Comparing Zero-Shot Large Language Model Prompting with Human Coding of Theory Concepts in Student Essays

Shelley Keith
University of Memphis
Memphis, TN, USA
srkeith@memphis.edu

Philip I. Pavlik Jr.
University of Memphis
Memphis, TN, USA
ppavlik@memphis.edu

Kristen L. Stives
Seton Hall University
South Orange, NJ, USA
kristen.stives@shu.edu

Laura Jean Kerr
Grand Valley State University
Allendale, MI, USA
kerrlaur@gvsu.edu

Recent studies have explored the cost and time benefits of using artificial intelligence (AI), particularly large language models (LLMs), in coding student essays. While these models show promise, not enough is understood about the factors that affect how their qualitative coding performance compares to human coding. This study examines coding accuracy for content errors in college student essays on criminological theories by comparing human-coded results with outputs from four LLMs. We evaluated human-AI correlations, AI error, and AI bias across four LLMs, five prompt types, three theory content coding dimensions, and four criminological theories. Results indicate that LLM choice significantly influenced human-AI correspondence, with Claude Sonnet 4 exhibiting the best overall performance and GPT 4.1 Mini the worst. Prompt type had minimal impact on performance. Across models, error rates were lowest when identifying whether students listed a concept, and highest when assessing whether definitions were correct. LLMs performed better on concise theories than on more complex ones. The code is available at https://github.com/imrryr/LLM-queries

Keywords: large language models, automated essay scoring, qualitative coding, prompt engineering, student writing, criminological theory

1. INTRODUCTION

When assessing student writing, instructors must consider two important dimensions: conceptual ways to measure student comprehension and the point value (score) to assign to each facet of an assignment. The first task often involves determining what constitutes a correct or incorrect answer, which may be thought of as the process of coding, in which instructors assess the extent of students' comprehension. The second task, scoring an assignment, involves assigning a quantitative value that represents the degree of student comprehension. Evaluating written assignments can be time-consuming and tedious, especially for instructors who aim to provide individualized feedback. Therefore, the aim of automated essay scoring (AES) is to reduce workloads for faculty, process high volumes of essays from high-stakes assessments, and provide consistency of evaluation for course credit (Kooli & Yusuf, 2025; Pack et al., 2024).

AES refers to the use of computer algorithms to evaluate and assign scores to written essays by analyzing features of the text, such as content relevance, organization, grammar, vocabulary, coherence, and sometimes argument quality, to approximate human rater judgments in a consistent and scalable way (Shermis & Wilson, 2021). The present study builds on deductive qualitative content analysis by applying predefined theory-based concepts to student texts in a manner consistent with traditional qualitative methods (McClure et al., 2024). Although traditional AES systems were not explicitly grounded in qualitative methods, both human coding and AES involve the systematic application of analytic criteria to text. While human coders apply these criteria based on professional judgment and expertise, AES systems make determinations algorithmically. In the context of theory-based evaluation of student essays, automated essay scoring can thus be understood as a specialized form of deductive coding. Contemporary AES frameworks simultaneously evaluate multiple dimensions and traits of writing quality, such as grammar, coherence, content relevance, and semantic accuracy, reflecting a shift from relying solely on holistic scoring approaches (Sun & Wang, 2024). While AES has historically relied on simpler computer algorithms to analyze and comment on student essays, the emergence of artificial intelligence (AI), particularly in the form of large language models (LLMs), offers a natural extension of AES capabilities (Mizumoto & Eguchi, 2023).

LLMs are transformer-based neural network models that can generate logically coherent text by training on large amounts of language data (Lo, 2023a; Mizumoto & Eguchi, 2023; Pack et al., 2024). Recent studies note the advantages of using LLMs in the deductive qualitative coding process or applying codes to data based on an established codebook (Kirstin et al., 2024; Jiang et al., 2021). Some evidence suggests that AIs can assist humans in coding data, such as identifying patterns more quickly (Lopez-Fierro & Nguyen, 2024), and for certain tasks, LLM coding is comparable to the level of consensus reached by human coders (Chew et al., 2023; Kirstin et al., 2024). Thus, the use of LLMs may be a beneficial tool to offset the time-consuming process of coding (Morgan, 2023). Indeed, for most coding tasks, AI completes them considerably faster than human coders (Chew et al., 2023; Morgan, 2023) and, in some cases, more accurately (McClure et al., 2024).

While LLMs show promise in automatically scoring and providing feedback for student essays and in coding qualitative data, the results depend on prompt engineering and the choice of AI model (Cain, 2024). Prompt engineering is the art and science of providing clear instructions to AI to produce desired outputs (Lo, 2023a). However, this field is still in its infancy, and to enable AI to evaluate essays effectively, studies are needed to examine how prompts affect how competing LLM models code essays compared to human coding.

Our approach builds on deductive qualitative content analysis, applying predefined theory concepts to student texts in a manner consistent with traditional qualitative methods (McClure et al., 2024), with applications to AES. While the procedure resembles deductive qualitative coding, the objective of the present study is evaluative scoring of theory-based concepts in student writing. Specifically, we translate deductive essay coding rubrics into zero-shot prompts with various levels of information. This paper explores how variations in these zero-shot prompts affect concordance between AI and human coding of criminological theory concepts from student essays. Our goal is to evaluate which model and prompt type best align with human coding. In addition, we examine how the LLMs perform compared to humans on coding dimensions (i.e., listing, attempting to define concepts regardless of accuracy, or correctly defining theoretical concepts) and whether alignment varies by the criminology theories discussed in the student papers. We assess alignment with humans on each of these factors through correlation, error rate, and bias measures.

2. Literature Review

2.1. History and Current Uses of Automated Essay Scoring

Ellis Page created the first AES system, Project Essay Grade (PEG), in the 1960s (Page, 1966). This system was designed to predict writing quality. It used proxy variables (e.g., average sentence length, number of paragraphs, essay length in words) to measure different writing constructs (Chung & O'Neil, 1997). The system was trained using human-scored essays to create prediction models based on correlation statistics (Dikli, 2006). Once trained, subsequent essays were scored based on these models. While the models were able to assess surface-level features, they could not evaluate content, and new prediction models were needed to analyze different essays (Chung & O'Neil, 1997). Since PEG, numerous AES systems have been developed in recent decades, with notable examples such as E-rater, Intelligent Essay Assessor, and Criterion (Dikli, 2006). AES systems have been created based on patterns and statistics, Natural Language Processing, and more recently, deep learning techniques (Ramesh & Sanampudi, 2022).

The key feature of each system is to grade (score) essays using input from previously scored essays. There are three major scoring approaches: holistic, analytical, and trait-based scoring. In a holistic approach, also considered the fastest and least expensive, many features of an essay are considered, but an overall score is generated based on the whole of the piece (Chung & O'Neil, 1997; Ke & Ng, 2019). An analytical approach to scoring includes a scale for each criterion, scores, and feedback for each criterion (Ke & Ng, 2019). Some essay grades or scores are calculated by summing the scale scores, while others use criterion weights. Finally, in trait-based scoring, a single trait or multiple traits of a text piece may be examined. The purpose of the writing determines the exact trait or traits examined and, in some cases, by the availability of raters, as multi-trait scoring is considered the most labor-intensive, time-consuming, and cost-intensive of the three scoring approaches (Ohta et al., 2018). Ideally, an AES system would yield the same overall score for an essay regardless of the scoring approach.

Establishing consistency in scoring between humans and AES systems is necessary before instructors can fully make use of these systems. Past research suggests AES systems can reach high levels of agreement with human raters (Lim et al., 2021; Yun, 2023). AES systems offer consistency in scoring for grading, objective assessment, the ability to detect plagiarism, and the ability to identify grammar, mechanics, and other linguistic features (Sevcikova, 2018). However, they have also been criticized for their emphasis on surface-level information and inability to evaluate content development (Sevcikova, 2018) or linguistic context (Atkinson & Palma, 2025). Additionally, these systems can be costly to run (Sevcikova, 2018). Finally, AES and human raters are also subject to reliability and validity concerns (Ramesh & Sanampudi, 2022). Some of these concerns may be addressed using LLMs.

The use of LLMs for AES is a more recent area of investigation (Mizumoto & Eguchi, 2023). AES systems were designed with a focus on essays to evaluate content mastery or grammar and mechanics, whereas LLMs were developed more broadly to assess and generate human language, answer questions, and complete other language-related tasks, offering a range of opportunities for learning (Kasneci et al., 2023). The response generated by an LLM depends on the model used and varies across tasks (Mansour et al., 2024; Shen et al., 2023).

Mizumoto and Eguchi (2023) provide an early test of the ability of the ChatGPT-3.5 text-davinci-003 model to automatically score TOEFL11, a test for non-native English speakers. TOEFL (Test of English as a Foreign Language) evaluates English proficiency and scores each essay as low, medium, or high. They find that a high degree of accuracy exists between the

benchmark scores of TOEFL11 and what ChatGPT-3.5 can produce, and that the inclusion of lexical features improves scoring accuracy (Mizumoto & Eguchi, 2023). A recent study from Pack et al. (2024) examined interrater reliability between humans and four LLMs: Google's PaLM, Claude 2, GPT-3.5, and GPT-4.0, using a sample of college admissions language proficiency essays following a 6-point scoring rubric. GPT-4.0 had the highest interrater reliability with humans among the four LLMs, but the authors note that in several circumstances, the LLMs did not perform as expected. For example, some LLMs sometimes added language to the essay and then scored it, while at other times they assigned scores that did not match the rubric used by human coders (Pack et al., 2024). Continuous testing of LLMs is needed to determine which LLM produces scores most similar to those of human coders, as new models are released very frequently.

2.2. Uses of LLMs in Qualitative Coding

While some aspects of essay scoring are purely quantitative, it also involves qualitative coding. Qualitative coding is the process used to discover patterns within information (Auerbach & Silverstein, 2003), in which codes/labels (e.g., words, phrases) are created by researchers to denote some aspect(s) of the data (Linneberg & Korsgaard, 2019). Using inductive reasoning, a researcher may develop codes as they emerge from the data, without being influenced by prior work (e.g., Barany et al., 2024). Alternatively, using deductive reasoning during the coding process allows a researcher to draw on previous research to create pre-defined codes to apply to the data (Linneberg & Korsgaard, 2019). When testing existing theory, deductive coding is advantageous because researchers can use predefined concepts and look for evidence that supports them (Williams & Moser, 2019).

Whether a researcher is manually sorting and organizing information into themes or using coding software, the coding process is time-consuming (Jiang et al., 2021). Handling large amounts of data can be burdensome and make analyses difficult (Jiang et al., 2021). One scoring method, primary trait scoring, is similar to qualitative, deductive coding in that it focuses on identifying the presence and quality of an essay's aspect or trait (Saunders, 1999). A trait of an essay is selected by determining the purpose or function of the writing assignment (Lloyd-Jones, 1977). Once the writing trait is defined, the coder examines the text for the presence and quality of the trait, which then determines the score applied to the essay.

More recently, LLMs have offered new opportunities to advance qualitative coding for research purposes, in essay scoring, and in theory-based concept identification, especially as LLMs can more readily analyze larger datasets than humans, enabling more generalizable results (Bano et al., 2024). LLMs can produce codes and scores comparable to those of human coders and can even suggest codes not originally created by human coders, which may be of higher quality (Barany et al., 2024). LLMs are a useful tool for qualitative coding, as these automated systems can be modified to code for a variety of constructs. LLMs, such as GPT, make coding decisions based on the input (e.g., a definition) they are given and may be less sensitive to the interpretation of constructs than human coders, who may bring their own interpretations (Liu et al., 2025). Providing detailed, context-rich information to AIs can improve the LLM's ability to code precisely (Zhang et al., 2024). Therefore, it is important to understand which information to provide the LLMs to produce the most accurate coding.

2.3. Prompt Engineering

LLMs can qualitatively code because they can generate coherent, contextually appropriate responses to user inputs or prompts (Cain, 2024) by learning language patterns from vast amounts

of textual data (Velásquez-Henao et al., 2023). Prompts should not be too complex, to ensure LLMs do not misunderstand the input, and thus, provide biased responses or hallucinations, i.e., made-up or incorrect information (Cain, 2024). Therefore, prompt engineering, "techniques and methods to design, write, and optimize instructions for LLMs," has emerged as an important field that guides users in obtaining accurate answers to their queries (Velásquez-Henao et al., 2023, p. 11).

Researchers must carefully design prompts based on their content knowledge to avoid outputs (responses) that are biased, incorrect, or misleading (Cain, 2024; Velásquez-Henao et al., 2023). Inputs to LLMs follow various prompt strategies, including zero-shot, one- or few-shot, and chain-of-thought prompts. In zero-shot prompts, the model is given a task with no labeled examples, relying solely on information learned during LLM training (Brown et al., 2020; Liu et al., 2025). Zero-shot prompts may still include detailed task instructions or conceptual definitions. In contrast, one-shot and few-shot prompts provide the model with one or more examples within the prompt, enabling in-context learning (Liu et al., 2025; Tripathi et al., 2025). Prompting strategies may also include chain-of-thought (CoT) instructions, in which the human instructs the LLM to communicate its reasoning step-by-step, thereby improving task performance and the interpretation of the output (Wei et al., 2022). CoT was developed to address more complex reasoning tasks such as solving detailed math problems (Tripathi et al., 2025).

