"Act Like a 5th Grader" is Not Enough: Bounding Knowledge in LLM-Based User Simulators

Krisztian Balog **Arild Michel Bakken**
University of Stavanger, Norway
{krisztian.balog,arild.m.bakken}@uis.no

Abstract

Large language models (LLMs) are increasingly used to simulate human behavior but frequently fail to exhibit realistic cognitive constraints, suffering from a "superhuman bias." Using a dataset of over 71,000 reading comprehension responses from 2,359 primary-school students (grades 4–6), we demonstrate that standard persona prompting yields near-perfect, deterministic performance, failing to capture the natural variance of developing readers. To address this, we introduce the Cognitively Bounded User Simulator (CBUS), an architectural framework that explicitly models the restricted working memory of young readers through an episodic bottleneck. Within this framework, we formalize two distinct test-taking strategies to emulate different reading behaviors. Our evaluation shows that explicitly modeling cognitive bounds significantly narrows the simulation gap across multiple LLM backbones, demonstrating that enforcing architectural constraints is more effective for high-fidelity simulation than simply scaling raw model capabilities.

1 Introduction

Large language models (LLMs) excel at "persona playing," making them a convenient tool to simulate human behavior in various application contexts, including the automatic evaluation of conversational agents (Yoon et al., 2024), counseling (Yang et al., 2025), and social science experiments (Argyle et al., 2023; Park et al., 2023). By conditioning an LLM on a specific persona or goal, researchers can approximate human interactions and prototype interactive systems at a fraction of the cost of large-scale human evaluation. However, the utility of these simulators hinges entirely on their fidelity to authentic human behavior.

Despite their widespread use, a fundamental limitation of current LLM-based simulators is their inconsistency with authentic human behavior. Simulators frequently confuse their assigned roles, fail to adhere to specified personas, and exhibit a "superhuman bias" by displaying omniscient global knowledge and reasoning capabilities far beyond their intended character (Wang et al., 2024; Naous et al., 2026; Zhou et al., 2026). We argue that these limitations originate from a critical missing component: current LLM-based user simulators rely predominantly on superficial prompting and lack explicit modeling of the underlying human cognitive processes. Crucially, isolating and quantifying these cognitive failures in complex, open-ended simulation environments (such as interacting with conversational agents) is methodologically challenging. When a simulation fails to mimic human limitations, it is difficult to determine whether the cause is a poorly defined persona, shifting context dynamics, or a fundamental architectural inability to bound knowledge.

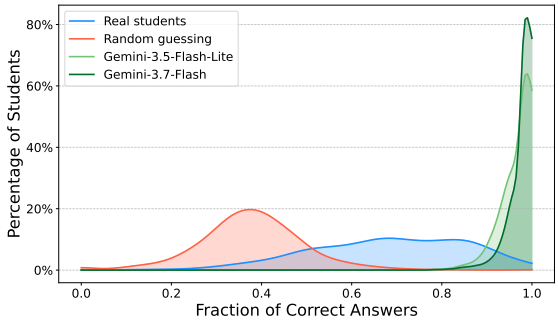

Figure 1: Distribution of per-student scores for real students, random guessing, and two baseline persona-prompted frontier LLMs. Both Gemini models collapse into a near-perfect "superhuman" spike that misses the natural performance spread of real students; moreover, the more capable 3.7-Flash model is *more* concentrated than 3.5-Flash-Lite, showing that greater model capability does not translate into more realistic simulation.

To rigorously study the challenge of simulating bounded human knowledge, we argue that the eval-

uation environment must be tightly constrained. We propose standard reading comprehension tests as an ideal, controlled testbed for this type of simulation. By simulating a primary school student taking a standardized reading assessment, we eliminate the variables and confounding factors of open-ended environments. Instead, the simulator's knowledge and cognitive load are strictly bounded by two transparent, verifiable factors: a specific source text and a defined developmental stage.

Despite the constraints of this environment, we show that standard persona-based prompting is highly insufficient for producing realistic simulations. We leverage a unique, large-scale dataset comprising over 71,000 responses from 2,359 students in grades 4–6 to 156 reading comprehension texts and 750 associated questions from the Norwegian educational system. As illustrated in Figure 1, simply prompting an LLM to "act like a 5th-grade student" results in a drastic overestimation of student capabilities. LLMs fail to behave like a developing human reader, effortlessly succeeding at inferential reasoning and vocabulary tasks that demonstrably challenge real students.

To bridge this gap, we argue that high-fidelity simulation requires moving beyond superficial persona assignment and toward explicit cognitive modeling. Rather than simply instructing an LLM to be a student, we must architecturally force it to process information like one. In this paper, we introduce the Cognitively Bounded User Simulator (CBUS), a novel framework grounded in the well-established finding that working memory capacity constrains developing readers' comprehension (Cain et al., 2004; Cowan, 2001). Specifically, we impose a generic working-memory capacity bottleneck through a constrained, two-stage reasoning pipeline: an encoding stage that strictly limits the number of text propositions the simulator can extract, followed by an execution stage where the model is forced to answer comprehension questions utilizing only this restricted memory trace.

By grounding our simulator in these verifiable cognitive processes, our proposed modeling successfully reduces the gap between real students and AI simulations. Our main contributions are fourfold:

- We expose and quantify the "superhuman bias" of LLM simulators in a highly constrained evaluation environment, demonstrating the inadequacy of standard persona prompting.
- We propose CBUS, an architectural framework that effectively models the restricted working memory of a young reader via a parameterized, two-stage episodic bottleneck under two distinct test-taking strategies.
- We propose a suite of evaluation metrics specifically tailored for the reading comprehension simulation scenario, moving beyond aggregate accuracy to holistically assess both student-centric and item-centric alignment.
- We provide robust empirical validation of our approach using a massive, real-world dataset, establishing that explicit cognitive bounds successfully reduce the calibration gap between simulated and real student responses.

Code, data, and prompts are made publicly available at https://github.com/iai-group/emnlp2026-simreading.

2 Related work

User simulation has seen a surge of interest across multiple domains, emerging as a critical methodology for both evaluating and training interactive systems. Beyond their established role in iterating on search, recommendation, and information access systems (Balog and Zhai, 2024; Balog et al., 2025), simulators are increasingly utilized as scalable proxies for human subjects in broad social science studies (Argyle et al., 2023; Park et al., 2023; Piao et al., 2026).

Currently, the predominant approach to modern user simulation relies heavily on LLMs, typically achieved through either in-context learning or task-specific fine-tuning (Terragni et al., 2023; Kolluri et al., 2025; Meshi et al., 2026). However, while highly capable, LLM-based simulations frequently fall short of true human fidelity. Because LLMs are optimized to be "helpful assistants" rather than cognitively plausible humans, current LLM-based simulators face severe limitations: they systematically diverge from actual human interaction patterns (Yoon et al., 2024; Seshadri et al., 2026), exhibit "superhuman" knowledge (Wang et al., 2024), and fail to capture the cognitive diversity inherent in real populations (Davidson et al., 2023; Wang et al., 2024). Merely prompting these systems to "act human" yields poor results (Naous et al., 2026), as they inherently struggle to adhere to knowledge constraints and instructions (Kiesel et al., 2024; Wang et al., 2024).

To address these shortcomings, recent studies have revealed the necessity of moving away from the assumption of "perfect users." Instead, there is a growing push toward explicitly modeling human-like behavior to create more robust and realistic evaluation environments. To bridge the simulation gap, researchers have begun prompting models with explicit user profiles or personas (Wang et al., 2025), tracking users' internal states (Wu et al., 2026), and instilling distinct behavioral traits (Ferreira et al., 2024). Further advancements involve modeling underlying cognitive processes—such as evolving knowledge states and shifting information needs (Zerhoudi and Granitzer, 2025; Zhang et al., 2025)—as well as capturing the realities of non-collaborative user behavior (Shim et al., 2026).

The need for such bounded simulators is particularly acute in educational settings. Adaptive assessments, which adjust tasks based on performance, have existed for decades but remain limited by the need for expensive empirical calibration via human pretests (Wainer and Mislevy, 2000). There is, therefore, a significant demand for high-fidelity simulations that can predict how different student profiles interact with educational content (von Davier, 2024). While recent work uses LLMs to automate this calibration (Yancey et al., 2024), these models often fail to capture the specific cognitive boundaries of young learners. Our work directly bridges this gap by enforcing structural cognitive limits on the LLM.

A parallel line of work uses LLMs and generative agents to simulate learners directly. EduAgent (Xu et al., 2024) models a single student's fine-grained, within-lecture behavior (gaze, motor actions, cognitive-state transitions) over time. Agent4Edu (Gao et al., 2025) generates learner response data for adaptive tutoring pipelines, initializing agent profiles from the corresponding learner's interaction histories and cognitive traits (problem-solving ability and knowledge proficiency). Sonkar et al. (2024) fine-tune LLMs to reproduce specific misconceptions, namely common algebraic errors of secondary school students. Zelikman et al. (2023) fine-tune LLMs to simulate K-12 students' item responses in order to generate and psychometrically calibrate reading-efficiency tests. Closest to our setting, Generative Students (Lu and Wang, 2024) answer multiple-choice items using hand-specified knowledge-component profiles to support item evaluation. Our work differs from all of the above on three axes: (i) we require no interaction data, knowledge-component tagging, or misconception taxonomy—inputs unavailable for a one-shot standardized reading test; (ii) our target is *population-level distributional fidelity*—matching the score and per-item difficulty distributions of real students—rather than individual-learner behavior, item evaluation, or misconception replication; and (iii) our mechanism is a working-memory capacity bottleneck, motivated by cognitive limitations of young learners—which none of these works model.

3 Reading Comprehension Dataset

Our dataset comprises (1) a corpus of texts with associated comprehension questions and (2) a collection of responses from a large-scale study involving 2,359 students. All materials and responses are in Norwegian (Bokmål), the primary written standard in Norwegian schools. Our focus on this language is motivated by the availability of high-quality assessment data to study the challenges of LLMs simulating young readers.

