big data and cognitive computing

MDPI

Systematic Review

Better Prompts, Better Usefulness: A Systematic Review and Experimental Evaluation of Structured Prompting Techniques in Large Language Models

Alessia Cantini and Andrea De Mauro *

Department of Business and Management, LUISS Guido Carli University, 00197 Rome, Italy; alessia.cantini@alumni.luiss.it

* Correspondence: ademauro@luiss.it

Abstract

Large Language Models (LLMs) have rapidly become central components of cognitive computing systems and AI-assisted knowledge work. However, the effectiveness of LLM-generated outputs depends not only on the model's capabilities but also on the structure of the prompts used to guide them. This study investigates how structured prompting techniques influence perceived output usefulness in business-oriented tasks. First, we conduct a systematic literature review following PRISMA guidelines to identify, classify, and synthesize existing prompt enhancement strategies. The review leads to the development of a taxonomy distinguishing task-alignment techniques (e.g., one-shot and few-shot prompting) from reasoning-transparency techniques (e.g., Chain-of-Thought prompting). Building on this taxonomy, we design a controlled experimental study in which knowledge workers evaluate LLM-generated outputs across analytical and summarization tasks. Using linear mixed-effects modeling, we assess the impact of prompting techniques and the moderating role of Generative AI usage frequency. Results indicate that structured prompting significantly increases perceived usefulness compared to baseline approaches, with the combination of example-based conditioning and explicit reasoning scaffolding yielding the highest evaluations. The moderating effect of usage frequency is not statistically significant, suggesting that the benefits of structured prompt design are robust across different experience levels. These findings position prompt structure as a practical cognitive interface mechanism and provide evidence-based guidelines for enhancing human–AI interaction in cognitive computing environments.

Keywords: large language models; prompt engineering; prompting techniques; human evaluation; systematic literature review

Academic Editor: Fabrizio Messina

Received: 25 March 2026
Revised: 14 June 2026
Accepted: 18 June 2026
Published: 6 July 2026

1. Introduction

The increasing integration of Artificial Intelligence into organizational workflows is transforming how work is planned and how employees carry out their tasks [1]. This transformation is not limited to automation in a narrow operational sense. Rather, it increasingly concerns the way organizations access, process, interpret, and mobilize knowledge across everyday activities. In this respect, AI is reshaping not only task execution but also the cognitive conditions under which work is performed, especially in contexts characterized by information abundance, time pressure, and the need for rapid judgment. Evidence shows that using Generative AI can boost productivity and improve output quality, while

also enhancing learning and providing a better work environment for employees [2]. These benefits are particularly relevant for knowledge-intensive settings, where employees are required to synthesize information, formulate interpretations, draft content, and support decisions under conditions of uncertainty. Nonetheless, research on human–AI interaction reveals that Generative AI systems may also introduce inefficiencies if users struggle to interpret and effectively use AI outputs. Challenges such as assessing responses, excessive dependence on the system, and workflow interruptions can hinder performance and introduce new risks [3]. Accordingly, the organizational value of Generative AI cannot be understood as a simple consequence of model capability alone. It also depends on whether users are able to interact with these systems in ways that produce outputs perceived as relevant, clear, reliable, and practically useful for the task at hand.

In this context, prompt engineering, defined as the craft of creating effective queries to improve AI responses, is crucial for maximizing the effectiveness of Generative AI systems [4]. Several studies explicitly state that well-designed prompts enhance performance and reliability while reducing hallucinations, establishing prompt engineering as a key method for improving generative models [5]. More broadly, prompt engineering can be understood as the main interaction mechanism through which users communicate task requirements, contextual constraints, and output expectations to Large Language Models (LLMs). Since these systems do not directly access user intent beyond what is expressed in natural language, the prompt becomes the primary vehicle through which human goals are translated into machine behavior. For instance, statistically significant differences have been observed between zero-shot and few-shot prompting strategies across multiple LLMs in controlled annotation and classification settings [6]. Similarly, research on multi-modal Large Language Models shows that variations in prompt design and input structuring across visual and textual modalities can substantially influence reasoning accuracy in experimental environments [7]. Taken together, these findings suggest that prompt design is not a marginal technical detail, but a central condition shaping the quality of AI-generated responses. Even relatively small changes in structure, specificity, or guidance may influence not only output correctness but also coherence, depth of reasoning, and alignment with users' actual needs.