Research has begun to examine different LLMs and various prompting strategies that yield the most accurate responses, especially as compared to human coding (Kim et al., 2023). Radford et al. (2019) demonstrated that large-scale language models like GPT-2 can perform a variety of language tasks in a zero-shot setting, marking a major shift in how researchers understood the flexibility and potential of unsupervised LLMs. However, zero-shot prompting may not always produce scores that match human ratings. For example, Johnson and Zhang (2024) used zero-shot scoring on 13,121 essays, representing eight different essay prompts, and asked GPT-4o to assign a numerical score on a scale of 1 to 6. This study found AI scores were 0.9 points lower on average than those assigned by humans and may have produced biased scores for Asian/Pacific Islander students who received even lower scores on average (Johnson & Zhang, 2024).

LLM performance may be influenced by whether an open-source or closed-source model is used and by the recency of the LLM's version. In a recent study, Seßler et al. (2025) compared five LLMs (GPT-3.5, GPT-4, GPT-o1, LLaMA 3-70B, and Mixtral 8x7B) to human coders using a zero-shot prompt on language-based and content-based criteria in student essays. The researchers found that closed-source models had higher overall correlations with human ratings and that these correlations increased with each successive model version. The models aligned more closely with human scoring on language-related essay criteria rather than content-related aspects. Of all the LLMs, the most recent version assessed, GPT-o1, had the highest correlation with humans in the overall category ($r = 0.74$) (Seßler et al., 2025). "These patterns suggest that the architectural and training differences between closed-source and open-source models significantly impact their evaluation strategies and reliability" (Seßler et al., 2025, p. 470).

When it comes to few-shot prompting, the level of specificity in an example may make a difference in the scores LLMs produce compared to humans. Yoshida (2024) found that GPT-3.5 models were more sensitive to the selection and order of examples than GPT-4, with GPT-4 performing well with zero-shot prompting. Whether including one or multiple examples, it is important for researchers to carefully construct their examples, as they can influence how well models perform (Yoshida, 2024). Liu et al. (2025) examined the potential of GPT-4 in coding qualitative data using four prompting strategies: zero-shot, few-shot with positive examples, few-shot with positive and negative examples, and few-shot with context. They found that few-

shot prompting generally improved how well the model performed in comparison to zero-shot prompting, but examples need to be broadly representative of the data, and ideas that may be difficult for humans to code are also hardest for GPT-4 to code (Liu et al., 2025).

McClure et al. (2024) used advances in few-shot and CoT prompting to assess how GPT-4 performed in coding high school students' English papers on basic recall, integrating new information, and generating new ideas, compared to human coders. They found that these refinements to prompt engineering increased accuracy in coding papers for recall and integration, and that both AI and humans struggled to identify the creation of new ideas (McClure et al., 2024). The authors conclude that LLM models such as GPT-4 offer advantages in efficiency and accuracy in scoring information with defined parameters, but human oversight is especially needed for coding more complex student writing tasks (McClure et al., 2024).

Mansour et al. (2024) used four prompts to assess how well GPT-3.5-turbo-0301 and Llama-2-13b-chat-hf performed in evaluating English student essays holistically and on specific traits. Specifically, they compared zero-shot prompting with and without additional instructions and persona, and one-shot prompting. While the performance of GPT-3.5-turbo-0301 improved with the inclusion of context in the prompt, it was less sensitive to changes in prompts than and Llama-2-13b-chat-hf, which exhibited more variable performance depending on the prompt and task type (Mansour et al., 2024).

The use of LLM-assisted content analysis (LACA) can refine prompts used for deductive coding. Chew et al. (2023) explain that this approach entails "(1) codebook co-development with an LLM and tests of validity; (2) tests of reliability between human coders and the LLM; and (3) replacement of manual coding with LLM coding for the final coded data set" (p. 3). They test the use of role-based prompts, chain-of-thought prompts, and zero-shot prompts on four datasets, including Tweets, paragraph-length texts, news articles, and reports, using GPT-3.5-Turbo, and found a high level of agreement between the LLM and human coders. Using LACA, they were able to better identify code sections with high disagreement between the LLM and humans. Thus, this approach can help improve coding for qualitative research and LLM performance (Chew et al., 2023).

In sum, while various prompting strategies, such as few-shot prompting, show improved performance on specific tasks, prior research demonstrates that model outputs can be sensitive to the selection of in-context examples, affecting scoring outcomes and concordance with human raters (Liu et al., 2025; Yoshida, 2024). Zero-shot prompting offers an advantage in that models can be compared more directly, and biases related to exemplar selection are reduced. Accordingly, a zero-shot prompting strategy provides a more reliable baseline for comparing LLM-based coding to human coding. The present study uses zero-shot prompting as the baseline while varying specific instructions or types of information provided to the LLM to assess how these prompt variations affect the results.

2.4. CURRENT STUDY

The goal of this study is to compare AI-generated coding of theoretical concepts in student papers with human-coded data using zero-shot prompts. Specifically, we vary both the LLM used and the instructions provided as part of zero-shot prompts to assess how these factors influence concordance with human coding. We compared four different LLMs—ChatGPT-4.1 Mini, ChatGPT-4.1 Full, Gemini 2.5 Pro, and Claude Sonnet 4. We also examine whether the coding dimensions of the tasks (i.e., listing, defining, and defining correctly) and the complexity of the theory affect agreement between LLMs and human coders. Additionally, we test potential interactions between prompt type and the LLM, and between prompt type and

the coding dimension. We highlight these possible interactions based on previous research, which found that LLMs perform differently depending on prompt type and that variations in prompts affect task accuracy (Mansour et al., 2024).

In text and discourse annotation, substantial debate exists over the appropriate agreement coefficient, including whether to use unweighted or weighted kappa and, if so, which weighting scheme (e.g. Hayes and Krippendorff, 2007). Because our design treats human ratings as a fixed reference and focuses on the magnitude and structure of disagreement rather than chance-corrected rater symmetry, we report correlation, error, and bias. Indeed, the selected metrics jointly capture systematic shift, magnitude of disagreement, and rank-order agreement, providing a more interpretable decomposition of the same underlying agreement structure. This provides more information than a single measure like alpha or kappa, which cannot distinguish constructs such as calibration error, where there is high correlation and high bias, or poor discrimination, where correlation is low, and bias is also low.

We test whether zero-shot prompts that include a variety of informational supports yield results with higher correlation, less bias, and fewer absolute errors than those of human coders. In addition, we expect that the LLMs will be closer to human coding when asked to identify whether concepts are listed, rather than when asked to identify whether concepts are defined or defined correctly. The theories examined, social control theory, self-control theory, deterrence theory, and rational choice theory, vary in the number and complexity of concepts. We expect that correspondence will be closer when the concepts from the theories are in "discrete semantic units" including single words for concepts and/or include fewer concepts (e.g., attachment) versus more complicated ideas (e.g., self-control is stable over the life-course once established), given that AI models likely have trouble identifying "semantically expansive categories" (Hila & Hauser, 2025, p. 283). For example, social control theory includes 6 distinct concepts, while self-control theory includes 13 key concepts. We suspect the AI will correlate with humans and show less bias and error when the theory includes more easily recognizable concepts, as in social control theory. The remaining two theories likely fall in the middle, with deterrence theory including distinct concepts but a large number of them, and rational choice theory including fewer concepts represented by discrete semantic units. Finally, we predict that more advanced and expensive LLMs will outperform less expensive versions due to their enhanced reasoning capabilities.

3. METHOD

3.1. SAMPLE AND PARTICIPANTS

The sample of papers for this study was drawn from two Criminological Theory courses completed in the Fall 2014 and Spring 2015 semesters at a large southeastern public university in the United States. The analytic sample comprises 39 students who completed 67 papers, reflecting the subset of papers selected for analysis from the larger pool of consenting participants. Criminological Theory was a required course for criminology majors and was designated a writing-intensive course. The original purpose of the data collection was to determine whether students improved in writing in the spring semester with the inclusion of a writing center embedded tutor (i.e., treatment group) in comparison to the preceding fall semester with limited involvement of the writing center (i.e., control group) (See Kastner et al., 2018). The data are a good fit for this study because students completed them before the advent of AI, ensuring their papers reflected their understanding of the theories based on the readings and lecture material.

In addition, the data remain relevant given that the theories covered in this course remain foundational in present-day undergraduate and graduate criminological theory courses.

A total of 20 students in Fall 2014 and 26 students in Spring 2015 consented to participate in the study. In the Fall 2014 semester, 60% of the participants were male, 75% were white, 20% were African American, and 5% were Hispanic. The Spring 2015 semester included 42% males, 65% white, 31% African American, and 4% Hispanic. In both Fall 2014 and Spring 2015, students were required to complete four of six possible short paper assignments. Two of these papers needed to be completed before the midterm, and two more before the final exam. In each paper, students were required to write between four and six pages describing and comparing two related criminological theories, designated as Theory A and Theory B for analysis.

For this study, two paper assignments were analyzed to assess how AI coding compares with human coding. These two paper assignments were selected because the highest number of students completed them for both semesters, yielding 33 papers for the third short paper assignment, in which Theory A referred to Social Control Theory and Theory B referred to Self-Control Theory, and 34 papers for the sixth short paper assignment, in which Theory A referred to Deterrence Theory and Theory B referred to Rational Choice Theory. The sample included 39 students, with 22 of these students choosing to complete both papers.

3.2. CODEBOOKS AND PROMPTS

3.2.1. Human codebooks

To compare students' understanding of criminology theories in the two courses, a codebook was created by a four-person research team including the course instructor, the course-embedded tutor affiliated with the Writing Center, the graduate student teaching assistant (TA) for the academic year of the data collection, and the TA from a previous section of the course. The codebook was constructed using concepts from each theory, drawing on the instructor's lectures and assigned readings. Each concept in the codebook was defined concisely, and corresponding textbook page numbers were provided with every code to offer clarity and reference context (See Appendix 1). To gauge students' level of understanding of the theories, the team coded each paper for whether: (1) the student listed the concept, (2) the student defined the concept, and (3) the student defined the concept correctly. The number of concepts listed, defined, and defined correctly was totaled for each theory within a student's paper and recorded in a database (See Appendix 1).

For a concept to be considered "listed," a student needed to include a word or phrase that referred to that specific idea. For a concept to be considered "defined," a student needed to attempt to explain the concept, regardless of the accuracy of the student's definition. Finally, for a concept to be considered correctly defined, a student needed to include an accurate definition or a quotation from the reading with an additional interpretation of its meaning.

After drafting the codebook and after each revision, research team members applied it to a student paper for comparison. When discrepancies in coding occurred, additional clarifications were added to the codebook to ensure valid and reliable coding. Over a few weeks, the codebook was refined and finalized to ensure interrater reliability. Refinements included clarifying the definitions of the concepts so that each coder could accurately and consistently identify when a concept was defined correctly.

After the codebook was finalized, the research team split into pairs: the course instructor and the embedded writing center tutor formed one pair, and the two TAs the other. Each pair independently coded a sample of student papers. The two teams compared their coding until both

teams consistently scored papers the same. The remaining papers were then divided among the two teams for final coding. Within each team, each paper was coded independently by team members to further ensure reliability. While unlikely given the previous norming process, if the two coders within a team diverged, consensus on coding was attained through social moderation (Herrenkohl & Cornelius, 2013). Unfortunately, discrepancies were not tracked, which prevents the calculation of inter-rater reliability.

3.2.2. LLM prompts

Recent innovations in prompt engineering informed the design and refinement of all prompts used in this study (Cain, 2024; Lo, 2023b; Park & Choo, 2024; Velásquez-Henao et al., 2023). Prompt engineering refers to the "process of writing, refining, and optimizing human-defined inputs to obtain high-quality desired outputs from generative AI models" (Park & Choo, 2024, p. 2). To create structured and effective prompts, we drew on the PARTS framework, which emphasizes the inclusion of key components: Persona, Aim, Recipients, Theme, and Structure (Google for Educators, 2024). Persona refers to assigning the LLM a role to understand the task's context better. Second, effective prompts include the Aim or goal of the task. An effective prompt may also designate the Recipient or audience so the LLM can tailor its response (not manipulated in this study). Next, LLMs may be more successful at tasks when provided with the Theme or additional information, tone, style, or restrictions required for the task (Park & Choo, 2024). Finally, a prompt should include the desired output format or Structure such as a code or table.

