, **while LLMs do not fully align with expert judgments, they demonstrate strong potential as estimators of cognitive complexity in RC items, especially in ES and 3-level TL classifications.** Notably, the competitive performance of open-source models relative to GPT-4o suggests that reliance on proprietary LLMs may not be necessary for analyzing cognitive complexity. While we employ simple prompting methods to better isolate inherent model capabilities, more advanced prompt engineering may yield further improvements.

Subtask	Gemma2		Mistral		Qwen2.5		GPT-4o	
	9B	27B	7B	24B	7B	32B	mini	-
1.1. Falsifiability Judgment	80.6	77.0	65.9	80.8	76.0	88.6	77.0	85.5
1.2. Evidence Sentence Counting	46.4	47.7	41.1	44.0	43.7	47.6	41.0	44.6
2.1. Inference Detection	72.9	75.3	65.4	82.4	66.6	81.6	81.4	79.8
2.2. Paraphrasing Detection	81.3	84.3	66.6	84.9	72.0	85.3	87.4	88.5
2.3. Phrase Reordering Detection	58.0	67.9	58.6	61.0	41.3	47.5	66.6	63.3

Table 4: LLM performance on subtasks measuring fine-grained abilities required for the main classification tasks. **Bolded** values denote the best model for each subtask.

6 Analysis

6.1 Fine-Grained Feature Analysis

To better understand the capabilities and limitations of LLMs, we decomposed the two classification tasks into a set of fine-grained subtasks. For each subtask, we constructed few-shot CoT prompts by adapting those used in the main experiments, with instructions tailored to the specific cognitive feature under evaluation. Results are reported in Table 4.

The ES classification task was divided into two core subtasks: **Subtask 1.1: Falsifiability Judgment** – determining whether a *Not True* statement is *False* (contradicted) or *Not Given* (lacking sufficient evidence). **Subtask 1.2: Evidence Sentence Counting** – identifying how many sentences are required to support or refute a statement. Most models performed reliably on falsifiability classification, but performance was substantially lower for evidence sentence counting. This suggests that while LLMs can distinguish between refuted and unsupported statements, they struggle to accurately identify the full set of sentences referenced by human annotators when solving the item.

The TL classification task was evaluated through a hierarchical breakdown of transformation types, allowing for a more detailed examination of how LLMs handle different forms of linguistic transformation: **Subtask 2.1. Inference Detection** – distinguishing inference-based statements from those explainable by surface-level transformations, such as paraphrasing or phrase reordering. **Subtask 2.2. Paraphrasing Detection** – identifying whether a statement is a lexical rephrasing or a verbatim restatement. **Subtask 2.3: Phrase Reordering Detection** – detecting reordering of words and phrases. According to the results, model performance showed distinct strengths and weaknesses across subtasks. GPT-4o-mini performed comparably to, or even better than, GPT-4o—especially on inference and phrase reordering detection, where

GPT-4o underperformed. Paraphrasing detection was handled well across models, while phrase reordering remained especially challenging.

Overall, LLMs showed consistent performance on falsifiability classification (Subtask 1.1) and paraphrase detection (Subtask 2.2), but struggled with evidence sentence counting (Subtask 1.2) and phrase reordering detection (Subtask 2.3). While models demonstrated stable answer prediction performance (as shown in Table 3), our analysis indicates that even when they solved items correctly, they often failed to explicitly capture the cognitive features underlying their reasoning process, pointing to a limitation in metacognitive awareness.

6.2 Error Analysis

We further examined the prediction tendencies of GPT-4o and the best-performing model in each task. Figure 2 shows the distribution of the number of evidence sentences predicted by Gemma2-27B and GPT-4o, compared to human-selected evidence. LLMs tend to select fewer sentences than humans, often defaulting to a single sentence as evidence. Table 5 reports the degree of alignment between the evidence sentences selected by each model and those identified by experts. As in the previous observation, recall scores were consistently lower than corresponding precision scores, indicating that models often fail to explicitly retrieve all the sentences required by humans to solve the item. One possible explanation is that LLMs—by encoding the entire passage before engaging in step-by-step reasoning—tend to assign insufficient attention to sentences containing seemingly minor details that nevertheless exert a significant influence on the answer choice.

Figure 3 presents the distribution of TL predictions for items with *single-sentence evidence*. LLMs consistently achieved higher precision for *WM* and *I* items than for other categories. However, they frequently misclassified *TWM* as *WM*, failing

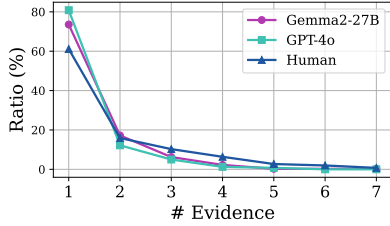


Figure 2: Distribution of the number of evidence sentences selected by LLMs and humans.