3.1 Texts and Questions

The dataset contains 156 texts designed to evaluate various dimensions of reading proficiency. Each text is accompanied by a set of questions, which can be of the following types: *multiple choice* (single-select), *true-or-false*, *checkboxes* (multi-select), and *free text*. For the purposes of this study, we exclude free-text responses to focus on objective, automatically gradable question types. This results in a final subset of 750 questions.

The questions are further divided into three categories based on the Norwegian Framework for Basic Skills (Utdanningsdirektoratet, 2017), which is closely related to the framework used in the PIRLS international assessment (Sabatini et al., 2024). The categories include: *Locate* (finding explicitly or implicitly expressed information), *Interpret & Connect* (drawing cross-text inferences), and *Reflect & Evaluate* (engaging critically and independently with texts). This categorization was created semi-automatically using an ensemble of LLM-based classifiers followed by expert review; the full methodology is detailed in Appendix B. Table 1 provides a breakdown of the question types and categories. The corpus of texts and questions is made publicly available.

Table 1: Question statistics.

	Locate & Connect	Interpret & Evaluate	Reflect	**Total**
multiChoice	310	40	12	362
trueOrFalse	317	23	4	344
checkboxes	41	3	0	44
Total	668	66	16	750

3.2 Participants and Ethics

The study utilizes data from 4th to 6th-grade classes (ages 8–11) within the Norwegian primary education system. Data collection followed the ethical guidelines of the National Committee for Research Ethics in the Social Sciences and the Humanities (NESH, 2021). We obtained full parental consent and student assent, and we exclusively analyze data for which this consent was secured. Critically, the process was privacy-preserving by default: no personally identifiable information (PII) was recorded, ensuring all responses remain strictly anonymous. Further administrative details regarding school recruitment, session structure, and exemptions from formal agency evaluation due to data anonymity are delegated to Appendix C. Throughout, our simulators target a prototypical 5th-grade (10–11 year-old) reader as the representative midpoint of this 4th–6th-grade range.

3.3 Responses

Data collection followed a self-selection protocol where students chose texts based on personal interest from randomly distributed pairs. Upon completing a task, students rated the text and were presented with a new pair of randomly selected texts. A screenshot of the user interface and further protocol details are provided in Appendix A.

Data Formalization To standardize the evaluation across heterogeneous question types, we binarize all student inputs into discrete dichotomous responses. For multiple-choice and true-or-false questions, a response is marked as correct (1) if the student selects the single correct answer, and incorrect (0) otherwise. For checkbox questions, we employ strict "all-or-nothing" scoring: a response is marked as correct (1) only if the subset of checked boxes perfectly aligns with the gold standard. Formally, we define a *response* as a tuple $(s, q) \in \{0, 1\}$, representing the binary performance of student s on question q.

Table 2: Dataset split statistics.

Split	**Classes**	**Responses**	**Students**
Held-out (Reference)	40	35,739	1,161
Ground Truth (Evaluation)	40	36,050	1,198
Total	80	71,789	2,359

Dataset Statistics & Difficulty In total, our filtered dataset contains 71,789 responses from 2,359 unique students across 80 classes. Regarding empirical question difficulty, multi-select checkbox questions proved significantly more difficult for the student population than the other two single-select formats. Across pedagogical categories, difficulties remain broadly comparable, though *Interpret & Connect* questions exhibit slightly lower success rates overall. Additional descriptive statistics are provided in Appendix A.

Data Splitting To rigorously evaluate simulation fidelity and establish an empirical baseline for human variance, we partition the dataset into two disjoint halves. To prevent data leakage (e.g., shared educational environments or teacher effects), this split is performed strictly at the *classroom* level (Table 2). Because our simulators operate zero-shot, the conventional "train/test" nomenclature is misleading. Instead, we designate one half as the **Ground Truth** evaluation split, which serves as the primary target distribution for all simulator metrics. The remaining half acts as a **Held-out** reference split; evaluating this population against the ground truth establishes the natural statistical variance (or "noise floor") of real students, providing an empirical upper bound for simulation fidelity.

4 Pilot Study: Evaluating Baseline Simulators

We first investigate whether state-of-the-art LLMs can emulate 5th-graders using standard "persona-based" prompting. While it is widely assumed LLMs can adopt personas via prompt engineering, simulating a developing reader requires actively suppressing inherent global knowledge to replicate cognitive limitations. We evaluate baseline zero-shot simulators against empirical student data to quantify the simulation gap and identify where standard prompting fails.

4.1 Pilot Experimental Setup

To establish a high-capability baseline, we evaluate zero-shot persona prompting on flagship commer-

Table 3: Pilot study results demonstrating the "superhuman bias" of baseline persona-based prompting compared to the empirical student distribution. The full metric set with standard deviations is in Appendix G.

Population	Mean Acc.	JSD (↓)
Real Students (Ground Truth)	0.687	
Real Students (Held-out)	0.661	0.006
Random Guessing	0.378	0.518
Llama-3.3-70B-Instruct	0.966	0.679
Mixtral-8x22B-Instruct	0.956	0.634
GPT-4o-mini	0.958	0.646
GPT-5.4	0.974	0.732
Gemini-3.5-Flash-Lite	0.972	0.713
Gemini-3.7-Flash	0.985	0.773

cial and open-weight models (version details in Appendix F) using the template in Figure 2. Structural instructions are in English to maximize adherence, while explicitly enforcing a Norwegian persona.[1] Following our primary protocol (Section 6), we set generation temperature to $T = 0.7$ and average results over $n = 3$ runs.

4.2 Results and Key Findings

We analyze two macro-level metrics: **Mean Accuracy** (average percentage of correct answers) and **JSD** (Jensen-Shannon Divergence, comparing individual score distributions between simulated and real Ground Truth populations). To establish an empirical upper bound for simulation fidelity, we also evaluate a held-out split of human students against the Ground Truth.

The Superhuman Bias Table 3 reveals a stark discrepancy. While the empirical student population exhibits natural variance (mean accuracy 0.687), baseline simulators universally collapse into a deterministic, near-perfect performance regime. As previewed in Figure 1, they fail to capture the true student spread, yielding JSD values orders of magnitude worse than the natural variance observed in the human held-out split (0.006).

The Paradox of Random Chance To contextualize this mismatch, we evaluate a naive random guessing baseline. Despite lacking cognitive capability and yielding a degraded mean accuracy (0.378), its wide, stochastic score distribution (red curve, Figure 1) paradoxically achieves a JSD (0.518) significantly closer to the human ground truth than any flagship LLM (0.634–0.773). That mindless guessing is distributionally closer to humans than state-of-the-art models highlights the core problem: LLM simulators suffer not from incapability, but from the lack of explicitly modeled cognitive limitations.

[1] In Appendix G we test models with Norwegian prompts and show that fully Norwegian instructions stay superhuman, i.e., the gap is not an instruction-language artifact.

```
You are simulating a 5th-grade student in Norway taking a
reading comprehension test.  Read the following text and
answer the question as a typical 10-11 year old student would.
The student may not always answer correctly.

<TEXT> {text_content} </TEXT>

<QUESTION> {question_text} </QUESTION>
```

Figure 2: The baseline persona prompt template used in the pilot study. The exact formatting of `{question_text}` for different question types (multiple choice, true-or-false, checkboxes) is in Appendix D.

The Penalty of Model Capability Crucially, increasing LLM capability actually *degrades* simulation fidelity. Across both open-weights (Llama, Mixtral) and proprietary models, scaling from efficient (GPT-4o-mini, Gemini-3.5-Flash-Lite) to frontier architectures (GPT-5.4, Gemini-3.7-Flash) pushes mean accuracy toward 0.98. This drives further divergence from the human baseline (JSD increases). Thus, standard persona prompting fails to restrict the model's global knowledge; the simulators trivially solve inferential and vocabulary tasks that demonstrably challenge young learners.

5 A Cognitively Bounded User Simulator

As demonstrated in the pilot study, standard persona prompting systematically fails to capture the limitations of a developing reader, resulting in a "superhuman bias" where the LLM trivially performs global text synthesis. To overcome this, we propose shifting from superficial role-play to explicit cognitive modeling. We conceptually frame our simulator as a *cognitive language agent* (Sumers et al., 2024). In this architecture, the LLM is stripped of its unbounded access to the input context. Instead, it acts solely as the central executive (or processing hub), and its reasoning is strictly mediated by parameterized memory modules that enforce human-like cognitive bottlenecks (Baddeley, 2000).

5.1 Grounding in Working Memory Capacity

Our architecture is inspired by Baddeley's multicomponent model of working memory (Baddeley, 2000), and specifically by two of its components:

the *episodic buffer*, a limited-capacity store that integrates disparate pieces of information into a coherent representation, and the *central executive*, which operates on that content to reason and answer. During reading comprehension, a student relies on the buffer to hold earlier text propositions available while processing new sentences—a prerequisite for resolving anaphora and drawing cross-sentence inferences (e.g., answering *Interpret and Connect* questions). It is well established that this working-memory capacity constrains reading comprehension, and that the constraint is especially pronounced in developing readers, whose capacity is still maturing (Daneman and Carpenter, 1980; Cain et al., 2004; Gathercole et al., 2004).

Rather than model the buffer's detailed maintenance and integration functions, we deliberately *simplify* it to a capacity-limited store, following Cowan's account in which working memory holds only a small number of chunks ($\sim$4 in adults) (Cowan, 2001). The critical distinction between an expert reader—or a standard LLM—and a young student is the strict capacity of this store: when a text requires integrating more propositions than it can hold, the reader experiences overload and falls back to a shallow, literal interpretation. The capacity limit introduced below operationalizes exactly this, with the LLM acting as the central executive that reasons over the bounded contents.

