AI Prompt Writing Rubric: A Validity and Reliability Study

Nurcan İnan
Marmara University, Educational Sciences ,Department Of Curriculum And Instruction, İstanbul, Türkiye
nrcinan@gmail.com
ORCİD: 0000-0003-3612-6011

Sibel Cengizhan
Marmara University, Educational Sciences ,Department Of Curriculum And Instruction, İstanbul, Türkiye
sibel@marmara.edu.tr
ORCİD:0000-0001-5862-2927

Seyfi Kenan
Marmara University, Educational Sciences ,Department Of Curriculum And Instruction, İstanbul, Türkiye
seyfi.kenan@marmara.edu.tr
ORCİD:0000-0002-3773-7693

ABSTRACT
This study introduces the development and validation of an analytical rubric designed to teach sixth-grade students how to write effective prompts. An initial draft rubric was developed based on a literature review on ChatGPT and prompt engineering, as well as the opinions of seven experts. The rubric was piloted with 32 sixth-grade students. We re-evaluated content validity, assessed construct validity through factor analysis, and measured internal consistency through Cronbach's alpha. During validation, four items were removed due to low common variance, and item 10 was excluded for redundancy. The final version demonstrated robust construct validity and internal consistency. Moreover, the Fleiss' kappa value of 0.29 showed fair to moderate interrater agreement. Implications for Practice or Policy: This section presents implications for educators, policymakers, students, and researchers: (1) Policymakers can create assessment tools aligned with AI-integrated curricula, using the developed rubric as a guide. (2) Educators can use the rubric for lesson planning, assessing prior knowledge, and measuring skill development. (3) Researchers may build foundational K-12 assessment studies based on this work. (4) Students can enhance their AI communication by writing clearer, more polite, and purposeful prompts, thereby improving their written expression and self-assessment skills.
Keywords: Artificial intelligence prompt, ChatGBT 3.5, K-12, Validity and reliability, Writing rubric

INTRODUCTION

The concept of intelligence

Intelligence has long been regarded as one of the core concepts that has captivated philosophers, psychologists, and scientists from Ancient Greece to the present day (Sternberg, 2005). Despite its enduring significance, there is still no consensus among researchers regarding the nature and scope of intelligence (Woodcock, R. W.,1990; Solso, 1995; Halonen & Santrock, 1996). While intelligence is often described as one's ability to adapt to their environment, solve problems, and learn from experience, how these processes function across different contexts remains a subject of ongoing debate (Sternberg, 2005). In this debate, various theorists proposed differing classifications of the core abilities that constitute intelligence. Spearman's Theory of General Intelligence, for example, claims that intelligence is a unified construct, with mental energy serving as the driving force behind all cognitive actions (Köksal, 2007). In contrast, Thorndike's (1920) Multifactor Theory conceptualizes intelligence as the capacity to effectively navigate novel situations and respond with appropriate solutions. Similarly, Sternberg's (1985) Triarchic Theory of Intelligence divides intelligence into three interrelated dimensions: analytical, creative, and practical. These frameworks offer more comprehensive perspectives on human intelligence and contribute significantly to our understanding of cognitive processes.

Research on intelligence has consistently highlighted problem-solving, decision-making, and environmental adaptability as core components of the construct (Sutarso, 1998; Budak, 2000; Rau, 2001). These components suggest that intelligence is not solely rooted in cognitive processes but is also closely linked to physical and emotional domains (Damasio, 1999). In this context, Piaget defined intelligence as one's ability to adapt to their environment and organize their thoughts and behaviors accordingly (Clark, 2019). By the 1980s, intelligence began to be understood as a measurable construct, often expressed through intelligence quotient (IQ) scores (Hoerr, 2000). Alfred Binet laid the foundation for the IQ concept by developing the first systematic intelligence test aimed at determining children's mental age (Myers, 1998). However, subsequent research demonstrated that IQ tests fall short of capturing the full scope of an individual's cognitive potential, emphasizing the need to

conceptualize intelligence as a multidimensional construct (Riggio et al., 2002). Similarly, Wechsler (1943) argued that intelligence assessments should encompass not only cognitive but also emotional and social components. Supporting this perspective, Gardner (1983; 1999) proposed the theory of multiple intelligences, contending that intelligence extends beyond mathematical and linguistic abilities and that individuals may exhibit exceptional strengths across various domains. These diverse efforts to understand human intelligence have laid the groundwork for developing models that mimic cognitive processes, ultimately enhancing machines' capacity to learn, solve problems, and adapt to their environments (McCarthy, 2007). Furthermore, these theoretical advancements have played a foundational role in shaping the field of artificial intelligence (AI). Notably, McCarthy (2007) redefined intelligence not as an exclusively human trait but as a phenomenon observable in some animals and even certain machines, thereby broadening the definition to include artificial entities.

Artificial intelligence

The concept of machine intelligence first entered academic discourse in 1950 through Alan Turing's seminal question, "Can machines think?", which laid the foundation for what became known as the Turing Test. This test was designed to evaluate whether a machine could exhibit human-like cognitive abilities. In the test, a human evaluator engaged in written communication with two participants– one human and one machine. If the evaluator could not reliably distinguish the responses from the machine from those from the human, the machine was considered to have the ability to think (Turing, 1950). Turing's work provided a theoretical framework that not only established the foundation of AI but also guided the development of modern AI systems. Over time, this framework has spurred practical applications of AI across a wide range of fields, including education, healthcare, finance, agriculture, industry, retail, security, transportation, logistics, law, and the creative industries. For example, an AI-based portable electronic assistive device was developed to support visually impaired individuals in navigating their environments independently (Shariff, 2020). Similarly, AI has been employed to detect safety and quality issues, while it has been used to automate personalized advertisement targeting (Davenport & Ronanki, 2021). In the field of education, particularly in language learning, AI applications have been developed for speech and pronunciation recognition, as well as for answering student queries, thus demonstrating the potential to support personalized learning processes (Hill et al., 2015). These diverse implementations have led to varying definitions of AI, tailored to the specific needs and perspectives of each discipline. Computer engineers, for instance, describe AI as a form of machine learning that utilizes artificial neural networks to apply algorithms to data for pattern recognition, decision-making, and predictive analysis (Say, 2018; Marr, 2020; Chivers, 2020; Yılmaz, 2022). In the field of medicine, scholars conceptualize AI within the framework of the "artificial human" (Aydın & Değirmenci, 2018), while philosophers explore its implications for consciousness, mind, and reasoning (Köse, 2022). Education researchers, on the other hand, describe AI in terms of its parallels with cognitive processes, focusing on a computer's ability to reason, solve problems, generalize, adapt, comprehend language, and make decisions in ways that resemble human cognition (Shidiq, 2023).