Evidence Selection	Precision	Recall	F1
Gemma2-27B	86.4	78.3	78.8
Mistral-24B	82.4	74.4	74.3
Qwen2.5-32B	85.4	76.9	77.4
GPT-4o	88.8	79.2	80.0

Table 5: Model performance on evidence sentence selection. For each instance, precision, recall, and F1 were computed by comparing the predicted and reference evidence sets over their union, and the final scores were obtained by averaging these instance-level values across the dataset.

to detect subtle reordering of phrases. Models also showed confusion between *P* and *TP*, often misclassifying them as each other or as *I*. These patterns align with the earlier results in Table 4, particularly for phrase reordering detection, where all models struggled. Representative error cases illustrating these confusions are analyzed in the case study.

6.3 Evaluating the Effect of Deep Reasoning on Cognitive Complexity Estimation

We investigated whether LLMs with advanced reasoning capabilities are also effective at analyzing cognitive complexity. For this experiment, we used Qwen3-32B (Team, 2025), a reasoning-specialized model that supports a “thinking mode” designed for deep reasoning and self-reflection. We compared the performance of Qwen3-32B in two settings— with and without thinking mode—on the cognitive difficulty classification tasks. As a baseline, we also included Qwen2.5-32B, a model from the same family but not specialized for reasoning.

In Table 6, we observed that Qwen3-32B in thinking mode underperformed its non-thinking-mode counterpart, achieving lower F1 scores across both classification tasks. These results suggest that advanced reasoning and self-reflection capabilities do not necessarily enhance a model’s ability to classify cognitive complexity features such as ES and TL. This may be because these tasks

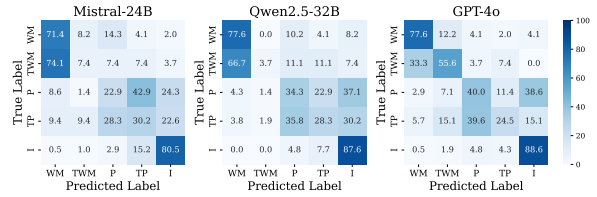


Figure 3: Distribution of TL labels predicted by LLMs for single-sentence evidence items.

Model	ES	TL (5-level)	TL (3-level)
Qwen2.5-32B	70.0	52.2	66.1
Qwen3-32B _{non-thinking}	65.4	58.6	78.2
Qwen3-32B _{thinking}	62.6	56.3	69.4

Table 6: Performance comparison of LLMs with and without deep-thinking capabilities.

do not require complex multi-step reasoning, but rather fine-grained categorization of human cognitive processes—something better handled through intuitive pattern recognition than abstract reasoning. The experimental results align with this interpretation and highlight the distinction between deep reasoning ability and the metacognitive awareness.

6.4 Case Study

Figure 4 presents error cases that illustrate common failure patterns in LLMs. We analyze two examples where GPT-4o incorrectly classifies items in the ES and TL classification tasks. In Case #1, the model misclassifies an item requiring *multi-sentence evidence* as *single-sentence evidence*. The passage involves multiple characters in dialogue, and correctly identifying the speaker of a specific utterance is essential for determining the truth of the statement. According to annotators, both Sentence (15) and Sentence (16) are required: Sentence (15) establishes that the subject is “Sadness,” while Sentence (16) describes her action. However, the model references Sentence (15) in its explanation and correctly links “Sadness” to the action in Sentence (16), yet asserts: “Sentence (16) explicitly mentions Sadness’s need to be by herself.” The case reveals a metacognitive failure: the model arrives at the correct factual judgment but fails to recognize the reasoning behavior it engaged in to reach that conclusion.

In Case #2, the item labeled as *TWM* was incorrectly classified as *WM*. A comparison with Sentence (4) reveals a reordering of content words caused by a shift to the passive construction. How-

<i>Case #1</i>	
<p>Passage: (1) Once upon a time, there was an island where all the feelings lived: Happiness, Sadness, Knowledge, and all of the others, including Love. (...) (15) Sadness was close by so Love asked, "Sadness, let me go with you." (16) "Oh...Love, I am so sad that I need to be by myself!" (17) Happiness passed by Love, too, but she was so happy that she did not even hear when Love called her. ... Statement: Sadness wanted to stay alone. Ground Truth ES: Multi-sentence evidence</p>	<p>GPT-4o's Response: 1. Understand the Statement - Statement: "Sadness wanted to stay alone." - Key claim: Sadness had a desire to be alone. 2. Retrieve Relevant Information from the Passage - Sentence (15): Love asks Sadness if she can go with her. - Sentence (16): Sadness responds, "Oh...Love, I am so sad that I need to be by myself!" 3. Assess Scope of Evidence - Sentence (16) directly supports the statement by indicating Sadness's desire to be alone. - <u>The sentence explicitly mentions Sadness's need to be by herself</u>, which aligns with the statement that Sadness wanted to stay alone. (...) Answer: Single</p>
<i>Case #2</i>	
<p>Passage: (1) Detective Wolf and Miss Fox (2) Detective Wolf has waited in the doorway all the morning. (3) He is going to meet Miss Fox. (4) She bought a bag of diamonds from Africa. (...) Statement: The diamonds were bought from Africa. Ground Truth TL: Transformed Word Matching</p>	<p>GPT-4o's Response: 1. Analyze the Statement - Content words: diamonds, bought, Africa 2. Locate Supporting Evidence in the Passage - Sentence (4): "She bought a bag of diamonds from Africa." 3. Compare the Statement and Evidence - All content words match exactly with the passage: "diamonds," "bought," "Africa." - <u>The order of content words is exactly the same as in the statement.</u> (...) Answer: Word Matching</p>