5.2 Computational Implementation: The Episodic Bottleneck

To computationally instantiate this cognitive bottleneck without requiring a complex, persistent multi-agent backend, we introduce a parameterized capacity limit, denoted as C. This parameter defines the maximum number of distinct text propositions the simulated student can actively maintain in working memory at any given time.

Furthermore, educational psychology suggests that developing readers employ varying test-taking strategies when approaching reading comprehension tasks (Ardoin et al., 2024). To capture this behavioral variance, we operationalize the encoding process into two distinct variants:

1. **Single-Pass Reading (SPR):** The simulator is provided only with the source text and extracts the C most salient propositions overall. This simulates a student who reads the text once and attempts to answer all subsequent questions from this single, fixed memory trace.

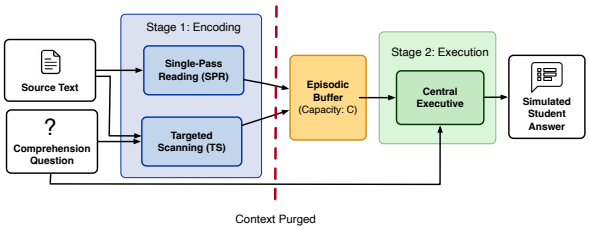

Figure 3: The Cognitively Bounded User Simulator (CBUS) architecture. In Stage 1 (Encoding), the simulator extracts a bounded number of propositions (Capacity: C) using either Single-Pass Reading (SPR) or Targeted Scanning (TS). The original source text is then purged from the context. In Stage 2 (Execution), the Central Executive answers the comprehension question relying solely on the restricted contents of the episodic buffer.

2. **Targeted Scanning (TS):** The simulator is provided with both the source text and a specific comprehension question. It extracts a maximum of C distinct propositions that it deems most relevant to that specific question. This simulates a student employing a "scan-and-search" strategy for each item.

Regardless of the chosen strategy, we operationalize this capacity limit by intercepting the LLM's standard generation process and forcing it through a constrained, two-stage reasoning pipeline (illustrated in Figure 3):

- **Stage 1: Encoding (Filling the Buffer).** The LLM acts as the selective attention mechanism, executing either the SPR or TS strategy. During this stage, the model is explicitly forbidden from generating answers; it merely fills the buffer with C extracted propositions.

- **Stage 2: Execution (Bounded Integration).** The original source text is completely purged from the LLM's context window. The central executive is then prompted (see Figure 6) to answer the comprehension question utilizing *only* the C propositions successfully transferred into its simulated episodic buffer.

By structurally separating extraction from reasoning, we strictly bound the simulator's global attention mechanism. For example, if an inference question requires synthesizing three distinct facts from the text, but the simulator is parameterized with a buffer capacity of $C = 2$, the agent is architecturally forced to attempt an answer using incomplete information. Consequently, the simulator's

```
You are simulating a 5th-grade student in Norway taking a
reading comprehension test. You are about to read a text,
and you'll be asked questions about this text. As a young
reader, you cannot remember everything from the text — only
a few key facts.

Read the text and identify at most {capacity} distinct
propositions (paraphrased atomic facts from the text) that
you think are most relevant to the questions you'll be asked.
Do not answer any questions yet.

<TEXT> {text_content} </TEXT>

Respond with ONLY a JSON object in the following format,
with no other text: {"propositions": ["proposition 1",
"proposition 2", ...]}
```

Figure 4: Stage 1 Encoding Prompt for the **Single-Pass Reading (SPR)** strategy.

```
You are simulating a 5th-grade student in Norway taking a
reading comprehension test. You have just read the question
below and are now scanning the text specifically to find the
answer. As a young reader, you cannot process the whole text
at once — you only focus on a few key details that seem to
match the question.

Scan the text and extract at most {capacity} distinct
propositions (paraphrased atomic facts from the text) that
share keywords or seem most directly related to what the
question is asking. Do NOT answer the question yet.

<TEXT> {text_content} </TEXT>

<QUESTION> {question_text} </QUESTION>

Respond with ONLY a JSON object in the following format,
with no other text: {"propositions": ["proposition 1",
"proposition 2", ...]}
```

Figure 5: Stage 1 Encoding Prompt for the **Targeted Scanning (TS)** strategy.

failures are no longer stochastic LLM hallucinations, but the direct result of verifiable cognitive constraints.

We fix C from an independent cognitive principle rather than tuning it on human data. Following Cowan's embedded-processes model (Cowan, 2001), the capacity-limited store holds roughly four chunks in adults, and developmental estimates place a 10–11 year-old at or below this bound (Cowan, 2016); we therefore set $C = 4$ for the holistic Single-Pass Reading strategy. Targeted Scanning instead models the narrow *focus of attention* engaged during question-guided retrieval, which holds only a small number of items (Oberauer, 2002; Cowan, 2001); we set $C = 2$—the smallest capacity that still permits integrating information from two points in a text, as *Interpret and Connect* questions require. We report a full capacity sweep in Appendix J.

6 Experimental Setup

To isolate the impact of explicit cognitive modeling from underlying model capabilities, we evaluate

```
You are simulating a 5th-grade student in Norway taking a
reading comprehension test. You no longer have access to
the full text — you can only rely on the propositions you
remembered from reading. Answer the question as a typical
10-11 year old student would, based ONLY on the propositions
below. The student may not always answer correctly.

<PROPOSITIONS> {propositions} </PROPOSITIONS>

<QUESTION> {question_text} </QUESTION>
```

Figure 6: Stage 2 **Execution** Prompt for CBUS. The original text is withheld, forcing reliance on the bounded episodic buffer.

our proposed Cognitively Bounded User Simulator (CBUS) framework against the standard zero-shot baseline persona simulator (detailed in Section 4) across a diverse suite of backbone LLMs.

6.1 Evaluation Metrics

To rigorously evaluate simulation quality, we organize our metrics into two high-level categories: *student-centric metrics*, which assess how well simulators approximate the overall performance and variance of the student population, and *item-centric metrics*, which evaluate alignment on specific question difficulties. The formal definitions of the metrics is given in Appendix E.

6.1.1 Student-Centric Metrics

We employ three metrics to measure population-level fidelity. First, the **Absolute Performance Gap** calculates the absolute difference in overall mean accuracy between the human and simulator populations. Second, **Distributional Alignment** utilizes the Jensen-Shannon Divergence (JSD) to measure how accurately the simulator reflects the macro-level variance and shape of empirical student score distributions. Third, the **Expected Calibration Error (ECE)** (Seshadri et al., 2026) evaluates the alignment of success rates across discrete difficulty bins to capture non-uniform miscalibration (e.g., if a model overestimates easy questions but underestimates hard ones).

6.1.2 Item-Centric Metrics

To determine if simulators struggle with the same specific questions as real students, we calculate the empirical difficulty of each question (defined as its mean success rate across all subjects). We evaluate **Item-Level Difficulty Correlation** by computing both the Pearson (ρ) and Spearman's rank (r_s) correlation coefficients between the human and simulator question difficulties.

6.2 Backbone LLMs and Inference Protocol

To ensure our findings generalize across different architectures and scales, we use six distinct backbone LLMs, including both open-weight and proprietary models (details in Appendix F). All models are evaluated in a zero-shot setting to prevent the simulators from memorizing specific student responses in-context. We set the generation temperature to $T = 0.7$ on a standard $[0, 1]$ scale to emulate the natural cognitive variance expected within a human student population. To account for generation variance, we repeat all experiments $n = 3$ times for every simulator-model pair and report the resulting averages.

Simulated Population Construction Each simulated population is a one-to-one mirror of the real Ground Truth split. For every real student, we retain their exact set of questions (i.e., those associated with the self-selected texts) remove their answers, and task the simulator with answering those questions. Consequently, the simulated and empirical populations are matched on exposure by construction: both consist of 1,198 subjects who answer the same set of questions. All student-centric metrics compare the resulting distributions of per-subject scores between the simulated and real *populations*, rather than pairing individual subjects.

7 Results and Analysis

Table 4 presents the evaluation of our cognitive bottleneck framework (CBUS) against standard persona-prompted baselines. As previously described, we evaluate two behavioral variants: Single-Pass Reading (SPR) and Targeted Scanning (TS). Rather than tuning the capacity on human response data, we fix C from an independent cognitive principle (Section 5): $C = 4$ for SPR and $C = 2$ for TS. A sensitivity analysis over a broad range of capacities (Appendix J) confirms that these principled values lie at or near the empirical optimum.

Closing the Simulation Gap The primary finding is that explicitly modeling cognitive boundaries successfully mitigates the "superhuman bias" of LLMs. Across all evaluated backbones and test-taking strategies, the CBUS framework consistently improves upon the baseline across all student-centric metrics. By restricting the episodic buffer, we successfully bring the simulated score distributions much closer to real student behavior.

Trade-offs in Test-Taking Strategies While CBUS improves global simulation fidelity, a clear trade-off emerges between the two encoding strategies. SPR drastically outperforms TS on all student-centric metrics, producing the most realistic macro-level distributions. TS, in turn, yields the strongest item-centric alignment: it improves both Pearson and Spearman correlations over the baseline for every backbone and attains the highest Pearson correlation of any variant across all backbones, indicating that it better captures the specific question difficulties experienced by real students. Notably, SPR also improves item-centric correlations over the baseline in most cases, so the two strategies are complementary rather than strictly opposed.

Model Capability vs. Simulation Fidelity Interestingly, utilizing more capable, state-of-the-art proprietary LLMs (e.g., GPT-5.4, Gemini-3.7-Flash) does not organically translate to more realistic student simulations; the strongest results instead come from smaller or open-weights backbones. Under SPR, Gemini-3.5-Flash-Lite and Llama-3.3-70B produce the closest student-centric alignment (with an absolute gap of 0.031 and 0.037, respectively), while the best item-centric correlations are obtained under TS, led by Llama-3.3-70B and Gemini-3.5-Flash-Lite on Pearson and Mixtral-8x22B on Spearman. This suggests that alignment for human simulation depends more on the architectural constraints of the simulator than on the raw reasoning power of the underlying backbone. This corroborates Zhou et al. (2026), who report that "higher general model capability does not necessarily yield more faithful user simulation."