The goal was to create prompts to be CLEAR: Concise (clear, brief), Logical (well-structured), Explicit (uses precise or specific language), Adaptive (customized to the specific tasks), and Restrictive (giving limits on the format, length, or scope of the task) (Lo, 2023b; Park & Choo, 2024). Relying on insights from Cain (2024) and Lo (2023a), prompts were created and evaluated based on deep subject-matter knowledge and critical thinking about the AI's responses to avoid hallucinations and inaccuracies. The goal of our prompts was to align them with the human codebook and ensure they were consistent and clear to the AI models. To confirm that the prompts were clear, well-structured, precise, tailored, and adequately restricted, meta-prompting techniques were used, in which prompts were refined for each theory's concepts based on input from ChatGPT-4o (Reynolds & McDonnell, 2021).

In line with best practices, our prompts (see Table 1) began with a zero-shot approach, asking the LLM to complete a basic task using its internal knowledge with minimal additional instructions (Velásquez-Henao et al., 2023). Each prompt variation added information to the zero-shot prompt. Each coding task included one of three Aims: to determine whether criminological theory concepts were listed (present or absent), defined (an attempt was made to define the concept), or defined correctly (the concept was correctly explained). Including the Aim of the task is essential for building on and refining the prompt (Park & Choo, 2024; Velásquez-Henao et al., 2023). This baseline condition was then extended across multiple instructional prompt variants that incrementally added guidance aligned with the PARTS framework.

Across experimental conditions, additional guidance was introduced using four separate modifications. In our first prompt variation, we include Persona or assign the AI the role of a graduate assistant grading student papers for an upper-division course. Another prompt embedded the theorists' names from the textbook readings within the zero-shot prompt (Theme). The next separate parameter included a definition of the concept (Theme). The final prompt included specific instructions for considering concepts from the theories as adequately listed, defined, or correctly defined after the original zero-shot prompt information was presented (Theme).

Finally, all prompts ended with Structure, which explicitly instructed the AI to provide a 1 if the concept was listed, defined, or correctly defined, depending on the task, and 0 if not.

While examining the effects of few-shot or chain-of-thought prompting is of theoretical interest, such analyses were beyond the scope of this study. We focused instead on instruction-based prompts in order to avoid additional complexity introduced by example creation or multistep reasoning about theoretical concepts in student papers. Constructing appropriate examples is nontrivial and may introduce biases through exemplar selection. Moreover, expert-generated examples may not align with student writing well, and student examples may be imprecise and highly variable. Relying on zero-shot prompting enhances the generalizability of the findings by ensuring that the results do not depend on nuances of the selected examples.

Table 1: AI prompt examples for "attachment concept" from social control theory (Theory A, Short Paper 3).

	Coding Task Dimensions		
Prompt Types	Lists	Defines	Defines Correctly
Zero-shot prompt	Analyze the following paper to determine whether it **lists** the concept of "attachment," specifically within the context of social control theory (also known as social bond theory), and not in reference to other theories.	Analyze the following paper to determine whether it **defines** the concept of "attachment," specifically within the context of social control theory (also known as social bond theory), and not in reference to other theories.	Analyze the following paper to determine whether it **defines correctly** the concept of "attachment," specifically within the context of social control theory (also known as social bond theory), and not in reference to other theories.
Persona and zero-shot prompt	You are a graduate teaching assistant grading upper-division student papers in a Sociology course for Criminology majors. **[and Zero-shot prompt].**	You are a graduate teaching assistant grading upper-division student papers in a Sociology course for Criminology majors. **[and Zero-shot prompt].**	You are a graduate teaching assistant grading upper-division student papers in a Sociology course for Criminology majors. **[and Zero-shot prompt].**
Zero-shot prompt with author embedded	Analyze the following paper to determine whether it **lists** the concept of "attachment," specifically within the context of the readings by Hirschi on social control theory (also known as social bond theory), and not in reference to other theories.	Analyze the following paper to determine whether it **defines** the concept of "attachment," specifically within the context of the readings by Hirschi on social control theory (also known as social bond theory), and not in reference to other theories.	Analyze the following paper to determine whether it **defines correctly** the concept of "attachment," specifically within the context of the readings by Hirschi on social control theory (also known as social bond theory), and not in reference to other theories.
Zero-shot prompt and definition	**[Zero-shot prompt and].** A correct definition is: Attachment	**[Zero-shot prompt and].** A correct definition is: Attachment	**[Zero-shot prompt and].** A correct definition is: Attachment

	Coding Task Dimensions		
Prompt Types	Lists	Defines	Defines Correctly
	refers to how much an individual likes, respects, or cares about conventional others. When the attachment bond is strong, individuals refrain from engaging in deviant behavior because they do not want to disappoint those they care about.	refers to how much an individual likes, respects, or cares about conventional others. When the attachment bond is strong, individuals refrain from engaging in deviant behavior because they do not want to disappoint those they care about.	refers to how much an individual likes, respects, or cares about conventional others. When the attachment bond is strong, individuals refrain from engaging in deviant behavior because they do not want to disappoint those they care about.
Zero-shot prompt and instructions	**[Zero-shot prompt and].** It is acceptable if the student uses a variation of the word (e.g., "attached") or refers to the idea without using the exact term, as long as it is clear that the reference is within the context of social control theory.	**[Zero-shot prompt and].** The concept is considered defined if the student elaborates on the meaning of the concept after mentioning it. This may include: - A paraphrased definition in their own words - Use of the root word in the explanation (e.g., "committed" for commitment) - A quotation that conveys the definition - A relevant example that illustrates the concept The definition does not need to be complete or correct — this task only checks for an attempt to define the concept.	**[Zero-shot prompt and].** Do NOT give credit if the student simply restates the concept using a version of the root word (e.g., “commitment means being committed”) without meaningful elaboration. Do NOT give credit if the student provides only a direct quote from a source without further explanation in their own words. Treat terms like "crime," "lawbreaking," "delinquency," and "deviant behavior" as interchangeable, as long as the meaning aligns with the concept. Do NOT penalize inconsistent use of pronouns. Focus only on whether the definition is correct within the context of social control theory (also known as social bond theory).
Note: Each prompt ended with "Return 1 if the concept is [listed; defined; defined correctly] in this context, and 0 if it is not. Return no additional text." (The paper followed immediately to complete the prompt).			

3.3. Overview of LLMs Tested

We selected four closed-source large language models (LLMs) to provide a broad comparison of widely used systems to human coding. These models offer advantages over open-source models due to their higher reported reliability (Mansour et al., 2024; Seßler et al., 2025). Our first LLM, GPT-4.1 Mini (OpenAI), was chosen to test whether an inexpensive LLM could perform accurately. The other three more powerful LLMs were chosen for their reasonable costs and because sources such as OpenRouter.ai suggest they are widely adopted. As of 30 July 2025, monthly usage was 1.93 trillion tokens for Claude Sonnet 4 (rank 1), 727 billion tokens for Gemini 2.5 Pro (rank 5), 171 billion tokens for GPT-4.1 (rank 13), and 161 billion tokens for GPT-4.1 Mini (rank 14) (OpenRouter, 2025).

GPT-4.1 Mini (OpenAI). Released 14 April 2025, this "middle-sized" variant of GPT-4.1 offers a one-million-token context window and costs $0.40 per-million input tokens and $1.60 per-million output tokens. It achieves 65% on GPQA (Graduate-Level Google-Proof Q&A) (llm-stats.com, 2025) and 23.6% on SWE-bench-Verified, a human-validated 500-task subset of Software Engineering-bench that measures how well a model fixes real GitHub bugs (OpenAI, 2025). GPT-4.1 Mini is included in our comparison to test whether increased error, poorer correlation, and greater bias occur in a less expensive system, thus highlighting the possible importance of using the more advanced models.

GPT-4.1 (full). Introduced the same day, the full model costs $2.00 per-million input tokens and $8.00 per-million output tokens, with the same context window of one million tokens. Performance scores are 66.3% GPQA (llm-stats.com, 2025) and 54.6% SWE-bench-Verified (OpenAI, 2025).

Gemini 2.5 Pro (Google). Launched 17 June 2025, pricing is $1.25 per-million input tokens and $10.00 per-million output tokens for a 1.05-million-token context window (OpenRouter.ai, 2025). Performance is 83% GPQA (llm-stats.com, 2025) and 63.8% SWE-bench-Verified (Kavukcuoglu, 2025). This model was included due to its recent popularity and high-performance scores. It is a reasoning model, which means it charges output tokens for internal discussions and problem-solving before it reaches a final answer.

Claude Sonnet 4 (Anthropic). Announced 22 May 2025, it costs $3.00 per-million input tokens and $15.00 per-million output tokens and accepts up to 200 K tokens of context. Scores are 83.8% GPQA (llm-stats.com, 2025) and 72.7% SWE-bench-Verified (OpenRouter.ai, 2025). This was our second reasoning model and was chosen to broaden our test with another company's LLM product.

3.4. Pre-processing and Metric Computation

Our analysis pipeline began with processing the anonymized student paper PDFs into text files, followed by applying each prompt to each LLM, for each theory, for each content dimension, and for each concept in the theory, resulting in 60,060 prompts (see Table 2). After summing the values for each concept, we arrived at a total of 8,040 data rows. We accessed LLMs via the convenient OpenRouter.ai website, which provides API-level control over prompting for most LLMs with unified billing. This enabled us to create the entire analysis pipeline that standardized the application of the AIs. Temperature was set to 0 to ensure the response was 0 or 1 and to reduce variability in responding (Lo, 2023b), though small variability may be introduced by other sources (Atil et al., 2024; Schmalbach, 2025). We will discuss this further in the limitations section. LLM ratings were merged with the human ratings for each LLM, prompt type,

dimension, and theory. Each rating was the sum of the scores for the concepts in that paper, prompt, dimension, and theory.

Table 2: Structure of the factors of analysis.

Theory	Students	LLMs	Prompt Dimensions	Prompt Types	Concepts	Total Prompts
A Paper 3	33	4	3	5	6	11880
B Paper 3	33	4	3	5	13	25740
A Paper 6	34	4	3	5	7	14280
B Paper 6	34	4	3	5	4	8160

After collapsing the data for each theory across concepts, we computed 3 dependent measures for our modeling: correlations between AI and humans, absolute error, and normalized sign bias. First, Pearson's *r* between LLM and human ratings was calculated for each of the 240 design cells (LLM × dimension × prompt_type × theory = 4 × 3 × 5 × 4 = 240), as described previously, paper and theory are collapsed into 4 theory categories. To support regression, r values were Fisher-Z transformed: $z_r = \text{atanh}(r)$. Second, for every student across all conditions, we computed total errors between the human and the LLM and modeled them as a binomial distribution over the total possible errors. Third, also for every student, in each cell of the design, we computed normalized signed bias = `(LLM - human) / maximum possible`, which ranges from −1 to 1: positive values indicate the model labeled more concepts than the human (overestimation), negative values indicate fewer (underestimation), and 0 means exact agreement.

3.5. Analysis Strategy

Our primary analyses included three regressions (shown in Table 3), a fixed-effects regression on the correlations for each of our 240 conditions, a mixed-effects Binomial regression to predict the error rate with student as the random factor, and a mixed-effects regression to predict the bias, again with student as the random factor. We considered main effects for LLM, prompt type, theory, and coding dimension (listed, defined, or correctly defined).

Table 3: Summary ANOVA table.

Outcome	Data	Model	Distribution	Random effects
Fisher Z of *r* (`z_r`)	240 obs	`llm_id + prompt_type + dimension + theory + llm:prompttype + prompttype:dimension`	Normal	None
Total Errors (`n choose k`)	8040 obs	`llm_id + prompt_type + dimension + theory + llm:prompttype + prompttype:dimension`	Binomial	Intercept for the student
Normalized Signed bias (`bias`)	8040 obs	`llm_id + prompt_type + dimension + theory + llm:prompttype + prompttype:dimension`	Normal	Intercept for the student

4. Results

4.1. CORRELATION ANALYSIS

This analysis examines how closely LLM ratings match the pattern of human ratings by calculating Pearson correlations within each of our 240 experimental conditions (4 LLMs × 5 prompt types × 3 dimensions × 4 theories). Higher values indicate stronger correlations between the LLM and human coding. We use the Fisher Z transformation to fit the distributional requirements of standard regression. For reference, consider that a Fisher Z of .9 corresponds to a correlation of .716, and a Fisher Z of .7 corresponds to a correlation of .604.

A four-factor ANOVA tested how Fisher Z–transformed LLM-to-human correlations varied across large language model (LLM; 4 levels), prompt type (5), coding dimension (3), and theory (4). The Fisher Z transformation allowed us to put the correlation on an unbounded normal scale, enabling standard regression. The model accounted for substantial variance, $R^2 = .63$ (adjusted $R^2 = .57$). Table 4 shows the ANOVA result for the regression.

Table 4: ANOVA for correlation of human to LLM.