Figure 4: Representative error cases with GPT-4o's responses.

ever, GPT-4o claims: "The order of content words in the evidence is exactly the same as in the statement." This suggests that the model sometimes fails to detect subtle syntactic transformations. These two cases together suggest that, even when LLMs provide seemingly coherent explanations, they may miss structural cues and fail to reflect on the reasoning processes underlying their own answers. Such blind spots underscore a persistent challenge of detecting and modeling the cognitive features underlying human problem-solving processes.

7 Conclusion

The cognitive complexity of the problem-solving process is a key factor for analyzing the prior difficulty of RC items, yet no scalable method currently exists for automatically measuring it. In this study, we investigated whether LLMs can predict the cognitive complexity of RC items through two cognitively grounded dimensions: Evidence Scope and Transformation Level. To this end, we constructed RECO—a benchmark dataset of RC items annotated along these two dimensions—and conducted a comprehensive evaluation of eight LLMs under diverse prompting and decoding configurations.

The results show that LLMs have strong potential as proxies for cognitive complexity estimation, with some open-source models—such as Qwen2.5-32B—achieving performance comparable to proprietary systems like GPT-4o. Nevertheless, LLMs are not fully aligned with human experts and exhibit limitations in their metacognitive awareness—particularly in detecting phrase reordering or identifying all necessary evidence from the passage.

We hope our findings encourage further research in item difficulty estimation and difficulty-controlled item generation—contributing to the development of more interpretable and cognitively aligned educational NLP systems.

Limitations

Label Imbalance and Data Scale. Our dataset exhibits label imbalance across cognitive dimensions, which is an inevitable outcome of annotating RC items randomly sampled from assessments. The scarcity of certain labels can be attributed to several factors: such item types may be more difficult to create, less emphasized in instructional practice, or underrepresented in the specific assessments from which our samples were drawn. However, supplementing these underrepresented categories would require additional large-scale annotation, which was infeasible under our budget constraints.

Moreover, constructing the dataset required costly expert annotation—three raters per item—which constrained our ability to perform supervised fine-tuning. Nevertheless, our dataset comprises 776 items, a substantially larger resource than the previously released CMCQRD dataset containing only 289 items (Mullooly et al., 2023). Our findings also suggest that larger models (24B–32B) exhibit more robust in-context classification than smaller ones (7B–9B), thus highlighting opportunities for future work on data augmentation and knowledge distillation to build smaller yet effective models.

Limited Coverage of RC Item Types. The two cognitive dimensions investigated—Evidence Scope and Transformation Level—do not generalize across all RC item types in capturing their cognitive complexity. For instance, questions targeting main ideas or author intent inherently require multi-sentence inference. In contrast, the dimensions we employ are most relevant to factual detail questions such as TFNG, MTF, and WH-questions. Because different RC item types involve distinct cognitive complexity factors, addressing all of them comprehensively would be beyond the scope of a single study. In this work, we therefore focus on two key dimensions best captured by factual detail questions, as this question format spans a wide range of cognitive complexity—from single-sentence word matching to multi-sentence inference. Our goal is to establish a foundation for future research on additional factors across a broader set of RC item types.

Limits of a Single Factor in Explaining Item Difficulty. Finally, the two dimensions studied here represent only a subset of the many factors influencing RC item difficulty. Establishing a direct or linear relationship between a single factor and difficulty would require controlling for all other variables, which is not the case in our dataset. For instance, a multi-sentence inference item may not necessarily be harder than a single-sentence paraphrasing item if the latter references a more complex passage. Thus, *these two factors alone cannot fully account for item difficulty*. This study instead focuses on testing whether LLMs can estimate cognitively grounded dimensions that have traditionally relied on human annotation.

Potential Data Contamination We constructed RECO exclusively from the test split of RACE++, with college-level items supplemented from the validation split to meet the required sample size. Nevertheless, we acknowledge that some LLMs may have incorporated the RACE++ test and validation splits into their pre-training corpora. Even so, we believe the impact on our evaluation is limited: while exposure to RACE++ in its original multiple-choice format may help models recall specific answers, it does not directly expose the cognitive-complexity labels required in RECO. Consistent with this, our results show that high RC performance does not translate into accurate cognitive complexity classification, suggesting that incidental memorization of RACE++ items does not confer

a meaningful advantage on our tasks.