The combination of high computational power, vast data volumes, and advanced machine learning algorithms has recently driven significant progress in AI-based technologies (Russell, 2021). Thus, the capabilities of artificial intelligence now extend far beyond language processing and decision-making; they encompass a broader spectrum that includes visual perception, learning, autonomous action, and even creative thinking (Altıntop, 2023). In parallel with this expansion, numerous models have emerged, many of which are trained using complex architectures such as artificial neural networks developed through deep learning techniques (Goodfellow et al., 2016). Among the most notable are large language models (LLMs) built on transformer-based architectures (Sutton& Baro, 2018), which have become closely associated with AI in the public imagination, evident in the widespread use of voice and text assistants (e.g., Siri, Alexa, and ChatGPT). Particularly, generative AI tools, including language models and visual content generators, have enhanced users' productivity and enabled them to perform various tasks more efficiently (Kutlucan & Seferoğlu, 2024). One such tool is Chatbot Generative Pre-Trained Transformer (ChatGPT), released by OpenAI at the end of 2022 (ExcelinEd, 2023). Free access and its ability to produce highly relevant responses to user prompts (Günbatar & Ağgün, 2024) have underscored the growing necessity for clear and effective communication in written interaction with AI. This form of communication is defined by the concept of a "prompt," which refers to the initial input text designed to elicit a specific response from a language model (Bea et al., 2024). In other words, a prompt functions as a guiding input that steers the output toward a particular task (Brown et al., 2020). In this regard, the technique known as "prompt engineering" assumes a key role in optimizing the performance of AI systems.

Prompt engineering refers to the strategic formulation of natural language inputs and the refinement of interactions with LLMs. Prompts—consisting of task instructions, input data, and the expected output format

serve as a critical interface between human intent and machine response during the inference phase. Among the overall prompt strategies outlined by OpenAI, several stand out: articulating tasks and expected outputs with precision, supplying reference texts to minimize hallucinated content, decomposing complex assignments into manageable components, and allowing the model time to engage in reasoning processes (OpenAI, n.d.). In this light, the crafting of effective prompts is increasingly recognized as a pivotal factor influencing both the relevance and accuracy of LLM-generated outputs. Systematic prompt design plays a crucial role in optimizing model performance and tailoring outputs to specific needs (Mesko, 2023; Chakraborty et al., 2024; Schulhoff et al., 2024). Moreover, prompt strategies may be categorized based on their underlying objectives—ranging from guiding model behavior without the need for retraining to fostering reasoning capabilities and mitigating the risk of misinformation (Chakraborty et al., 2024). Anchored in these strategic insights, we developed a series of prompt steps and identified related behavioral subcomponents. Accordingly, the prompt dimensions and their corresponding behavioral indicators were structured as follows:

1.Clarity and Precision in Task Definition (a) Clearly articulates their requests when defining the task to a language-based AI. (b)Breaks down the task into small, manageable components as concise sentences.
2. Profile Creation (a) Creates a level-appropriate profile to guide the AI's responses.
3. Grammar Usage and Expression (a) Adheres to grammatical rules. (b) Provides accurate commands when defining the task for the language-based AI. (c) Enhances the effectiveness of the prompt by using adjectives, conjunctions, and adverbs correctly.
4. Politeness and Professional Tone (a) Employs polite expressions when interacting with the language-based AI. (b) Adopts a tone consistent with the defined task and user profile.
5. Output Refinement and Attention to Detail (a) Evaluates the generated output. (b) Writes follow-up prompts with refined details to complete or improve nuanced aspects of the output.
6. Critical Evaluation and Verification of Information (a) Critically evaluates the accuracy and reliability of the information provided by the language-based AI.

Measuring these dimensions and behavioral indicators, alongside evaluating students' academic performance, plays a critical role in analyzing the effectiveness of instructional processes (Ministry of National Education [MoNE], 2024). In this regard, the development of a valid and reliable tool for assessing prompt writing skills is considered essential. Guided by this need, the present study aims to develop a rubric capable of validly and reliably assessing the prompt writing competencies of sixth-grade students.

Rubrics, recognized as a specialized form of checklist, serve not only to articulate the ideal qualities of a completed task to students but also to provide teachers with a structured mechanism for assessing and scoring student performance in detail. Rubrics outline a set of specific criteria, each accompanied by clearly defined performance levels, which makes them effective tools not only for evaluating performance but also for tracking student progress and guiding learning (Brookhart, 2013). With access to these scoring criteria, students are better equipped to understand the expectations for improving their future work (Moskal & Leydens, 2000). In addition, rubrics enable individuals to assess their own performance (Mertler, 2001; Oakleaf, 2009), while offering both teachers and students constructive feedback on their strengths and weaknesses (Hall & Salmon, 2003). Teachers can also systematically monitor learners' progress toward achieving educational goals through rubric-based assessment (Arter, 2002).

Rubrics are often classified into two main categories: holistic and analytic. While holistic rubrics provide an overall assessment of a student's work (Taggart et al., 1998), analytic rubrics examine student performance in detail, aligned with specific learning objectives, and are regarded as one of the most effective and widely used assessment tools (Reeves, 2011). Due to their structured and targeted nature, analytic rubrics are often seen as more practical and comprehensive than holistic ones (Amanvermez İncirkuş & Beyreli, 2019). In this sense, we adopted an analytic rubric in this study to assess sixth-grade students' written communication skills using language-based AI tools, specifically ChatGPT. In developing the rubric, we followed the steps outlined by Goodrich Andrade (2001), including identifying the assessment criteria, defining the levels of performance, and consulting expert opinions to ensure the validity and clarity of the rubric. Then, we performed a comprehensive analysis to assess the validity and reliability of this rubric.

Literature review

In the relevant literature, a plethora of studies focus on forming a solid theoretical foundation for analysis by scrutinizing the role of ChatGPT and similar AI tools in education, the significance of prompt instruction, the function of rubrics in educational settings, and the necessity of evaluating prompts. In the realm of education, a growing body of research has emerged regarding the integration of language-based AI tools, with increasing attention being paid to the pedagogical applications of ChatGPT. Recent studies have explored the multifaceted role of ChatGPT in instructional contexts. For example, Qureshi (2023) noted that universities attempt to

leverage technology to enhance student learning and that instructors serve as facilitators in this process. Similarly, Mhlanga (2023) reported that ChatGPT is currently under examination for a variety of academic purposes, including instruction delivery, language learning, educational feedback, and assessment. Advances in AI models enable the development of innovative educational applications capable of fundamentally transforming learning experiences. Chiu et al. (2024) examined the educational potential of ChatGPT and other AI technologies, highlighting their capacity to personalize student learning and enhance engagement. Moreover, AI tools support teachers in tailoring lessons to individual needs and hold promises for fostering personalized learning experiences in university settings (Vatansever, 2024). Xu (2022) further emphasized that AI can serve as a valuable aid in problem-solving and rapid-response scenarios, particularly within STEM (Science, Technology, Engineering, Mathematics) education. In their 2024 bibliometric analysis on ChatGPT-based learning, Ching-Yi Chang and colleagues found that its use was most concentrated in educational technology, English language learning, and STEAM (Science, Technology, Engineering, Art & Mathematics)-related studies.

Previous research on prompt engineering indicated that the quality of prompts significantly influences the quality of output generated by LLMs for relevant tasks (Knoth et al., 2024). In their study exploring the relation between AI literacy and prompt engineering skills among university students, Knoth et al. (2024) presented experimental evidence showing that advanced prompt engineering fosters the production of higher-quality LLM outputs and enhances users' ability to harness the potential of such technologies more effectively. Prompt writing is positioned as a measurable skill that distinguishes individuals who can productively utilize LLMs from those who struggle to generate desired outcomes. Another study highlighted that while individuals without expertise in AI can engage in prompt engineering, their limited understanding of LLM capabilities and the tendency to mimic human-human instructions hinder systematic progress (Zamfirescu-Pereira et al., 2023). Similarly, relevant research showed that students often approach LLM-based AI systems as though interacting with a human, using socially desirable phrases such as 'Hello!' and 'Thank you.' This behavior is attributed to the human-like interfaces and conversational abilities of LLMs, which lead users to anthropomorphize these systems (Bewersdorff et al., 2025). Collectively, the previous research underscores the significance of equipping individuals, particularly K-12 students, with prompt engineering skills.