This point is especially important in business contexts. In organizational practice, users do not typically rely on Generative AI only for highly standardized or benchmark-like tasks. Instead, they often use these systems for open-ended, cross-functional, and cognitively varied activities, such as interpreting indicators, summarizing documents, drafting communications, or generating first-pass analyses. In such settings, the quality of interaction becomes decisive because the usefulness of an output cannot be reduced to technical accuracy alone. A response may be formally plausible yet still fail to support the user if it is too generic, poorly structured, insufficiently contextualized, or difficult to act upon. From this perspective, prompt engineering matters because it shapes not only what the model produces but also how effectively the output can be incorporated into real work practices.

Prompt design, therefore, should be seen as more than a technical optimization technique. It also represents an emerging organizational capability that mediates human–AI collaboration. If Generative AI is increasingly embedded in routine knowledge work, then the ability to formulate prompts that guide the model appropriately becomes part of the broader set of skills required to extract value from these systems. This is particularly relevant because many business users are not AI specialists and interact with LLMs through natural language rather than through formal programming or model tuning. As a result, prompting is one of the most accessible yet consequential levers through which organizations can influence the practical usefulness of Generative AI.

Despite this growing evidence, existing studies predominantly investigate prompting techniques in domain-specific settings and highly technical tasks. For instance, Nugroho and Shaferi [8] analyze how prompt engineering affects ChatGPT's stock recommendations in the Indonesian energy sector, while J. Wang [9] examines the role of prompting strategies in AI-generated business communication messages. Other contributions provide rigorous comparisons of prompting techniques, yet these are largely confined to software engineering, coding, structured data processing, or multimodal benchmarks [10–13]. While these studies offer valuable insights into the performance of different prompting strategies, they also reveal an important limitation in the current literature. Much of the available evidence remains fragmented across specific tasks, technical domains, and evaluation settings. This makes it difficult to build a cumulative understanding of how prompt enhancement techniques affect the usefulness of outputs in routine business situations that require interpretation, synthesis, and judgment rather than narrowly bounded technical execution. As a result, there remains limited systematic evidence on how different prompting techniques influence the usefulness of outputs in cross-domain business tasks that reflect everyday organizational knowledge work. Cross-domain applications have been identified as a promising direction for advancing prompt engineering research [14]. This gap carries practical implications. Organizations increasingly experiment with Generative AI across diverse functions, yet managers and employees still lack clear, evidence-based guidance on which prompting strategies are more appropriate for different forms of knowledge work and why.

To address this gap, our study develops an operational framework of prompt enhancement techniques based on a systematic review and evaluates their impact on perceived usefulness in routine business-related tasks. More specifically, the study adopts a usage-centred perspective on prompt engineering. Rather than treating prompts merely as technical inputs, we conceptualize prompt enhancement techniques as structured interaction strategies that can shape the relevance, clarity, and actionability of AI outputs in organizational settings. By focusing on analysis and synthesis activities within an enterprise context, the study moves beyond narrowly defined technical benchmarks and examines how knowledge workers evaluate structured prompting strategies in cognitively demanding, real-world scenarios. This choice is theoretically and managerially relevant. Analysis and synthesis tasks are common across business functions, are less constrained than purely factual retrieval, and more directly reflect the kinds of interpretive work for which employees increasingly turn to Generative AI. By studying these tasks, the paper aims to provide insight into how prompt structure influences perceived usefulness in settings that are closer to everyday organizational practice than to laboratory-style benchmarks.