Acknowledgments

This research was supported by Culture, Sports and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sports and Tourism in 2025 (Project Name: Development of an AI-Based Korean Diagnostic System for Efficient Korean Speaking Learning by Foreigners, Project Number: RS-2025-02413038, Contribution Rate: 45%); by the IITP (Institute of Information & Communications Technology Planning & Evaluation) - ITRC (Information Technology Research Center) grant funded by the Korea government (Ministry of Science and ICT) (IITP-2026-RS-2024-00437866, Contribution Rate: 45%); and by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH), Contribution Rate: 10%). We also thank Jonghwi Kim for valuable feedback on this paper.

References

- Samah AlKhuyaey, Floriana Grasso, Terry R Payne, and Valentina Tamma. 2024. Text-based question difficulty prediction: A systematic review of automatic approaches. *International Journal of Artificial Intelligence in Education*, 34(3):862–914.
- Richard C Anderson. 1972. How to construct achievement tests to assess comprehension. *Review of educational research*, 42(2):145–170.
- Luca Benedetto, Giovanni Aradelli, Paolo Cremonesi, Andrea Cappelli, Andrea Giussani, and Roberto Turin. 2021. On the application of transformers for estimating the difficulty of multiple-choice questions from text. In *Proceedings of the 16th workshop on innovative use of NLP for building educational applications*, pages 147–157.
- Benjamin S Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, David R Krathwohl, et al. 1956. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York.
- John R Bormuth, John Manning, Julian Carr, and David Pearson. 1970. Children’s comprehension of between-and within-sentence syntactic structures. *Journal of educational psychology*, 61(5):349.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda

- Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Matthew Byrd and Shashank Srivastava. 2022. Predicting difficulty and discrimination of natural language questions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 119–130.
- Inn-Chull Choi and Youngsun Moon. 2020. Predicting the difficulty of efl tests based on corpus linguistic features and expert judgment. *Language Assessment Quarterly*, 17(1):18–42.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.
- Andreea Dutulescu, Stefan Ruseti, Mihai Dascalu, and Danielle Mcnamara. 2024. How hard can this question be? an exploratory analysis of features assessing question difficulty using llms. In *Proceedings of the 17th International Conference on Educational Data Mining*, pages 802–808.
- Susan E Embretson and C Douglas Wetzel. 1987. Component latent trait models for paragraph comprehension tests. *Applied psychological measurement*, 11(2):175–193.
- Roy Freedle and Irene Kostin. 1991. The prediction of gre reading comprehension item difficulty for expository prose passages for each of three item types: Main ideas, inferences and explicit statements. *ETS Research Report Series*, 1991(2):i–53.
- Ronald K. Hambleton and Ronald W. Jones. 1993. [Comparison of classical test theory and item response theory and their applications to test development](#). *Educational Measurement: Issues and Practice*, 12(3):38–47.
- Fu-Yuan Hsu, Hahn-Ming Lee, Tao-Hsing Chang, and Yao-Ting Sung. 2018. Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Information Processing & Management*, 54(6):969–984.
- Zhenya Huang, Qi Liu, Enhong Chen, Hongke Zhao, Mingyong Gao, Si Wei, Yu Su, and Guoping Hu. 2017. Question difficulty prediction for reading problems in standard tests. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.
- Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.
- Dorit Hutzler, Esther David, Mireille Avigal, and Rina Azoulay. 2014. Learning methods for rating the difficulty of reading comprehension questions. In *2014 IEEE International Conference on Software Science, Technology and Engineering*, pages 54–62. IEEE.
- Yoshee Jain, John Hollandner, Amber He, Sunny Tang, Liang Zhang, and John Sabatini. 2025. Exploring the potential of large language models for estimating the reading comprehension question difficulty. In *International Conference on Human-Computer Interaction*, pages 202–213. Springer.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Radhika Kapoor, Sang T Truong, Nick Haber, Maria Araceli Ruiz-Primo, and Benjamin W Domingue. 2025. Prediction of item difficulty for reading comprehension items by creation of annotated item repository. *arXiv preprint arXiv:2502.20663*.
- Miyoung Ko, Sue Park, Joonsuk Park, and Minjoon Seo. 2024. Hierarchical deconstruction of llm reasoning: A graph-based framework for analyzing knowledge utilization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4995–5027.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794.
- John P Lalor, Hao Wu, and Hong Yu. 2019. Learning latent parameters without human response patterns: Item response theory with artificial crowds. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4249–4259.
- Yichan Liang, Jianheng Li, and Jian Yin. 2019. A new multi-choice reading comprehension dataset for curriculum learning. In *Asian Conference on Machine Learning*, pages 742–757. PMLR.
- Adian Liusie, Vatsal Raina, Andrew Mullooly, Kate Knill, and Mark JF Gales. 2023. Analysis of the cambridge multiple-choice questions reading dataset with a focus on candidate response distribution. *arXiv preprint arXiv:2306.13047*.

- Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.
- Frederic M. Lord. 1980. *Applications of Item Response Theory To Practical Testing Problems*. Routledge.
- Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 16–27.
- Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai. 2014. *Automated evaluation of text and discourse with Coh-Matrix*. Cambridge University Press.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Andrew Mullooly, Øistein Andersen, Luca Benedetto, Paula Buttery, Andrew Caines, Mark JF Gales, Yasin Karatay, Kate Knill, Adian Liusie, Vatsal Raina, et al. 2023. The cambridge multiple-choice questions reading dataset.
- Irina Pandarova, Torben Schmidt, Johannes Hartig, Ahcène Boubekki, Roger Dale Jones, and Ulf Brefeld. 2019. Predicting the difficulty of exercise items for dynamic difficulty adaptation in adaptive language tutoring. *International Journal of Artificial Intelligence in Education*, 29(3):342–367.
- Gi-Pyo Park. 2004. Comparison of 12 listening and reading comprehension by university students learning english in korea. *Foreign Language Annals*, 37(3):448–458.
- Jae-Woo Park, Seong-Jin Park, Hyun-Sik Won, and Kang-Min Kim. 2024. Large language models are students at various levels: Zero-shot question difficulty estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8157–8177.
- Elaheh Rafatbakhsh and Alireza Ahmadi. 2023. Predicting the difficulty of efl reading comprehension tests based on linguistic indices. *Asian-Pacific Journal of Second and Foreign Language Education*, 8(1):41.
- Vatsal Raina and Mark Gales. 2024. Question difficulty ranking for multiple-choice reading comprehension. *arXiv preprint arXiv:2404.10704*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.
- Ana-Cristina Rogoz and Radu Tudor Ionescu. 2024. Unibucllm: Harnessing llms for automated prediction of item difficulty and response time for multiple-choice questions. *arXiv preprint arXiv:2404.13343*.
- Gemma Team. 2024a. *Gemma*.
- Qwen Team. 2024b. *Qwen2.5: A party of foundation models*.
- Qwen Team. 2025. *Qwen3 technical report*. Preprint, arXiv:2505.09388.
- Masaki Uto, Yuto Tomikawa, and Ayaka Suzuki. 2023. Difficulty-controllable neural question generation for reading comprehension using item response theory. In *Proceedings of the 18th workshop on innovative use of NLP for building educational applications (BEA 2023)*, pages 119–129.
- Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Mayi Xu, Yongqi Li, Ke Sun, and Tiejun Qian. 2024. Adaption-of-thought: Learning question difficulty improves large language models for reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5468–5495.
- Ying Xu, Dakuo Wang, Mo Yu, Daniel Ritchie, Bingsheng Yao, Tongshuang Wu, Zheng Zhang, Toby Jia-Jun Li, Nora Bradford, Branda Sun, et al. 2022. Fantastic questions and where to find them: Fairytaleqa—an authentic dataset for narrative comprehension. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–460.
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380.
- Ya Zhou and Can Tao. 2020. Multi-task bert for problem difficulty prediction. In *2020 international conference on communications, information system and computer engineering (cisce)*, pages 213–216. IEEE.

1. Evidence Mapping
Identify all sentences necessary to verify the factuality of the option.
[Passage]

On a small farm in Mexico, there are no schools.
 A bus is the school!
 The driver of the bus is the teacher!
 It is a school bus, but it doesn't take children to school.
 It just goes round from place to place, and sometimes it comes to this farm.
 The bus will stay here for three months.
 The farmers call it a school on wheels.
 Every time the bus comes, the farmers come running to it, shouting and laughing.
 They warmly welcome the school bus!
 When the bus is on the farm, in the morning, the teacher teaches the small children.
 In the afternoon, the bigger children come to have their lessons because they must work in the morning.
 At night, the fathers and mothers come to school.
 They want to learn, too.
 How the farmers hope that some day they can have a real school on their farm!

