

Article

A Metacognitive Blind Spot: Student Comprehension, AI Reliance, and the Conceptual Difficulty Gap

Igor Crk *  and Eren Gultepe 

Department of Computer Science, Southern Illinois University Edwardsville, Edwardsville, IL 62026, USA;
egultep@siue.edu

* Correspondence: icrk@siue.edu

Abstract

This exploratory study investigates the relationship between student metacognition, use of artificial intelligence, and empirical performance within a multidisciplinary course on AI. Using data from a sample of college-age, full-time undergraduate students (averaging 18 participants per assessment) enrolled in an in-person junior seminar at a Midwestern U.S. university, we correlate student self-assessments with standard readability metrics (e.g., Flesch–Kincaid), L2SCA metrics, and objective assessment outcomes, analyzing how learners evaluate their own comprehension and how they deploy AI tools in response to the complexity of 21 reading assignments over 8 weeks. We find that students' perceptions of linguistic difficulty correlate with classical readability scores, but their perceptions do not predict their success nor does their engagement with assistive AI. The results suggest that students utilize generative AI tools as a habitual baseline rather than a strategic response to difficult material. We argue that while students can identify surface-level linguistic friction, they fail to recognize deep conceptual hurdles, leading to a false sense of mastery that neither their intuition nor their AI assistants appear to mitigate. We propose that a quantifiable metric, the Conceptual Difficulty Gap (CDG), may be useful for identifying a class of texts that syntactically appear to be simple, but consistently trigger performance failures. Crucially, we uncover a possible metacognitive blind spot: student self-ratings of difficulty are negatively correlated with this gap, implying that student assessments of difficulty are not based on actual conceptual difficulty. Furthermore, self-reported AI reliance shows no correlation with the gap, indicating that students may not be strategically deploying generative AI tools to mitigate conceptual difficulty.

Keywords: generative AI; AI reliance; reading comprehension; readability metrics; illusion of knowing



Academic Editor: Diego Vergara

Received: 16 May 2026

Revised: 3 July 2026

Accepted: 6 July 2026

Published: 10 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons](#)

[Attribution \(CC BY\) license](#).

1. Introduction

As generative AI tools become ubiquitous in higher education, the nature of reading and preparation may be fundamentally shifting. Today's hybrid students no longer appear to simply engage with a text, but instead may be relying on a workflow that combines reading with AI-assisted synthesis, or one that avoids reading the assigned texts altogether. This introduces a critical new variable to pedagogical research: does the availability of AI change how students perceive difficulty, and do they deploy these tools strategically to bridge gaps in understanding? This paper explores these questions within a rigorous interdisciplinary AI course, with readings spanning classical philosophy, computing history, and evolutionary biology, where the intrinsic density of the reading material may serve as

a stress test for both student comprehension and AI capability. With a modest enrollment and the specialized nature of the course, this study is an exploratory investigation intended to pilot a novel metric rather than provide definitive predictive generalizability.

A fundamental paradox of today's education is that while the varied learning environments provide students with unprecedented autonomy, they may also simultaneously be short-circuiting the cognitive architecture required to navigate them (Azevedo & Dever, 2022). For example, generative AI can produce highly fluent summaries that remove syntactic complexity but inadvertently sabotage the students' metacognitive monitoring that should trigger strategic learning interventions. When reliance on generative AI becomes habitual rather than strategic, recent evidence suggests that it may induce a state of mental laziness and demotivation for independent study (Sanz-Tejeda et al., 2025), as students reflexively lean on generative AI instead of dedicating the cognitive effort to overcome the friction encountered in complex texts. The missing effort is a necessity for deep comprehension.

While classic quantitative readability metrics and students themselves can identify surface-level indicators of linguistic friction (such as dense syntax or rare vocabulary), both are blind to accessible prose hiding profoundly abstract concepts. Coupled with an over-reliance on generative AI, the resulting illusion of mastery may leave students completely unaware of their own comprehension failures until they encounter objective assessment. We argue that in a hybrid learning environment, one which incorporates generative AI as an assistive learning technology, the primary challenge is no longer access to information, but the cultivation of metacognitive judgment required to know when that information has actually been understood. We believe this work to be complementary to the goal of AI transitioning from content delivery to becoming an external regulator of learning by accurately detecting, tracking, and scaffolding a learner's metacognitive and self-regulated learning states (Azevedo & Wiedbusch, 2023).

2. Related Work

The integration of artificial intelligence (AI) tools into education has gained significant attention, particularly in enhancing students' reading comprehension and its measurable impact on academic performance. Recent studies highlight both the promise and complexity of this integration across various educational settings (Chen et al., 2020; Giannakos et al., 2025).

However, the ubiquity of generative AI in educational environments has the potential to alter how students process information and evaluate their own performance. In this vein, recent work moves beyond earlier research into task completion and surface-level performance metrics and emphasizes a widening gap between performance gains and learning outcomes. A primary mechanism underlying this divergence is the "illusion of fluency", wherein students mistake the ease of processing highly polished, externally generated AI content for authentic internal comprehension (Bjork et al., 2013). When a student encounters a text with simple syntax but abstract concepts, the ease of decoding the words creates a "fluency heuristic" that suppresses metacognitive monitoring (Glenberg et al., 1982). Very recent work observed that when text revision or content parsing feels effortless, students experience diminished motivation to cross-reference primary sources or interrogate structural claims, termed "metacognitive laziness" (Fan et al., 2025). While these studies establish the presence of a fluency heuristic, there remains an empirical gap in how this phenomenon is quantitatively modeled relative to specific text architectures, i.e., whether a lack of syntactic resistance, encountered through use of LLMs as an assistive tool for reading comprehension, creates a false sense of mastery and leaves the conceptual gap unnoticed until assessment.

