

Please quote as: Reinhard, P., Sajzev, V., Li, M. M. & Leimeister, J. M. (2026). Evaluating and Improving Prompt Quality in LLM-Based Assistants: A Synthesis of Criteria and Indicators. Hawaii International Conference on System Sciences (HICSS), Maui, Hawaii, USA.

Evaluating and Improving Prompt Quality in LLM-Based Assistants: A Synthesis of Criteria and Indicators

Philipp Reinhard
University of Kassel
philipp.reinhard@uni-kassel.de

Vladimir Sajzev
University of Kassel
vladimirsajzev@gmail.com

Mahei Manhai Li
University of Kassel; University of
St.Gallen
mahei.li@unisg.ch

Jan Marco Leimeister
University of Kassel; University of
St.Gallen
janmarco.leimeister@unisg.ch

Abstract

Generative AI (GenAI) assistants, particularly large language models (LLMs), are gaining increasing relevance across domains. The quality of outputs generated by these systems is highly contingent on the input prompts, giving rise to new professional roles such as prompt engineers. In this study, we systematically examine evaluation criteria and optimization methods that can improve prompt quality. Drawing on a systematic literature review, we identify key criteria, including clarity, accuracy, and precision, and initial measurement techniques. In addition, we synthesize common optimization methods such as iterative refinement and shot-based prompting. Our work contributes to the growing efforts to standardize the evaluation and improvement of prompts in interactions with LLM-based assistants, thereby fostering a more rigorous and coherent understanding of the prompt quality construct.

Keywords: Large Language Models, Prompt Engineering, Prompt Quality, Evaluation Criteria

1. Introduction

Since OpenAI's release of ChatGPT in late 2022, generative AI, particularly large language model (LLM)-based assistants, has seen rapid adoption. LLMs are trained on vast data to generate coherent text and other content (e.g., images, videos) (Feuerriegel et al., 2024). Prominent systems such as DALL·E 2 and GPT-4 are reshaping communication and collaboration (Nah et al., 2023). GenAI extends beyond language to encompass multimodal content, rendering it increasingly valuable across industries (Feuerriegel et al., 2024; Schneider et al., 2024). It is reshaping communication and workflows in both the private and public sectors, with applications in

knowledge-intensive fields like IT support (Reinhard et al., 2024) and everyday tasks such as note-taking and time management (Knoth et al., 2024).

Using GenAI effectively requires careful design of input instructions, commonly referred to as *prompts* (White et al., 2023). Prompts act as commands that significantly influence the quality and relevance of the output (Knoth et al., 2024). Even slight modifications can lead to vastly different responses. Prompt engineering involves deliberately crafting and refining prompts to guide AI behavior, such as adopting specific styles or answering complex queries, and is increasingly being viewed as a potential new approach to interacting with and programming AI systems (Denny et al., 2024). The quality of the output is closely tied to the precision of the prompt, making prompt engineering essential for the effective use of AI (Velásquez-Henao et al., 2023), in a manner comparable to how data quality shapes overall AI performance in recommender systems (Li et al., 2024; Reinhard et al., 2023).

Building on the premise that higher-quality prompts yield more effective GenAI outputs, we examine *prompt quality* as a key determinant in human-AI delegation. While prior work stresses precise prompting, the attributes, operationalization, and measurement of prompt quality remain unclear and unstandardized (Schulhoff et al., 2024). Prior work has explored the development of prompt pattern catalogs (White et al., 2023) and instructional scaffolds such as worked examples (Tolzin et al., 2024) to support the acquisition of prompt engineering skills. However, the field still lacks systematic models for evaluating or improving prompt quality, as well as a theoretically grounded understanding of its constituent dimensions.

Against this backdrop, the present study aims to identify and structure core evaluation criteria and

optimization strategies through a systematic literature review (Webster and Watson 2002). To guide this investigation, we pose the following research question: *In what ways can the quality of prompts be improved and evaluated, and what criteria and indicators are considered when assessing prompts?*

We address this research question by analyzing 21 studies on prompt quality to derive evaluation criteria and aggregate optimization methods for prompt engineering. The paper first outlines the conceptual foundations of large language models, the role of prompts, and challenges in prompt engineering, before introducing our systematic literature review methodology. Findings include core criteria, potential indicators and measures, and a classification of optimization techniques. Building on these insights, we discuss implications for prompt design and GenAI performance and propose five avenues for future research.

2. Conceptual background

2.1 Large language models

LLMs have evolved rapidly, giving rise to new architectures and techniques for their effective use (Minaee et al., 2024). This study focuses specifically on LLMs, as prompts are text-based and therefore particularly relevant for analysis within this domain. LLMs process textual input through multiple stages. First, tokenization breaks the input into manageable units known as tokens, which may represent entire words, subwords, or syntactic components depending on the model (Zhao et al., 2023). Next, these tokens are transformed into numerical vectors through embedding techniques, allowing the model to represent their meaning within a semantic space (Singh et al., 2024). Based on this representation, the model then predicts the most probable next token using mechanisms such as self-attention (Zhao et al., 2023; Vaswani et al., 2017). In the final step, decoding, the model selects a token from the probability distribution. This is guided by decoding strategies such as greedy decoding or sampling, which favor either the most likely or a more diverse and creative continuation.