Option A. The children and their parents on the farm all come to the bus school to learn. (True)

2. Reasoning Complexity Measurement
Identify the level of reasoning required to determine the factuality of the option.

(i) When a single sentence alone is sufficient to determine the factuality of the option,
 Word Matching Transformed Word Matching
 Paraphrase Transformed Paraphrase Inference

(ii) When multiple sentences from the passage are required together to determine the factuality of the option,
 Word Matching Paraphrase Inference

3. Additional Survey
Do you agree that this option is True?
 Yes No
 I am confident that my response is correct.
 Strongly agree Somewhat agree Somewhat disagree disagree

1. Evidence Mapping
Identify all sentences necessary to verify the factuality of the option.
[Passage]

On a small farm in Mexico, there are no schools.
 A bus is the school!
 The driver of the bus is the teacher!
 It is a school bus, but it doesn't take children to school.
 It just goes round from place to place, and sometimes it comes to this farm.
 The bus will stay here for three months.
 The farmers call it a school on wheels.
 Every time the bus comes, the farmers come running to it, shouting and laughing.
 They warmly welcome the school bus!
 When the bus is on the farm, in the morning, the teacher teaches the small children.
 In the afternoon, the bigger children come to have their lessons because they must work in the morning.
 At night, the fathers and mothers come to school.
 They want to learn, too.
 How the farmers hope that some day they can have a real school on their farm!
 NO EVIDENCE

Option B. A school bus is a real school for farmers' children. (Not True)

2. Reasoning Complexity Measurement (NO EVIDENCE selected: skip this section.)
Identify the level of reasoning required to determine the factuality of the option.

Revise only the portion of the option that conflicts with the passage. Apply minimal edits, and make use of wording from the evidence sentence whenever possible.

Assess the reasoning complexity based on your revised option.

(i) When a single sentence alone is sufficient to determine the factuality of the option,
 Word Matching Transformed Word Matching
 Paraphrase Transformed Paraphrase Inference

(ii) When multiple sentences from the passage are required together to determine the factuality of the option,
 Word Matching Paraphrase Inference

3. Additional Survey
Do you agree that this option is Not True?
 Yes No
 I am confident that my response is correct.
 Strongly agree Somewhat agree Somewhat disagree disagree

Figure 5: Example annotation sheets for *True* (left) and *Not True* (right) statements. For *Not True* statements, annotators could mark “No Evidence” when the passage lacked sufficient information (corresponding to the *Insufficient Evidence* category). Annotators were also asked to create a minimally revised *True* version of each *False* statement to enable assessment of its Transformation Level.

A Data Annotation Details

We recruited three experts via Upwork³, informing them in advance that the task involved data collection for research purposes and that their anonymity would be guaranteed. They independently labeled 238 MTF items (952 statements) across six batches. Annotators were compensated per item according to the education level of the source test: \$1.20 for middle school, \$1.50 for high school, and \$2.00 for college-level items, with additional payment for training and revision.

Figure 5 illustrates examples of the annotation sheet used by annotators. For each statement, annotators were shown the passage, the statement, and its factuality label. Annotators first identified the sentence(s) necessary to verify the statement’s factuality (Evidence Scope) and then determined its Transformation Level based on the lexical and structural relationship between the statement and the selected evidence. Because TL is defined only for *True* statements, annotators revised each *False* statement—excluding those labeled as *Insufficient Evidence*—into its closest *True* version before assigning a TL label.

To further ensure data quality, we added two additional survey questions. First, annotators were asked whether they agreed with the provided factuality label, and items with disagreement were discarded. Second, they reported their confidence in the labels they assigned; although no low-confidence items were reported, any such cases would have been excluded. Finally, items flagged as potentially problematic or ethically inappropriate were also removed.

Figure 6 presents inter-annotator agreement for both annotation dimensions. Among the 898 items retained after filtering based on annotators’ survey responses, 86.41% received the same label from at least two annotators, while 53.23% achieved full agreement among all three. Although the annotation

³<https://www.upwork.com>

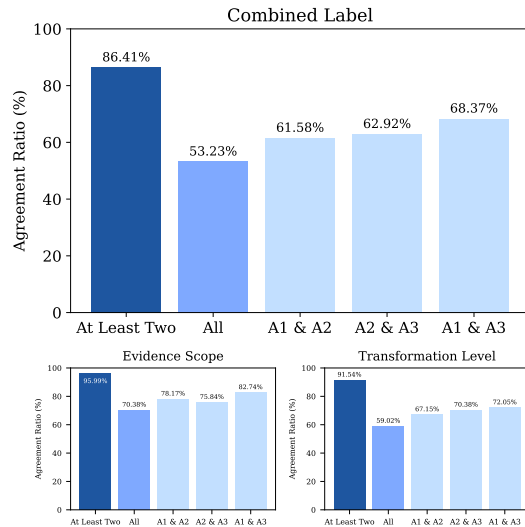


Figure 6: Inter-annotator agreement across labeling dimensions. Agreement ratios are shown for (top) combined labels of two dimensions. “At Least Two” indicates majority agreement among annotators, while “All” requires unanimous agreement. Pairwise agreements between annotators (A1, A2, A3) are also reported. For TL agreement, multi-evidence items labeled as *word matching* or *paraphrasing* were considered equivalent to *transformed word matching* and *transformed paraphrasing*, respectively.