Effect	df_1, df_2	F	p	partial η^2
LLM	3, 207	15.91	< .001	0.19
Prompt type	4, 207	1.77	0.136	0.03
Dimension	2, 207	65.10	< .001	0.39
Theory	3, 207	44.23	< .001	0.39
LLM × Prompt type	12, 207	1.13	0.339	0.06
Prompt type × Dimension	8, 207	2.58	0.010	0.09

Agreement with human scores differed across models (partial $\eta^2 = .19$). Coding dimension of listing, defining, or correctly defining yielded distinct correlation levels (partial $\eta^2 = .39$). Correlations varied across the four criminological theories (partial $\eta^2 = .39$). Prompt type showed no overall main effect (partial $\eta^2 = .03$).

4.1.1. Correlation Interactions

Figure 1 shows the interaction between coding dimension and prompt type. Taken together, these complementary slices show that the dimension effect (listing > defining > correctly defining) is highly stable, whereas prompt effects are subtle. Most notably, the definition prompts may hurt performance when a definition is required, and neither the instruction prompts nor the definition prompts were helpful for the listing task.

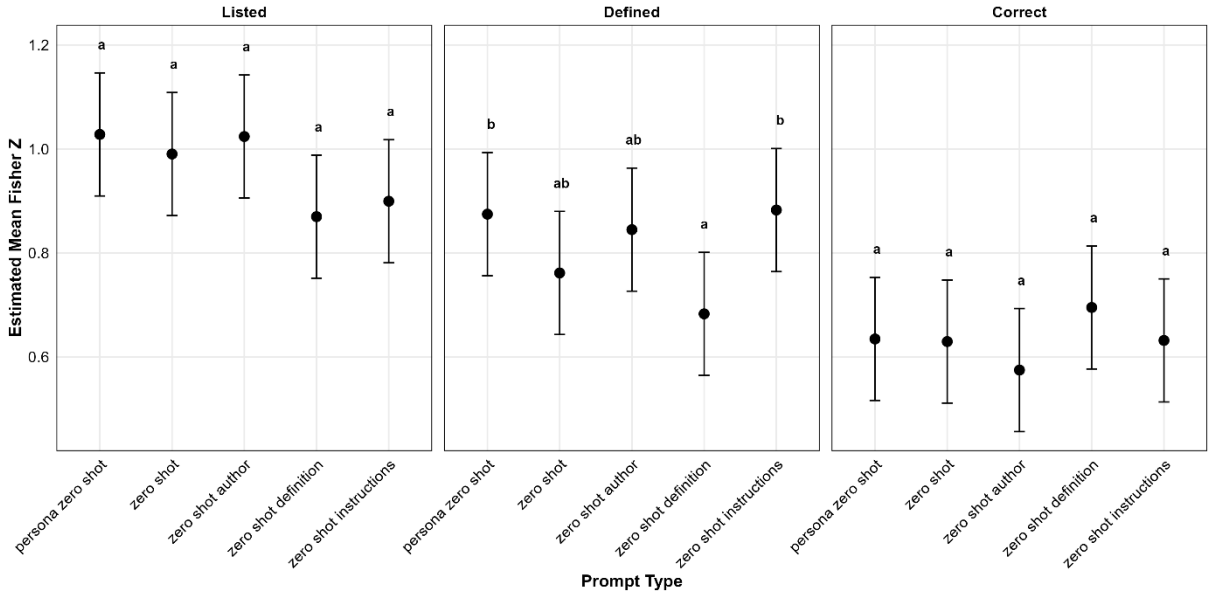

Figure 1: Coding dimension by prompt type correlation interaction. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Šidák-adjusted (used for multiple independent comparison families) pairwise comparisons within each prompt type; means sharing the same letter within a panel are not significantly different ($\alpha = 0.05$).

4.1.2. Correlation Main effects

Figure 2 offers a concise visual summary of estimated marginal means, with compact letter display indicating significant groupings.

LARGE LANGUAGE MODEL. GPT 4.1 Full and Claude 4 Sonnet did not differ, $\Delta z = 0.02$, $p = .88$. Gemini 2.5 Pro did differ from GPT 4.1 Full ($\Delta z = 0.10$, $p = .015$) but did not differ from Claude 4 Sonnet ($\Delta z = 0.08$, $p = .105$). GPT4.1 Mini was significantly worse than all other models (all Δzs > 0.10, ps <.007), standing alone in the lowest tier.

PROMPT TYPE. All five prompt styles shared the same compact letter display letter; no pairwise contrast was significant (largest $|\Delta z| = 0.10$, all ps > .077).

CODING DIMENSION. Listing surpassed defining ($\Delta z = 0.15$, $p < .001$) and correctly defining ($\Delta z = 0.33$, $p < .001$); defining also exceeded correctly defining ($\Delta z = 0.18$, $p < .001$).

THEORY. The largest contrast was Theory A Paper 3 versus Theory A Paper 6 ($\Delta z = 0.37$, $p < .001$). All remaining differences were significant (*ps < .006*).

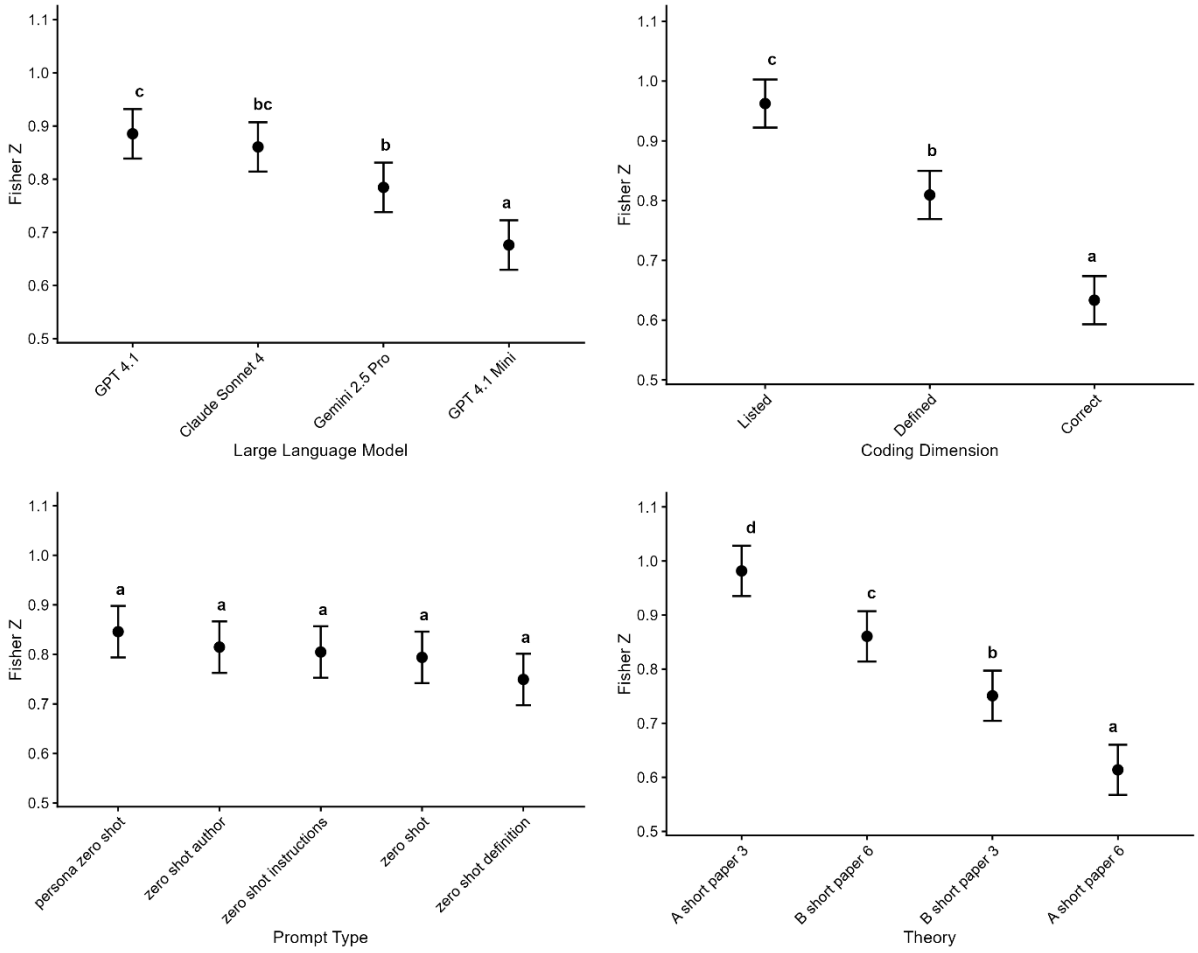

Figure 2: Estimated marginal means for each factor in the correlation regression model. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Tukey-adjusted pairwise comparisons; means sharing the same letter are not significantly different ($\alpha = 0.05$).

4.2. Error Rate Analysis

This analysis quantifies the rate of disagreement between LLMs and humans by modeling the probability that a given concept will be scored differently. It measures how far humans were from the LLM's total response for the condition. We use a binomial model because our outcome is a proportion (total errors divided by total concepts), which accounts for the fact that papers differ in the number of concepts. The logit values represent the log-odds of error; more negative values indicate lower error rates—for example, a logit of -2 corresponds to approximately 12% error probability per concept, meaning a paper with 6 concepts would have about 0.72 expected errors.

A binomial mixed effects regression predicting total *absolute error* included fixed effects for large language model (4), prompt type (5), coding dimension (3), criminological theory (4), and the two interactions (LLM × prompt type, prompt type × dimension). A random intercept for students captured repeated responses ($N = 8040$ observations from 39 students). We modeled normalized absolute error with a binomial GLM because it is a proportion—errors in human-AI matching divided by the total number of concepts in the paper. The logit parameter for each

value represents the probability that each concept has an error, so lower values are better. Table 5 shows the regression result.

Table 5: ANOVA for the absolute error binomial model

Effect	*df*	χ^2	*p*	Deviance Explained
LLM	3	979.10	< .001	0.057
Prompt type	4	25.08	< .001	0.001
Dimension	2	1750.46	< .001	0.101
Theory	3	1448.43	< .001	0.084
LLM × Prompt type	12	90.64	< .001	0.005
Prompt type × Dimension	8	55.90	< .001	0.003

Student-level variability was captured by the random effect (SD = .412). Fixed effect McFadden's R^2 was .157. Fixed + Random effect McFadden's R^2 was .192. McFadden's R^2 measures the total proportion of deviance explained by the entire model compared to a null model. In contrast, deviance-based effect sizes measure the unique proportion of deviance explained by each factor within a given model.

4.2.1. Error Rate Interactions

In Figure 3, below, we can see that Claude Sonnet 4 produced the lowest error, with excellent performance for zero-shot, zero-shot persona, and instructions. In contrast, Gemini 2.5 Pro did poorly with the instructions. While GPT 4.1 Full performed better than GPT 4.1 Mini, it was noteworthy that both were more consistent in their lack of interaction with the prompt types.

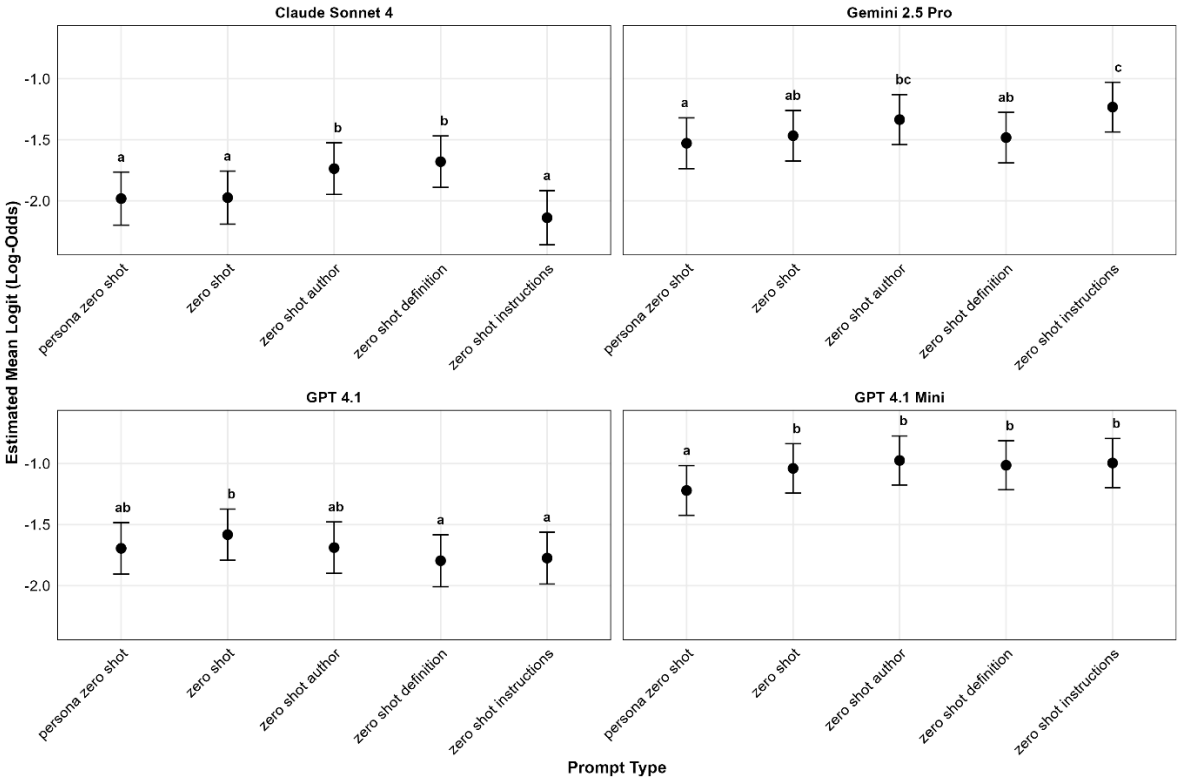

Figure 3: LLM by prompt type error rate interaction. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Šidák-adjusted (used for multiple independent comparison families) pairwise comparisons within each prompt type; means sharing the same letter within a panel are not significantly different (α = 0.05).