2.2 Prompts: Challenges and impact

Prompts are composed by users and elicit corresponding outputs from the model, which may take the form of text, images, audio, video, or combinations thereof (Schulhoff et al., 2024). The ability to issue instructions in natural language enables

intuitive and accessible interaction. Prompts play a pivotal role in determining the quality of generated output, and formulating prompts is increasingly regarded as an acquirable skill that improves with experience (Eager & Brunton, 2023; He et al., 2024).

Imprecise or poorly formulated prompts can result in inaccurate or irrelevant outputs. In this context, multimodal inputs offer potential advantages by providing additional contextual information (Clarísó & Cabot, 2023). Ambiguously phrased prompts can yield varying interpretations, often leading to divergent outputs (Schulhoff et al., 2024). Moreover, factors such as prompt phrasing, content structure, and sequence of instructions significantly influence output quality. Marked performance differences are observed when prompts are issued in low-resource languages compared to English (Schulhoff et al., 2024). Additionally, hallucinations, where the model generates plausible yet factually incorrect responses, pose a critical challenge. Addressing this issue requires careful, prompt evaluation and design (Annapureddy et al., 2024). Well-crafted prompts have been shown to mitigate hallucinations and enhance model performance (Chen et al., 2023).

Despite growing but rather fragmented knowledge about the characteristics of effective prompts, systematic, particularly quantitative, approaches to measuring prompt quality remain underdeveloped. This calls for an aggregated analysis of evaluation criteria and indicators for assessing prompt quality.

2.3 Prompt engineering

Prompt engineering refers to the process of designing, refining, and optimizing prompts to guide LLM-based assistants (Velásquez-Henao et al., 2023). Put differently, it's the practice aimed at enhancing prompt quality. Its primary objective is to convey user intent as precisely as possible to the model in order to elicit relevant, accurate, and useful outputs. A well-structured prompt can significantly enhance the relevance and quality of generated content (Velásquez-Henao et al., 2023). Consequently, the ability to craft effective prompts has emerged as a core competency in the use of GenAI systems (Eager & Brunton, 2023).

However, prompt engineering goes beyond merely adapting the clarity or conciseness of a prompt; it also includes more complex methods and structures to facilitate LLMs' performance. Prompt engineering encompasses a variety of strategies and techniques that enable targeted interaction with language models (Annapureddy et al., 2024). Among these is the chain-of-thought prompting technique, which guides the

model through structured intermediate reasoning steps to solve complex tasks (Schulhoff et al., 2024). This approach has proven especially effective in logical and mathematical problem-solving. Beyond such applications, prompt engineering enables adaptation to diverse use cases, ranging from creative text generation to data-driven analytical tasks.

3. Methodology

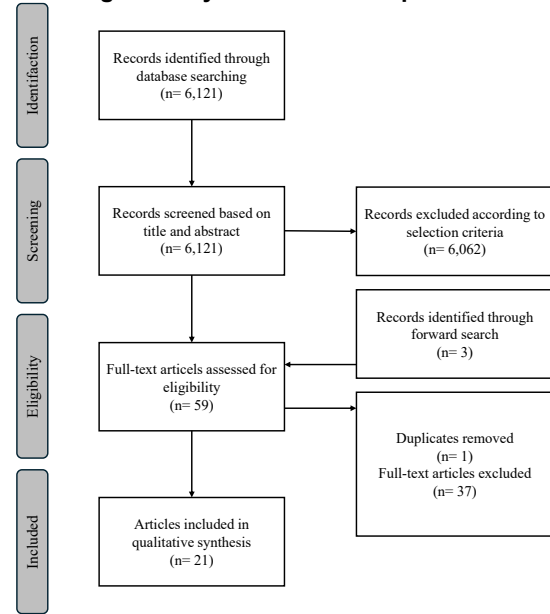
Before conducting the systematic literature review (Webster & Watson, 2002), we developed initial search terms (e.g., *Prompt Quality*, *Prompt, GenAI, Evaluation Methods*) combined with the Boolean operator AND. A preliminary Google Scholar search yielded seven relevant articles, which informed refinement of the keywords while keeping them broad. The final search string again employed AND to combine terms.

(Generative AI OR LLM) AND Prompt AND (Quality OR Evaluation Methods OR Pattern)

To refine the scope of the review, we defined inclusion and exclusion criteria in advance and did not impose any limits on publication date. We included studies evaluating prompt quality or proposing prompt-improvement methods for LLM tasks, provided they offered operational definitions, indicators, procedures, or clear prompt structures. We prioritised peer-reviewed journals and conferences, admitting select directly relevant preprints provided the full text was accessible via the publisher or our university library, and the paper was in English. Duplicates were removed. Based on these criteria, we searched the ScienceDirect database, the ACM Digital Library, and the Arxiv database, which were chosen for their technical orientation and relevance to LLMs and prompt engineering. Search terms were combined using Boolean operators and, where useful, quotation marks to increase precision and reduce irrelevant matches. Screening proceeded in two stages: first, titles were screened for thematic fit; then, abstracts were screened for key terms and operational detail. Eligible items then entered abstract screening and full-text assessment. We excluded works that lacked an operationalisation of prompt quality or actionable improvement procedures; discussed general AI without prompt-specific measures; or lacked sufficient methodological detail (e.g. no criteria or indicators). Consequently, many records were rejected at the title or abstract stage due to an insufficient focus on evaluation criteria or prompt improvement approaches.