Although several studies examined the educational applications of ChatGPT at the K-12 level, research specifically focused on assessing students' prompt writing skills remains notably scarce. Among the limited body of work, Kobara et al. (2024) developed analytic rubrics to assess learning outcomes within AI education programs for K-12 students and analyzed the reliability of these rubrics through an AI-focused educational workshop. The findings revealed that, while the rubric demonstrated internal consistency, its inter-rater reliability remained poor. The study highlights the need for more tailored rubrics and the identification of appropriate methods for assessing learning outcomes.

Ultimately, the analytic rubric developed in the present study (Appendix A) is expected to contribute to the evaluation of prompt writing skills at the K-12 level. Moreover, we anticipate that while both researchers and practitioners can reap such a context-specific analytic assessment tool, our findings will guide future rubric development efforts in this emerging area.

THE STUDY

In this study, we employed a methodological design to ensure the validity and reliability of the rubric developed. During the 2023–2024 academic year, we administered the draft version of the rubric to 32 sixth-grade students attending a private school located on the European side of Istanbul. Two teachers and one researcher then evaluated the rubric. In the development and validation phases of the rubric that assesses the commands from students during their interactions with ChatGPT 3.5 and their prompt writing steps, we adhered to the following steps (Taggart et al., 1998; Brookhart, 2013; Schoepp et al., 2018; Amanvermez İncirkuş & Beyreli, 2019; Bhatnagar et al., 2021):

Reviewing the literature on prompt engineering and ChatGPT

Initially, we identified the key components of prompt writing based on the existing literature on prompt engineering and ChatGPT (Spasić & Janković, 2023; Vairamani & Nayyar, 2024; Öz, 2024; Lindley & Whitham, 2024; Li & Klabjan, 2024; Ein-Dor et al., 2024; OpenAI, n.d.; Prompting Guide, 2025). In AI-mediated writing tasks, the process of generating commands is grounded in a conceptual formula that consists of three main elements: the task, the instructions, and the role (John, 2023). The 'task' refers to clearly and explicitly stating what the prompt aims to achieve. In the pilot study, an example of such a task was having students instruct a language-based AI model to compose a poem. 'Instructions' denote the steps the AI model is expected to follow in executing the task, while the 'role' defines the persona/profile the model should adopt

while generating the text (Melanson & Maman, 2024). In this study, we noted that students asked ChatGPT to adopt the perspective of a sixth-grade student and to produce poems suited to that level. Accordingly, as prompt writing skills need to be designed in alignment with students' language and written communication abilities, we identified appropriate criteria and tasks for the assessment process.

Creating a draft rubric

In constructing the dimensions of the prompt rubric, we utilized both native language learning outcomes and Information and Communication Technologies (ICT) curriculum objectives as foundational criteria, as these are considered prerequisites for the language and written communication skills underpinning prompt writing steps. We tried to ensure that the tasks expected to be carried out using AI tools for text generation were aligned with the topics covered in class. Accordingly, key aspects of native language competencies (e.g., grammar, spelling conventions, and expression) were incorporated into the rubric as sub-dimensions. In addition, we included complete and error-free writing, as well as the use of polite language, in the draft rubric as sub-dimensions and observable behaviors to be assessed (Oz, 2023). The draft rubric consists of 7 dimensions and includes 11 behavioral indicators evaluated across four performance levels: Excellent, Proficient ,Basic, and Needs Improvement. We structured the drafting process of the rubric around the following steps to ensure its validity and reliability.

FINDINGS

Calculating content validity rates and content validity indices through expert opinions

To ensure the content validity of the draft rubric, we refined the items based on feedback from seven experts employed in various schools and universities. These experts included a computer science teacher, a native language teacher, an academic specializing in computer science, two curriculum and instruction specialists, and two experts in educational measurement and evaluation. In this process, we employed the Lawshe technique. According to Lawshe (1975), the minimum content validity ratio (CVR) value for an instrument evaluated by seven experts should be 0.99. Accordingly, we calculated the CVR value to be 1.00 for our draft rubric and made several revisions to some rubric items in line with expert opinions. For example, the indicator initially worded as "Defines the task to ChatGPT clearly and explicitly" was revised by one expert as "Clearly articulates their requests while defining the task to a language-based AI tool." Another expert recommended including adverbs of frequency in the indicator (e.g., always, usually). Moreover, we separated items involving multiple actions (e.g., "Always breaks down the task into small, manageable components in concise sentences and creates an appropriate profile to complete the task") into distinct behaviors based on expert opinions, resulting in an increased number of indicators. In addition, descriptions under the "Grammar Usage" and "Output Refinement and Attention to Detail" dimensions were rewritten to ensure clarity and precision.

Following these revisions, the total number of behavioral indicators increased to 14. We also restructured the performance levels based on expert recommendations and finalized them as "Excellent," "Proficient," "Partially Proficient," and "Not Proficient." The findings related to Lawshe's analysis are presented in Table 1.

Table 1. Findings of the CVR values

	Relevant	Useful/Need to be revised	Irrelevant	Total number of experts (N)	Responses to "Relevant" Items (NG)	CVR	Decision
Item 1	4	3	0	7	7	+1.00	Accept
Item 2	5	2	0	7	7	+1.00	Accept
Item 3	6	1	0	7	7	+1.00	Accept
Item 4	5	2	0	7	7	+1.00	Accept
Item 5	6	1	0	7	7	+1.00	Accept
Item 6	7	0	0	7	7	+1.00	Accept
Item 7	5	2	0	7	7	+1.00	Accept
Item 8	5	2	0	7	7	+1.00	Accept
Item 9	6	1	0	7	7	+1.00	Accept
Item 10	6	1	0	7	7	+1.00	Accept
Item 11	7	0	0	7	7	+1.00	Accept

Pilot study and administering the rubric

We carried out the pilot study for the draft rubric with the designated sample. A total of 32 sixth-grade students received a brief introduction to ChatGPT in their ICT class and were subsequently asked to complete a poetry-writing task using the platform. To ensure that participants were adequately prepared to complete the task, we

administered a Turkish language test to assess their basic language skills before the pilot study. This test helped identify deficiencies in students' language abilities. After providing the necessary support to address these gaps, we conducted the pilot study to further test the validity and reliability of the rubric. Prior to implementation, ethical approval forms were obtained from the Ministry of National Education, the participating school, the affiliated university, and the students' parents.

In the pilot study, students were allowed to complete the task based on their current written communication skills, without prior instruction on how to interact effectively with a language-based AI tool. The poetry-writing task allowed for natural language use; therefore, students could engage authentically with the AI. The evaluation involved three raters: the researcher, a Turkish language teacher, and an ICT teacher. Each student was assigned a single task, resulting in a total of 32 rubric-based assessments. Following the evaluation, we received no revision suggestions from the raters.