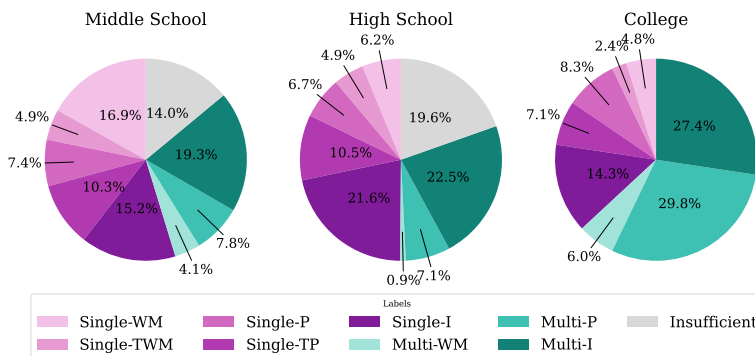


Figure 7: Distribution of combined difficulty labels in the RECO dataset across educational levels.

process was guided by structured label definitions, it required close examination of lexical, syntactic, and inferential relationships between statements and evidence, making it more demanding than typical classification tasks. In some cases, minor oversights—such as missing subtle paraphrases or nuanced details—led to disagreements. To mitigate their impact on data quality, we removed items for which all three annotators provided different labels. For items with partial agreement (i.e., two matching labels and one dissenting), the authors manually reviewed all annotations and resolved discrepancies by cross-referencing annotators’ justifications with the passage content. After this adjudication process, we obtained 776 annotated TFNG items, which we refer to as RECO

B Data Analysis

Figure 7 presents the distribution of cognitive labels in RECO across different educational levels. *Word-matching* items with *single-sentence evidence*—representing the lowest cognitive complexity—are more prevalent in middle school exams but decrease substantially at higher levels. Conversely, *multi-sentence* and *inference*-based items occur more frequently in high school and college assessments. These patterns suggest that cognitive complexity varies considerably even among items within the same educational level or format, implying that educational level alone is insufficient for fine-grained difficulty analysis.

Figure 8 illustrates the maximum gap between successive evidence sentences, as identified by the

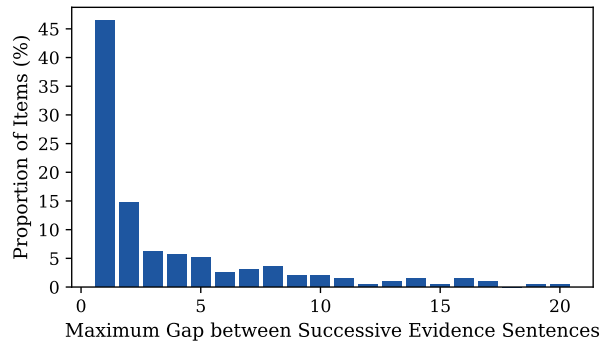


Figure 8: Distribution of the maximum gap between successive evidence sentences for multiple-sentence evidence items. For each item, the gap is measured as the largest distance between any two consecutive evidence sentences in the passage.

raters, for multi-sentence evidence items. According to the chart, approximately 45% of these items have evidence sentences that are adjacent in the passage, whereas about 55% require comprehension across sentences that are farther apart. This indicates that many items demand integration of information from non-contiguous parts of the passage, rather than relying solely on one or two consecutive sentences.

C Experimental Details

In our experiments, we used open-source models from Hugging Face⁴ with the following model names:

- Gemma2-9B: google/gemma-2-9b-it
- Gemma2-27B: google/gemma-2-27b-it
- Mistral-7B: mistralai/Mistral-7B-Instruct-v0.3
- Mistral-24B: mistralai/Mistral-Small-24B-Instruct
- Qwen2.5-7B: Qwen/Qwen2.5-7B-Instruct
- Qwen2.5-32B: Qwen/Qwen2.5-32B-Instruct
- Qwen3-32B: Qwen/Qwen3-32B

In addition, we used GPT-4o and GPT-4o-mini via the OpenAI API⁵, with model versions gpt-4o-2024-08-06 and gpt-4o-mini-2024-07-18, respectively.

For self-consistency decoding Wang et al. (2022), we used the default hyperparameter values of each model that the authors defined, especially for top- p , top- k , and temperature. All experiments were conducted using two NVIDIA A100 40GB GPUs.

D Supplementary Micro-F1 Results

Tables 7, 8, and 9 present the Micro-F1 results corresponding to Tables 3, 4, and 6, respectively.