The review was carried out between mid-September and the end of October 2024. According to Figure 1, a total of 25 publications from arXiv, 14 from the ACM Digital Library, and 17 from ScienceDirect were included in the screening process. In the forward citation search, three additional relevant studies were identified from references in three core publications. This resulted in a total of 59 publications included in the abstract screening phase.

Figure 1: Systematic review process



Following a detailed paper evaluation, 38 publications were excluded, leaving 21 studies deemed relevant for further analysis. These were systematically examined and synthesized in a concept matrix (Webster & Watson, 2002). The goal was to extract and organize evaluation criteria and optimization approaches for prompts from the existing literature. The identified evaluation criteria were categorized based on the data quality dimensions proposed by Wang and Strong (1996). These categories served as the analytical framework for assessing prompt quality.

4. Results

This chapter presents the key findings of the systematic literature review, which aimed to identify scholarly publications that define criteria for evaluating prompt quality and categorize prompt optimization methods. The results are organized as follows: Section 4.1 presents and classifies the identified optimization techniques; Section 4.2 outlines core evaluation criteria for prompt quality; and Section 4.3 associates these criteria with relevant indicators and measurement approaches. Collectively,

these findings establish a structured foundation for addressing our research question and contribute to a more comprehensive understanding of prompt quality within GenAI systems.

4.1 Optimization methods

As part of the literature review, five overarching categories (see Appendix) of optimization methods were identified that can be applied to enhance prompt quality.

Shot-based methods. This category includes zero-shot, few-shot prompting, and example sequencing. Zero-shot prompting refers to solving a task without providing the model with any examples, while few-shot prompting involves supplying a limited number of examples to guide the model’s behavior (Chen et al., 2023; Zhao et al., 2023). The latter has been shown to be particularly effective in steering the model toward desired outcomes (Knoth et al., 2024). The order in which examples are presented also significantly influences model performance (Schulhoff et al., 2024).

Chain-of-thought reasoning. This method encourages the model to solve tasks through a step-by-step reasoning process rather than generating a direct final answer. It enhances the model’s ability to reason, improves interpretability, and contributes to more accurate outputs (Chen et al., 2023; Liu et al., 2024).

Iterative refinement. Cited in 17 of the 21 reviewed publications, this method involves a multi-stage process in which prompts and responses are continuously revised and improved (Krishna et al., 2024; Patel & Parmar, 2024). Iterative refinement can reduce hallucinations and increase the consistency of outputs (Zhao et al., 2023).

Role-based prompting. In this approach, the model is assigned a specific role (e.g., professor, historian) to provide more context-aware and higher-quality responses (Schulhoff et al., 2024; Chen et al., 2023). Adopting a predefined role allows the model to simulate domain-specific expertise.

Sequential prompting. This technique breaks down complex tasks into discrete steps, with each prompt building upon the previous one. The goal is to increase accuracy, comprehensibility, and transparency in the model’s reasoning process (Chen et al., 2023). Chain-of-thought (CoT) elicits intermediate reasoning steps within a single prompt and single completion. In contrast, sequential prompting decomposes the task into multiple prompts whose inputs and outputs are chained and can be verified between steps. CoT is suitable for single-shot complex reasoning, whereas sequential prompting is preferable for multi-step workflows that need

decomposition, checkpoints, and error control. In short, CoT captures an internal reasoning trace, while sequential prompting externalizes the process into controllable stages.

4.2 Evaluation criteria for prompt quality

In addition, the systematic literature review synthesized six core criteria to evaluate prompt quality (Table 1 and Table 2): (1) *clarity*, (2) *accuracy*, (3) *precision*, (4) *relevance*, (5) *completeness*, and (6) *conciseness*. These dimensions build on the data quality framework proposed by Wang and Strong (1996) as a foundational anchor in terms of overarching dimensions of intrinsic, contextual, and representational data quality, supplemented by domain-specific considerations relevant to LLM-based assistants.

(1) Clarity. The additional criterion of clarity was emphasized in 18 of the reviewed publications, making it the most frequently cited dimension of prompt quality. Clarity refers to the unambiguity and comprehensibility of a prompt’s phrasing, ensuring that the model can interpret the intended task without confusion (Zhao et al., 2023). While related to (2) *accuracy*, which concerns the factual correctness and alignment of a prompt with its intended task, clarity focuses on how clearly and explicitly the task is communicated. A prompt may be factually accurate yet still unclear if phrased in a vague, ambiguous, or overly complex manner. To promote clarity, effective prompts include well-defined goal statements, clearly specified input and output formats, and unambiguous response expectations (Chen et al., 2023). As such, clarity plays a foundational role in guiding the model’s understanding and reducing the risk of misinterpretation, especially in complex or multi-step tasks

(2) Accuracy. Among the intrinsic criteria, accuracy was emphasized in 16 of the 21 reviewed publications. It refers to the extent to which a prompt is factually correct and aligned with the intended task (Kim & Lu, 2024). Accuracy can be enhanced through techniques such as verification questions that validate factual correctness (Sahoo et al., 2024) or through context-enriching follow-up queries that clarify user intent (White et al., 2023). Accordingly, the literature reveals several technical approaches that support users in crafting more accurate prompts. A closely related concern is ambiguity in user prompts, which can significantly constrain both prompt accuracy and the reliability of the generated outputs (Zhao et al., 2023; Annapureddy et al., 2024). Ambiguity often results from vague or underspecified prompts, increasing the likelihood of misinterpretation by the model.