Establishing the construct validity of the rubric

We examined the clustering of behavioral indicators in the rubric using factor analysis. Specifically, we employed an exploratory factor analysis (EFA) with principal component analysis and orthogonal rotation (varimax) to determine the construct validity of this multi-dimensional scoring rubric and to uncover its factorial structure. We chose the principal component analysis as it is one of the most commonly used and practical methods in practice, while the varimax rotation was selected based on the assumption that there would be no correlation between factors.

There were no missing values in the dataset. We also calculated z-scores to detect outliers, and data within the range of $-3 < z < 3$ were retained for analysis. Among the 210 cases, we excluded nine outliers exceeding this range. In addition, the following four items with communalities of 0.5 or below were excluded from the analysis: "Providing examples in the prompt," "Specifying key terms (e.g., topic or main idea) that align with the task," "High-level alignment between the output and the prompt," and "Adhering to grammar rules." To assess the suitability of the dataset for factor analysis, we examined the Kaiser-Meyer-Olkin (KMO) value and Bartlett's test of sphericity. The KMO value was found to be .740, indicating a moderate level of sampling adequacy. Bartlett's test of sphericity produced a statistically significant result, $\chi^2(45) = 1085.443$, $p < .001$, confirming that the data exhibited multivariate normality (Büyüköztürk, 2022). We also examined the scree plot to explore the dimensionality of the items analyzed through factor analysis (Figure 1).

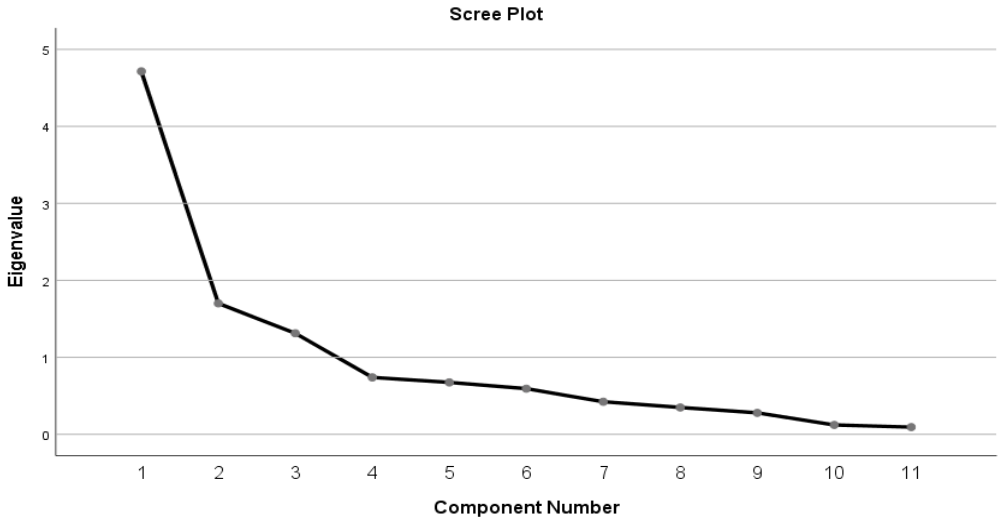

Figure 1. Scree plot

The rubric produced two factors. While the first factor accounted for 42.74% of the total variance, the second factor explained 17.08%. Table 2 presents the distribution of items across the factors along with their corresponding factor loadings.

Table 2. Factorial Structure of the Rubric, Item Factor Loadings, and Item Contributions to Common Variance

Items	Factor 1	Factor 2	Contribution to Common Variance
Item 1	0.84		0.70
Item 4	0.72		0.54
Item 9	0.88		0.78
Item 11	0.88		0.78
Item 12	0.76		0.64
Item 13	0.76		0.59
Item 6		0.89	0.79
Item 14		0.89	0.79
Item 8	0.58		0.35
Item 10		-0.98	0.84
Eigenvalue	4.26	2.54	
Variance Explained	42.74	17.08	
Total Variance Explained		59.82	

Table 2 presents the distribution of the items across two factors, along with their factor loadings, contributions to common variance, eigenvalues, and explained variance. Factor loadings indicate the strength of the relations between items and the corresponding factors. High factor loadings (.70 and above) suggest that an item is strongly represented within that factor (Deniz, 2021). Accordingly, Item 1 demonstrated a strong association with Factor 1, with a factor loading of .84, making it one of the key components of this dimension. Item 4 also loaded significantly on Factor 1 with a loading of .72. Items 9 and 11 both showed high loadings f .88 on Factor 1, further emphasizing their salience within this dimension. Items 12 and 13 lso loaded strongly on Factor 1, each with a loading of .76, reflecting their close alignment with this dimension. Items 6 and 14 both loaded on Factor 2 with a factor loading of .89, indicating their centrality in this dimension. Item 10 exhibited a negative factor loading of −.98 on Factor 2, suggesting that it represents an opposing component within this factor. Finally, Item 8 had a factor loading of .58 on Factor 1, indicating a relatively lower contribution to this factor.

In this study, only factors with eigenvalues > 1 were retained in the analysis. Accordingly, the factor analysis yielded the eigenvalues for these two factors to be 4.26 and 2.54, respectively. Factor 1 accounted for 42.74% of the total variance, and Factor 2 explained 17.08%, indicating that together, the two factors explained 59.82% of the total variance. This proportion is generally considered acceptable in factor analysis and suggests that the scale has an adequately supported construct validity (Field, 2018).

Establishing the reliability of the rubric

We calculated Cronbach's alpha coefficients to assess the internal consistency of the two dimensions of the rubric. A coefficient closer to 1.00 indicates a higher degree of internal consistency among the items (Kula & Mor, 2016). Accordingly, the coefficient was found to be .89 for the seven-item Factor 1 and .76 for the two-item Factor 2. As both values exceed the commonly accepted threshold of .70, the internal consistency of each factor can be considered adequate.

We performed an item analysis to examine the relation between Items 8 and 10. Exploring inter-item correlations basically aims to bring further evidence to the internal consistency of a measurement tool and explore the association between items (Kilmen, 2022). suggesting that participants scoring highly on one item also tended to score highly on the other. This finding also implies a genuine relation between these items rather than a random association. Given the size of the dataset (n = 201) for this analysis, this finding further contributed to the reliability of the rubric. Correlations between items are often used to assess conceptual overlap; while excessively high correlations (e.g., r > .80) may indicate redundancy, values between .40 and .70 are generally considered acceptable in scale development research (Deniz, 2021). Hence, Item 8 was retained in the rubric despite its relatively modest contribution to Factor 1. However, due to its strong inverse loading and limited theoretical alignment, Item 10—related to the use of polite language—was excluded from the final version of the rubric.

Calculating the inter-rater reliability of the rubric

Following the examination of the internal consistency of the pilot-tested draft rubric, we evaluated its inter-rater reliability using Fleiss' kappa coefficient. Fleiss' kappa is a statistical measure used to evaluate the level of agreement among multiple raters when classifying a common set of items or individuals into specific categories (Fleiss et al., 2003). However, each rating criterion inherently requires its own sub-criteria (Bıkmaz Bilgen & Doğan, 2017). Thus, the evaluation involves a progression from numerical judgments to performance-based

interpretation, ultimately leading to the measurement of consensus. In this context, assessing the kappa coefficient serves as a response to the question: "To what extent did the raters agree?" Accordingly, we calculated the inter-rater reliability coefficient using the statsmodels.stats.interrater.fleiss_kappa() function in Python. The resulting Fleiss' kappa value was 0.29, suggesting a fair to moderate level of agreement among raters (Sim & Wright, 2005).