Other evidence along the same lines suggests that, because large language models produce linguistically cohesive, highly structured, and authoritative text, they generate a false sense of correctness and depth (Kumar et al., 2026). Similarly, evidence suggests that

students who rely on generative AI tend to produce less sophisticated scientific arguments and engage less in reflective inquiry (Stadler et al., 2024) and that generative AI tools may inhibit critical engagement and can potentially lead to long-term over-reliance on the tool and diminished skill for independent problem-solving (Lee et al., 2025). The existing literature leaves open the question of the intentionality of this reliance, i.e., whether students treat AI as a strategic cognitive scaffold for text density or as a non-strategic baseline.

Students relying on AI-generated summaries and study notes perceived reading comprehension tasks as easier, but demonstrated worse long-term retention than peers who effortfully generated their own notes (Kreijkes et al., 2025). Over-reliance on conversational models has also been shown to result in lower performance on subsequent assessments, attributed to students bypassing various active retrieval practices like self-quizzing, which are essential to learning (Barcaui, 2025).

3. Materials and Methods

Study participants were 22 college-age adults, generally in their third year of study, as full-time students who were enrolled in an in-person seminar course at a Midwestern University in the United States that met once a week for 16 weeks. Data was collected in the first 9 weeks of the course, which met once per week in the Fall of 2025. On average, 18 students participated in any given week. The data consists of a brief Likert-scale questionnaire and a set of multiple choice questions (MCQs) assessing comprehension for each week's assigned readings. Questionnaire and comprehension MCQs were completed by each student anonymously at the start of each week's meeting. No personally identifiable information was collected, anticipating that collecting such information would have skewed the data by students under-reporting their reliance on LLMs during the week's preparations for class. However, this decision made longitudinal analyses of the individual respondents impossible. Participation was zero-stakes and optional. Instructors used the week's MCQs only as an gauge of overall engagement with the material.

3.1. About the Course

The course is a semester-long seminar integrating classical philosophy, the history of computing, and evolutionary biology in a multidisciplinary framework for building an intuitive understanding of Artificial Intelligence and developing knowledge to assess the field's most controversial claims, such as those made about conscious AI and Artificial General Intelligence. The required course readings formed the basis for guided discussions in the first eight weeks of the sixteen week course. The weekly readings which were assessed at the start of each week's class meeting are detailed in Table 1.

Table 1 describes texts used to build the course's philosophical foundations, beginning with Descartes, who posits the mind as a distinct "thinking thing", Locke, who argues for the mind as a blank slate populated by experience, Hume, who deconstructs the self into a "bundle of perceptions", Hegel, who explores the development of Spirit and History, and Nietzsche, who critiques the intellect and truth as anthropomorphic metaphors. These inquiries into the nature of mind transition into the technical realization of machine intelligence through Ceruzzi's history of digital computing, alongside foundational AI texts by Turing, who establishes the "Imitation Game" (a.k.a. the Turing Test), McCarthy, who defines the field in the Dartmouth proposal, and Minsky, who explores heuristic programming and problem-solving. Bridging these domains, Dennett applies an evolutionary lens to explain how "competence without comprehension" leads to intelligent design, while Churchland critiques classical functionalism in favor of neurobiology. Readings from Baars, Lane, and Metzinger offer contemporary scientific theories of consciousness, ranging from

the “theater” metaphor and Global Workspace Theory, to the biochemical roots of feeling and the self as a representational “Ego Tunnel”.

Table 1. Course readings by week. Parenthesized numbers identify that week’s text. MCQs are numbered by week, text identifier, and sequence number; e.g., 412 identifies the second question about Dennett in week 4.

Week	Label	Source
1	descartes (1)	Descartes, René. <i>Meditations on first philosophy</i> (1641). <i>Meditation One: Concerning Those Things that can be Called into Doubt</i> , 6 pages.
1	locke (2)	Locke, John. <i>An Essay Concerning Human Understanding</i> (1690). <i>Book II, Chapter I, Of Ideas in General, and Their Original</i> , 8 pages.
1	hume (3)	Hume, David. <i>A Treatise of Human Nature</i> (1739). <i>Section VI, Of Personal Identity</i> , 7 pages.
2	dennett_pt1 (1)	Dennett, Daniel. <i>From Bacteria to Bach and Back</i> (2017). Chapters 1–5. About 100 pages.
3	churchland (1)	Churchland, Paul M. <i>Functionalism at 40</i> (2005). 18 pages.
3	nietzsche (2)	Nietzsche, Friedrich. <i>On Truth and Lies in a Nonmoral Sense</i> (1873). 12 pages.
3	hegel (3)	Hegel, G.W.F. <i>Selection from Reason in History: A General Introduction to the Philosophy of History</i> (1837). 4 pages.
4	dennett_pt2 (1)	Chapters 6–11. About 110 pages.
5	ceruzzi_pt1 (1)	Ceruzzi, Paul E. <i>Computing, A Concise History</i> (2012). Chapters 1–3. About 90 pages.
5	turing (2)	Turing, Alan M. <i>Computing Machinery and Intelligence</i> (1950). 28 pages.
5	mccarthy (3)	McCarthy, John, et al. <i>A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence</i> (1955). 12 pages.
6	minsky (1)	Minsky, Marvin. <i>Steps Toward Artificial Intelligence</i> (1961). 23 pages.
6	ceruzzi_pt2 (2)	Chapters 4 and 5. 39 pages.
6	dennett_pt3 (3)	Chapter 12. About 28 pages.
7	ceruzzi_pt3 (1)	Chapters 6 and 7. 39 pages.
7	mccarthy2 (2)	McCarthy, John. <i>What is Artificial Intelligence?</i> (2007). 14 pages.
7	dennett_pt4 (3)	Chapter 13. About 40 pages.
8	baars (1)	Baars, Bernard J. <i>On Consciousness—Science and Subjectivity</i> (2021). <i>Part 2: A Scientific Approach to Consciousness</i> . 55 pages.
8	dennett_pt5 (2)	Chapter 14. About 30 pages.
8	lane (3)	Lane, Nick. <i>Life Ascending</i> (2009). (<i>Chapter 9</i>) <i>Consciousness</i> . 13 pages.
8	metzinger (4)	Metzinger, Thomas. <i>The Ego Tunnel</i> (2009). Introduction-Chapter 2. About 36 pages.