Table 1: Evaluation criteria

Authors (Year)	Intrinsic			Contextual		Representational
	Accuracy	Clarity	Precision	Relevance	Completeness	Conciseness
Balepur et al. (2024)	x					
Chang et al. (2024)	x	x	x	x	x	
Chen et al. (2023)	x	x	x	x		
Eager & Brunton (2023)		x	x		x	
He et al. (2024)	x	x	x			
Kim et al. (2024)		x		x		
Knoth et al. (2024)		x		x		
Korzynski et al. (2023)	x	x				
Lo (2023)		x	x	x		x
Patel & Parmar (2024)	x	x	x	x		x
Robertson et al. (2024)		x				x
Sahoo et al. (2024)	x	x	x			
Schulhoff et al. (2024)	x	x	x	x		
Singh et al. (2024)	x			x		
Velásquez-Henao et al. (2024)	x	x	x	x	x	
White et al. (2023)	x	x	x		x	x
Ye et al. (2024)	x				x	
Yu-Te Lee et al. (2024)	x	x			x	
Zequan et al. (2024)	x	x	x			
Zhao et al. (2023)	x	x				
Total	15	17	11	9	6	4

Furthermore, the sequence and framing of examples within a prompt have also been shown to influence the accuracy of model responses, as they shape the model’s internal representation of task expectations (Schulhoff et al., 2024).

(3) Precision. The criterion of precision was addressed in 11 of the 21 reviewed publications. Precision refers to the degree of specificity in a prompt’s formulation, aiming to reduce ambiguity and guide the model more directly toward the desired output (Chen et al., 2023). Often discussed in conjunction with *clarity* (Ekin, 2023; Lo, 2023), precision differs in its focus on the level of detail and exactness in language, rather than general comprehensibility. While clarity ensures that the prompt is understandable and unambiguous, and accuracy ensures the factual correctness and task alignment, precision emphasizes how narrowly and explicitly the prompt defines what is expected from the model.

(4) Relevance. The contextual dimension of relevance is also frequently discussed in the literature. Relevance refers to the degree to which a prompt is appropriate and aligned with the requirements of the

given task or goal (Chang et al., 2024). Nine studies emphasize the importance of tailoring prompts to specific task contexts, highlighting that generic or misaligned prompts often lead to suboptimal outcomes (Kim et al., 2024). To support systematic assessment, the LLM-EVAL framework proposed by Lin and Chen (2023) provides a structured methodology for evaluating a prompt’s relevance, incorporating criteria such as task fit and contextual adequacy.

(5) Completeness. Completeness refers to the extent to which a prompt contains all information necessary for the model to correctly interpret and execute the task. This criterion was addressed in four of the reviewed studies, with White et al. (2023) and Faragó (2023) emphasizing that completeness can be improved through follow-up prompts or iterative refinement strategies. Incomplete prompts, by contrast, often lead to vague or incorrect outputs, underscoring the importance of ensuring that all relevant contextual and task-specific elements are explicitly stated, which emphasizes the relatedness to (2) relevance.

(6) Conciseness. The representational criterion of conciseness was examined in four of the reviewed publications. Conciseness involves reducing prompts to their essential elements while retaining the necessary information for task execution, thereby promoting both clarity and computational efficiency. Excessively long prompts may introduce noise, dilute the intended meaning, or lead the model to misallocate attention across less relevant components. The CLEAR framework aligns with this principle by encouraging users to prioritize core content and eliminate redundancy to optimize model performance (Lo, 2023).

Table 2. Definition and scope of the criteria

Criteria	Description
<i>Clarity</i>	The unambiguity and comprehensibility of a prompt’s phrasing ensure the model can interpret the intended task without confusion (Zhao et al., 2023).
<i>Accuracy</i>	The extent to which a prompt is factually correct and aligned with the intended task (Kim & Lu, 2024).
<i>Precision</i>	The degree of specificity in a prompt’s wording to reduce ambiguity and guide the model toward the desired output (Chen et al., 2023).
<i>Relevance</i>	The extent to which a prompt is appropriate and aligned with the requirements of the given task or goal (Chang et al., 2024).
<i>Completeness</i>	The degree to which a prompt includes all information necessary for correct interpretation and execution by the model (Eager and Brunton, 2023).
<i>Conciseness</i>	The reduction of a prompt to its essential elements while preserving required information, promoting clarity and model efficiency.

Overall, the findings reveal that a range of interrelated criteria contribute to prompt quality. Their

combined application provides a foundation for the systematic improvement and potential standardization of prompt engineering practices.

To demonstrate practical use of our framework, we present a single-shot prompt (Figure 2) for a common analytical task and briefly show how its design encodes the core quality criteria.