The study also contributes by combining two complementary perspectives. First, it systematically reviews the literature in order to identify, classify, and synthesize the main prompt enhancement techniques discussed in prior research. Second, it empirically examines whether these techniques make a difference when evaluated by users in realistic business scenarios. This combination allows the paper to connect the fragmented technical literature on prompting with a human-centred assessment of usefulness. In doing so, it positions prompt structure as a relevant mechanism of human–AI coordination in knowledge work, rather than as a purely experimental variable isolated from organizational use.

Research on AI-employee collaboration emphasizes that the effectiveness of AI systems depends on employees' skills and experience in working with them [15]. Building on this perspective, we examine whether users' familiarity with AI, operationalized as frequency of use, influences their evaluation of different prompting strategies. Specifically, we test whether usage frequency moderates the relationship between prompt enhancement techniques and their perceived usefulness in business-related tasks. This additional focus is

critical because widespread exposure to Generative AI does not necessarily imply a better understanding of how to use it effectively. Frequent users may become more comfortable with AI systems, but they may not automatically develop greater sensitivity to differences in prompt quality or greater ability to recognize the value of more structured prompting strategies. Examining the frequency of use as a moderating factor, therefore, helps clarify whether the benefits of prompt enhancement are contingent on prior familiarity or remain robust across different levels of user experience.

From these premises, we articulate our research questions:

RQ1. What prompt enhancement techniques reported in the literature improve output usefulness in business settings, and how do they contribute to general business enquiries?

RQ2. In the context of business-related tasks, to what extent do different prompting techniques influence perceived usefulness, and is this relationship moderated by users' frequency of Generative AI usage?

This paper is structured as follows: Section 2 presents the theoretical framework, including key concepts related to Generative AI, Large Language Models, and prompt engineering in a business setting. Section 3 describes the research design and methodology. Sections 4 and 5 showcase the findings from the systematic literature review and the experimental evaluation. The discussion of these results is in Section 6. Sections 7 and 8 explore their managerial and theoretical implications. Finally, Section 9 addresses study limitations and future research directions, while Section 10 provides the conclusions.

2. Related Work

2.1. Generative AI and Large Language Models (LLMs)

Artificial Intelligence (AI) encompasses computational systems capable of performing tasks associated with human intelligence. Within AI, Machine Learning (ML) enables models to learn from data rather than relying on predefined rules [16]. While traditional ML approaches are predominantly discriminative and focused on prediction tasks, recent advances in deep learning have enabled deep generative models (DGMs), which learn the underlying probability distribution of data and generate novel outputs [17].

Generative Artificial Intelligence (GenAI) refers to systems capable of producing new content such as text, images, or code. Large Language Models (LLMs) represent a prominent class of generative models based on transformer architectures. These models convert textual input into dense vector representations through tokenization and embeddings, capturing semantic relationships in continuous vector spaces [18,19]. Self-attention mechanisms allow scalable modeling of contextual dependencies [18].

LLMs are typically trained using self-supervised objectives such as next-token prediction, enabling them to generate contextually coherent text conditioned on input prompts [19]. Despite their strong generative capabilities, their internal representations remain largely opaque, leading to black-box characterizations [20]. Because task specification occurs entirely through natural-language input, prompts constitute the primary mechanism through which users control and adapt LLM behavior.

2.2. GenAI in Business Contexts

The diffusion of Generative Artificial Intelligence (GenAI) has accelerated its adoption in organizational environments, particularly for knowledge-intensive and task-oriented activities. Unlike traditional automation systems, GenAI supports content generation, reasoning, and ideation, thereby augmenting human cognitive work [21].

From an information systems perspective, GenAI should be conceptualized as a socio-technical system embedded within organizational workflows rather than as an isolated analytical tool. Feuerriegel et al. [22] distinguish model, system, and application-level

perspectives, arguing that business value emerges from the interaction among technical capabilities, system design, and contextual implementation.

Existing research predominantly examines domain-specific applications—such as marketing, healthcare, education, and software development—often focusing on narrowly defined tasks [21]. While these studies provide localized insights, they offer a limited understanding of cross-functional usage patterns and generalizable value mechanisms. The fragmentation between model-centric and application-centric research further complicates systematic comparison across contexts [22].