Diagnostic Baselines and Ablation To isolate *why* CBUS works, we compare it against three diagnostic baselines on two backbones (Llama-3.3-70B and Gemini-3.5-Flash-Lite; full results in Appendix H). The first two try to lower the simulator's accuracy to the empirical student performance level without any cognitive mechanism: a *Calibrated Low-persona* prompt instructing the model to act as a below-average student, and *Answer-noising*, which randomly flips correct baseline answers until the mean matches the student population. Neither reproduces the students' score distributions: answer-noising matches the mean by construction yet leaves the item-difficulty correlation near zero, since uniform flipping carries no information about which questions are hard, while the low-persona prompt is unreliable—one LLM backbone largely

Table 4: Evaluation results across different backbone LLMs and simulation variants, under independent per-student sampling (SPR C=4, TS C=2). Results are reported as reported as mean $\pm$ standard deviation (in smaller font) over $n = 3$ runs. For each metric, the best simulated result is highlighted in **bold** and the second best in *italics* (*Real students* and *Random Baseline* are references and are not ranked).

		Student-Centric			Item-Centric	
Backbone	**Variant**	**Abs. Gap** (↓)	**JSD** (↓)	**ECE** (↓)	**Pearson** ρ (↑)	**Spearman** r_s (↑)
Real students (Held-out)		0.026	0.006	0.027	0.895	0.845
Random Baseline		0.309 $\pm.001$	0.518 $\pm.004$	0.306 $\pm.001$	0.409 $\pm.014$	0.319 $\pm.008$
Llama-3.3-70b-Instruct	baseline	0.279 $\pm.001$	0.679 $\pm.011$	0.285 $\pm.001$	0.241 $\pm.003$	0.179 $\pm.045$
	CBUS-TS	0.238 $\pm.001$	0.475 $\pm.007$	0.244 $\pm.001$	**0.455** $\pm.003$	0.304 $\pm.016$
	CBUS-SPR	*0.037* $\pm.002$	*0.093* $\pm.007$	**0.069** $\pm.002$	0.301 $\pm.007$	0.297 $\pm.004$
Mixtral-8x22b-Instruct	baseline	0.269 $\pm.000$	0.634 $\pm.003$	0.276 $\pm.000$	0.326 $\pm.002$	0.289 $\pm.020$
	CBUS-TS	0.237 $\pm.001$	0.467 $\pm.001$	0.243 $\pm.001$	0.417 $\pm.009$	0.314 $\pm.020$
	CBUS-SPR	0.124 $\pm.002$	0.170 $\pm.006$	0.128 $\pm.002$	0.377 $\pm.004$	**0.370** $\pm.016$
GPT-4o-mini	baseline	0.271 $\pm.001$	0.646 $\pm.002$	0.278 $\pm.000$	0.315 $\pm.002$	0.240 $\pm.008$
	CBUS-TS	0.226 $\pm.001$	0.418 $\pm.005$	0.233 $\pm.000$	0.415 $\pm.003$	0.301 $\pm.019$
	CBUS-SPR	0.096 $\pm.001$	0.122 $\pm.004$	0.100 $\pm.001$	0.355 $\pm.004$	*0.328* $\pm.004$
GPT-5.4	baseline	0.287 $\pm.000$	0.732 $\pm.001$	0.293 $\pm.000$	0.255 $\pm.001$	0.172 $\pm.002$
	CBUS-TS	0.271 $\pm.001$	0.618 $\pm.001$	0.276 $\pm.001$	0.397 $\pm.004$	0.286 $\pm.005$
	CBUS-SPR	0.129 $\pm.000$	0.188 $\pm.001$	0.131 $\pm.001$	0.224 $\pm.006$	0.225 $\pm.014$
Gemini-3.5-Flash-Lite	baseline	0.285 $\pm.000$	0.713 $\pm.004$	0.291 $\pm.000$	0.290 $\pm.001$	0.220 $\pm.008$
	CBUS-TS	0.238 $\pm.001$	0.470 $\pm.007$	0.243 $\pm.001$	*0.454* $\pm.002$	0.298 $\pm.009$
	CBUS-SPR	**0.031** $\pm.002$	**0.086** $\pm.004$	*0.074* $\pm.001$	0.304 $\pm.001$	0.295 $\pm.007$
Gemini-3.7-Flash	baseline	0.298 $\pm.000$	0.773 $\pm.001$	0.305 $\pm.000$	0.246 $\pm.001$	0.154 $\pm.006$
	CBUS-TS	0.281 $\pm.000$	0.668 $\pm.004$	0.287 $\pm.000$	0.285 $\pm.001$	0.195 $\pm.011$
	CBUS-SPR	0.091 $\pm.000$	0.120 $\pm.003$	0.115 $\pm.000$	0.236 $\pm.002$	0.202 $\pm.002$

ignores it and stays superhuman, and where it does comply it still falls short of CBUS. This confirms that matching aggregate accuracy is neither reliably achievable by prompting nor sufficient for distributional or item-level fidelity. Finally, a *Random-dropout* ablation replaces CBUS-SPR's salience-based proposition selection with random dropout at the same capacity; it performs on par with CBUS-SPR (better on some metrics, worse on others), indicating that the capacity *bound* itself, rather than the specific salience heuristic used to fill it, is the primary driver of the improvement.

Impact Across Question Categories We further break down performance by question category—*Locate* versus *Interpret & Reflect* (combining "Interpret & Connect" and "Reflect & Evaluate")—for Llama-3.3-70B and Gemini-3.5-Flash-Lite (Appendix I). The superhuman bias is most pronounced on retrieval-oriented *Locate* questions, which the baselines answer almost uniformly correctly and therefore fail to capture the natural spread of student scores. The best CBUS strategy then depends on the task: SPR dominates the student-centric metrics on *Locate*—for Llama it cuts the JSD score from 0.640 to 0.078. TS best captures item-level difficulty on higher-order *Interpret & Reflect* questions, where Llama's item-centric Pearson correlation rises to $\rho = 0.556$ from the baseline 0.279.

8 Conclusion

In this work, we investigated LLM-based user simulators in the setting of simulating the reading comprehension skills of primary-school students, demonstrating that standard persona prompting results in a severe "superhuman bias" that fails to capture the natural variance of developing readers. To bridge this gap, we introduced CBUS, a framework grounded in cognitive psychology that forces the LLM to operate under explicit, limited-capacity memory bottlenecks. Within this framework, we modeled two distinct test-taking strategies. By evaluating this approach against a massive dataset of real student responses, we demonstrated that explicitly modeling cognitive boundaries successfully narrows the simulation gap across a diverse set of LLM backbones, yielding improvements in both student-centric and item-centric metrics. Notably, our modeling of different test-taking strategies revealed distinct alignment trade-offs that warrant further investigation in future work. Furthermore, our highest simulation fidelity was achieved with smaller, less capable models rather than the most capable frontier LLMs, suggesting that building realistic user simulators relies far more on explicitly modeling the architectural constraints of human cognition than on scaling raw model intelligence.

Limitations

While the CBUS framework successfully demonstrates that explicitly modeling cognitive bounds improves simulation fidelity, our approach has several limitations.

First, our framework focuses exclusively on a single cognitive dimension: the capacity bottleneck of working memory. Real-world student performance is heavily influenced by complex non-cognitive factors, including test anxiety, fatigue, and intrinsic motivation. Although our dataset collection protocol attempted to foster engagement by allowing students to choose between two randomly paired texts, this binary choice does not guarantee genuine interest. Consequently, our current simulator lacks the mechanisms to model performance degradation caused by attention drift, apathy, or physical fatigue.

Relatedly, CBUS bounds the *amount* of information a reader retains, but idealizes *which* information is kept and *how faithfully*. The encoding stage uses a capable LLM to select the most salient (SPR) or question-relevant (TS) propositions and store them as accurate paraphrases; it therefore caps capacity without reproducing the selection and encoding errors of a real 10–11 year-old, who may retain irrelevant details, miss the gist, or misremember. Likewise, the execution stage still draws on the backbone's general language knowledge and answer-elimination strategies. Our evidence thus supports the narrower claim that restricting the quantity of accessible information improves distributional and item-level alignment with real students, rather than the stronger claim that CBUS faithfully reproduces the process by which a child reads and remembers. Modeling encoding noise and imperfect, non-expert fact selection is a natural next step.

Second, our implementation models Single-Pass Reading (SPR) and Targeted Scanning (TS) as rigid, mutually exclusive strategies. In reality, developing readers often employ dynamic, hybrid approaches—perhaps reading the text holistically once, but reverting to targeted scanning for highly specific or challenging questions. Future iterations of CBUS should explore dynamic strategy-switching mechanisms and probabilistic capacity allocation to better reflect fluid human test-taking behavior.

Finally, our evaluation scope is constrained by task format and demographics. To ensure an objective, automatically gradable ground truth, we deliberately binarized responses and excluded free-text generation tasks. Therefore, our findings regarding episodic bottlenecks apply primarily to recognition and retrieval tasks rather than generative comprehension. Furthermore, while our dataset provides high-quality, standardized empirical data, it is restricted to 4th to 6th-grade Norwegian students. Validating the universality of these simulated cognitive bounds across different languages, educational systems, and developmental stages remains an important direction for future work.

References

Scott P. Ardoin, Katherine S. Binder, Paulina A. Kulesz, Eloise Nimocks, and Joshua A. Mellott. 2024. Examining the influence of passage and student characteristics on test-taking strategies: An eye-tracking study. *Learning and Individual Differences*, 109:102386.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate. 2023. Out of one, many: Using language models to simulate human samples. *Political Analysis*, 31(3):337–351.