Comprehension MCQs were constructed to assess understanding of concepts communicated by the readings and independently validated to ensure that they were accurate and well-composed according to Haladyna’s multiple-choice item writing guidelines (Haladyna et al., 2002). Additionally, the weekly questionnaire included a brief survey regarding the preparations for the week’s class consisting of four parts:

1. Choosing an appropriate ratio of reading/AI use for the week overall (reading 100% and no help from AI 0%; reading 75% and some help from AI 25%; reading 25% and a lot of help from AI 75%; not reading 0% and only using AI 100%; and no preparation).
2. Self-rated comprehension of the week’s readings (not at all; somewhat; mostly; completely).
3. Self-rated difficulty of each of week’s readings (very easy; easy; neutral; difficult; very difficult). One rating was provided for each text assigned for that week.
4. Choosing an appropriate ratio of reading/AI use for each of the week’s readings (using the same scale as part 1).

3.2. Identifying “Fair” MCQs

Since we intentionally did not collect personal identifiers, we cannot calculate discrimination indices for this dataset. Instead, we rely on a fairness metric combining item difficulty (percent correct responses) and normalized Shannon Entropy.

As a measure of response distribution, we use Shannon Entropy, H , computed for an MCQ with n options as shown in Equation (1).

$$H = - \sum_{i=1}^n (p_i * \log_2(p_i)) \quad (1)$$

where p_i is the fraction of students choosing option i .

It follows that maximum entropy ($H_{max} = \log_2(n)$) for an MCQ with four options occurs when all choices are equally likely and $H_{max} = 2$, and for MCQs with five options $H_{max} = 2.322$. Since our MCQs may have four or five options, we calculate the normalized Shannon Entropy (H_{norm}), as in Equation (2).

$$H_{norm} = \frac{H}{H_{max}} \quad (2)$$

High entropy may indicate that a question is unfair or ambiguous, while low entropy may indicate that a question is too easy or obvious. However, high entropy may also indicate guessing, while low entropy alone could indicate that a question is trivial (all answer the same correct option) or flawed (all answer the same incorrect option). To establish that a question is “fair”, we combine entropy and item empirical difficulty (percent correct responses) in Equation (3) and remove all low-scoring MCQs (*Fairness Score* < 0.3) from subsequent analysis.

$$\text{Fairness Score} = H_{norm} * (1 - p_{correct}) \quad (3)$$

3.3. Readability and Complexity

We compute traditional readability and L2SCA metrics for all course texts to establish a basis for comparison with empirical difficulty and the collected questionnaire data. The course texts are detailed in Table 1. After conversion to plain text files and cleanup of various conversion artifacts, we use a collection of Python scripts to calculate the traditional metrics: Flesch, Flesch–Kincaid, Gunning–Fog, SMOG, Coleman–Liau, ARI, and Dale–Chall.

L2SCA metrics are additionally computed with the L2 Syntactic Complexity Analyzer 4.2.2 (Lu, 2010). The structures recognized and counted by L2SCA, as well as the syntactic complexity indices computed using these, are shown in Table 2.

Table 2. L2SCA metrics.

Metric	Description
W	Count of words.
S	Count of sentences.
VP	Count of verb phrases (verb and its arguments except the subject of an independent clause or coordinate clause).
C	Clauses (subject and a verb).
T	Count of T-units (minimally terminable units, i.e., shortest grammatically allowable sentence)
DC	Count of Dependent Clauses
CT	Complex T-Units (main clause, all subordinate clauses, and non-clausal modifiers)
CP	Coordinate Phrases (grammatically equivalent words, phrases, or clauses connected by a coordinating conjunction)
CN	Complex Nominals (single units consisting of nouns and modifiers, which can be nouns or adjectives)
MLS	Mean length of sentence
MLT	Mean length of T-Unit
MLC	Mean length of clause
C/S	Clauses per sentence
VP/T	Verb phrases per T-unit
C/T	Clauses per T-unit
DC/C	Dependent clauses per clause
DC/T	Dependent clauses per T-unit
T/S	T-units per sentence
CT/T	Complex T-unit ratio
CP/T	Coordinate phrases per T-unit
CP/C	Coordinate phrases per clause
CN/T	complex nominals per T-unit
CN/C	complex nominals per clause

4. Results

We used 28 MCQs to evaluate comprehension across 8 assessments in a course with 22 enrolled students and an average of 18 students assessed each week, resulting in 486 MCQ responses in total. Since individuals could not be identified, discrimination indices cannot

be computed. Instead, we rely on empirical measures of difficulty and Shannon entropy to eliminate unfair MCQs.