DISCUSSION

In the literature, the study by Mott et al. (2003) highlighted the significant role of analytic rubrics in enhancing the reliability and validity of writing assessments, emphasizing their effectiveness in evaluating both written and visual narratives. Similarly, Dimopoulos et al. (2013) demonstrated that Learning Analytics-enhanced Rubrics (LAe-R) provide more comprehensive and data-driven assessments by integrating traditional rubric structures with learning analytics, thereby enabling more objective teacher evaluations and deeper insights into student development. Rayon et al. (2014) likewise underscored the importance of enriched rubrics in competency-based assessment by illustrating how student performance data can be systematically analyzed through rubric-based frameworks. In a related vein, Kocakülah (2021) developed a rubric to assess pre-service teachers' problem-solving skills and found that rubric-based assessment supported consistent scoring, accurately captured performance, and positively contributed to academic achievement. Collectively, these studies position rubrics as foundational tools in data-driven and structured assessment practices within education.

Validity and reliability studies on rubrics, alongside research evaluating student performance through rubric-based assessments, consistently emphasize the role of rubrics not only as reliable measurement instruments but also as pedagogical frameworks that guide instruction, clarify learning objectives, and minimize bias in assessment. Rubrics facilitate the integration of assessment with instruction and promote student engagement by supporting self- and peer-assessment practices. Moreover, they are widely regarded as authentic assessment tools that can be embedded in real-world problem-solving tasks (Mertler, 2001; Oakleaf, 2009; Petropoulou, 2011; Brookhart, 2013). In this respect, rubric development research assumes a critical role in the effective integration of artificial intelligence applications into educational contexts.

Within the context of AI-supported education, acquiring prompt-writing skills for effective communication with language-based AI systems and framing these skills within an analytic rubric may facilitate learning processes and address individual learner needs. Although a substantial body of research has explored AI-supported educational practices—such as enhancing motivation, engagement, classroom participation, academic achievement, and decision-making—relatively limited attention has been devoted to the assessment of AI-related competencies through performance-based tools (Arndt, 2023; Khan et al., 2023; Liu et al., 2020; Paek & Kim, 2021; Kumar & Raman, 2022; Winkler & Soellner, 2018; Yılmaz et al., 2021). While previous studies support the use of rubrics for evaluating student performance, they have predominantly focused on conceptual or outcome-oriented dimensions of learning rather than the procedural skills involved in interacting with AI systems. For instance, Kobara et al. (2024) developed analytic rubrics to assess learning outcomes in K–12 AI education; however, their findings revealed low inter-rater reliability, suggesting the need for more context-sensitive and skill-specific assessment criteria. In this regard, the rubric developed in the present study offers a valuable contribution by targeting prompt-writing skills as a measurable and assessable component of AI-supported learning at the K–12 level.

Previous research on assessing learning effectiveness in K–12 artificial intelligence education has largely adopted a knowledge-oriented perspective, focusing on students' conceptual understanding of AI, including dimensions such as understanding of AI, methods of AI use, AI–human relationships, and AI ethics (Kobara et al., 2024). These approaches primarily evaluate students' abilities to explain, relate, and generalize AI-related concepts. In contrast, the rubric developed in the present study addresses a complementary and relatively underexplored dimension of AI education by focusing on students' prompt-writing performance during interactions with language-based AI systems. Rather than assessing what students know about AI, this rubric operationalizes how effectively they can communicate with AI to obtain meaningful and accurate outputs. Accordingly, the emphasis shifts from conceptual AI literacy to an interactional, procedural, and performance-based conception of AI literacy. While previous rubrics evaluate students' explanations of AI usage and ethical considerations, the present rubric captures micro-level writing behaviors that directly shape AI-generated responses, such as clarity and precision in task definition, decomposition of complex requests, construction of appropriate user or role profiles, grammatical accuracy and expressive richness, explicit specification of tone, and systematic output refinement through follow-up prompts. Furthermore, the "Critical Evaluation and Verification of Information" dimension extends beyond general ethical awareness by assessing students' active evaluation of the accuracy and reliability of AI-generated content. By evaluating students' engagement in iterative prompt–response cycles rather than isolated responses, the rubric aligns closely with authentic

classroom practices involving generative AI. Taken together, the rubric does not aim to replace existing AI education assessment tools but rather to complement them by bridging the gap between knowing about AI and using AI productively.

LIMITATIONS

This study has several limitations that should be acknowledged. First, the rubric was validated with a relatively small sample of 32 sixth-grade students from a single private school in Istanbul, which limits the generalizability of the findings to different school types and educational contexts. Second, the study was conducted within a single grade level and instructional setting; therefore, the applicability of the rubric to other grade levels or subject areas remains to be examined. Third, although inter-rater agreement was assessed using two teachers and one researcher, including a larger and more diverse group of raters could provide stronger evidence for scoring consistency. Finally, the rubric was developed based on student interactions with ChatGPT 3.5 within a specific instructional design; thus, its use with other generative AI models or alternative instructional approaches requires further validation.

CONCLUSIONS

In this study, we developed and validated an analytic rubric designed to assess sixth-grade students' prompt-writing skills for effective and accurate communication with ChatGPT. The final rubric consists of five primary dimensions and nine items, and evidence for its validity and reliability was obtained through expert review, exploratory factor analysis, and internal consistency analyses. The factor analysis revealed a robust two-factor structure with satisfactory eigenvalues, explained variance, and factor loadings, while communality values indicated that the extracted factors adequately represented the items. Reliability analyses demonstrated acceptable internal consistency for both factors, and inter-rater agreement, assessed via Fleiss' kappa, indicated a moderate level of consistency among raters. Overall, the findings suggest that the analytic rubric constitutes a valid and reliable tool for evaluating K–12 students' prompt-writing performance and holds promise as a benchmark for AI-integrated educational assessment practices. In addition, the rubric supports the development of students' written communication skills by encouraging clarity, precision, appropriate tone, and purposeful interaction with language-based AI systems.

IMPLICATIONS AND FUTURE RESEARCH

Future research may benefit from employing additional methodological approaches, such as confirmatory factor analysis, reliability generalization, and consistency checking, to further strengthen the evidence for the rubric's validity and reliability (Kamış & Doğan, 2017). Comparative studies examining alternative assessment tools—such as holistic rubrics, checklists, or sequential rating scales—may also provide insights into the most effective strategies for evaluating prompt-writing skills. To enhance inter-rater consistency in classroom implementations, structured teacher training and norming processes are recommended prior to rubric use. Moreover, the analytic rubric developed in this study may be adapted for broader user groups and extended to assess prompt writing for diverse generative AI tasks, including visual or video content generation, thereby supporting future research and practice in this rapidly evolving field.