In Figure 4 below, we see some indication that providing definitions or instructions reduces errors for correct definitions, but may not be advantageous in other cases, with definitions, author, or instructions indicating that they impair the detection of listing relative to the other categories.

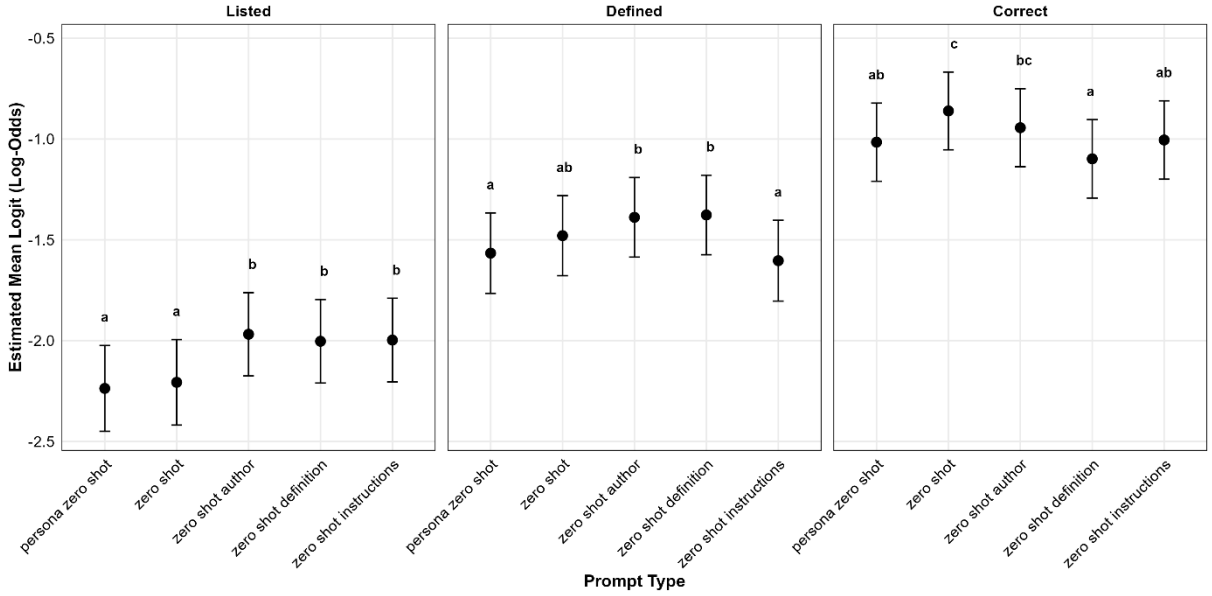

Figure 4: Prompt type by coding dimension error rate interaction. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Šidák-adjusted (used for multiple independent comparison families) pairwise comparisons within each prompt type; means sharing the same letter within a panel are not significantly different (α = 0.05).

4.2.2. Error Rate Main Effects

Figure 5 shows the estimated marginal means for the logit score. In a binomial model—the same framework used in ordinary logistic regression—we predict the probability p that each concept is scored as an error. The model first expresses this probability on the log-odds (logit) scale: the intercept gives the log odds of an error for the reference combination of factors, and each categorical coefficient is added to or subtracted from that baseline whenever its level is present. After all relevant coefficients are summed, the total is converted back to a probability with the logistic transformation shown earlier.

LARGE LANGUAGE MODEL. Estimated marginal means showed clear differences among the four LLMs. Claude Sonnet 4 (logit = -1.90, SE = 0.07, $\hat{p}$ = .13) had less error than GPT 4.1 Full (-1.71, 0.07, .15), Gemini 2.5 Pro (-1.41, 0.07, .20), and GPT 4.1 Mini (-1.05, 0.07, .26); GPT 4.1

Full outperformed Gemini 2.5 Pro and GPT 4.1 Mini; Gemini 2.5 Pro surpassed GPT 4.1 Mini. All Tukey-adjusted pairwise tests were significant ($|t| \geq 6.25$, $p<.001$).

PROMPT TYPE. Estimated marginal means showed modest variation across prompt conditions. Tukey adjusted significance emerged only for persona zero shot vs. zero shot author (estimate = -0.17, $t = 5.14$, $p< .001$), persona zero shot vs. zero shot definition (-0.11, $t = 3.36$, $p=.007$), and zero shot instructions vs. zero shot author (0.11, $t = 3.05$, $p=.019$). All remaining prompt error rate contrasts were nonsignificant ($|t| \leq 2.65$, $p\geq.062$).

CODING DIMENSION. Estimated marginal means varied markedly across the three dimensions. Listing showed the lowest error rate (logit = -2.08, $SE = 0.07$, $\hat{p} = .11$), followed by defining (-1.48, 0.07, .19) and correctly defining (-0.99, 0.07, .27). Tukey adjusted pairwise tests indicated all ps<.001.

THEORY. Estimated marginal means differed sharply across the four short paper conditions. Theory A Paper 3 yielded the lowest error rate (logit = -2.21, $SE = 0.07$, $\hat{p} =.10$), less than Theory B Paper 6 (-1.46, 0.07, .19), Theory A Paper 6 (-1.39, 0.07, .20), and Theory B Paper 3 (-1.01, 0.07, .27). Tukey adjusted pairwise tests showed all differences highly significant, except Theory B Paper 6 and Theory A Paper 6, which were not significantly different ($p = .156$).

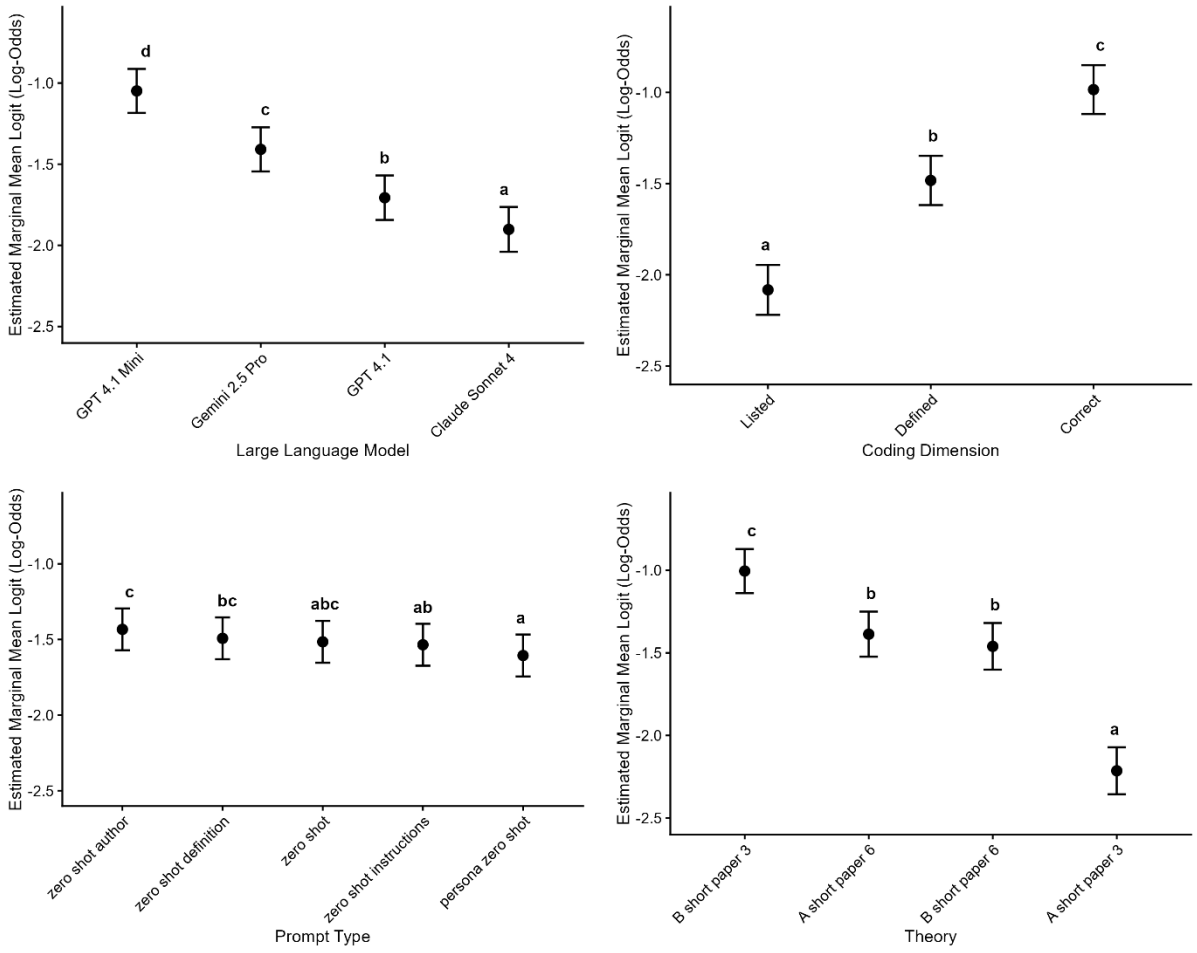

Figure 5: Estimated marginal means for each factor in the binomial mixed-effects model of error rate. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Tukey-adjusted pairwise comparisons; means sharing the same letter are not significantly different ($\alpha = 0.05$).

4.3. Bias Analysis

This analysis examines systematic tendencies in how LLMs score relative to humans, revealing whether models consistently over- or under-identify concepts across the coding dimension. Normalized bias scores range from -1 to 1, where negative values indicate the LLM identified fewer concepts than humans, positive values indicate more concepts identified, and values near zero suggest balanced scoring. Unlike absolute error, which captures any disagreement, bias reveals directional patterns that might indicate how LLMs interpret ambiguous cases differently from human coders.

A standard mixed-effects regression predicting *normalized bias* included fixed effects for large language model (4), prompt type (5), coding dimension (3), criminological theory (4), and the two interactions (LLM × prompt type, prompt type × dimension). A random intercept for students captured student-level variability in bias results ($N = 8040$ observations). Table 6 displays the result.

Table 6: ANOVA for mixed-effects bias model.

Effect	F	df_1	df_2	p	partial η^2
LLM	578.36.80	3	8000.6	< .001	0.178
Prompt type	41.90	4	8000.6	< .001	0.021
Dimension	941.96	2	8000.6	< .001	0.191
Theory	417.42	3	8025.2.3	< .001	0.135
LLM × Prompt type	19.46	12	8000.6	< .001	0.028
Prompt type × Dimension	5.00	8	8000.6	< .001	0.005

Student-level variability was captured by the random effect (*SD* .0904). Marginal R^2 (fixed effects) was 0.3594, and Conditional R^2 (fixed + random) was 0.463.

4.3.1. Bias Interactions

The interaction plot for the LLM by prompt type (Figure 6) mostly suggests effects for definition prompts, which produce low bias for Claude Sonnet 4, even below the optimal value of 0. Instruction prompts tended to increase bias in both reasoning models, Claude Sonnet 4 and Gemini.