Figure 1 shows the empirical difficulty (calculated as $1 - p_{correct}$) and fairness scores for each MCQ. Of the 28 MCQs, we identify 9 with Fairness Scores less than 0.3, which are potentially poorly posed or contain ambiguous or misleading options. By setting the threshold at 0.3, we ensure that the questions that remain are statistically stable: they are neither so easy that they generate no meaningful data (e.g., everyone answered correctly) nor are they so defective that they introduce random statistical noise. For example, if a question has moderate difficulty, where $1 - p_{correct} = 0.40$ (or 60% of students answered correctly), it requires a normalized entropy of at least 0.75 ($0.75 * 0.40 = 0.30$) to keep, and we consider 0.75 a reasonable spread among the remaining distractors that indicates neither total confusion nor a single broken choice.

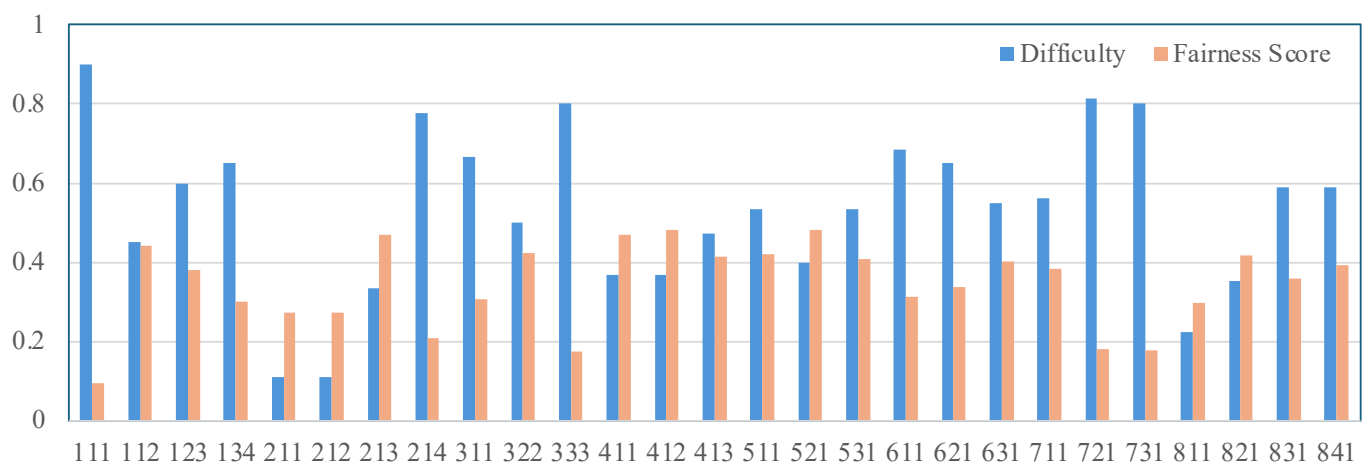


Figure 1. Difficulty, calculated as $1 - p_{correct}$, and Fairness Score for each deployed MCQ. MCQs are identified by unique identifiers constructed from week of assignment, reading number, and a sequence number to account for multiple MCQs assessing the same reading. For example, MCQ 214 is the fourth question assessing comprehension of text dennett_pt1 (1) assigned for week 2. See Table 1 to identify the texts associated with the first two digits of MCQ identifiers.

We also remove any data where the students reported no preparation for the week's discussion (i.e., 0% effort in reading and 0% effort in AI-assisted reading). The following analysis is based on the remaining 267 rows of data.

4.1. Correlational Analyses

All Spearman's rank correlation results are represented comprehensively in Figure 2. It can be observed that the strongest correlations overall are due to trivial redundancy in the collected measures. For example, counts of Words (W), Sentences (S), Verb Phrases (VP), Clauses (C), and T-units (T) are all highly correlated with each other since they all scale with the text. Similar redundancy is present in the calculations of readability metrics like Flesch–Kincaid and Gunning–Fog, resulting in trivial correlations.