EFERENCES

Altıntop, M. (2023). Yapay zekâ/akıllı öğrenme teknolojileriyle akademik metin yazma: ChatGPT örneği. *Süleyman Demirel Üniversitesi Sosyal Bilimler Enstitüsü Dergisi, 46*, 186–211. https://dergipark.org.tr/tr/pub/sbe/issue/79677/1254533

Amanvermez İncirkuş, F., & Beyreli, L. (2019). A rubric for assessing critical thinking skills through narrative texts. *Journal of Mother Tongue Education, 7*(3), 597–629. https://doi.org/10.16916/aded.553569

Andrade, H. G. (2001). The effects of instructional rubrics on learning to write. *Current Issues in Education, 4*(1), 1–39.

Arndt, H. (2023). AI and education: An investigation into the use of ChatGPT for systems thinking. *arXiv.* https://doi.org/10.48550/arXiv.2307.14206

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom.* Corwin Press.

Aydın, İ. H., & Değirmenci, C. H. (2018). *Yapay zekâ.* Girdap Yayınları.

Bae, J., Kwon, S., & Myeong, S. (2024). Enhancing software code vulnerability detection using GPT-4o and Claude-3.5 Sonnet: A study on prompt engineering techniques. *Electronics, 13*(2657). https://doi.org/10.3390/electronics13132657

Bertalan, M. (2023). Prompt engineering as an important emerging skill for medical professionals: Tutorial. *Journal of Medical Internet Research, 25,* e50638. https://doi.org/10.2196/50638

Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., Bannert, M., Kasneci, E., Kasneci, G., Zhai, X., & Nerdel, C. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. *Learning and Individual Differences, 118,*

102601. https://doi.org/10.1016/j.lindif.2024.102601
Bhatnagar, R., Tanguay, C. L., Sullivan, C., & Many, J. E. (2021). Observation of field practice rubric: Establishing content validity and reliability. *Georgia Educational Researcher, 18*(2), Article 1. https://doi.org/10.20429/ger.2021.180201
Brookhart, S. (2013). *How to create and use rubrics for formative assessment and grading.* ASCD Member Book. ISBN 978-1-4166-1507-1
Brown, T., 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., Wu, J., Winter, C., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33,* 1877–1901. http://dx.doi.org/10.48550/arXiv.2005.14165
Bıkmaz Bilgen, Ö., & Doğan, N. (2017). Puanlayıcılar arası güvenirlik belirleme tekniklerinin karşılaştırılması [The comparison of interrater reliability estimating techniques]. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi, 8*(1), 63–78. https://doi.org/10.30831/akukeg.1027601
Büyüköztürk, Ş. (2022). *Sosyal bilimler için veri analizi el kitabı: İstatistik, araştırma deseni, SPSS uygulamaları ve yorum* [Data analysis handbook for social sciences: Statistics, research design, SPSS applications, and interpretation] (29. bs.). Pegem Akademi.
Chakraborty, O., Sahoo, A., Panda, R., & Das, A. (2024). XPL: A cross-model framework for semi-supervised prompt learning in vision-language models. *Transactions on Machine Learning Research, 6.* https://openreview.net/forum?id=oxAZv3QD6M
Chang, C.-Y., Chen, I.-H., & Tang, K.-Y. (2024). *Roles and Research Trends of ChatGPT-based Learning: A Bibliometric Analysis and Systematic Review*. *Educational Technology & Society, 27*(4), 471-486.
Chen, E., Wang, D., Xu, L., Cao, C., Fang, X., & Lin, J. (2024). A systematic review on prompt engineering in large language models for K-12 STEM education. *arXiv.* https://doi.org/10.48550/arXiv.2410.11123
Chiu, T. K. F. (2024). A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT. *Educational Technology Research and Development, 72,* 2401–2416. https://doi.org/10.1007/s11423-024-10366-w
Chivers, T. (2023). *Artificial intelligence does not hate you.* Mundi Publishing.
Clark, R. M. (2019). *Intelligence analysis: A target-centric approach* (6th ed.). CQ Press.
Damasio, A. (1999). *Descartes' error.* Varlık Publishing.
Davenport, T. H., & Ronanki, R. (2021). Artificial intelligence for the real world. In *Harvard Business Review Artificial Intelligence Book* (pp. 7–29). Optimist Publishing.
Deniz, K. Z. (Ed.). (2021). *İSTATİSTİKOLAY 2 - Çok değişkenli istatistik.* Nobel Akademi Yayıncılık.
Dimopoulos, I., Petropoulou, O., & Retalis, S. (2013). Assessing students' performance using the learning analytics enriched rubrics. In *LAK '13: Proceedings of the Third International Conference on Learning Analytics and Knowledge* (pp. 195–199). https://doi.org/10.1145/2460296.2460335
Ein-Dor, L., Toledo-Ronen, O., Spector, A., Gretz, S., Dankin, L., Halfon, A., Katz, Y., & Slonim, N. (2024). Conversational prompt engineering. *arXiv.* https://doi.org/10.48550/arXiv.2408.04560
ExcelinEd. (2023). *ChatGPT and education: FAQs for state policymakers.* ExcelinEd Policy Toolkit. https://excelined.org/wp-content/uploads/2023/05/c3_DigitalPolicy_OnePager_ChatGPT-FAQs_2023.pdf
Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). Sage Publications.
Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Wiley-Interscience.
Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* Basic Books.
Gardner, H. (1999). *Multiple intelligences: Interviews and articles.* BZD Publishing.
Gazzaniga, M. S. (1992). *Nature's mind: The biological roots of thinking, emotions, sexuality, language, and intelligence.* Basic Books Press.
Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning.* MIT Press.
Günbatar, M. S., & Ağgün, N. (2024, September 4–7). Educational applications of artificial intelligence (ChatGPT). In *Proceedings of the 16th National Science and Mathematics Education Congress (UFBMEK 2024)* (pp. 996–997). Van Yüzüncü Yıl University.
Hall, E. W., & Salmon, S. J. (2003). Chocolate chip cookies and rubrics: Helping students understand rubrics in inclusive settings. *Teaching Exceptional Children, 35*(4), 8–13. http://dx.doi.org/10.1177/004005990303500401
Halonen, J. S., & Santrock, J. W. (1996). *Psychology: Contexts of behavior* (2nd ed.). Brown & Benchmark Publishers.
Hill, J., Ford, W. R., & Farreras, I. G. (2015). Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in Human Behavior, 49,* 245–250. https://doi.org/10.1016/j.chb.2015.02.026
Hoerr, T. (2000). *Becoming a multiple intelligences school.* ASCD.