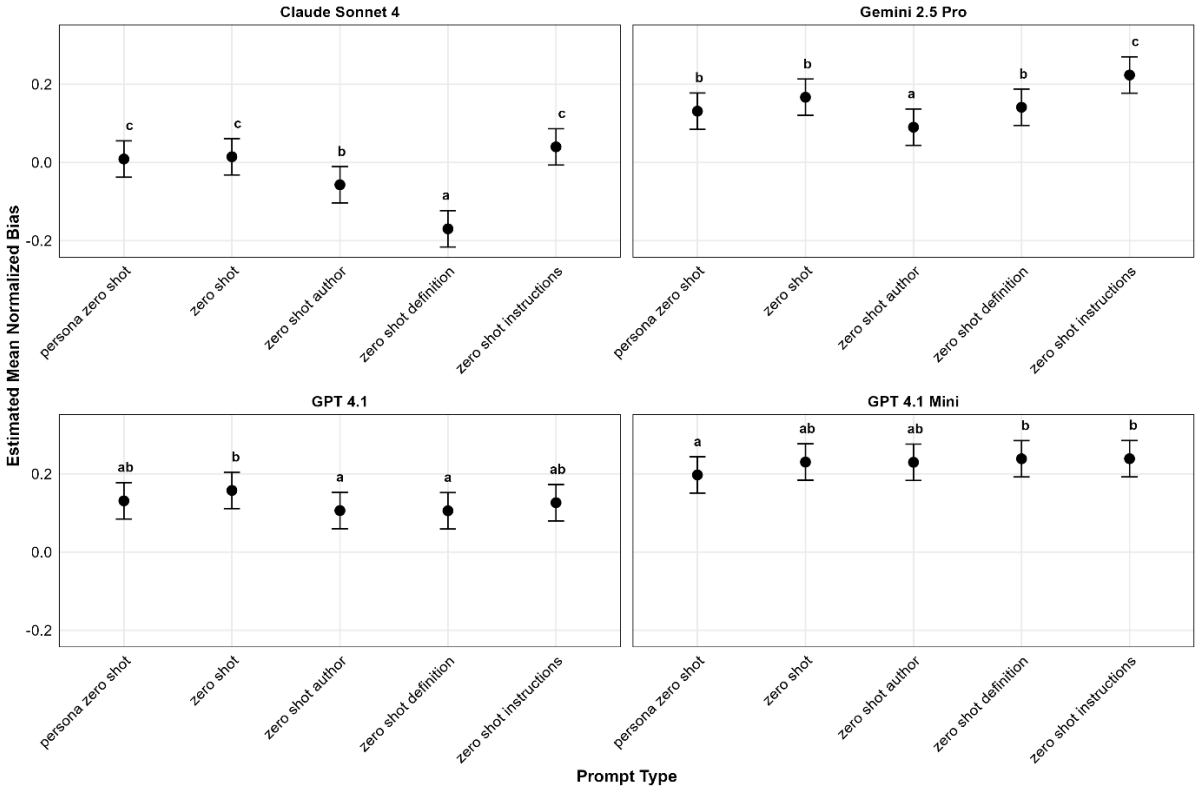

Figure 6: LLM by prompt type bias interaction. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Šidák-adjusted (used for multiple independent comparison families) pairwise comparisons within each prompt type; means sharing the same letter within a panel are not significantly different ($\alpha = 0.05$).

In the interaction between prompt and dimension, shown below in Figure 7, effects were rather weak, as might be expected with less than 1% of the deviance explained by this interaction.

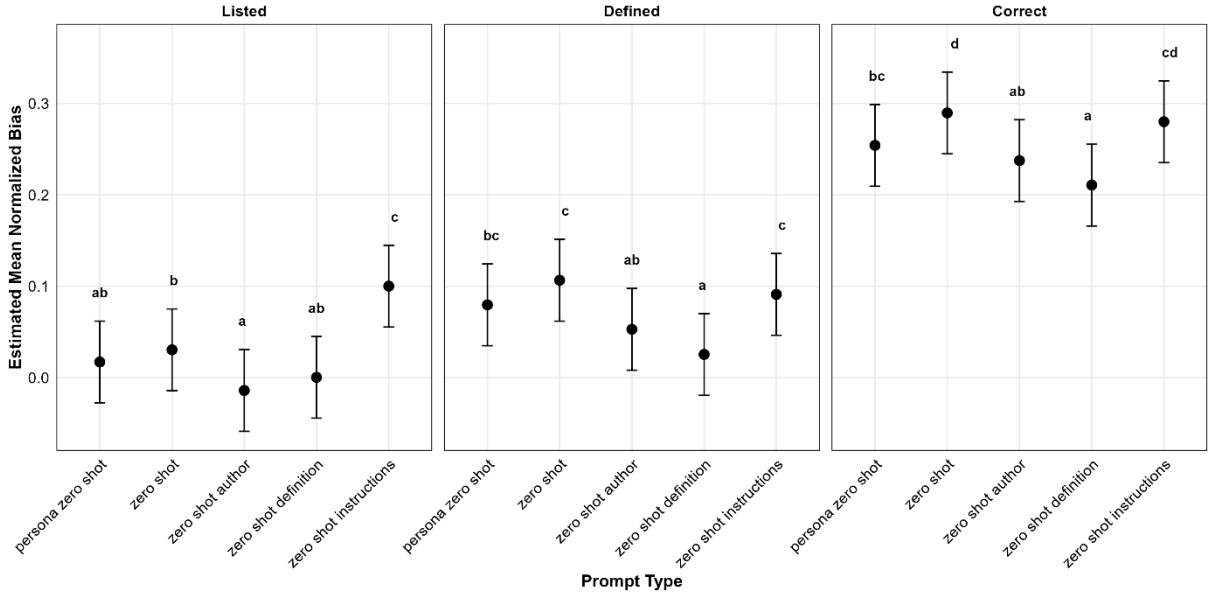

Figure 7: Coding dimension by prompt type bias interaction. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Šidák-adjusted (used for multiple independent comparison families) pairwise comparisons within each prompt type; means sharing the same letter within a panel are not significantly different ($\alpha = 0.05$).

4.3.2. Bias Main effects

Figure 8 shows the estimated marginal means for bias scores for the main effects.

LARGE LANGUAGE MODEL. Estimated marginal means revealed a clear gradient in bias across models. Claude Sonnet 4 showed the *lowest* bias (M = -0.03, SE = 0.015, 95 % CI [-0.06, 0.00]), significantly below GPT 4.1 Full (0.13, 0.015), Gemini 2.5 Pro (0.15, 0.015), and GPT 4.1 Mini (0.23, 0.015); all Tukey adjusted comparisons with Claude Sonnet 4 were highly reliable ($|t| \geq 24.61$, $p < .001$). GPT 4.1 Mini was the most biased, exceeding Gemini by 0.08 and GPT-4.1 Full by 0.10 ($|t| \geq 11.90$, $p < .001$). Gemini displayed slightly more bias than GPT-4.1 ($\Delta = 0.03$, $t = 3.89$, $p < .001$).

PROMPT TYPE. Estimated marginal means showed a clear gradient in bias across prompt types. Zero-shot instructions ($M = 0.157$, $SE = 0.015$, 95 % CI [0.126, 0.188]) and zero-shot (0.142, 0.015) produced the greatest bias, persona zero shot was intermediate (0.117, 0.015), and zero shot author (0.092, 0.015) and zero shot definition (0.079, 0.015) showed the least.

CODING DIMENSION. Estimated marginal means revealed a strong effect of dimension on bias: correctly defining showed the highest bias ($M = 0.254$, $SE = 0.015$, 95 % CI [0.225, 0.284]), defining was intermediate (0.071, 0.015), and listing the lowest (0.027, 0.015). Tukey-adjusted pairwise comparisons confirmed that each level differed significantly (correctly defining > defining , $\Delta = 0.183$, $t = 32.9$; correctly defining > listing, $\Delta = 0.228$, $t = 40.85$; defining > listing, $\Delta = 0.044$, $t = 7.95$; all ps < .001).

THEORY. Estimated marginal means differed markedly across theory conditions. Theory B Paper 3 showed the highest mean bias ($M = 0.254$, $SE = 0.015$, 95 % CI [0.224, 0.285]), followed by Theory A Paper 6 (0.106, 0.015), Theory A Paper 3 (0.071, 0.015), and Theory B Paper 6 (0.038, 0.015). Tukey-adjusted pairwise tests confirmed that every adjacent step in this ranking was reliable ($|t| \geq 4.91$, $p < .001$), with the largest gap between Theory B Paper 3 and Theory B Paper 6 ($\Delta = 0.216$, $t = 32.14$).

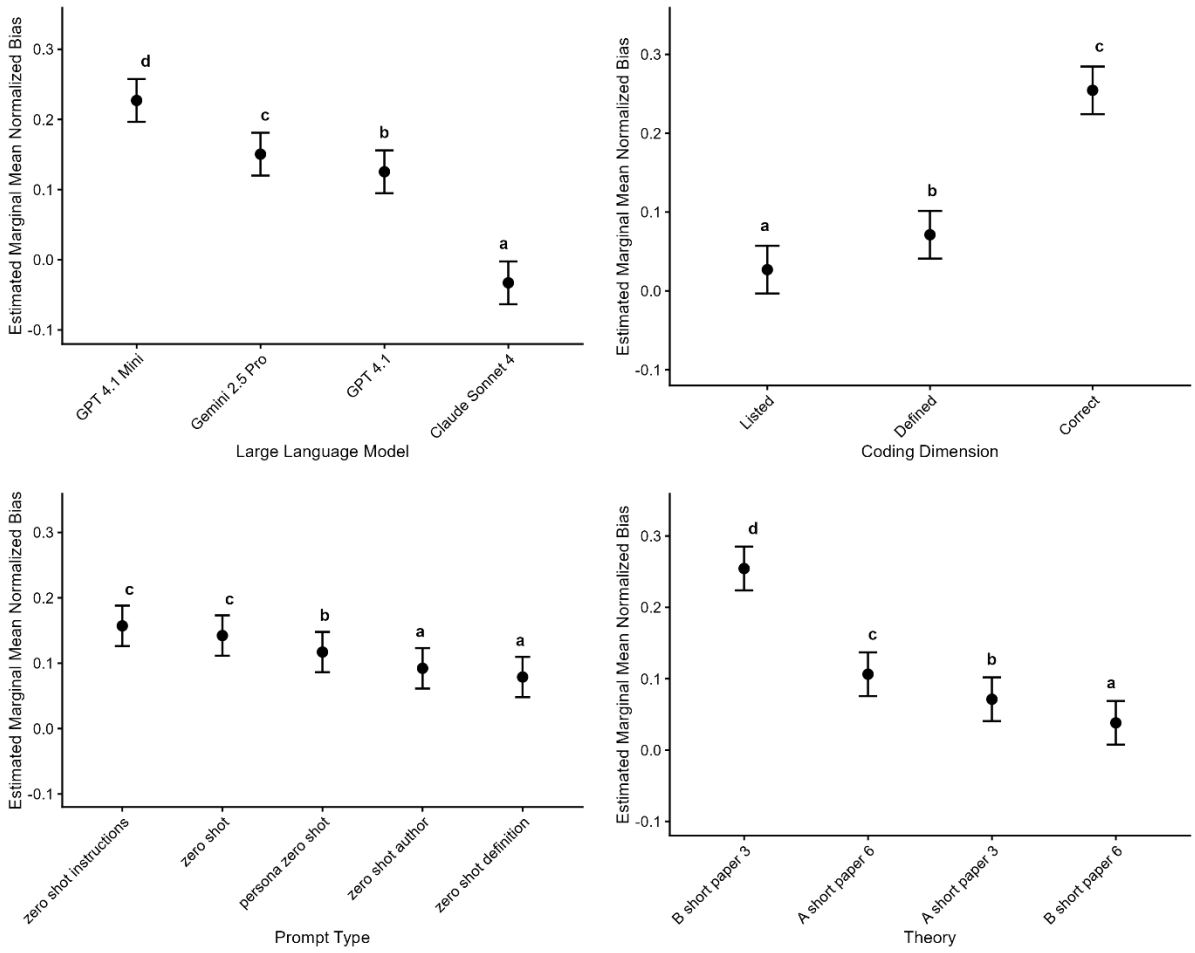

Figure 8: Estimated marginal means for each factor in the linear mixed-effects model of bias. Points represent estimated marginal means with 95% confidence intervals (error bars). Letters above each point indicate statistical groupings using Tukey-adjusted pairwise comparisons; means sharing the same letter are not significantly different ($\alpha = 0.05$).

5. Discussion

This study examined correlations, errors, and biases of LLMs in comparison to human coding across four LLMs, five prompt variants, three theory-based coding dimensions, and four criminological theories. The novel comparison methods, which decompose agreement into three factors, provide a more nuanced understanding of results than simpler single-value agreement statistics. For example, according to the correlation measures, there is no evidence of a difference between Sonnet 4 and GPT 4.1 Full in Figure 2; however, for bias, there was significantly less for Sonnet 4 in Figure 8. Such nuances allow comparisons across LLM conditions that may reveal practically important differences. For example, if the coding we used was part of a scoring tool, having low bias is critically important.