Figure 2: Illustrative example

Model input (Prompt):	
Write a 110–130 word executive brief on “Q2 revenue variance for RetailCo (year-over-year)”. Use only the provided data table . Structure the output as one paragraph with (1) one-sentence key finding, (2) two brief drivers(e.g., price/mix, volume, or channel shift), (3) one actionable recommendation, and (4) one KPI to monitor next quarter.	
Explanation of the criteria:	
CLARITY:	The task, source, audience tone, and required components are explicit.
ACCURACY:	Using the term ‘year-over-year’ conveys the intended task with greater accuracy.
PRECISION:	The fixed word range and exact component counts reduce interpretation space and stabilize outputs.
RELEVANCE:	‘Use only the provided data table’ anchors every claim in the given numbers and prevents external additions.
COMPLETENESS:	The combination of finding, drivers, action, and KPI covers diagnosis, cause, and follow-up measurement.
CONCISENESS:	Short-sentence style and the word count keep the prompt concise.

4.3 Measuring prompt quality criteria

The identified evaluation criteria serve as the foundation for the subsequent analysis of associated indicators (Table 3). These indicators further elaborate the criteria and offer a basis for the potential measurement of prompt quality. The literature review revealed a lack of standardized qualitative or quantitative measurement approaches. In response, this chapter develops an initial set of methodological propositions to guide the systematic assessment of prompt quality based on our review. Drawing on established practices in evaluating LLM behavior (e.g., Manakul et al., 2023), it repurposes these measurement approaches, such as factuality and self-consistency, for the context of prompt evaluation.

Clarity can be operationalized using indicators such as ambiguity and uncertainty. Zhao et al. (2023) observe that ambiguous prompts often lead to inaccurate responses. Ambiguity can be assessed by systematically reviewing prompts for potential sources of confusion or by adapting evaluation methods originally developed for LLMs, such as uncertainty estimates (Zhang & Choi, 2023), to quantify the level of ambiguity. In addition, Nouri et al. (2021) propose a taxonomy of clarity flaws that can be applied by human annotators to assess prompt quality. Finally, readability scores offer a complementary perspective by providing a quantifiable indication of a prompt’s linguistic clarity.

In relation to the criterion of *accuracy*, the indicator of correctness is defined as the requirement that generated outputs be factually accurate and reliable, as emphasized by Chang et al. (2024). In addition, the use of verification questions emerges as a complementary indicator, identified in prior research (He et al., 2024) as a strategy to detect and reduce inaccuracies in generated content.

The criterion *precision* is linked to the indicator level of detail, which reflects the extent to which a prompt includes specific information. A higher level of detail can improve the accuracy of model outputs (Chang et al., 2024; Zhao et al., 2023). One way to measure precision is to calculate the proportion of detailed elements relative to the total word count of the prompt. One approach is the concept of concrete and abstract words from the field of linguistics (Brysbaert et al., 2014).

From the criterion of *relevance*, the indicator of context sensitivity is derived. Chang et al. (2024) emphasize that prompts should provide sufficient contextual information to ensure that the model can accurately interpret and execute the task. One possible measurement approach involves expert review of the prompt to evaluate the presence and adequacy of relevant contextual elements. Beyond this, relevance remains closely tied to the specific task context and is comparatively underexplored in the existing literature.

Table 3: Measurement methods

Criteria	Indicators	Measurement methods
Clarity	<i>Ambiguity and uncertainty quantification</i>	Self-consistency (Manakul et al., 2023), scalar uncertainty estimates (Zhang and Choi, 2023)
	<i>Identification of clarity flaws</i>	Questionnaires include aspects like difficult wording, solution format, and resources. (Nouri et al., 2021)
	<i>Readability scores</i>	E.g., Flesch reading ease score
Accuracy	<i>Verification questions</i>	Context-specific verification questions derived from a connected knowledge base. (He et al., 2024)
	<i>Correctness</i>	Factuality scores (Chang et al., 2024)
Precision	<i>Concreteness ratings</i>	Score of concrete and abstract words (Brysbaert et al., 2014)
Relevance	<i>Context sensitivity</i>	Rating the appropriateness of the given context
Completeness	<i>Prompt components</i>	Verb, focus, context, focus and condition, alignment, etc. (Eager and Brunton 2023)
	<i>Prompt templates</i>	Context-specific templates used to evaluate completeness (Ye et al., 2024)
Conciseness	<i>Information density</i>	Relevant words in relation to the total number of words
	<i>Formality</i>	Formality scores (Pavlick and Tetreault, 2016)

For the criterion of *completeness*, no singular quantitative indicators or measurement methods could be extracted from the literature. However, research proposes prompt components (e.g., verb, focus, context, etc.) (Eager and Brunton, 2023) and prompt templates (Ye et al., 2024) as guides for improving and also measuring prompt quality (Knoth et al., 2024) in terms of completeness. The operationalization of completeness remains highly dependent on specific tasks and contexts, thereby limiting the generalizability of associated indicators. The criterion of *conciseness* is represented by the indicator information density. Lo (2023) argues that unnecessary information should be removed in order to guide the model more effectively. Information density may be calculated by dividing the number of relevant words by the total word count of the prompt. Other linguistic measures, such as formality (Pavlick and Tetreault, 2016), can be utilized to operationalize conciseness. Collectively, the presented criteria and indicators offer an initial systematic foundation for the analysis and improvement of prompts.

5. Discussion

Prompt evaluation criteria are essential for assessing prompt quality and improving GenAI outputs. The systematic literature review highlights clarity, accuracy, and precision as the most frequently cited criteria. *Clarity* helps prevent misinterpretation and supports more accurate internal processing by the model. Two indicators, unambiguity and comprehensibility, are associated with clarity, although no standardized measurement approaches exist for either, highlighting the need for further research into objective measurement methods.