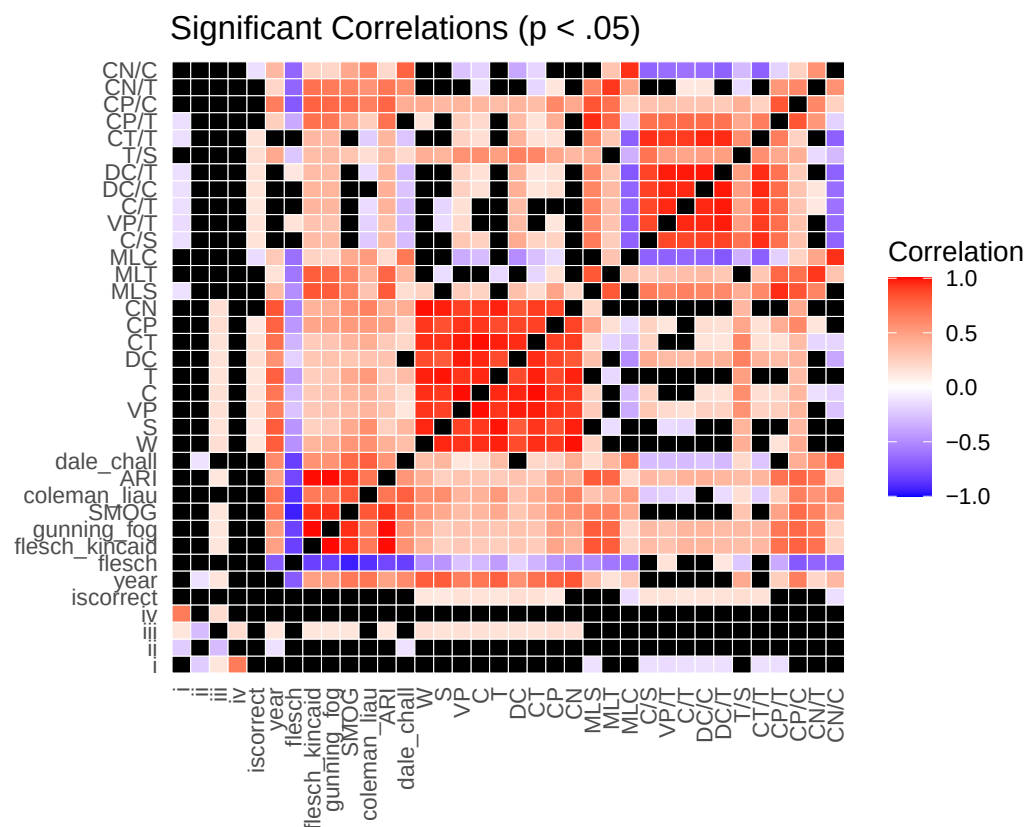


Figure 2. Pairwise Spearman's Rank Correlation indices for all student assessments of own preparation and reading difficulty, performance on assessment questions, and computed textual complexity metrics. Correlations that did not reach statistical significance at $p < 0.05$ and which have an undefined p -value (i.e., the diagonal, where there is zero variance) are masked.

4.1.1. Correlations with MCQ Difficulty

Correlations between student performance and the various textual features are generally weak, suggesting that no single traditional metric, like Flesch–Kincaid or Gunning–Fog, nor any of the more recent L2SCA metrics, strongly predict student performance on their own.

The highest positive correlations to student performance are in the raw count metrics, like Dependent Clauses (DC, $r \approx 0.16$, $p = 0.004$), T-units per sentence (T/S, $r \approx 0.16$, $p = 0.004$) and Verb Phrases (VP, $r \approx 0.16$, $p = 0.010$). These suggest that as the number of dependent clauses or verb phrases increases, the proportion of correct answers increases. In other words, students were more likely to answer correctly when reading longer texts, containing more clauses and sentences.

Highest negative correlations are associated with Mean Length of Clause (MLC, $r \approx -0.13$, $p = 0.014$) and Complex Nominals per Clause (CN/C, $r \approx -0.12$, $p = 0.025$), suggesting that as clauses become longer and more dense with complex nominals, student performance on associated MCQs decreases. In other words, students were less likely to answer correctly when faced with syntactically dense texts: high MLC suggests long strings of thought that require greater attention to process, and high CN/C is a hallmark of academic writing which tends to more densely pack abstract concepts into fewer grammatical units.

All four pretest questionnaire items (i, ii, iii, and iv) show negligible or weak correlations to the correctness of student responses and none reach significance.

4.1.2. Weekly AI Reliance

Students relied more on AI for texts that were syntactically simpler, e.g., shorter sentences and fewer clauses. The strongest correlations for Item 1 are negative with metrics

of sentence density. Students relied on AI more when number of clauses per sentence was lower (C/S, $r \approx -0.16$, $p = 0.021$), when there were fewer verb phrases per T-Unit (VP/T, $r \approx -0.16$, $p = 0.015$), and when the ratio of dependent clauses was lower (DC/C, $r \approx -0.15$, $p = 0.030$).

4.1.3. Understanding

Item 2 correlated negatively with Dale–Chall Readability Scores ($r \approx -0.12$, $p = 0.045$), meaning that students reported better understanding of the texts when they contained more familiar (easier) vocabulary.

4.1.4. Perceived Difficulty

Students' perceptions of difficulty align somewhat with objective complexity metrics. Their self-ratings of text difficulty (3) correlate most significantly with standard readability formulas like Flesch–Kincaid ($r \approx 0.16$, $p = 0.035$), Gunning–Fog ($r \approx 0.16$, $p = 0.031$), SMOG ($r \approx 0.15$, $p = 0.034$), and ARI ($r \approx 0.16$, $p = 0.038$).

Students also rated longer texts as more difficult. Item 3 correlates positively with Word Count (W, $r \approx 0.21$, $p = 0.005$) and Sentence Count (S, $r \approx 0.20$, $p = 0.016$).

4.1.5. Text-Specific AI Reliance

There is no significant correlation between the objective difficulty of a text and the self-reported amount of reliance on AI usage. Students were just as likely to lean heavily on AI for a relatively easy text as they were for a dense philosophical treatise. This suggests that AI assistance may have become habitual, decoupled from actual pedagogical need, rather than a scaffold for building knowledge.

4.2. Conceptual Difficulty Gap

As indicated by our analysis, MLC and CN/C are the strongest predictors of text difficulty, having a negative correlation to average performance on the MCQ associated with each text. To attempt to determine how the self-assessments in the pretest questionnaire relate to performance, we compute a novel measure termed the Conceptual Difficulty Gap based on MLC (CDG_{MLC}) and CN/C (CDG_{CNC}).