Of the two factors under the researchers' control, prompt type and LLM type, one of the most robust findings was the large difference between the LLMs in terms of correlation, error, and bias. These differences indicated that the researchers' choice of LLM is considerably more

important than the prompt type. Claude Sonnet 4 produced the strongest results overall, with Gemini 2.5 Pro producing less impressive results than GPT 4.1 Full, and GPT 4.1 Mini exhibiting significantly lower correlation, higher error, and greater positive bias. The weak performance of Gemini 2.5 Pro is particularly striking in light of the respective costs of the prompts for each LLM (see Table 7). Table 7 shows that the total prompt length was approximately 28.3 million tokens, and the reasoning models added considerable thinking tokens before the single output token. One direction for future research is to consider cost efficiency more explicitly; in contrast, we focused on dependent measures with no cost adjustment.

Table 7: Total costs by LLM for the prompts used.

LLM	Total cost	Tokens
Gemini 2.5 Pro	$178.13	43.5 million
Claude Sonnet 4	$169.18	35.9 million
GPT 4.1 Full	$56.77	28.3 million
GPT 4.1 Mini	$11.34	28.3 million

Another key finding from this study is that LLM performance varied greatly for the coding dimension. Listing concepts correlated most highly between the LLMs and humans (r = .75 for listing, r = .67 for defining, and r = .56 for correct response verification), suggesting that for more superficial coding dimensions, accuracy and reliability are high, and these degrade with the complexity of the judgement. While it was expected that including additional information and instructions in the prompts would increase the correlations between LLM and human coding for the additional coding dimensions, i.e., defining and defining concepts correctly, variations in prompts had little effect. In fact, in the case of zero-shot definitions and zero-shot instructions, the ability to accurately identify listed theories declines, and the ability to identify defined terms correctly improves only slightly. These findings align with previous research showing that, for well-defined tasks, zero-shot prompts can yield high agreement between humans and AI (Liu et al., 2025; McClure et al., 2024; Radford et al., 2019; Reynolds & McDonnell, 2021).

Despite the overall small effect sizes for the main effects of prompt type, prompt type was highly interactive. Each LLM reacts to the prompt type differently: Claude Sonnet 4's bias decreased with zero-shot authors and zero-shot definitions, whereas bias increased for Gemini 2.5 Pro with zero-shot instructions, indicating that AIs may perform better with different prompts, which aligns with prior research (Mansour et al., 2024). Inspecting the pattern of significant results across the interactions between prompt and LLM more closely in Figures 3 and 6 provides evidence that the Sonnet model drives much of the interaction effect with the LLM, showing greater variability by prompt for error and bias. The prompt type was also interactive with the dimension in some results, but the pattern is less clear. In summary, it appears that definition prompts show some evidence of being better when correctness needs to be judged, but this reverses when judgments are simpler.

In addition, we found that the LLMs aligned less closely with humans on certain theories, particularly those that included numerous, more nuanced concepts or semantically expansive categories, leading to more errors. For example, social control theory (A short paper 3), which LLMs performed better on than self-control theory (B short paper 3), includes 6 concisely defined concepts. In contrast, self-control theory includes 13 concepts whose ideas overlap and are less easily defined (See Appendix 1). These results are similar to Pacchioni et al. (2025), who found that ChatGPT's ability to identify the presence of concepts from Sykes and Matza's criminological theory, the techniques of neutralization, was unequal. Pacchioni et al. (2025)

argue that some concepts have "more precise and formal definitions," thus facilitating the model's identification of them (p. 10). While beyond the scope of this study, future studies should expand on prompt types to include few-shot approaches, providing examples of the concepts to test for improvements in accuracy, especially for defining and correctly defining concepts.

5.1. Practical Implications

In the education sector, LLMs have been used to provide personalized feedback and valuable learning experiences for both students and instructors (Hadi et al., 2023; Wang et al., 2023). Additionally, LLMs trained on local content can serve as personal tutors, tailoring content, exercises, and examples to the student's level of progress (Baidoo-Anu & Ansah, 2023; Kasneci et al., 2023). Findings from our study can guide students in prompting AIs to confirm whether they have included all relevant concepts in their papers. Given the level of bias that LLMs demonstrated in defining and correctly defining concepts, more modifications may be needed to the prompts used, and careful consideration needs to be given to the choice of LLM before students can rely on this to determine the quality of their work.

Our data have implications for the use of AI in collaboration with writing centers. The data from this study included one course with an embedded writing center tutor (treatment group). While students made fewer theory content errors in the treatment class, compared to the "control" class, and reported increases in writing ability over the course of the semester in the treatment course, these resources may not always be available for undergraduate courses (Kastner et al., 2018; Keith et al., 2020). Therefore, training non-embedded tutors from the writing center with no prior knowledge of LLMs on the use of AIs and prompting techniques may be beneficial. However, writing centers and tutors have vacillated in their acceptance and use of LLMs, and as their availability increases, writing center staff would benefit from establishing best practices for AIs (Essid & Cummins, 2025). Prompt engineering is a type of digital literacy; therefore, tutors must be able to support students in engineering prompts similarly to how they share their expertise in citation style formatting. Stakeholders are calling for tutors to be trained in prompt literacy, LLM limitations, and best practices (Essid & Cummins, 2025). Tutors are already trained to help students convey their points while meeting assignment requirements. However, they will need extensive training in recognizing LLM-generated text to support students in refining and editing LLM output to develop their points and establish their author voice (Coetzer & van Aardt, 2024). Concerning the data from this study, trained writing center tutors in prompt engineering can facilitate the use of LLMs to help students identify areas of improvement, especially as it relates to identifying specific concepts from papers

In addition to training writing center staff, students may benefit from AI literacy (Kim et al., 2025). Students will need content knowledge to discern the quality of AI-generated responses and to compare them to the content, which is a process of critical thinking (Cain, 2024). Generative AI can be used to facilitate student comprehension of course objectives such as those in a criminological theory course. In addition, it can be used to help students think critically about theoretical concepts and to synthesize and refine assumptions and ideas. When students prompt an LLM to produce information about a theory and compare its output to the original theory text, they may move beyond comprehension to fulfill course objectives, such as synthesizing and evaluating concepts (Kim et al., 2025). In sum, instructors, students, and writing center staff can use findings from this study to guide improved student understanding of criminology theories.

5.2. LIMITATIONS AND FUTURE WORK

The results from this study make it clear that human-AI collaborations are advantageous. In the same way that human qualitative coding and analyses need to be transparent, LLM decisions should be reviewed by humans, given the varying levels of bias identified in this study and the potential for hallucinations in AI models (Cain, 2024; Jiang et al., 2021). Future studies should address this study's limitations by including audits of human coding for accuracy and reliability. In this study, coding decisions were made by consensus after a norming process. However, it is unclear whether the two teams of researchers would reach the same conclusions on the same papers, given that the work was divided between them. In addition, among human coders of qualitative data, consensus is not always clear-cut and may be flexible, depending on who is leading the project and/or who has more experience (Jiang et al., 2021). Indeed, in this study, one team included the course instructor, who likely had more influence in reaching consensus when team members diverged in coding.

While we set the temperature parameter to 0 to minimize output variability, recent research demonstrates that LLM API outputs are not perfectly deterministic even under these conditions (Atil et al., 2024; Schmalbach, 2025). Individual prompts can exhibit stochastic variation due to floating-point precision, hardware differences across servers, and non-deterministic GPU operations. However, our large sample size (approximately 15,000 prompts per model) provides statistical robustness against trial-level noise; assuming that such variability is random rather than systematic, the law of large numbers ensures that our aggregate findings would be stable across runs. A more nuanced concern is the possibility of systematic changes in model implementations over time or across different API endpoints. Such implementation-level variation is difficult to detect or control in API-based research, as researchers are practically dependent on proprietary provider systems that may undergo updates without notification. We cannot rule out the possibility that our results reflect specific model versions or server configurations that were active during our data collection period, though we think this is unlikely. Future work validating LLM-based coding approaches should include replication studies across different time periods and fixed providers to assess the practical significance of these reproducibility concerns. This was beyond the scope of the current work.

Future studies should expand the prompt types, including additional examples and the sequential addition of information, to determine which steps improve agreement with human coding. Future research should explore a chain-of-thought approach in which each prompt is provided to the AI, which reanalyzes the same paper, allowing the model to reevaluate its initial response (Wei et al., 2022). Additionally, LLMs could be used to assess students' comprehension of source-based writing assignments, in which students must integrate multiple sources of information (McCarthy et al., 2022). Although the present study focused only on student comprehension from their use of a primary textbook, assessing student comprehension through writing assignments that integrate multiple sources would be a valuable area to explore.

This work contributes to the literature by evaluating multiple coding dimensions using a predefined codebook to determine which factors affect agreement between LLMs and humans. Overall, this study demonstrates that the choice of LLM matters greatly for human-AI concordance. In addition, this study found weak effects for the importance of prompt type, and in some cases, that additional information led to more bias. Finally, LLMs performed best on simpler tasks, such as identifying whether a concept was listed, and when the theories included few, concise concepts.

Declaration of Generative AI Software Tools in the Writing Process

During the preparation of this work, the author(s) used OpenAI o3 and GPT-4o in the preliminary interpretation of statistical outputs (Methods and Results sections) and in wording refinement throughout the manuscript to clarify model findings and improve clarity and concision. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

Atil, B., Chittams, A., Fu, L., Ture, F., Xu, L., and Baldwin, B. 2024. LLM Stability: A detailed analysis with some surprises. *arXiv:2408.04667v1 [cs.CL]*

Atkinson, J., and Palma, F. 2025. An LLM-based hybrid approach for enhanced automated essay scoring. *Scientific Reports 15*, 14551.

Auerbach, C., and Silverstein, L. B. 2003. *Qualitative data: An introduction to coding and analysis*. NYU Press.

Baidoo-Anu, D., and Ansah, L. O. 2023. Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI 7*, 1, 52-62.

Bano, M., Hoda, R., Zowghi, D., and Treude, C. 2024. Large language models for qualitative research in software engineering: Exploring opportunities and challenges. *Automated Software Engineering 31*, 1, 1-12.

Barany, A., Nasiar, N., Porter, C., Zambrano, A. F., Andres, A. L., Bright, D., Shah, M., Liu, X., Gao, S., Zhang, J., Mehta, S., Choi, J., Giordano, C., and Baker, R. S. 2024. ChatGPT for dducation research: exploring the potential of large language models for qualitative codebook development. In *Proceedings of the International Conference on Artificial Intelligence in Education,* A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt Eds., 134-149.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. 2020. Language models are few-shot learners. *Advances in neural information processing systems 33*, 1877-1901.

Cain, W. 2024. Prompting change: Exploring prompt engineering in large language model AI and its potential to transform education. *TechTrends 68*, 47-57.

Chew, R., Bollenbacher, J., Wenger, M., Speer, J., and Kim, A. 2023. LLM-Assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv 2306.14924*.

Chung, K. W. K., and O'Neil, H. F. 1997. Methodological approaches to online scoring of essays. National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA, 1-35.

Coetzer, Z., and van Aardt, P. 2024. Unsilencing the student voice: Detecting and addressing ChatGPT-generated texts presented as student-authored texts at a university writing centre. *International Journal of Critical Diversity Studies 6*, 2, 151-179.

Dikli, S. 2006. An overview of automated scoring of essays. *Journal of Technology, Learning and Assessment*.

GOOGLE FOR EDUCATORS. 2024. Generative AI for educators. Google LLC. https://grow.google/ai-for-educators/

ESSID, J., AND CUMMINS, C. 2025. A future for writing centers? Generative AI and what students are saying. *The Peer Review* 9, 2

HADI, M. U., AL-TASHI, Q., QURESHI, R., SHAH, A., MUNEER, A., IRFAN, M., ZAFAR, A., SHAIKH, M. B., AKHTAR, N., AL-GARADI, M. A., WU, J., AND MIRJALILI, S. 2023. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea preprints 1*, 3, 1-26.

HAYES, A. F., AND KRIPPENDORFF, K. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures 1*, 1, 77-89.

HERRENKOHL, L. R., AND CORNELIUS, L. 2013. Investigating elementary students' scientific and historical argumentation. *The Journal of the Learning Sciences 22*, 413-461.

HILA, A., AND HAUSER, E. 2025. Assessing the reliability of large language models for deductive qualitative coding: A comparative intervention study with ChatGPT. *Proceedings of the Association for Information Science & Technology 62*, 1, 275-285.

JIANG, J. A., WADE, K., FIESLER, C., AND BRUBAKER, J. R. 2021. Supporting serendipity: Opportunities and challenges for human-AI collaboration in qualitative analysis. In *Proceedings of the ACM on Human Computer Interaction*, 1-23.

JOHNSON, M., AND ZHANG, M. 2024. Examining the responsible use of zero-shot AI approaches to scoring essays. *Scientific Reports 14*, 1, 30064.

KASNECI, E., SESSLER, K., KÜCHEMANN, S., BANNERT, M., DEMENTIEVA, D., FISCHER, F., GASSER, U., GROH, G., GÜNNEMANN, S., HÜLLERMEIER, E., KRUSCHE, S., KUTYNIOK, G., MICHAELI, T., NERDEL, C., PFEFFER, J., POQUET, O., SAILER, M., SCHMIDT, A. SEIDEL, T., … KASNECI, G. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences 103*, 102274.