A defining characteristic of GenAI deployment is AI–human collaboration. Rather than replacing decision-makers, GenAI systems typically augment human capabilities by supporting problem-solving, synthesis, and decision processes [21]. However, organizational adoption introduces risks including hallucinations, bias, privacy concerns, and governance challenges [21].

Given the interaction-driven nature of GenAI systems, value creation depends on how effectively users articulate task requirements. This makes prompting a central mechanism in the use of organizational AI.

2.3. Prompt Engineering

Prompt engineering refers to the structured design and iterative refinement of prompts used to guide LLM outputs. Rather than modifying model parameters, prompting enables task adaptation of pre-trained models through carefully constructed instructions [23]. Conceptually, prompt engineering is an iterative cycle of formulation, evaluation, and refinement [24]. It represents an interaction design process rather than purely a technical optimization procedure.

A key advantage of prompting is task flexibility. A single base model can perform diverse tasks—such as summarization, reasoning, question answering, or code generation—through prompt variation alone [23]. However, LLMs are highly sensitive to prompt wording and structure. Small modifications can lead to substantial performance differences, and evaluations based on single prompt templates may be misleading [25]. Prompt effectiveness depends on contextual alignment between task characteristics, prompt structure, and model architecture [26]. Moreover, structured and context-rich prompts are associated with higher perceived usefulness in real-world applications [27].

Despite growing research interest, prompt engineering lacks standardized terminology and unified conceptual frameworks [23,24]. Nevertheless, recurring structural patterns have emerged across studies.

Prompt Enhancement Techniques

Prompt enhancement techniques refer to structured patterns used to systematically guide LLM outputs [28]. These techniques operationalize user intent through explicit instructions, contextual information, and output constraints.

Effective prompts typically prioritize clarity, explicit task specification, and balanced contextualization. Overly vague instructions or excessive information can reduce output quality. Furthermore, minor structural changes may substantially affect results due to the probabilistic nature of LLM generation [28].

Although prompting strategies often develop heuristically in practice, the literature increasingly attempts to formalize and categorize these techniques to enable systematic comparison and evaluation across contexts.

3. Materials and Methods

This study uses a two-stage research approach. RQ1 is examined through a systematic literature review and a qualitative content analysis to identify and categorize prompting techniques in business settings. RQ2 is investigated through an experimental evaluation of the usefulness of different prompting techniques and their combinations on real-world business prompts.

3.1. Systematic Literature Review (SLR)

To address RQ1, the proposed research conducted a systematic review in accordance with the PRISMA Guidelines. The objective was to identify prompt enhancement techniques that have been reported to improve the usefulness and effectiveness of large language model outputs in business-related contexts. Given the heterogeneity of tasks and evaluation settings, findings were synthesized qualitatively. The outcome of the review is a taxonomy of prompting techniques, classified by their primary reported effects.

3.1.1. Search Strategy

Literature paper retrieval was conducted using Scopus and Web of Science. An initial exploratory search focused on business-related studies to identify the prompting techniques discussed in business contexts. Based on insights from this phase, a second search was conducted to further examine these techniques in empirical studies that evaluate their effects on output quality and effectiveness across different tasks and contexts. Keywords related to prompting techniques, large language models and generative artificial intelligence, and evaluation dimensions were combined using the Boolean operators "AND" and "OR" and applied to titles, abstracts, and keywords. Table 1 specifies the query utilized for the two searches.

Table 1. Databases' Search Strategy Queries.