Accuracy (15 mentions) and *precision* (11 mentions) also emerge as highly relevant, as users expect correct and specific responses. However, the associated indicators, such as correctness, comprehensibility, and level of detail, likewise lack concrete operational measurement methods. *Relevance* is discussed in nine publications, but its dependence on specific tasks limits its generalizability. *Completeness* and *conciseness* are mentioned less frequently and may be considered subdimensions of clarity and precision. However, the inclusion of conciseness as a desirable criterion is not without challenges, since more elaborate or verbose prompts are often necessary to elicit higher-quality responses from LLMs. Again, no suitable indicators or defined metrics were identified. Despite widespread references to core criteria, a standardized evaluation framework is still lacking. A key rationale for this limitation lies in the inherently interactional and task-

contingent nature of prompt engineering. Unlike data quality, which is often treated as an objective property, the effectiveness of prompts is shaped by linguistic, pragmatic, and contextual factors. Nonetheless, clarity, accuracy, and precision may be considered foundational pillars for developing a prompt evaluation framework that can support systematic prompt optimization.

Iterative refinement is the most frequently cited optimization method, offering strong improvements through repetition and feedback. Other techniques, such as chain-of-thought prompting, role-based prompting, and shot-based methods, are also discussed, though less often. Combined with evaluation criteria, these approaches provide practical ways to enhance prompt quality. Ultimately, prompt effectiveness depends not only on formulation but also on applied strategies, with user competency remaining a key success factor. Prompt engineering is thus a learnable skill for improving AI interaction.

Overall, two central approaches emerge for systematically enhancing prompt quality. First, the application of well-defined evaluation criteria can guide the formulation of prompts that are clear, precise, and goal-oriented. These criteria help structure the input and improve its content quality. Second, prompt engineering methods - such as role prompting, chain-of-thought prompting, or few-shot learning - offer concrete strategies to boost model performance. The choice of method should be aligned with the type of prompt and intended outcome. These two approaches can be applied independently, but also complement each other effectively. For instance, prompts may be crafted in accordance with criteria such as clarity, relevance, and precision, while additionally leveraging techniques like role prompting to further enhance the model's output. Such combinations can meaningfully contribute to performance optimization and unlock new potential for the targeted use of GenAI systems.

To better understand the processes involved in giving instructions and to relate prompt engineering to established methods of instruction giving and delegation, we reviewed literature from domains such as education, communication, and professional environments (Haug et al. 2015). We aimed to uncover conceptual parallels between instructional design in these fields and prompt formulation for GenAI systems. Providing instructions is particularly relevant in the field of leadership (Dulewicz and Higgs 2005), crowdsourcing (Nouri et al. 2021), industrial work (Haug 2015), and education (Mok and Staub 2021). The analysis revealed initial thematic overlaps, particularly concerning the clarity and completeness of language. Both domains stress the importance of

clear, precise, and contextually relevant formulations. However, the scarcity of systematic academic studies limits the ability to conduct a rigorous comparison across fields. In conclusion, the methods of crafting effective instructions, whether in crowdsourcing environments or in routine business delegation, appear to share key characteristics with prompt engineering, yet remain underexplored in scholarly research.

Our research is not without limitations. Given the nascent and rapidly evolving nature of prompt engineering and prompt quality, the body of peer-reviewed literature remains limited. Consequently, our review includes preprints and other non-journal sources to capture the field's dynamic developments. Furthermore, the formulation of instructions is highly contingent on the discipline, the nature of the tasks, data characteristics, and modalities. Our study acknowledges these contingencies but does not provide an in-depth analysis of the boundary conditions of the criteria or their task-dependent nature. Accordingly, ensuring prompt quality and measuring it accurately and continuously is effortful due to the context-specific characteristics of prompt quality. While Wang and Strong's (1996) data quality framework provides a useful foundation, it does not capture the task-contingent and interactional qualities of prompt formulation. IS research offers additional theoretical resources for conceptualizing prompt quality, such as task-technology fit (Goodhue & Thompson, 1995) or affordance theory (Markus & Silver, 2008). These perspectives highlight that formulating high-quality prompts is not only about clarity and correctness but also about alignment with task demands and the socio-technical context in which human-AI collaboration unfolds.

6. Avenues for future research

We identify five avenues for future research on prompt quality and engineering, outlining critical directions for LLM research and the emerging role of prompt engineering.

Future research avenue 1: Towards a scale for prompt quality. While our analysis revealed recurring evaluation criteria and dimensions related to the informational quality of instructions, we identified a persistent lack of measurable indicators and a comprehensive framework for both qualitative and quantitative analysis of prompt quality. Future research should aim to develop and empirically test measurement scales that enable tracing the quality of GenAI outputs back to prompt quality and prompt engineering capabilities. Greater emphasis should be placed on validating and refining the identified evaluation criteria and their associated indicators. The

exploration of human-LLM interaction relies on the ability to evaluate prompt behavior, contributing to the development of prompt analytics as a new research domain. Moreover, longitudinal research could examine how prompting strategies and evaluation criteria evolve over time, as both LLMs and our understanding of them continue to develop rapidly.