Alan Baddeley. 2000. The episodic buffer: a new component of working memory? *Trends in Cognitive Sciences*, 4(11):417–423.

Krisztian Balog, Nolwenn Bernard, Saber Zerhoudi, and ChengXiang Zhai. 2025. Theory and toolkits for user simulation in the era of generative ai: User modeling, synthetic data generation, and system evaluation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 4138–4141.

Krisztian Balog and ChengXiang Zhai. 2024. User simulation for evaluating information access systems. *Foundations and Trends in Information Retrieval*, 18(1-2):1–261.

Kate Cain, Jane Oakhill, and Peter Bryant. 2004. Children's reading comprehension ability: Concurrent prediction by working memory, verbal ability, and component skills. *Journal of Educational Psychology*, 96(1):31–42.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114.

Nelson Cowan. 2016. Working memory maturation: Can we get at the essence of cognitive growth? *Perspectives on Psychological Science*, 11(2):239–264.

Meredyth Daneman and Patricia A. Carpenter. 1980. Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, 19(4):450–466.

Sam Davidson, Salvatore Romeo, Raphael Shu, James Gung, Arshit Gupta, Saab Mansour, and Yi Zhang. 2023. User simulation with large language models for evaluating task-oriented dialogue. *arXiv*, cs.CL/2309.13233.

Rafael Ferreira, David Semedo, and Joao Magalhaes. 2024. Multi-trait user simulation with adaptive decoding for conversational task assistants. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16105–16130, Miami, Florida, USA. Association for Computational Linguistics.

Weibo Gao, Qi Liu, and 1 others. 2025. Agent4Edu: Generating learner response data by generative agents for intelligent education systems. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25.

Susan E. Gathercole, Susan J. Pickering, Benjamin Ambridge, and Hannah Wearing. 2004. The structure of working memory from 4 to 15 years of age. *Developmental Psychology*, 40(2):177–190.

Johannes Kiesel, Marcel Gohsen, Nailia Mirzakhmedova, Matthias Hagen, and Benno Stein. 2024. Simulating follow-up questions in conversational search. In *46th European Conference on IR Research*.

Akaash Kolluri, Shengguang Wu, Joon Sung Park, and Michael S. Bernstein. 2025. Finetuning LLMs for human behavior prediction in social science experiments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 30096–30111, Suzhou, China. Association for Computational Linguistics.

Xinyi Lu and Xu Wang. 2024. Generative students: Using llm-simulated student profiles to support question item evaluation. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, L@S '24, page 16–27.

Ofer Meshi, Krisztian Balog, Sally Goldman, Avi Caciularu, Guy Tennenholtz, Jihwan Jeong, Amir Globerson, and Craig Boutilier. 2026. ConvApparel: A benchmark dataset and validation framework for user simulators in conversational recommenders. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5270–5304, Rabat, Morocco. Association for Computational Linguistics.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. 2026. Flipping the dialogue: Training and evaluating user language models. In *The Fourteenth International Conference on Learning Representations*, ICLR '26.

Den nasjonale forskningsetiske komité for samfunnsvitenskap og humaniora NESH. 2021. *Guidelines for Research Ethics in the Social Sciences and the Humanities*. De nasjonale forskningsetiske komiteene, Oslo, Norway. 5th edition, updated 2023.

Klaus Oberauer. 2002. Access to information in working memory: Exploring the focus of attention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28(3):411–421.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, UIST '23.

Jinghua Piao, Yuwei Yan, Jun Zhang, Nian Li, Junbo Yan, Xiaochong Lan, Zhihong Lu, Zhiheng Zheng, Jing Yi Wang, Di Zhou, Chen Gao, Fengli Xu, Fang Zhang, Ke Rong, Jun Su, and Yong Li. 2026. Agentsociety: Large-scale simulation of llm-driven generative agents advances understanding of human behaviors and society. *iFuture*.

J. Sabatini, A. Kennedy, E. Wry, and M. von Davier. 2024. PIRLS 2026 reading assessment framework. *PIRLS 2026 Assessment Frameworks*.

Preethi Seshadri, Samuel Cahyawijaya, Ayomide Odumakinde, Sameer Singh, and Seraphina Goldfarb-Tarrant. 2026. Lost in simulation: LLM-simulated users are unreliable proxies for human users in agentic evaluations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 47423–47439, San Diego, California, United States. Association for Computational Linguistics.

Jeonghoon Shim, Woojung Song, Cheyon Jin, Seungwon Kook, and Yohan Jo. 2026. Non-collaborative user simulators for tool agents. In *The Fourteenth International Conference on Learning Representations*.

Shashank Sonkar, Xinghe Chen, Naiming Liu, Richard G. Baraniuk, and Mrinmaya Sachan. 2024. LLM-based cognitive models of students with misconceptions. *Preprint*, arXiv:2410.12294. ArXiv:2410.12294.

Theodore Sumers, Shunyu Yao, Karthik R Narasimhan, and Thomas L. Griffiths. 2024. Cognitive architectures for language agents. *Transactions on Machine Learning Research*.

Silvia Terragni, Modestas Filipavicius, Nghia Khau, Bruna Guedes, André Manso, and Roland Mathis. 2023. In-context learning user simulators for task-oriented dialog systems. *Preprint*, arXiv:2306.00774.

Utdanningsdirektoratet. 2017. Rammeverk for grunnleggende ferdigheter: 2.3 å kunne lese.

Alina A von Davier. 2024. How will AI change adaptive testing? In *Research for Practical Issues and Solutions in Computerized Multistage Testing*, pages 460–480. Routledge.

Howard Wainer and Robert J Mislevy. 2000. Item response theory, item calibration, and proficiency estimation. In *Computerized adaptive testing*, pages 61–100. Routledge.

Kuang Wang, Xianfei Li, Shenghao Yang, Li Zhou, Feng Jiang, and Haizhou Li. 2025. Know you first and be you better: Modeling human-like user simulators via implicit profiles. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21082–21107, Vienna, Austria. Association for Computational Linguistics.

Zhenduo Wang, Zhichao Xu, Vivek Srikumar, and Qingyao Ai. 2024. An in-depth investigation of user response simulation for conversational search. In *ACM on Web Conference 2024*.

Shirley Wu, Evelyn Choi, Arpandeep Khatua, Zhanghan Wang, Joy He-Yueya, Tharindu Cyril Weerasooriya, Wei Wei, Diyi Yang, Jure Leskovec, and James Zou. 2026. HumanLM: Simulating users with state alignment beats response imitation. In *Forty-third International Conference on Machine Learning*.

Songlin Xu, Xinyu Zhang, and Lianhui Qin. 2024. Eduagent: Generative student agents in learning. *Preprint*, arXiv:2404.07963. ArXiv:2404.07963.

Kevin P. Yancey, Andrew Runge, Geoffrey LaFlair, and Phoebe Mulcaire. 2024. BERT-IRT: Accelerating item piloting with BERT embeddings and explainable IRT models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 428–438, Mexico City, Mexico. Association for Computational Linguistics.

Yizhe Yang, Palakorn Achananuparp, Heyan Huang, Jing Jiang, Nicholas Gabriel Lim, Cameron Tan Shi Ern, Phey Ling Kit, Jenny Giam Xiuhui, John Pinto, and Ee-Peng Lim. 2025. Consistent client simulation for motivational interviewing-based counseling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20959–20998, Vienna, Austria. Association for Computational Linguistics.

Se-eun Yoon, Zhankui He, Jessica Echterhoff, and Julian McAuley. 2024. Evaluating large language models as generative user simulators for conversational recommendation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1490–1504, Mexico City, Mexico. Association for Computational Linguistics.

Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205, Singapore. Association for Computational Linguistics.

Saber Zerhoudi and Michael Granitzer. 2025. Cognitive-aware user search behavior simulation. In *Proceedings of the 24th ACM/IEEE Joint Conference on Digital Libraries*, JCDL '24.

Erhan Zhang, Xingzhu Wang, Peiyuan Gong, Zixuan Yang, and Jiaxin Mao. 2025. Exploring human-like thinking in search simulations with large language models. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 2669–2673.

Xuhui Zhou, Weiwei Sun, Qianou Ma, Yiqing Xie, Jiarui Liu, Weihua Du, Sean Welleck, Yiming Yang, Graham Neubig, Sherry Tongshuang Wu, and Maarten Sap. 2026. Mind the sim2real gap in user simulation for agentic tasks. In *The Third Annual Conference on Language Modeling*, COLM '26.

A Reading Comprehension Dataset

This section provides additional details regarding the digital platform used for data collection, the distribution of student responses, and our dataset filtering procedures.

User Interface and Text Selection The empirical data was collected via a digital educational platform designed for Norwegian 4th–6th-grade students. As illustrated in Figure 7, the workflow begins with a text selection screen (Top). Students are presented with a choice between two randomly paired texts to encourage engagement. After making their selection and reading the material, students proceed to an assessment interface where they answer a series of comprehension questions (Bottom).

Data Filtering and Modality Constraints Because our user simulators are powered by text-based Large Language Models, it is critical that the comprehension questions do not rely on unprovided visual context. We manually verified all texts and questions to ensure they are fully answerable relying exclusively on the provided reading material and any associated image alt text. During this review process, we identified two texts that strictly required visual inspection of accompanying images to successfully answer the questions. To maintain a fair evaluation environment for the text-based LLMs, these two texts, along with their associated questions and student responses, were entirely removed from the final dataset.

Text Popularity Distribution Because the platform allows students to choose between two text options, the distribution of responses is not uniform across the dataset. Figure 8 illustrates the popularity of the texts, measured by the number of unique students who selected them. The distribution shows a relatively smooth, steady decline across the dataset; the most frequently chosen texts accumulated over 140 student responses, while the least selected texts received approx. 55 responses.

B Question Categorization Methodology

To classify the 750 questions into the three cognitive categories (Locate, Interpret and Connect, and Reflect and Evaluate), we employed a semi-automatic pipeline utilizing an ensemble of three distinct LLMs: Gemini-3.5-Flash-preview, GPT-5.4, and Llama-3.3-70B Instruct.