CDG_{MLC} computes the difference of the standardized z-scores of the MCQ's empirical hardness, as how often students failed compared to the average, and of the syntactic density of the associated text, as the length of clauses in the associated text compared to the average. A positive gap is seen for texts with relatively simple syntax but with content that is abstract or otherwise conceptually challenging for the students (e.g., Descartes' "I think, therefore I am"). A negative gap indicates a dense text structure that was nonetheless understood well by students.

$$CDG_{MLC,i} = \frac{\mu_p - p_i}{\sigma_p} - \frac{MLC_i - \mu_{MLC}}{\sigma_{MLC}} \quad (4)$$

CDG_{CNC} computes the difference of the standardized z-scores of the MCQ's empirical hardness and of the information density of the associated text, as in how dense is the text with complex noun phrases compared to the average. A positive gap is seen for texts that are written more simply than the average (less information-dense), but MCQ performance indicates difficulty for the students. Therefore, a positive gap may indicate that the difficulty lies not with the decoding of technical terms, but rather with the relationships between concepts. A negative gap means that the text is thick with jargon and noun phrases, but student performance indicates that the narrative made the complex ideas accessible.

$$CDG_{CNC,i} = \frac{\mu_p - p_i}{\sigma_p} - \frac{(CN/C)_i - \mu_{CNC}}{\sigma_{CNC}} \quad (5)$$

The findings shown in Figure 3 support the presence of a conceptual difficulty gap that aligns with expectations, as students typically anecdotally report difficulties with Descartes, Locke, and Metzinger, for example, while Dennett masterfully couches scientific rigor in conversationally accessible prose and metaphor. While students and readability formulas agree on what looks difficult (long sentences, big words), both fail to account for deceptively simple texts—like Descartes’ meditations or Nietzsche—where the prose is accessible but the concepts are abstract. It is in this gap that student performance collapses, unassisted by non-strategic use of AI. If valid, the CDG would support our premise that that in a hybrid learning environment, the primary challenge is no longer access to information, but the cultivation of the metacognitive judgment required to know when that information has actually been understood.

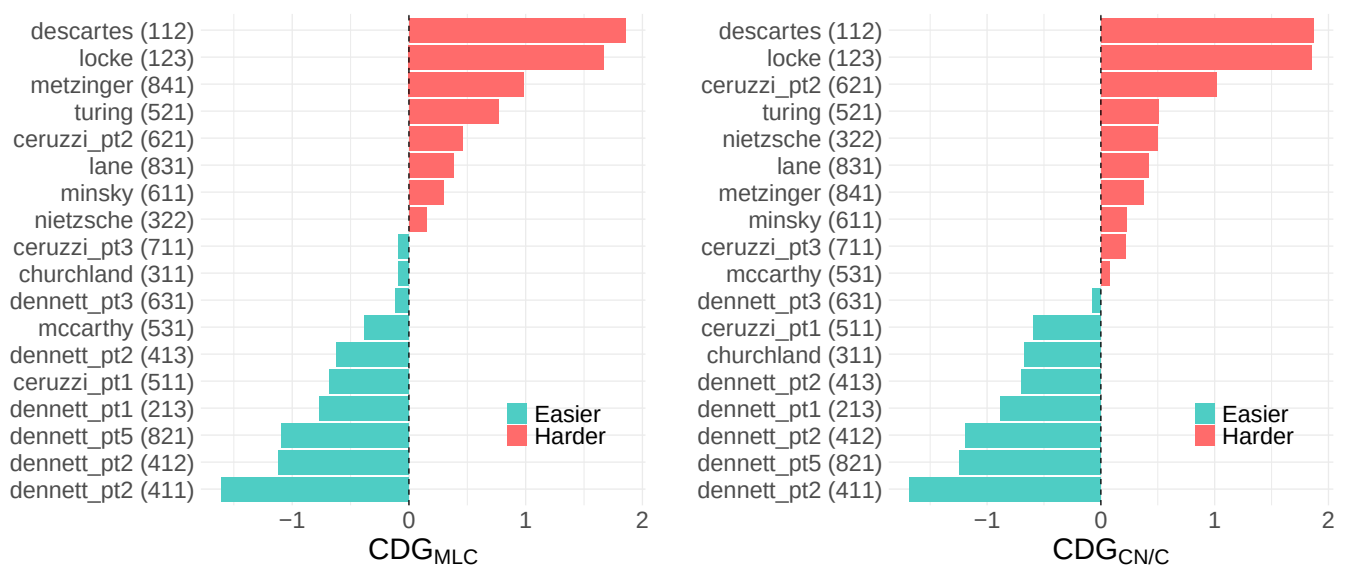


Figure 3. Conceptual Difficulty Gap (CDG) using Mean Length of Clause (MLC) and Complex Nominals per Clause (CN/C).

4.2.1. Statistical Validation of CDG

To show that the CDG reliably captures a unique dimension of text difficulty that other metrics do not, we apply a Spearman correlation analysis to establish the existence of any relationship between self-reported difficulty ratings and CDG. We find moderate negative correlation which was statistically significant between the average class rating of difficulty and CDG_{CNC} ($r \approx -0.53$, $p = 0.022$), and a weak negative correlation between individual ratings of difficulty and CDG_{CNC} which was statistically significant ($r \approx -0.14$, $p = 0.03$). Similarly, we observe a moderate negative correlation between the average class rating of difficulty and CDG_{MLC} ($r \approx -0.42$, $p = 0.087$). This marginal significance is reflective of the low statistical power inherent to our text-level sample size ($N \approx 18$ analyzed readings), emphasizing the exploratory nature of these initial metric validation tracking efforts.