John, I. (2023). *The art of asking ChatGPT for high-quality answers: A complete guide to prompt engineering techniques.* Nzunda Technologies Limited. ISBN 9781234567890
Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed.). Prentice Hall.
Khan, B., & Angadi, G. R. (2023). Artificial intelligence integration into school education: A review of Indian and foreign perspectives. *Millennial Asia.* https://doi.org/10.1177/09763996231158229
Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6,* 100225. https://doi.org/10.1016/j.caeai.2024.100225
Kobara, T., Saito, D., Washizaki, H., & Fukazawa, Y. (2024). Work in progress: Rubrics to assess learning effectiveness of artificial intelligence education for K-12. In *2024 IEEE World Engineering Education Conference (EDUNINE)* (pp. 1–4). https://doi.org/10.1109/EDUNINE60625.2024.10500453
Kocakülah, A. (2022). Development and use of a rubric to assess undergraduates' problem solutions in physics. *Participatory Educational Research, 9*(3), 362–382. http://dx.doi.org/10.17275/per.22.71.9.3
Kula Kartal, S., & Mor Dirlik, E. (2016). The historical development of the concept of validity and the most preferred method for reliability: Cronbach's alpha coefficient. *Abant İzzet Baysal University Journal of Faculty of Education, 16*(4), 1865–1879.
Kumar, V. V. R., & Raman, R. (2022). Student perceptions on artificial intelligence (AI) in higher education. In *2022 IEEE Integrated STEM Education Conference (ISEC)* (pp. 1–7). IEEE. https://doi.org/10.1109/ISEC51903.2022.9745086
Kurzweil, R. (2005). *The singularity is near: When humans transcend biology.* Viking Press.
Kutlucan, E., & Seferoğlu, S. S. (2024). The use of artificial intelligence in education: A KEFE and PEST analysis of ChatGPT. *Turkish Journal of Educational Sciences, 22*(2), 1059–1083. https://doi.org/10.37217/tebd.1368821
Köksal, A. (2007). Üstün zekalı çocuklarda duygusal zekâyı geliştirmeye dönük program geliştirme çalışması [A program development study for developing emotional intelligence in gifted children] (Doktora tezi, İstanbul Üniversitesi, Eğitim Bilimleri Enstitüsü).
Köse, U. (2022). *Yapay zeka felsefesi.* Doğu Kitapevi.
Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*(4), 563–575. https://psycnet.apa.org/doi/10.1111/j.1744-6570.1975.tb01393.x
Li, H., & Klabjan, D. (2024). Reverse prompt engineering. *arXiv.* https://doi.org/10.48550/arXiv.2411.06729
Li, J., Jangamreddy, N. K., Hisamoto, R., Bhansali, R., Dyda, A., Zaphir, L., & Glencross, M. (2024). AI-assisted marking: Functionality and limitations of ChatGPT in written assessment evaluation. *Australasian Journal of Educational Technology, 40*(4), 56–72. https://doi.org/10.14742/ajet.9463
Lindley, J., & Whitham, R. (2024). From prompt engineering to prompt craft. *arXiv.* https://doi.org/10.48550/arXiv.2411.13422
Liu, J., Chang, H., Forrest, J. Y.-L., & Yang, B. (2020). Influence of artificial intelligence on technological innovation: Evidence from the panel data of China's manufacturing sectors. *Technological Forecasting and Social Change, 158,* 120142. https://doi.org/10.1016/j.techfore.2020.120142
Marr, B. (2021). *The artificial intelligence revolution.* Optimist Publishing Group.
McCarthy, J. (2007, November 12). What is artificial intelligence? *Stanford University.* http://www-formal.stanford.edu/jmc/
Melanson, J., & Maman, B. (2024). *ChatGPT complete guide: Learn Midjourney, DALL-E 2 & more.* Self-published.
Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical Assessment, Research & Evaluation, 7*(25). https://doi.org/10.7275/gcy8-0w24
Mesko, B. (2023). The impact of multimodal large language models on health care's future. *Journal of Medical Internet Research, 25,* e52865. https://doi.org/10.2196/52865
Mhlanga, D. (2023). Open AI in education, the responsible and ethical use of ChatGPT towards lifelong learning. *SSRN Electronic Journal.* https://dx.doi.org/10.2139/ssrn.4354422
Ministry of National Education [Millî Eğitim Bakanlığı]. (2024). Regulation on measurement and evaluation of the Ministry of National Education: Part one, preliminary provisions [Millî Eğitim Bakanlığı ölçme ve değerlendirme yönetmeliği: Birinci bölüm, genel hükümler]. Ankara, Turkey.
Mitchell, T. (1997). *Machine learning.* McGraw-Hill.
Mollick, E., & Mollick, L. (2023). *The AI classroom: The ultimate guide to artificial intelligence in education.* Jossey-Bass.
Moskal, B. M., & Leydens, J. A. (2000). Scoring rubric development: Validity and reliability. *Practical Assessment, Research & Evaluation, 7*(10).
Mott, M. S., Etsler, C., & Drumgold, D. (2003). Applying an analytic writing rubric to children's hypermedia "narratives." *Early Childhood Research & Practice, 5*(1). https://doi.org/10.1177/2158244012445584
Myers, D. G. (1998). *Psychology* (5th ed.). Worth Publishers.

Oakleaf, M. (2009). Using rubrics to assess information literacy: An examination of methodology and interrater reliability. *Journal of the American Society for Information Science and Technology, 60*(5), 969–983. https://doi.org/10.1002/asi.21030
OpenAI. (n.d.). Prompt engineering. https://platform.openai.com/
Oz, H. (2023). *ChatGPT & artificial intelligence: Prompt engineering from zero to mastery* [Online course]. Udemy. https://www.udemy.com/course/chatgpt-ai-verimliligi-artrmak-icin-prompt-muhendisligi/
Paek, S., & Kim, N. (2021). Analysis of worldwide research trends on the impact of artificial intelligence in education. *Sustainability, 13*(14), 7941. https://doi.org/10.3390/su13147941
Petropoulou, O., Vassilikopoulou, M., & Retalis, S. (2011). Enriched assessment rubrics: A new medium for enabling teachers to easily assess students' performance when participating in complex interactive learning scenarios. *Operational Research, 11*(2), 171–186. http://dx.doi.org/10.1007/s12351-009-0047-5
Pranab, S., Ayush, K. S., Sriparna, S., Vinija, J., Samrat, M., & Aman, C. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv.* https://doi.org/10.48550/arXiv.2402.07927
Prompting Guide. (2025). *Prompting guide.* OpenAI. https://cookbook.openai.com/
Qureshi, B. (2023). Exploring the use of ChatGPT as a tool for learning and assessment in undergraduate computer science curriculum: Opportunities and challenges. *arXiv.* https://doi.org/10.48550/arXiv.2304.11214
Rau, W. A. (2001). The relationship of emotional intelligence test scores to job performance evaluation scores in the management group of a health care organization (Unpublished doctoral dissertation). Medical University of South Carolina.
Rayon, A., Guenaga, M., & Núñez, A. (2014). Supporting competency assessment through a learning analytics approach using enriched rubrics. In *TEEM '14: Proceedings of the Second International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 291–298).
Reeves, A. R. (2011). *Where great teaching begins: Planning for student thinking and learning.* ASCD Press.
Riggio, R. E., Murphy, S. E., & Pirozzolo, F. J. (Eds.). (2002). *Multiple intelligences and leadership.* Lawrence Erlbaum Associates.
Russell, S., & Norvig, P. (2021). *Artificial intelligence: A modern approach* (4th ed.). Pearson.
Sander, S., Michael, I., Nishant, B., Konstantine, K., Amanda, L., Chenglei, S., Yinheng, L., Aayush, G., HyoJung, H., Sevien, S., et al. (2024). The prompt report: A systematic survey of prompting techniques. *arXiv.* https://doi.org/10.48550/arXiv.2406.06608
Say, C. (2018). *Artificial intelligence in 50 questions.* Bilim ve Gelecek Library.
Schoepp, K., Danaher, M., & Ater Kranov, A. (2018). An effective rubric norming process. *Practical Assessment, Research, and Evaluation, 23*(1), 11. https://doi.org/10.7275/z3gm-fp34
Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., ... & Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques. *arXiv.* https://doi.org/10.48550/arXiv.2406.06608
Shariff, A. F., B. S., A., Radhika, M., & S, S. (2020). Blind assistance using artificial intelligence. *International Journal of Engineering Applied Sciences and Technology, 5*(3), 396–401. http://dx.doi.org/10.33564/IJEAST.2020.v05i03.063
Shidiq, M. (2023). The use of artificial intelligence-based ChatGPT and its challenges for the world of education: From the viewpoint of the development of creative writing skills. In *Proceedings of the International Conference of Education, Society and Humanity, 1*(1).
Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257–268. [PMID: 15733050]
Solso, R. L. (1995). *Cognitive psychology* (4th ed.). Allyn and Bacon.
Spasić, A. J., & Janković, D. S. (2023). Using ChatGPT standard prompt engineering techniques in lesson preparation: Role, instructions, and seed-word prompts. In *Proceedings of the International Conference on Emerging eLearning Technologies and Applications (ICEST).* IEEE. https://doi.org/10.1109/icest58410.2023.10187269
Sternberg, R. J. (1985). *Human abilities: An information processing approach.* W. H. Freeman & Co Press.
Sternberg, R. J. (2004). North American approaches to intelligence. In R. J. Sternberg (Ed.), *International handbook of intelligence* (pp. 411–444). Cambridge University Press.
Sternberg, R. J. (2005). *Intelligence and intelligence testing.* Cambridge University Press.
Sutarso, P. (1998). Gender differences on the emotional intelligence inventory (Unpublished doctoral dissertation). University of Alabama.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
Taggart. (1999). *Rubrics: A handbook for construction and use.* R&L Education.
Thorndike, E. L. (1920). Intelligence and its uses. *Harper's Magazine, 140,* 227–235.
Tom, B. 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. *arXiv.* https://arxiv.org/abs/2005.14165