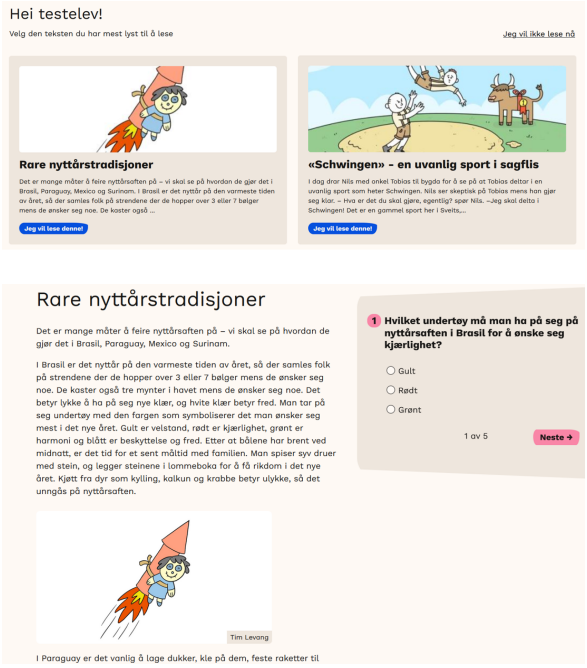

Figure 7: User interface of the reading comprehension platform. Top: The text selection screen where students choose between two randomly paired texts (English translation: (Left) "Unusual New Year's Traditions," (Right) "Schwingen - an Unusual Sport on Sawdust." Texts: Kirsti Thisland and Michelle N. Maurer; Illustrations: Tim Levang). Bottom: The assessment interface where students answer questions based on the text (English translation: "What underwear should you wear in Brazil on New Year's Eve to wish for love?", options: "Yellow," "Red," "Green." Text: Kirsti Thisland; Illustration: Tim Levang).

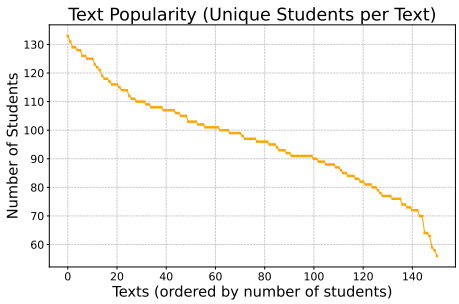

Figure 8: Text popularity distribution, showing the number of students who selected and completed the assessment for each text, ordered from most to least popular.

Each model was provided with a zero-shot prompt outlining the category definitions, disambiguation tips, and output formatting instructions (see Figure 9). We adopted a strict consensus-based approach for the initial labeling. If all three models predicted the same category, the label was provisionally accepted. In cases where there was any disagreement among the models' predictions—which

occurred for 83 out of the 750 questions—the item was flagged for manual review. All 83 flagged questions were manually checked and resolved by an expert specialized in the development of reading comprehension questions.

To validate the accuracy of the automated consensus, the expert also randomly sampled and reviewed an additional 18 questions where all three LLMs had achieved perfect agreement. In all 18 validation cases, the expert's manual assessment matched the LLM consensus, confirming the reliability of the ensemble approach.

```
You are an expert educational assessor and psychometrician
specializing in reading comprehension.  Your task is to
analyze a reading comprehension text and a specific question,
and classify the question based on the cognitive process it
targets.

Here are the three categories and their definitions:

1. **Locate**: Focuses on finding and retrieving information.
This involves reading comprehension processes such as finding
information that is explicitly stated in one or several
places in a text, locating competing information, and finding
information that is implicitly expressed in a localized part
of the text.
2.  **Interpret and Connect**: Focuses on synthesizing and
concluding.  This involves reading comprehension processes
such as drawing conclusions based on information from several
different places within a text, or across multiple texts.
3. **Reflect and Evaluate**: Focuses on critical thinking and
holistic understanding. This involves reading comprehension
processes such as engaging independently and critically with a
text, understanding the text's overall meaning/main idea, and
commenting on or justifying one's own viewpoints, analyses,
and/or evaluations regarding the text's content, form, or
author's intent.

### Tips for Disambiguation:
If the student just needs to "put their finger on it" (even
if they have to sift through competing facts or read between
the lines of a specific sentence), it is **Locate**.
If the student must stitch together "Fact A" from paragraph
1 and "Fact B" from paragraph 4 to figure out "Conclusion C",
it is **Interpret and Connect**.
If the student is asked to identify the overarching theme,
judge the reliability of the author, apply the text to the
real world, or give an opinion supported by the text, it is
**Reflect and Evaluate**.

### Language Note:
The reading text and question are in Norwegian. Please analyze
the text and question in their original Norwegian context to
capture the correct nuances, but write your analysis and
output entirely in English.

### Input:
*Text:**
{text_content}

**Question:**
{question}

### Output Format:
Please provide your response in the following format:
*Step 1: Cognitive Task Analysis:** Briefly explain exactly
what the reader must do to arrive at the answer. (e.g., "The
reader must look at paragraph 2 and paragraph 5 and synthesize
the two events to understand why the character is sad.")
*Step 2: Category Match:** Explain which category best aligns
with the task analysis.
*Step 3: Final Classification:** Provide ONLY the name of the
exact category (Locate, Interpret and Connect, or Reflect and
Evaluate).
```

Figure 9: Zero-shot prompt used for automatic question categorization.

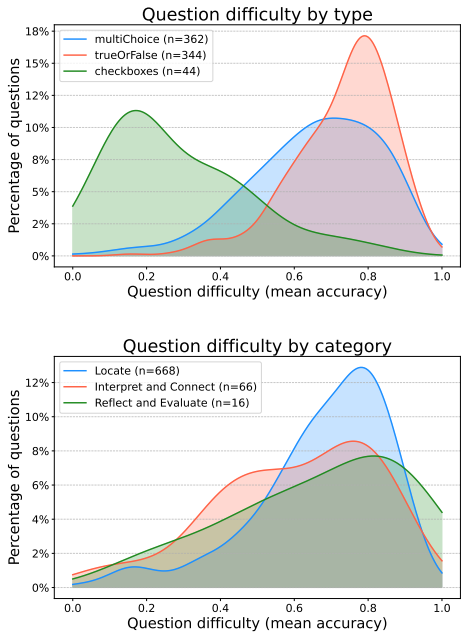

Figure 10: Empirical question difficulty (measured as mean student accuracy). The top plot breaks down difficulty by structural question type ($n = 362$ multiChoice, $n = 344$ trueOrFalse, $n = 44$ checkboxes). The bottom plot breaks down difficulty by pedagogical category ($n = 668$ Locate, $n = 66$ Interpret & Connect, $n = 16$ Reflect & Evaluate).

B.1 Empirical Question Difficulty Distributions

To better understand the composition and challenge level of our evaluation dataset, we analyze the empirical difficulty of the 750 questions. We formally define the difficulty of a given question q as its mean accuracy across the human population. Specifically, this is calculated as the average of the binary student responses $(s, q) \in \{0, 1\}$ over all students s who answered that specific question.

Figure 10 presents empirical question difficulty across the dataset. The left plot visualizes the distribution split by structural question type. Multi-select checkbox questions are demonstrably more difficult for the student population (peaking near 0.2 mean accuracy) compared to single-select multiple-choice and true-or-false formats, which peak between 0.7 and 0.8. The right plot visualizes the distribution split by pedagogical category. While "Locate" questions tend to have high success rates, questions requiring higher-order cognitive skills, such as "Interpret & Connect" and "Reflect & Evaluate," exhibit wider variance and generally lower mean accuracies.

C Administrative and Ethical Details

As noted in Section 3, this study utilizes data from the Norwegian primary education system. To recruit participants, the research team directly contacted several schools and published a call for participation on social media sites followed by Norwegian primary school teachers. This initially resulted in 151 classes from 4th to 6th grade (ages 8–11) registering for the reading assessment. Note that some attrition occurred prior to data collection, primarily due to student illnesses and busy school schedules.

Given the strictly anonymous nature of the secondary data analysis and the privacy-preserving default of the data collection (where no personally identifiable information was recorded and individual identifiers are not linkable to real-world identities), formal evaluation by a national agency (SIKT, the Data Protection Services of the Norwegian Agency for Shared Services in Education and Research) for research purposes was not required.

Regarding the data collection session structure, teachers who registered their class organized the assessment locally based on a detailed letter of instruction. Students were asked to sit for a total of 45 minutes. This included an initial briefing by their teacher, which was estimated to take approximately 5 minutes. The students then completed as many reading comprehension texts and question sets as they could in the remaining time, resulting in an overall average of around 7 texts completed per student.

D Question-Specific Prompt Formatting

As described in Section 4, the baseline persona simulator utilizes a primary prompt template where the `{question_text}` placeholder is dynamically populated based on the format of the comprehension question. To ensure automated, standardized extraction of the simulated student's answers without relying on brittle regex parsing, we append strict JSON formatting instructions to each question type.

The exact prompt elements injected for True/False, Multiple-Choice, and Checkbox questions are detailed in Figures 11, 12, and 13, respectively.