4.2.2. Sensitivity of CDG to Fairness Score

To determine whether the correlation between the self-reported difficulty ratings and CDG is robust across a range of Fairness Score values, we performed a brief sensitivity analysis by varying the Fairness Score threshold. Table 3 shows that the correlations hold outside of the 0.3 threshold that was used throughout the analysis presented here. There was no significant shift in CDG values regardless of the threshold setting; i.e., the texts that had a positive CDG at 0.3 remained positive, and those with a negative CDG remained negative.

Table 3. Spearman correlations of the average class ratings of difficulty and CDG metrics, with p -values for Fairness Score Thresholds (FST) ranging from 0.25 to 0.35. N is the number of rows of data remaining after the threshold-based filtering.

	FST	N	r	p
CDG_{MLC}	0.25	334	−0.46	0.029
	0.30	267	−0.42	0.087
	0.35	226	−0.51	0.054
CDG_{CNC}	0.25	334	−0.51	0.014
	0.30	267	−0.53	0.022
	0.35	226	−0.58	0.022

5. Discussion

The unique pedagogical landscape created by a course that covers a variety of sources, from Turing to Descartes, means that some readings are linguistically dense but conceptually straightforward, while others are syntactically simple but philosophically opaque. Standard readability metrics, which rely on sentence length and syllable counts, catch the former but miss the latter. This study addresses the need for a new metric capable of quantifying student performance in a high-concept environment. Overall, the results could indicate a confidence–competence gap, as there was no correlation between a student’s self-assessed comprehension and their actual test performance. Furthermore, their feeling of comprehension was unrelated to the complexity of the text. This may indicate a metacognitive failure; i.e., students may conflate the readability of an AI-generated digest with their own understanding of the underlying concepts.

We propose the Conceptual Difficulty Gap as a quantifiable measure of this difficulty. The metric calculates the standardized difference (z-score) between a text’s linguistic readability and its empirical failure rate on assessments. In terms of the classic Construction-Integration (CI) model (Kintsch, 1998), a positive gap may arise from a cognitive mismatch between the surface code (syntax and vocabulary of the text) and the situation model (mental representation of its meaning). When a text features low linguistic complexity, such as short sentences and familiar words, the reader experiences high processing fluency at the surface code level. The effortless decoding triggers a misleading fluency heuristic, which suppresses the student’s metacognitive monitoring and creates an “illusion of knowing”. However, because the text’s underlying content is highly abstract and requires complex inferential reasoning, the student fails to construct a valid situation model. So, a positive CDG may be a quantifiable outcome of deceptively smooth reading experiences due to clear AI summaries leading to increased failure rates on well-posed comprehension assessments.

CDG-Guided Intervention

The CDG score is intended as an explanatory tool which can guide pedagogical interventions based on the expected causes of comprehension failure. For example, once established, the CDG can be used to map the difficulty of course texts for future incarnations of the same course. The main idea behind the CDG as an indicator of metacognitive blind spots is that the CDG is directional: a high negative gap appears to contain Dennett’s dense but conceptually accessible narrative, while a high positive gap appears to contain texts with deceptively simple syntactic complexity which masks abstractions of considerable complexity, like Descartes. A negative Spearman correlation between the CDG and self-reported difficulty ratings would indicate that as the invisible conceptual danger of a text increases, the perception of difficulty decreases.

High positive CDG scores appear to identify texts like Descartes, or Locke, where syntax is simple, but failure rates are high. Possible interventions might include Socratic

seminars or comprehension checks that precede the readings and ensure adequate preparation before the readings are attempted.

High negative CDG scores identify texts like Dennett, which likewise deal with complex concepts, but assume far less prior knowledge and as a result employ detailed narrative metaphors that increase the likelihood of comprehension. Possible interventions may include restructuring the course with these texts as the conceptual backbone to guide students through complex systems of thought.

In general, high CDG scores identify texts for which students could be guided to use LLMs to explicitly map the logical dependencies between the premises instead of simply producing summaries. Additionally, students could be required to perform brief in-person pre-reading tasks that highlight the upcoming abstract hurdles and alert students to potential fluency illusions.

6. Limitations

The lack of personal identifiers on comprehension assessments prevented longitudinal tracking of student performance and therefore the use of methods usually employed in evaluating question quality, like discrimination indices. However, we believe that the metrics used are adequate in evaluating question fairness on a per-question basis based on student performance.

The results presented here may be specific to a high-concept, multidisciplinary seminar and may vary from results of more technical or procedural courses. Furthermore, a key limitation of this study is its constrained statistical power. While the absolute Spearman correlation coefficients provide an indicator of observed effect sizes, the underlying text-level sample size means that the findings may be characterized as exploratory insights that require scaling in larger multi-section studies to achieve robust confirmatory power.

Additionally, because the CDG metric relies on empirical assessment failure ($1 - p_{correct}$), a high positive CDG could plausibly manifest due to poorly written assessment questions rather than conceptual difficulty, though we mitigate this by filtering out items with low fairness and entropy scores. The metric also assumes a baseline level of effort, meaning that a widespread lack of motivation or presence of widespread academic dishonesty could inflate failure rates, in which case the CDG may be a reflection of engagement rather than text complexity; we mitigate this with a zero-stakes, anonymous assessment, thought to isolate authentic reading behaviors.