Phase	Query
Exploratory	("prompt engineering" OR "prompt design" OR "prompt enhancement" OR "prompting technique*" OR "prompt strategy" OR "prompt optimisation") AND ("generative AI" OR "large language model*" OR "LLM" OR "GPT" OR "ChatGPT" OR "foundation model*" OR "text generation") AND ("marketing" OR "business application" OR "business")
Final	TITLE ("prompting techniques" OR "zero-shot" OR "one-shot" OR "few-shot" OR "chain-of-thought" OR "role prompt" OR "self-consistency" OR "chain-of-verification") AND TITLE-ABS-KEY ("large language model*" OR "LLM*" OR "generative AI") AND TITLE-ABS-KEY ("quality" OR "relevance" OR "usefulness" OR "effectiveness" OR "efficiency" OR "accuracy" OR "clarity" OR "reliability" OR "decision-making")

Note: * indicates a wildcard character used in the search strategy to capture multiple word variations and endings of a root term.

3.1.2. Inclusion/Exclusion Criteria

Based on the chosen protocol, the eligibility criteria were set before screening to ensure the relevance and quality of the included studies. The inclusion criteria were as follows:

- Peer-reviewed journal articles indexed in Scopus and/or Web of Science;
- Studies published in English;
- Studies in which prompting techniques for large language models represent the primary focus of the investigation;
- Studies that explicitly conceptualize prompting as a methodological element, by defining, analyzing, evaluating, or comparing prompting approaches (e.g., zero-shot, few-shot, chain-of-thought, and their combinations).

Exclusion criteria included:

- Non-peer-reviewed or non-academic publications;
- Studies in which prompting is only mentioned incidentally or used implicitly without methodological discussion;
- Studies focused exclusively on model architecture, training procedures or benchmark optimization without substantive analysis of prompting techniques.

3.1.3. Screening Method

The study selection process adhered to PRISMA guidelines [29]. Initially, 327 records were identified from various databases. During an initial filtering stage, 130 records that were clearly irrelevant to the review were removed. Specifically, studies that used the term "zero-shot" to describe model evaluation after training, rather than inference-time prompt design, were excluded.

After removing duplicates, titles and abstracts were screened according to the predefined criteria, leading to the exclusion of 115 additional studies. A full-text assessment led to the exclusion of five more papers. Studies were included if they explicitly isolated and evaluated the effects of prompting techniques on task performance. Studies proposing automated or adaptive prompting approaches were kept only if baseline prompting techniques were clearly identified and evaluated independently. After excluding 20 studies because of missing data or unreported key outcomes, 53 studies [9,12,30–80] remained eligible and were included in the systematic review and qualitative content analysis.

The PRISMA flow diagram detailing the selection process, adapted from Page MJ et al. [29], is presented in Figure 1.