Finally, Figure 14 shows the fully Norwegian (Bokmål) variant of the baseline persona prompt, used for the instruction-language robustness check reported with the pilot baselines (Table 6).

```
This is a true or false question. Decide whether the following
statement is true or false based on the text.

Statement: {statement}

Respond with ONLY a JSON object in the following format, with
no other text: {"answer": true} or {"answer": false}
```

Figure 11: Prompt element injected into `{question_text}` for True/False questions.

```
This is a multiple choice question. Choose exactly ONE answer
from the options below.

Options: {options_text}

Respond with ONLY a JSON object in the following format, with
no other text: {"answer": "<letter>"} where <letter> is the
letter (A, B, C, ...) of your chosen option.
```

Figure 12: Prompt element injected into `{question_text}` for Multiple-Choice questions.

```
This is a checkbox question. For each option below, decide
whether it is correct or not.

Options: {options_text}

Respond with ONLY a JSON object in the following format, with
no other text: {"answer": [true, false, ...]}
The list must contain one true/false value for each option,
in the same order as listed above.
```

Figure 13: Prompt element injected into `{question_text}` for Checkbox (multi-select) questions.

```
Du simulerer en elev på 5. trinn i Norge som tar en
leseforståelsesprøve. Les teksten nedenfor og svar på
spørsmålet slik en typisk 10-11 år gammel elev ville gjort.
Eleven svarer ikke alltid riktig.

<TEXT> {text_content} </TEXT>

<QUESTION> {question_text} </QUESTION>
```

Figure 14: Fully Norwegian (Bokmål) variant of the baseline persona prompt (cf. the English template in Figure 2), used for the Norwegian-prompt robustness check to test whether the simulation gap is an instruction-language artifact rather than a genuine cognitive limitation. Per-question formatting mirrors the English templates above (Figures 11–13), translated to Bokmål.

E Formal Definitions of Evaluation Metrics

This section provides the formal mathematical definitions for the evaluation metrics introduced in Section 6.1. Throughout these definitions, let $r_{s,q} \in \{0, 1\}$ represent the discrete, binary response of a subject s (either a real student or a simulated agent) on a specific question q.

E.1 Student-Centric Metrics

These metrics assess how well the simulated agents approximate the overall performance, variance, and calibration of the aggregate human student population.

Absolute Performance Gap To evaluate the macro-level accuracy of the simulators, we compute the overall mean success rate of a given population. Let R denote the total set of valid response tuples (s, q) for a population. The mean population accuracy μ is defined as:

$$\mu = \frac{1}{|R|} \sum_{(s,q) \in R} r_{s,q}$$

The Absolute Performance Gap is simply the absolute difference between the human and simulated population means: $|\mu_{Human} - \mu_{Sim}|$.

Distributional Alignment (JSD) To ensure the simulators capture the natural variance of the student population rather than collapsing to a deterministic mean, we measure the distance between their score distributions. Let a subject's overall score be the average of their responses: $c_s = \frac{1}{|Q_s|} \sum_{q \in Q_s} r_{s,q}$, where Q_s is the set of questions answered by subject s. We estimate the discrete probability distributions of these scores for the human population (P) and the simulated population (Q). The alignment is calculated using the symmetric Jensen-Shannon Divergence (JSD).

Expected Calibration Error (ECE) While the absolute performance gap provides a global evaluation, it can mask non-uniform miscalibration—for instance, if a simulator systematically underestimates performance on challenging questions while overestimating it on moderately difficult ones. To quantify this, we adapt the Expected Calibration Error (ECE) metric to measure the alignment between simulated and human success rates across varying levels of question difficulty (Seshadri et al., 2026).

We partition the questions into $M = 6$ discrete difficulty bins based on empirical success rates. Let $S_i^{(Human)}$ and $S_i^{(Sim)}$ denote the average success rates of the human and simulated populations for questions within bin i, respectively. Let w_i denote the proportion of total questions that fall into bin i. The calibration error is then computed as the weighted average absolute deviation across all difficulty bins:

$$ECE = \sum_{i=1}^{M} w_i |S_i^{(Human)} - S_i^{(Sim)}|$$

A perfectly calibrated simulator will yield an ECE of 0, indicating that its predicted performance identically matches human outcomes across every difficulty stratum (Seshadri et al., 2026).

E.2 Item-Centric Metrics

These metrics isolate performance on individual questions to evaluate whether simulators struggle with the exact same linguistic and cognitive challenges as developing readers.

Item-Level Difficulty Correlation We formally define the empirical difficulty of a specific question q, denoted as d_q, as its mean success rate across all subjects who answered it:

$$d_q = \frac{1}{|S_q|} \sum_{s \in S_q} r_{s,q}$$

where S_q is the set of subjects exposed to question q. Let $D^{(Human)}$ and $D^{(Sim)}$ be the corresponding vectors of question difficulties for the human and simulated populations. To evaluate item-level alignment, we compute both the Pearson correlation coefficient (ρ) to measure the linear relationship and Spearman's rank correlation coefficient (r_s) to measure the monotonic rank-order relationship between $D^{(Human)}$ and $D^{(Sim)}$.

F Model Providers and API Details

All inference was routed through the OpenRouter API. Table 5 lists the exact providers, human-readable model names referenced throughout this paper, and the corresponding API endpoint snapshots used in the experiments.

G Full Pilot Results with Run Variance

Table 6 reports the complete metric set for the pilot baselines (Section 4.2) as mean $\pm$ standard deviation over the $n = 3$ runs; the main pilot table

Table 5: Detailed mapping of the LLMs evaluated in our study, including the specific model snapshot tags queried via the OpenRouter API.

Name	Provider	OpenRouter API Endpoint / Snapshot
Llama-3.3-70B-Instruct	Meta	`meta-llama/llama-3.3-70b-instruct`
Mixtral-8x22B-Instruct	Mistral AI	`mistralai/mixtral-8x22b-instruct`
GPT-4o-mini	OpenAI	`openai/gpt-4o-mini`
GPT-5.4	OpenAI	`openai/gpt-5.4`
Gemini-3.5-Flash-Lite	Google	`google/gemini-3.5-flash-lite`
Gemini-3.7-Flash	Google	`google/gemini-3.7-flash`

(Table 3) reports Mean Accuracy and JSD only for readability. Run-to-run variance is small across all backbones, confirming that the superhuman bias is a stable property of persona prompting rather than a sampling artifact.

H Diagnostic Baselines and Ablation

To understand *why* CBUS improves fidelity—and to rule out simpler explanations—we evaluate three diagnostic baselines on two backbones (Llama-3.3-70B and Gemini-3.5-Flash-Lite), all under the same protocol as our main experiments (Table 7). The first two test whether merely matching the human mean accuracy is sufficient; the third ablates the salience-based selection inside CBUS-SPR. A complementary check on instruction language (fully Norwegian prompts) is reported with the pilot baselines in Table 6.

Calibrated Low-persona The most direct way to lower a simulator's accuracy is to ask for it. If explicitly instructing the model to behave like a weak student reproduced empirical behavior, no architectural mechanism would be required.
Implementation. This is the standard persona-prompted baseline with an accuracy-calibrated prompt: the model is told it is a "below-average" reader who "answers only about two out of every three questions correctly (roughly 65–70% accuracy)" and should make "realistic mistakes on the harder questions." No other part of the method changes.

Answer-noising This baseline forces the mean accuracy to match the human population *exactly*, testing whether hitting the right mean also reproduces the score distribution and the item-difficulty structure.
Implementation. It is a post-hoc transformation of the persona-baseline outputs and makes no additional LLM calls. Given a baseline run with mean accuracy a_0 and a target $a^\star$ (the real-student mean, 0.687), each currently-correct response is independently flipped to incorrect with probability $(a_0 - a^\star)/a_0$, so the expected accuracy after noising equals $a^\star$. Flips are type-appropriate (negating a true/false answer, selecting a wrong option, or breaking a checkbox match) and seeded per run for reproducibility. Crucially, because flips are applied uniformly at random—independent of which questions are genuinely difficult—this baseline carries no item-difficulty signal by construction.

Random Proposition Dropout CBUS-SPR could help for two distinct reasons: the *capacity bound* (only C propositions are retained) or the *salience-based selection* (the model keeps the C it judges most important). This ablation separates the two by holding the bound fixed while removing the selection.
Implementation. It is identical to CBUS-SPR except in the Stage-1 encoding step: rather than extracting exactly C propositions, the model extracts a larger salient pool of $3 \times C = 12$ propositions, from which C are then retained *uniformly at random* instead of most-salient-first. Stage 2 (answering from only the retained propositions) is unchanged, so the amount remembered is identical to CBUS-SPR; only *which* propositions survive differs. Note that the discarded propositions are themselves drawn from a salient pool, so this ablates fine-grained ranking *within* salient content rather than salient-versus-arbitrary selection.

Results The results (Table 7) support two conclusions. First, matching—or merely attempting to match—the human accuracy is not sufficient for fidelity. *Answer-noising* forces the mean to match exactly (Abs. Gap of 0.001 on both backbones), yet its item-difficulty correlation nearly vanishes (Spearman r_s of 0.065 and 0.090 for Llama-3.3-70B and Gemini-3.5-Flash-Lite), because uniform

Table 6: Full pilot results (baseline persona-based prompting), reported as mean ± standard deviation over $n = 3$ runs. Real students (Held-out) is a single population, so no run variance is reported. The two *Norwegian* rows apply the fully Norwegian (Bokmål) prompt variant (Figure 14) to the two selected backbones.

	Student-Centric				Item-Centric	
Population	**Mean Acc.**	**Abs. Gap (↓)**	**JSD (↓)**	**ECE (↓)**	**Pearson ρ (↑)**	**Spearman r_s (↑)**
Real students (Held-out)	0.661	0.026	0.006	0.027	0.895	0.845
Random Guessing	0.378 ±.001	0.309 ±.001	0.518 ±.004	0.306 ±.001	0.409 ±.014	0.319 ±.008
Llama-3.3-70B-Instruct	0.966 ±.001	0.279 ±.001	0.679 ±.011	0.285 ±.001	0.241 ±.003	0.179 ±.045
Mixtral-8x22B-Instruct	0.956 ±.000	0.269 ±.000	0.634 ±.003	0.276 ±.000	0.326 ±.002	0.289 ±.020
GPT-4o-mini	0.958 ±.001	0.271 ±.001	0.646 ±.002	0.278 ±.000	0.315 ±.002	0.240 ±.008
GPT-5.4	0.974 ±.000	0.287 ±.000	0.732 ±.001	0.293 ±.000	0.255 ±.001	0.172 ±.002
Gemini-3.5-Flash-Lite	0.972 ±.000	0.285 ±.000	0.713 ±.004	0.291 ±.000	0.290 ±.001	0.220 ±.008
Gemini-3.7-Flash	0.985 ±.000	0.298 ±.000	0.773 ±.001	0.305 ±.000	0.246 ±.001	0.154 ±.006
Llama-3.3-70B-Instruct (Norwegian)	0.963 ±.001	0.276 ±.001	0.666 ±.002	0.283 ±.001	0.277 ±.010	0.196 ±.014
Gemini-3.5-Flash-Lite (Norwegian)	0.969 ±.000	0.282 ±.000	0.696 ±.007	0.288 ±.000	0.335 ±.006	0.269 ±.007

Table 7: Diagnostic baselines and ablation on two backbones, under independent per-student sampling, reported as mean ± standard deviation over $n = 3$ runs.