KASTNER, S., KEITH, S., KERR, L. J., STIVES, K. L., KNIGHT-RORIE, W., FORSYTHE, K., FEW, K., JEN, C., AND LOCKHART, J. M. 2018. RAD collaboration in the writing center: An impact study of course-embedded writing center support on student writing in a criminological theory course. *Praxis: A Writing Center Journal 15*, 3, 34-53.

KAVUKCUOGLU, K. 2025, March 25. Gemini 2.5: Our most intelligent AI model. *The Keyword*, https://blog.google/technology/google-deepmind/gemini-model-thinking-updates-march-2025/.

KE, Z., AND NG, V. 2019. Automated essay scoring: A survey of the state of the art. *IJCAI 19*, 6300–6308.

KEITH, S., STIVES, K. L., KERR, L. J., AND KASTNER, S. 2020. The role of academic background and the writing centre on students' academic achievement in a writing-intensive criminological theory course. *Educational Studies 46*, 2, 154-169.

KIM, J., PARK, S., JEONG, K., LEE, S., HAN, S. H., LEE, J., AND KANG, P. 2023. Which is better? Exploring prompting strategy for LLM-based metrics. *arXiv preprint arXiv 2311.03754*.

KIM, J., YU, S., LEE, S.-S., AND DETRICK, R. 2025. Students' prompt patterns and its effects in AI-assisted academic writing: Focusing on students' level of AI literacy. *Journal of Research on Technology in Education*, 1-18.

KIRSTIN, E., BUCKMANN, A., MHAIDLI, A., AND BECKER, S. 2024. Decoding complexity: Exploring human-AI concordance in qualitative coding. *arXiv 2403.06607*.

KOOLI, C., AND YUSUF, N. 2025. Transforming educational assessment: Insights into the use of ChatGPT and large language models in grading. *International Journal of Human–Computer Interaction 41*, 5, 3388-3399.

LIM, C. T., BONG, C. H., WONG, W. S., AND LEE, N. K. 2021. A comprehensive review of automated essay scoring (AES) research and development. *Pertanika Journal of Science and Technology 29*, 3, 1875 - 1899.

LINNEBERG, M. S., AND KORSGAARD, S. 2019. Coding qualitative data: A synthesis guiding the novice. *Qualitative Research Journal 19*, 3, 259-270.

LIU, X., ZAMBRANO, A. F., BAKER, R. S., BARANY, A., OCUMPAUGH, J., ZHANG, J., PANKIEWICZ, M., NASIAR, N., AND WEI, Z. 2025. Qualitative coding with GPT-4: Where it works better. *Journal of Learning Analytics 1*, 1, 1-10.

LLOYD-JONES, R. 1977. Primary trait scoring. In *Evaluating writing: Describing, measuring, judging,* C. R. Cooper, and L. Odell, Ed. National Council of Teachers of English, 33-66.

LO, L. S. 2023a. The art and science or prompt engineering: A new literacy in the information age. *Internet References Services Journal 27*, 4, 203-210.

LO, L. S. 2023b. The clear path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship 49*, 4, 102720.

LOPEZ-FIERRO, S., AND NGUYEN, H. 2024. Making human-AI contributions transparent in qualitative coding. In *Proceedings of the 17th International Conference on Computer-Supported Collaborative Learning-CSCL 2024,* International Society of the Learning Sciences, 3-10.

MANSOUR, W., ALBATARNO, S., ELTANBOULY, S., AND ELSAYED, T. 2024. Can large language models automatically score proficiency of written essays? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024),* N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue Eds., 2777-2786.

MCCARTHY, K. S., YAN, E. F., ALLEN, L. K., SONIA, A. N., MAGLIANO, J. P., AND MCNAMARA, D. S. 2022. On the basis of source: Impacts of individual differences on multiple-document integrated reading and writing tasks. *Learning and Instruction 79, 101599*.

MCCLURE, J., SMYSLOVA, D., HALL, A., AND JIANG, S. 2024. Deductive coding's role in AI vs. human performance. In *Proceedings of the 17th International Conference on Educational Data Mining (EDM-2024 poster session),* B. Paaßen, and C. D. Epp Eds., 809-813.

MIZUMOTO, A., AND EGUCHI, M. 2023. Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics 2*, 2, 100050.

MORGAN, D. L. 2023. Exploring the use of artificial intelligence for qualitative data analysis: The case of ChatGPT. *Journal of Qualitative Methods 22*, 1-10.

OHTA, R., PLAKANS, L. M., & GEBRIL, A. 2018. Integrated writing scores based on holistic and multi-trait scales: A generalizability analysis. *Assessing Writing 38*, 21-36.

OPENAI 2025, April 14. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/

OPENROUTER 2025, July 30. OpenRouter. https://openrouter.ai/

PACCHIONI, F., FLUTTI, E., CARUSO, P., FREGNA, L., ATTANASIO, F., PASSANI, C., COLOMO, C., AND TRAVAINI, G. 2025. Generative AI and criminology: A threat or a promise? Exploring the potential and pitfalls in the identification of Techniques of Neutralization *PLOS ONE 20*, 4, 1-15.

PACK, A., BARRETT, A., AND ESCALANTE, J. 2024. Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence 6*, 100234, 1-9.

PAGE, E. B. 1966. The imminence of… grading essays by computer. *The Phi Delta Kappa 47*, 5, 238-243.

PARK, J., AND CHOO, C. 2024. Generative AI prompt engineering for educators: Practical strategies. *Journal of Special Education Technology 40*, 3, 411-417.

RADFORD, A., WU, J., CHILD, R., LUAN, D., AMODEI, D., AND SUTSKEVER, I. 2019. Language models are unsupervised multitask learners. *OpenAI blog 1*, 8, 9.

RAMESH, D., AND SANAMPUDI, S. K. 2022. An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review 55*, 2495-2527.

REYNOLDS, L., AND MCDONELL, K. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI Conference on human factors in computing systems,* Y. Kitamura, A. Quigley, K. Isbister, and T. Igarashi Eds., 1-7.

SAUNDERS, P. I. 1999. Primary trait scoring: A direct assessment option for educators. In *Proceedings of the National Council of Teachers of English Annual Convention*.

SCHMALBACH, V. 2025. Does temperature 0 guarantee deterministic LLM outputs? https://www.vincentschmalbach.com/does-temperature-0-guarantee-deterministic-llm-outputs/

SEßLER, K., FÜRSTENBERG, M., BÜHLER, B., AND KASNECI, E. 2025. Can AI grade your essays? A comparative analysis of large language models and teacher ratings in multidimensional essay scoring. In *Proceedings of the LAK25: The 15th International Learning Analytics and Knowledge Conference*, 462-472.

SEVCIKOVA, B. L. 2018. Human versus automated essay scoring: A critical review. *Arab World English Journal 9*, 2, 157-174.

SHEN, X., CHEN, Z., BACKES, M., AND ZHANG, Y. 2023. In ChatGPT we trust? Measuring and characterizing the reliability of ChatGPT. *arXiv 2304.08979*.

SHERMIS, M. D., AND WILSON, J. 2021. Introduction to Automatic Essay Evaluation. In *The Routledge International Handbook of Automated Essay Evaluation* M. D. Shermis, and J. Wilson, Eds. Routledge, 3-22.

SUN, K., AND WANG, R. 2024. Automatic essay multi-dimensional scoring with fine-tuning and multiple regression. *arXiv 2406.01198*.

TRIPATHI, S., ALKHULAIFAT, D., LYO, S., SUKUMARAN, R., LI, B., ACHARYA, V., MCBETH, R., AND COOK, T. S. 2025. A hitchhiker's guide to good prompting practices for large language models in radiology. *Journal of the American College of Radiology 22*, 7, 841-847.

VELÁSQUEZ-HENAO, J. D., FRANCO-CARDONA, C. J., AND CADAVID-HIGUITA, L. 2023. Prompt engineering: A methodology for optimizing interactions with AI-language models in the field of engineering. *Dyna 90*, 230, 9-17.

WANG, W., HADDOW, B., BIRCH, A., AND PENG, W. 2023. Assessing the reliability of large language model knowledge. *arXiv 2410.0124*.

WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., ICHTER, B., XIA, F., CHI, E. H., LE, Q. V., AND ZHOU, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems 35*, 24824-24837.

WILLIAMS, M., AND MOSER, T. 2019. The art of coding and thematic exploration in qualitative research. *International Management Review 15*, 1, 45-55.

YOSHIDA, L. 2024. The impact of example selection in few-shot prompting on automated essay scoring using GPT models. In *Proceedings of the International Conference on Artificial Intelligence in Education* A. M. Olney, I. Chounta, Z. Liu, O. C. Santos, and I. I. Bittencourt Eds. Cham: Springer Nature Switzerland, 61-73.

YUN, J. 2023. Meta-analysis of inter-rater agreement and discrepancy between human and automated English essay scoring. *English Teaching 78*, 3, 105-124.

ZHANG, L., WU, H., HUANG, X., DUAN, T., AND DU, H. 2024. Automatic deductive coding in discourse analysis: an application of large language models in learning analytics. *arXiv preprint arXiv 2410.0124*.

APPENDIX 1. HUMAN CODEBOOK

Concepts for Social Control Theory/Social Bond Theory (Hirschi, 1969, p. 216-223)	**Listed**	**Defined**	**Defined Correctly**
1. *Attachment:* Like, respect, or care about others (p. 217-218)			
2. *Direct control*: A person watches behavior and administers punishment when deviant behavior is perceived (p. 218)			
3. *Indirect control:* A person does not participate in deviant behavior because they don't want to disappoint others (p. 218)			
4. *Commitment:* Stake in conformity – obey laws and rules because a person fears consequences either immediate (disappointing others) or anticipated (future plans) (p. 218-219)			
5. *Involvement*: Time invested in conventional activities (p. 219-220)			
6. *Beliefs:* Acceptance of rules and laws and belief that a person should follow them or the ability to rationalize that breaking the law is not unacceptable (p. 220-221)			
Total Number of Concepts	**/6**	**/6**	**/6**

Concepts for A General Theory of Crime/Self-Control Theory (Gottfredson and Hirschi, 1990, p. 225-236)	**Listed**	**Defined**	**Defined Correctly**
Development of self-control: 1. Early in life and stable(p. 227-228)			
Manifestations of low self-control: 2. Not only crime, but accidents, relationship problems, job problems, etc.			
Elements of self-control: 3. Immediate gratification (p. 228-229)			
4. Exciting, risky, thrilling (risk taking)			
5. Few or little long-term benefits (short-sighted)			
6. Little skill or planning			
7. Pain or discomfort for the victim (insensitive)			
8. Self-control: "impulsive, insensitive, physical (as opposed to mental), risk taking, short sighted, and nonverbal" (229)			
Child-rearing and self-control: 9. *Attachment of parent and child (bonding):* love, caring, feelings of affection towards the child (p. 233-236)			

Concepts for A General Theory of Crime/Self-Control Theory (Gottfredson and Hirschi, 1990, p. 225-236)	Listed	De-fined	Defined Correctly
10. *Parental supervision:* external and internal controls of behavior, monitoring			
11. *Recognition of deviant behavior:* perception of deviant behavior			
12. *Punishment of deviant acts:* Curtailing behavior by punishing effectively but not too harsh or too lenient			
13. *Parental criminality:* Parents lacking in self-control do not raise children/socialize children well.			
Total Number of Concepts	**/13**	**/13**	**/13**

Concepts for Deterrence Theory (Stafford & Warr, 1993, p. 394-399)	Listed	De-fined	Defined Correctly
1. *General Deterrence/Indirect Punishment*- See others get punished so you fear getting caught and punished			
2. *Specific/Direct Deterrence*- you refrain from offending because you were caught and punished			
3. *Indirect Punishment Avoidance*- you commit crime because you see your friends avoid getting caught			
4. *Direct Punishment Avoidance*- you commit crime because you frequently avoid getting caught			
5. *Certainty*- likelihood of getting caught (p. 395)			
6. *Severity*- harshness of sanctions (p. 395)			
7. *Celerity*- swiftness			
Total Concepts	**/7**	**/7**	**/7**

Concepts for Rational Choice Theory (Cornish & Clarke, 1986, p. 400 - 405)	Listed	De-fined	Defined Correctly
1. Crime is rational but it is l*imited or bounded rationality*- Rational choice limited by time, faulty info, estimation of benefits, and characteristics of offender			
2. Crime-specific approach is needed - Different crimes meet different needs and the way information is handled varies among offenses (p. 401)			
3. *Criminal involvement*- processes through which individuals choose to become initially involved in particular forms of crime, to continue, or desist (p.401)			
4. *Criminal events*- commission of a specific crime based on immediate circumstances (p. 401)			
Total Concepts	**/4**	**/4**	**/4**