Future research avenue 2: Understanding of context-contingent prompt quality. Future research should focus on defining and operationalizing context-specific prompt quality, recognizing that effective prompting is highly dependent on the domain, task, and user intent. While existing studies often assess the output quality of LLMs, they rarely establish a causal link between the input prompt and the resulting performance, limiting our understanding of prompt effectiveness. To advance this area, researchers must investigate how prompt quality varies across contexts and develop methods for tracing outcomes back to prompt characteristics. Moreover, greater attention is needed to evaluate the transferability of prompt quality criteria across different application domains and LLM configurations, ensuring broader applicability and robustness of findings.

Future research avenue 3: Approaches of clarification and disambiguation. We found that reducing ambiguity is a recurring theme in the literature on prompt quality. The inherent lack of specificity in human language and the potential ambiguity of user intent pose significant challenges for LLM performance. While some studies aim to develop ambiguity-aware models to enhance intent recognition, unclear user prompts continue to lead to misunderstandings and reduce the effectiveness of human-AI interaction. This highlights the need for further research into approaches that clarify and disambiguate prompts by helping users express their intent more clearly.

Future research avenue 4: System-guided prompt refinement. Future research should explore system-guided approaches to prompt refinement that support users in crafting clearer, more effective prompts. This includes developing LLM-based mechanisms capable of identifying gaps, ambiguities, or inefficiencies in prompts and triggering appropriate refinement strategies. At this stage, tools such as ChatGPT or Langsmith can be leveraged to enhance prompt quality. Interactive and proactive agents that help users and prompt engineers resolve vague instructions can significantly enhance the accuracy and reliability of LLM outputs. A particularly promising direction involves integrating evaluation frameworks directly into GenAI systems, enabling real-time analysis and automated assessment of prompt quality within the interaction process itself.

Future research avenue 5: Theoretical understanding of instruction giving. A deeper theoretical understanding of instruction giving is essential yet currently underrepresented in prompt engineering research. Our approach linked prompt quality to established dimensions of data quality, such as intrinsic, contextual, and representational factors, to conceptualize how instructions shape LLM outputs. Future work should examine the mechanisms through which prompt quality influences model behavior, as well as the abilities, skills, and experiences that contribute to effective prompt engineering. Adopting an interdisciplinary perspective, drawing from fields such as communication, education, and cognitive science, can reveal important parallels between human-to-human delegation and prompting LLMs.

7. Conclusion

Our study offers the first structured synthesis of criteria and methods for prompt optimization, providing a foundation for improving prompt design and GenAI performance. While research has emphasized engineering techniques, we highlight a gap in evaluating prompt quality and propose key criteria with potential indicators, though standardized methods remain absent. A cross-domain perspective on instruction design highlights overlaps with delegation and instruction-giving, opening promising directions for future research on prompt analytics and prompt quality.

8. References

- Annapureddy, R., Fornaroli, A., & Gatica-Perez, D. (2024). Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice*, (07/2024), 1–22.
- Balepur, N., Gu, F., Ravichander, A., Feng, S., Boyd-Graber, J., & Rudinger, R. (2024). Reverse Question Answering: Can an LLM Write a Question so Hard (or Bad) that it Can't Answer?. *arXiv preprint arXiv:2410.15512*.
- Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods*, 46, 904–911.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., ... & Xie, X. (2024). A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3), 1–45.
- Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). Unleashing the potential of prompt engineering in large language models: a comprehensive review. *arXiv preprint arXiv:2310.14735*.
- Clariso, R., & Cabot, J. (2023). Model-driven prompt engineering. In *Proceedings of the 26th International Conference on Model Driven Engineering Languages and Systems*, 26, 47–54.
- Denny, P., Leinonen, J., Prather, J., Luxton-Reilly, A., Amarouche, T., Becker, B. A., & Reeves, B. N. (2024, March). Prompt Problems: A new programming exercise for the generative AI era. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1* (pp. 296–302).
- Dulewicz, V., & Higgs, M. (2005). Assessing leadership styles and organisational context. *Journal of managerial psychology*, 20(2), 105–123.
- Eager, B., & Brunton, R. (2023). Prompting higher education towards AI-augmented teaching and learning practice. *JUTLP*, 20(5), 1–19.
- Ekin, S. (2023). Prompt engineering for ChatGPT: A quick guide to techniques, tips, and best practices. *Authorea Preprints*.
- Faragó, D. (2023). Engineering a reliable prompt for generating unit tests: Prompt engineering for QA & QA for prompt engineering. *Softwaretechnik-Trends*, 43(3), 10–14.
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business & Information Systems Engineering*, 66(1), 111–126.
- Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS quarterly*, 213–236.
- Haug, A. (2015). Work instruction quality in industrial management. *International Journal of Industrial Ergonomics*, 50, 170–177.
- He, J., Wang, X., Liu, S., Wu, G., Silva, C., & Qu, H. (2024). POEM: Interactive Prompt Optimization for Enhancing Multimodal Reasoning of Large Language Models. *arXiv preprint arXiv:2406.03843*.
- Kim, T. S., Lee, Y., Shin, J., Kim, Y. H., & Kim, J. (2024). Evallm: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*.
- Kim, M., & Lu, X. (2024). Exploring the potential of using ChatGPT for rhetorical move-step analysis: The impact of prompt refinement, few-shot learning, and fine tuning. *Journal of English for Academic Purposes*, 71, 1–15.
- 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.
- Korzynski, P., Mazurek, G., Krzyrkowska, P., & Kurasinski, A. (2023). Artificial intelligence prompt engineering as a new digital competence: Analysis of generative AI technologies such as ChatGPT. *Entrepreneurial Business and Economics Review*, 11(3), 25–37.
- Krishna, S., Agarwal, C., & Lakkaraju, H. (2024). Understanding the effects of iteration prompting on truthfulness. *arXiv preprint arXiv:2402.06625*.
- Lee, S. Y.-T., Bahukhandi, A., Liu, D., & Ma, K.-L. (2024). Towards dataset-scale and feature-oriented evaluation of text summarization in large language model