⁴<https://huggingface.co>

⁵<https://openai.com>

Method	#Demo	Gemma2		Mistral		Qwen2.5		GPT-4o	
		9B	27B	7B	24B	7B	32B	mini	-
<i>Reading Comprehension</i>									
CoT	1	85.1	89.2	63.7	79.7	80.9	88.2	87.0	89.4
<i>Evidence Scope Classification [Human: 87.9]</i>									
SP	0	48.8	55.2	43.0	58.8	46.0	56.4	51.8	57.8
	1	48.8	<u>60.0</u>	45.8	<u>59.8</u>	49.8	<u>59.8</u>	<u>55.2</u>	60.8
	6	<u>50.4</u>	53.6	<u>49.4</u>	57.0	<u>51.8</u>	58.4	54.6	<u>65.7</u>
CoT	0	60.0	62.7	21.3	63.7	59.2	70.5	66.3	72.5
	1	<u>64.5</u>	<u>70.1</u>	53.6	66.3	58.4	73.1	<u>68.9</u>	75.5
	6	62.9	69.1	<u>55.6</u>	<u>68.5</u>	<u>60.6</u>	70.7	66.7	69.5
CoT (SC)	0	61.7	67.9	19.5	68.3	60.4	73.1	67.3	72.9
	1	67.7	72.3	57.8	71.1	61.9	72.9	72.3	<u>74.1</u>
	6	63.9	70.7	56.6	68.9	60.8	71.5	69.3	72.9
<i>(3-level) Transformation Level Classification [Human: 85.9]</i>									
SP	0	57.9	55.9	46.8	73.7	<u>62.4</u>	67.5	67.0	<u>70.7</u>
	1	55.2	58.3	58.4	<u>75.4</u>	58.8	<u>69.0</u>	67.8	69.7
	8	<u>58.3</u>	64.3	55.7	73.5	54.7	60.0	59.4	64.9
CoT	0	53.1	58.8	59.0	<u>76.9</u>	68.6	72.4	67.8	76.6
	1	50.9	56.8	48.6	<u>75.2</u>	<u>76.2</u>	<u>74.7</u>	71.9	78.1
	8	<u>57.1</u>	<u>61.3</u>	<u>62.8</u>	75.4	66.0	73.2	73.6	75.4
CoT (SC)	0	58.6	54.6	68.5	83.2	76.5	78.3	69.6	74.1
	1	54.6	58.0	<u>52.2</u>	<u>67.9</u>	<u>72.3</u>	<u>76.2</u>	<u>71.5</u>	<u>74.3</u>
	8	64.2	<u>62.7</u>	64.1	76.9	66.8	73.3	71.4	71.8
<i>(5-level) Transformation Level Classification [Human: 83.5]</i>									
SP	0	36.8	38.8	28.4	52.0	43.6	51.2	51.2	<u>54.8</u>
	1	<u>42.0</u>	42.4	34.0	<u>57.2</u>	<u>47.6</u>	<u>52.0</u>	<u>50.8</u>	52.8
	8	35.6	49.6	<u>40.0</u>	56.4	42.8	43.2	48.8	52.4
CoT	0	34.0	43.2	38.8	<u>59.2</u>	42.8	<u>59.6</u>	52.8	66.0
	1	37.6	38.4	32.0	54.0	<u>53.6</u>	58.4	56.0	67.2
	8	<u>42.8</u>	<u>47.6</u>	<u>48.0</u>	56.4	48.0	55.6	58.8	59.2
CoT (SC)	0	37.6	44.0	44.8	67.2	46.0	66.0	54.8	60.4
	1	43.2	45.2	39.2	53.6	56.8	63.2	55.6	<u>63.6</u>
	8	48.0	<u>48.4</u>	50.0	58.8	45.2	59.2	<u>56.4</u>	58.8

Table 7: Micro-F1 scores on the RC task and the ES and TL classification tasks. Greedy decoding is the default inference method, and SC indicates that self-consistency decoding is used. **Bolded** values denote each model’s best score per task; underlined values indicate the best score across demonstration settings.

Subtask	Gemma2		Mistral		Qwen2.5		GPT-4o	
	9B	27B	7B	24B	7B	32B	mini	-
1.1. Falsifiability Judgment	82.8	82.1	71.4	81.7	77.5	90.1	81.3	87.8
1.2. Evidence Sentence Counting	68.7	74.3	59.2	69.4	68.5	73.8	67.0	71.9
2.1. Inference Detection	73.4	75.8	65.5	82.4	68.7	81.7	81.4	80.0
2.2. Paraphrasing Detection	81.9	84.9	67.3	85.9	72.4	86.4	88.4	88.9
2.3. Phrase Reordering Detection	59.5	68.0	58.8	61.4	51.0	54.3	66.7	65.4

Table 8: Micro-F1 scores on subtasks that measure fine-grained abilities required for the main classification tasks. **Bolded** values denote the best model for each subtask.

Model	ES	TL	TL
		(5-level)	(3-level)
Qwen2.5-32B	70.5	59.6	72.4
Qwen3-32B _{non-thinking}	67.9	63.2	78.9
Qwen3-32B _{thinking}	65.5	60.4	71.8

Table 9: Comparison of Micro-F1 scores for LLMs with and without deep-thinking capabilities.