7. Future Work

The Conceptual Difficulty Gap may be a useful tool for guiding targeted scaffolding for instruction, which would form a basis for further empirical study of student outcomes. An example of targeted scaffolding might include a text difficulty heatmap for the course. The heatmap would associate a text's expected difficulty, based on the CDG, with specific pedagogical interventions, for example, additional engagement with abstract concepts of the text before it is assigned, where the pre-reading seminar activity would be commensurate with the difficulty gap: more time preparing for a reading of Descartes and Locke, less or none for Dennett. The easier texts, like Dennett, could be used to navigate jargon and foundational concepts without losing the conceptual thread. Another possible intervention is clear directions in the deliberate use of AI for achieving useful structural decomposition of the texts and in alerting students to the risks of the fluency illusion with the use of calibration quizzes to be completed along with the reading.

We plan to make use of these interventions and reassess correlations of student self-assessments and comprehension quiz performance. It is expected that interventions will likely improve student outcomes and narrow the CDG gap for the same set of texts. A

follow-up study with a revised course in which the suggested interventions will be applied will also provide an evaluation of the predictive and prescriptive validity of the CDG.

8. Conclusions

Our analysis of student performance data yields several interesting findings regarding student comprehension, the nature of students' reliance on AI in preparation for class discussion, and the predictive failure of the students' self-assessments. Our data suggests a metacognitive blind spot in which students' self-reported perception of difficulty is negatively correlated with the Conceptual Gap. In other words, the higher the conceptual danger of a text, the safer students felt while reading it. Because they did not experience linguistic friction, they assumed they understood the material, indicating a fluency illusion that led to lower test scores.

We also investigated whether students used AI tools to bridge this gap. If students were strategically optimizing their learning, we would expect AI usage to increase alongside conceptual difficulty. Instead, we found zero correlation between the Conceptual Gap and the intensity of AI assistance. This observation suggests that students deployed AI as a habitual baseline or a tool for attempting to grasp the meaning of longer texts, rather than as a targeted intervention for conceptual density.

Overall, these findings suggest that the primary risk in modern AI education is not that the material is too hard to read, but that its difficulty is invisible to both the student and their tools. The Conceptual Difficulty Gap may provide a signal that can alert educators to the invisible hurdles where intervention is most needed.

Author Contributions: Conceptualization, I.C.; methodology, I.C. and E.G.; software, I.C.; validation, I.C. and E.G.; formal analysis, I.C.; investigation, I.C.; resources, I.C.; data curation, I.C.; writing—original draft preparation, I.C.; writing—review and editing, I.C. and E.G.; visualization, I.C.; supervision, I.C.; project administration, I.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. The Institutional Review Board of Southern Illinois University Edwardsville (Protocol #3598, 27 May 2026) has determined this project is Not Human Subjects Research as it does not meet the definition of “human subject” under 45 CFR 46 §46.102(e)(1).

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because the data are part of an ongoing study. Requests to access the datasets should be directed to corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Azevedo, R., & Dever, D. (2022). Metacognition in multimedia learning. In *The Cambridge handbook of multimedia learning* (pp. 132–142). Cambridge University Press.
- Azevedo, R., & Wiedbusch, M. (2023). Theories of metacognition and pedagogy applied to AIED systems. In *Handbook of artificial intelligence in education* (pp. 45–67). Edward Elgar Publishing.
- Barcaui, A. (2025). ChatGPT as a cognitive crutch: Evidence from a randomized controlled trial on knowledge retention. *Social Sciences & Humanities Open*, 12, 102287. [CrossRef]
- Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology*, 64, 417–444. [CrossRef] [PubMed]
- Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, 8, 75264–75278. [CrossRef]

- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530.
- Giannakos, M., Azevedo, R., Brusilovsky, P., Cukurova, M., Dimitriadis, Y., Hernandez-Leon, D., Järvelä, S., Mavrikis, M., & Rienties, B. (2025). The promise and challenges of generative AI in education. *Behaviour & Information Technology*, 44(11), 2518–2544. [\[CrossRef\]](#)
- Glenberg, A. M., Epstein, W., & Wilkinson, T.-S. (1982). The calibration of comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 8(6), 585.
- Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, 15(3), 309–333. [\[CrossRef\]](#)
- Kintsch, W. (1998). *Comprehension: A paradigm for cognition*. Cambridge University Press.
- Kreijkes, P., Kewenig, V., Kuvalja, M., Lee, M., Hofman, J. M., Vitello, S., Sellen, A., Rintel, S., Goldstein, D. G., & Rothschild, D. (2025). Effects of LLM use and note-taking on reading comprehension and memory: A randomised experiment in secondary schools. *Computers & Education*, 243, 105514. [\[CrossRef\]](#)
- Kumar, S., Mikayelyan, A., & Vorfolomeyeva, O. (2026). Fluency illusion: A review on influence of ChatGPT in classroom settings. *Information*, 17(3), 299. [\[CrossRef\]](#)
- Lee, H.-P., Sarkar, A., Tankelevitch, L., Drosos, I., Rintel, S., Banks, R., & Wilson, N. (2025). The impact of generative AI on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems* (pp. 1–22). Association for Computing Machinery.
- Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4), 474–496. [\[CrossRef\]](#)
- Sanz-Tejeda, A., Domínguez-Oller, J. C., Baldaquí-Escandell, J. M., Gómez-Díaz, R., & García-Rodríguez, A. (2025). The impact of generative AI on academic reading and writing: A synthesis of recent evidence (2023–2025). In *Frontiers in education* (Vol. 10, p. 1711718). Frontiers Media SA.
- Stadler, M., Bannert, M., & Sailer, M. (2024). Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160, 108386. [\[CrossRef\]](#)

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.