- prompts. *IEEE Transactions on Visualization and Computer Graphics*.
- Li, M. M., Reinhard, P., Peters, C., Oeste-Reiss, S., & Leimeister, J. M. (2024). A value co-creation perspective on data labeling in hybrid intelligence systems: A design study. *Information Systems*, 120, 102311.
- Lin, Y.-T., & Chen, Y.-N. (2023). LLM-Eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*
- Liu, Z., Wang, L., & Li, X. (2024). The method of utilizing prompt engineering to train large language models for the intelligent expression of affair-elements. *Procedia Computer Science*, 24, 1179–1184.
- Lo, L. S. (2023). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship*, 49, 1–3.
- Manakul, P., Liusie, A., & Gales, M. J. (2023). Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*
- Markus, M. L., & Silver, M. S. (2008). A foundation for the study of IT effects: A new look at DeSanctis and Poole's concepts of structural features and spirit. *Journal of the Association for Information systems*, 9(10), 5.
- Minaee, S., et al. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*
- Mok, S. Y., & Staub, F. C. (2021). Does coaching, mentoring, and supervision matter for pre-service teachers' planning skills and clarity of instruction? A meta-analysis of (quasi-) experimental studies. *Teaching and Teacher Education*, 107, 103484.
- Nah, F. F.-H., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304.
- Nouri, Z., Gadiraju, U., Engels, G., & Wachsmuth, H. (2021). What is unclear? computational assessment of task clarity in crowdsourcing. In *Proceedings of the 32nd ACM conference on hypertext and social media* (pp. 165-175).
- Patel, H., & Parmar, S. (2024). Prompt engineering for large language model. *ResearchGate*, *Research proposal*. <https://doi.org/10.13140/RG.2.2.11549.93923>
- Pavlick, E., & Tetreault, J. (2016). An empirical analysis of formality in online communication. *Transactions of the association for computational linguistics*, 4, 61-74.
- Reinhard, P., Li, M. M., Dickhaut, E., Peters, C., & Leimeister, J. M. (2023). Empowering recommender systems in ITSM: a pipeline reference model for AI-based textual data quality enrichment. In *International Conference on Design Science Research in Information Systems and Technology* (pp. 279-293).
- Reinhard, P., Li, M., Peters, C., & Leimeister, J. M. (2024). Generative AI in customer support services: a framework for augmenting the routines of frontline service employees. In *Hawaii International Conference on System Sciences*.
- Robertson, J., Ferreira, C., Botha, E., & Oosthuizen, K. (2024). Game changers: A generative AI prompt protocol to enhance human-AI knowledge co-construction. *Business Horizons*, 67(5), 499-510.
- Sahoo, P., et al. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. *arXiv:2402.07927*.
- Schneider, J., Meske, C., & Kuss, P. (2024). Foundation models: a new paradigm for artificial intelligence. *Business & Information Systems Engineering*, 66(2), 221-231.
- Schulhoff, S., et al. (2024). The Prompt Report: A systematic survey of prompting techniques. *arXiv preprint arXiv:2406.06608v3*.
- Singh, A., Ehtesham, A., Gupta, G. K., Chatta, N. K., Kumar, S., & Khoei, T. T. (2024). Exploring prompt engineering: A systematic review with swot analysis. *arXiv preprint arXiv:2410.12843*.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS 30)*, 5998–6008
- Tolzin, A., Knoth, N., & Janson, A. (2024). Leveraging prompting guides as worked examples for advanced prompt engineering strategies. *Proceedings of the International Conference on Information Systems (ICIS 2024)*
- Velásquez-Henao, J. D., Franco-Cardona, C. J., & Cadavid-Higuaita, L. (2023). Prompt Engineering: a methodology for optimizing interactions with AI-Language Models in the field of engineering. *Dyna*, 90(SPE230), 9-17.
- Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(Spring), 5–33.
- Webster, J., & Watson, R. T. (2002). Analyzing the past to prepare for the future: Writing a literature review. *MIS Quarterly*, 26(2), xiii–xxiii.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... & Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Ye, Q., Axmed, M., Pryzant, R., & Khani, F. (2024). Prompt engineering a prompt engineer. *arXiv preprint arXiv:2311.05661*.
- Zequan, L., Longhao, W., & Xingsen, L. (2024). The method of utilizing prompt engineering to train large language models for the intelligent expression of affair-elements. *Procedia Computer Science*, 242, 1179-1184.
- Zhang, M. J., & Choi, E. (2023). Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469*.
- Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*, 1(2).