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Vairamani, A. D., & Nayyar, A. (2024). *Prompt engineering: Empowering communication.* CRC Press.

Vatansever, A. N. (2024). Üniversite öğrencilerinin yapay zekâ kavramına ilişkin metaforları ve görüşleri üzerine karşılaştırmalı nitel bir araştırma [A comparative qualitative study on university students' metaphors and opinions regarding the concept of artificial intelligence] (Yüksek Lisans tezi, Marmara Üniversitesi, Eğitim Bilimleri Enstitüsü). Council of Higher Education Thesis Center.

Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine, 37*(5), 360–363. PMID: 15883903

Wechsler, D. (1943). Non-intellective factors in general intelligence. *The Journal of Abnormal and Social Psychology, 38*(1), 101–103. https://doi.org/10.1037/h0060613

Winkler, R., & Soellner, M. (2018). Unleashing the potential of chatbots in education: A state-of-the-art analysis. *Academy of Management Proceedings, 2018*(1), 15903. https://doi.org/10.5465/AMBPP.2018.15903

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8*(3). https://doi.org/10.1177/073428299000800303

Xu, W., & Ouyang, F. (2022). The application of AI technologies in STEM education: A systematic review from 2011 to 2021. *International Journal of STEM Education, 9*(59). https://doi.org/10.1186/s40594-022-00377-5

Yılmaz, Y., Uzelli Yılmaz, D., Yıldırım, D., Akın Korhan, E., & Özer Kaya, D. (2021). Opinions of faculty of health sciences students on artificial intelligence and its use in healthcare [Yapay zekâ ve sağlıkta yapay zekânın kullanımına yönelik Sağlık Bilimleri Fakültesi öğrencilerinin görüşleri]. *Süleyman Demirel University Journal of Health Sciences, 12*(3), 297–308. https://doi.org/10.22312/sdusbed.950372

Yılmaz, A. (2021). *Yapay zeka* [E-book]. KODLAB Yayınları.

Zamfirescu-Pereira, J. D., Wei, H., Xiao, A., Gu, K., Jung, G., Lee, M. G., Hartmann, B., & Yang, Q. (2023). Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. In *Proceedings of the 2023 ACM Designing Interactive Systems Conference (DIS '23)* (pp. 2206–2220). Association for Computing Machinery.

Appendix A

AI prompt writing rubric

Dimensions	Excellent (4)	Proficient (3)	Partially Proficient (2)	Not proficient (1)	SCORE
Clarity and Precision in Task Definition	Always articulates their requests clearly when defining the task to a language-based AI.	Usually articulates their requests clearly when defining the task to a language-based AI.	Sometimes articulates their requests clearly when defining the task to a language-based AI.	Never articulates their requests clearly when defining the task to a language-based AI.	
	Always breaks down their requests into small, manageable components as concise sentences when defining the task to a language-based AI.	Usually breaks down their requests into small, manageable components as concise sentences when defining the task to a language-based AI.	Sometimes breaks down their requests into small, manageable components as concise sentences when defining the task to a language-based AI.	Enter their requests in a single, long, and hard-to-understand sentence when defining the task.	
Profile Creation	Always create a level-appropriate profile to complete the task and add details that elaborate on the task.	Usually create a level-appropriate profile to complete the task and add some details that elaborate on the task.	Does not create a profile when defining the task, but the task itself is clear and comprehensible.	Does not create a profile when defining the task and fails to express the task clearly and comprehensibly.	
Grammar Usage and Expression	Always provides accurate commands when defining the task for the language-based AI.	Usually provides accurate commands when defining the task for the language-based AI.	Sometimes provides accurate commands when defining the task for the language-based AI.	Provides commands using only a single word or phrase without any action-oriented element when defining the task for the language-based AI.	
	Enhances the effectiveness of the prompt by always using adjectives, conjunctions, and adverbs correctly.	Enhances the effectiveness of the prompt by usually using adjectives, conjunctions, and adverbs correctly.	Sometimes uses adjectives, conjunctions, and adverbs correctly. The request clarity is poor.	Never uses adjectives, conjunctions, and adverbs when defining the task.	
	Always adopts a tone consistent with the defined task. Ensures a more accurate output by clearly stating this tone in the prompt (e.g., casual-informal tone or formal-academic tone).	Usually adopts a tone consistent with the defined task.	Sometimes adopts a tone consistent with the defined task.	Never adopts a tone when defining the task.	
Output Refinement and Attention to Details	Always evaluates the generated output.	Usually evaluates the generated output.	Sometimes evaluates the generated output.	Never evaluates the generated output.	
	Always writes follow-up prompts with refined details to complete or improve nuanced aspects of the output.	Usually writes follow-up prompts with refined details to complete or improve nuanced aspects of the output.	Sometimes writes follow-up prompts with refined details to complete or improve nuanced aspects of the output.	Never writes follow-up prompts with refined details. Uses the output as it is without any refinement.	

Critical Evaluation and Verification of Information	Always critically evaluates the accuracy and reliability of the information provided by the language-based AI.	Usually critically evaluates the accuracy and reliability of the information provided by the language-based AI.	Sometimes critically evaluates the accuracy and reliability of the information provided by the language-based AI.	Never critically evaluates the accuracy and reliability of the information provided by the language-based AI.	