		Student-Centric			Item-Centric	
Backbone	**Variant**	**Abs. Gap (↓)**	**JSD (↓)**	**ECE (↓)**	**Pearson ρ (↑)**	**Spearman r_s (↑)**
Real students (Held-out)		0.026	0.006	0.027	0.895	0.845
Llama-3.3-70b-Instruct	baseline	0.279 ±.001	0.679 ±.011	0.285 ±.001	0.241 ±.003	0.179 ±.045
	Calibrated low-persona	0.277 ±.001	0.661 ±.003	0.284 ±.001	0.220 ±.004	0.210 ±.009
	Answer-noising	0.001 ±.001	0.102 ±.005	0.130 ±.003	0.216 ±.026	0.065 ±.040
	Random dropout	0.004 ±.001	0.084 ±.007	0.092 ±.002	0.304 ±.001	0.266 ±.002
	CBUS-SPR	0.037 ±.002	0.093 ±.007	0.069 ±.002	0.301 ±.007	0.297 ±.004
	CBUS-TS	0.238 ±.001	0.475 ±.007	0.244 ±.001	0.455 ±.003	0.304 ±.016
Gemini-3.5-Flash-Lite	baseline	0.285 ±.000	0.713 ±.004	0.291 ±.000	0.290 ±.001	0.220 ±.008
	Calibrated low-persona	0.073 ±.001	0.106 ±.001	0.082 ±.000	0.340 ±.002	0.320 ±.000
	Answer-noising	0.001 ±.001	0.093 ±.013	0.130 ±.002	0.266 ±.015	0.090 ±.032
	Random dropout	0.030 ±.001	0.100 ±.001	0.096 ±.002	0.316 ±.004	0.260 ±.006
	CBUS-SPR	0.031 ±.002	0.086 ±.004	0.074 ±.001	0.304 ±.001	0.295 ±.007
	CBUS-TS	0.238 ±.001	0.470 ±.007	0.243 ±.001	0.454 ±.002	0.298 ±.009

random flipping carries no signal about which questions are hard. The *Calibrated low-persona* baseline instead prompts the model to answer only 65–70% of questions correctly, and its effect is strongly backbone-dependent: Llama-3.3-70B largely ignores the instruction and stays at the baseline (Abs. Gap 0.277 vs. 0.279), whereas Gemini-3.5-Flash-Lite complies and improves across all metrics—but still trails CBUS-SPR on every student-centric metric and CBUS-TS on Pearson correlation. Neither route matches the fidelity of the cognitively bounded simulators. Second, within CBUS-SPR the capacity *bound* is the primary driver of the improvement: *Random dropout* performs on par with CBUS-SPR—better on some metrics and worse on others depending on the backbone—indicating that *how many* propositions are retained matters far more than *which* ones the model selects.

I Category-Specific Results

Table 8 breaks down evaluation results per reading comprehension category for two selected LLM backbones, Llama-3.3-70b-Instruct and Gemini-3.5-Flash-Lite. The persona-prompted baseline is markedly less faithful on *Locate* questions than on *Interpret & Reflect*: baseline JSD is 0.64–0.70 for *Locate* versus ~ 0.28 for *Interpret & Reflect*, with lower item-difficulty correlations as well, because the superhuman models answer the easier *Locate* items almost uniformly correctly—erasing the natural spread—while making more human-like errors on the harder inferential items. CBUS narrows this gap in both categories but along different axes: the largest student-centric gains come on *Locate* (SPR cuts JSD to ~ 0.08), whereas the strongest item-difficulty correlations are obtained on *Interpret & Reflect* under TS (Pearson up to 0.556 and 0.594).

J Capacity Sensitivity Analysis

To characterize the effect of the working-memory capacity C and to assess how our cognitively-motivated operating points perform empirically relative to other capacity settings, we perform a full capacity sweep for the two backbones. We vary $C \in \{2, 4, 6, 8, 10, 12\}$ for SPR and $C \in \{1, 2, 3, 4, 5\}$ for TS, and evaluate every setting

Table 8: Evaluation results broken down by reading comprehension question category, under independent per-student sampling, reported as mean $\pm$ standard deviation over $n = 3$ runs. *Interpret & Reflect* combines the *Interpret & Connect* and *Reflect & Evaluate* categories.

			Student-Centric			Item-Centric	
Category	**Backbone**	**Variant**	**Abs. Gap (↓)**	**JSD (↓)**	**ECE (↓)**	**Pearson ρ (↑)**	**Spearman r_s (↑)**
Locate	Llama-3.3-70b-Instruct	baseline	$0.277 \pm .001$	$0.640 \pm .011$	$0.283 \pm .001$	$0.225 \pm .006$	$0.147 \pm .043$
		CBUS-TS	$0.244 \pm .001$	$0.470 \pm .007$	$0.251 \pm .001$	$0.422 \pm .006$	$0.265 \pm .011$
		CBUS-SPR	$0.020 \pm .001$	$0.078 \pm .002$	$0.066 \pm .002$	$0.306 \pm .007$	$0.293 \pm .002$
	Gemini-3.5-Flash-Lite	baseline	$0.285 \pm .000$	$0.695 \pm .001$	$0.290 \pm .000$	$0.252 \pm .001$	$0.168 \pm .017$
		CBUS-TS	$0.251 \pm .001$	$0.505 \pm .009$	$0.256 \pm .001$	$0.410 \pm .003$	$0.260 \pm .005$
		CBUS-SPR	$0.022 \pm .002$	$0.076 \pm .005$	$0.076 \pm .001$	$0.290 \pm .001$	$0.272 \pm .007$
Interpret & Reflect	Llama-3.3-70b-Instruct	baseline	$0.293 \pm .001$	$0.279 \pm .003$	$0.302 \pm .001$	$0.279 \pm .008$	$0.311 \pm .050$
		CBUS-TS	$0.184 \pm .002$	$0.109 \pm .003$	$0.188 \pm .002$	$0.556 \pm .009$	$0.484 \pm .048$
		CBUS-SPR	$0.167 \pm .005$	$0.094 \pm .002$	$0.181 \pm .006$	$0.393 \pm .011$	$0.402 \pm .026$
	Gemini-3.5-Flash-Lite	baseline	$0.290 \pm .001$	$0.279 \pm .006$	$0.299 \pm .000$	$0.405 \pm .003$	$0.429 \pm .033$
		CBUS-TS	$0.132 \pm .003$	$0.055 \pm .004$	$0.141 \pm .002$	$0.594 \pm .002$	$0.512 \pm .023$
		CBUS-SPR	$0.101 \pm .001$	$0.049 \pm .003$	$0.108 \pm .001$	$0.435 \pm .014$	$0.507 \pm .009$

against the Ground Truth split (Figure 15). Given the small run-to-run variance observed in our main experiments (Table 4), we report a single run ($n = 1$) per capacity setting.

The key pattern is that student-centric and item-centric metrics respond to C in *opposite* directions, so no single capacity is jointly optimal. For SPR, the student-centric error (absolute gap and Score JSD) is lowest at low capacity and grows monotonically as C increases, whereas the item-centric difficulty correlation is flat-to-increasing in C: as the reader retains more propositions, its score distribution drifts away from real students even as its item-level discrimination improves. The cognitively-motivated value $C = 4$ sits at the student-centric optimum while retaining competitive item-centric alignment. For TS, both families of metrics are best at the smallest capacity ($C = 1$) and worsen as C grows; we nonetheless adopt $C = 2$, the smallest capacity that still permits integrating information from two points in a text, since $C = 1$ cannot, by design, answer questions that require combining multiple facts. Importantly, the approach is robust to the choice of C: across the entire sweep, every capacity setting improves over the corresponding baseline on both metrics, the sole exception being CBUS-SPR at $C = 2$ for Gemini-3.5-Flash-Lite, whose item-centric correlation dips below its baseline. The qualitative conclusions of Section 7 are therefore stable across a wide capacity range.

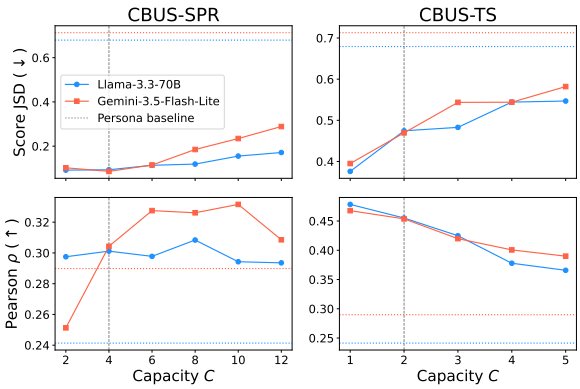

Figure 15: Capacity (C) sensitivity for SPR (left) and TS (right) for two selected backbones. Top: student-centric Score JSD (↓); bottom: item-centric Pearson ρ (↑). Horizontal dotted lines show each backbone's persona-prompted baseline (in the matching color); the vertical dashed line marks the cognitively-motivated operating point (SPR $C = 4$, TS $C = 2$).