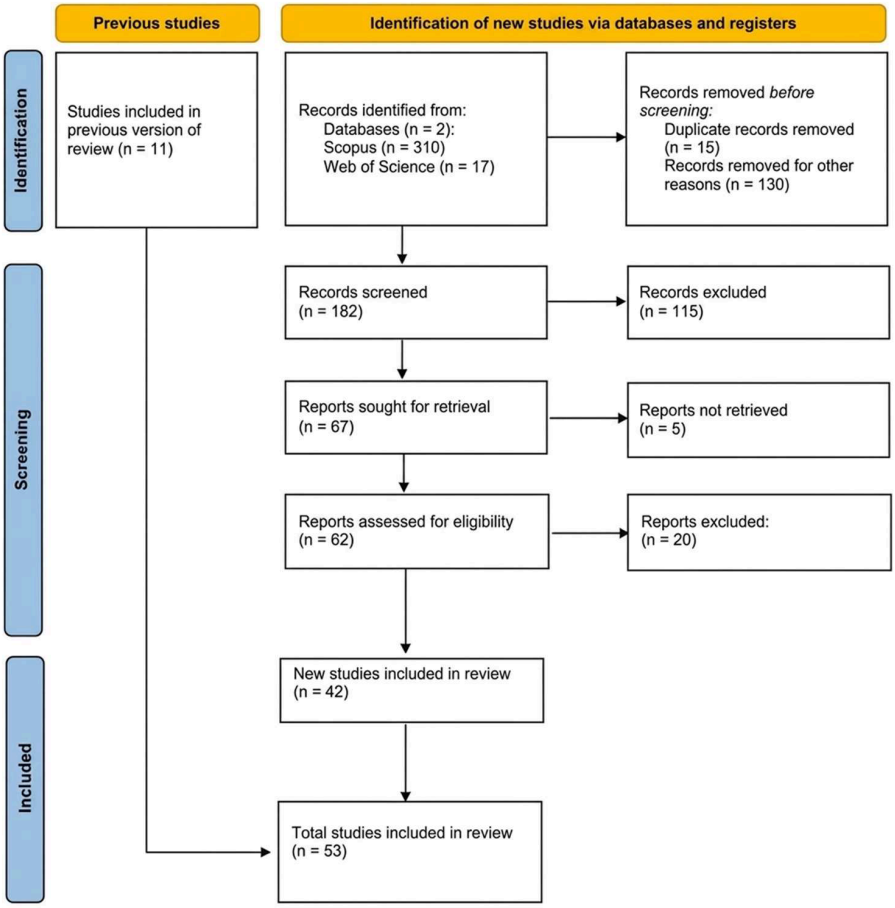

Figure 1. PRISMA 2020 flow diagram for updated systematic reviews which included searches of databases and registers only, adapted from Page MJ et al. [29].

3.1.4. Data Extraction

Data were systematically extracted using a structured table that included the prompting technique investigated, the application context, the evaluation approach, and the reported effects. When multiple techniques were analyzed within a study, each was recorded separately. Quantitative metrics were collected when available; however, no statistical aggregation was performed due to heterogeneity across tasks and evaluation settings. The complete data extraction table is provided in Appendix B.

3.1.5. Quality Assessment

An inductive qualitative content analysis was performed following Mayring's framework; see the main steps in Figure 2, [81]. This approach enables the systematic synthesis of heterogeneous empirical evidence and supports category development grounded in the reviewed material [82]. Coding focused on text segments explicitly addressing prompting strategies, their application context, and reported effects. Categories were developed iteratively to capture recurring functional patterns across tasks and domains, and refined in a pilot phase to ensure coherence and alignment with the research questions.

The unit of analysis was the prompting technique rather than the individual study, allowing multiple observations per article. In the final stage, categories were grouped into higher-order outcome dimensions forming the proposed taxonomy. The completed PRISMA 2023 Checklist is provided in the Supplementary Materials.

3.2. Experimental Design

This study uses a quantitative experimental design to examine how different prompting enhancement techniques affect perceived usefulness in common business tasks. The experiment included human evaluation to see how changes in prompting setups influence users' perceptions. The independent variable (IV) was defined by systematically changing the prompting techniques applied to the same business tasks. The dependent variable (DV) is Perceived Usefulness, measured at two related but separate levels: (1) Perceived Output Usefulness and (2) Perceived Prompt Quality. Additionally, the Frequency of GenAI Use was included as a moderating variable to account for differences in familiarity and exposure to AI systems.

A mixed-subjects design was implemented: participants evaluated multiple AI-generated outputs but were not exposed to all prompting conditions. This approach reduces fatigue and potential anchoring effects while preserving internal validity and ecological realism. Human evaluation was conducted through a structured online survey administered via Qualtrics XM (Qualtrics International Inc., Provo, UT, USA; https://www.qualtrics.com; accessed on 15 January 2026). The survey included an introduction, demographic questions, instructions and two randomized evaluation blocks corresponding to the selected business tasks. Randomization was uniformly distributed to ensure balanced exposure across prompting techniques.

The reliability and validity of the measurement instruments were assessed in Section 4 using Cronbach's alpha.

3.2.1. Identification of Business Use Cases

To ensure ecological validity, preliminary data were collected through a Microsoft Forms questionnaire distributed within a global organization. Respondents reported business-related tasks commonly performed using Generative AI tools such as ChatGPT and Microsoft Copilot, along with the prompts typically used. This exploratory phase identified four recurring business use cases: review (text enhancement), analysis, synthesis,

and general information retrieval. Table 2 presents the identified Generative AI use cases in Business.

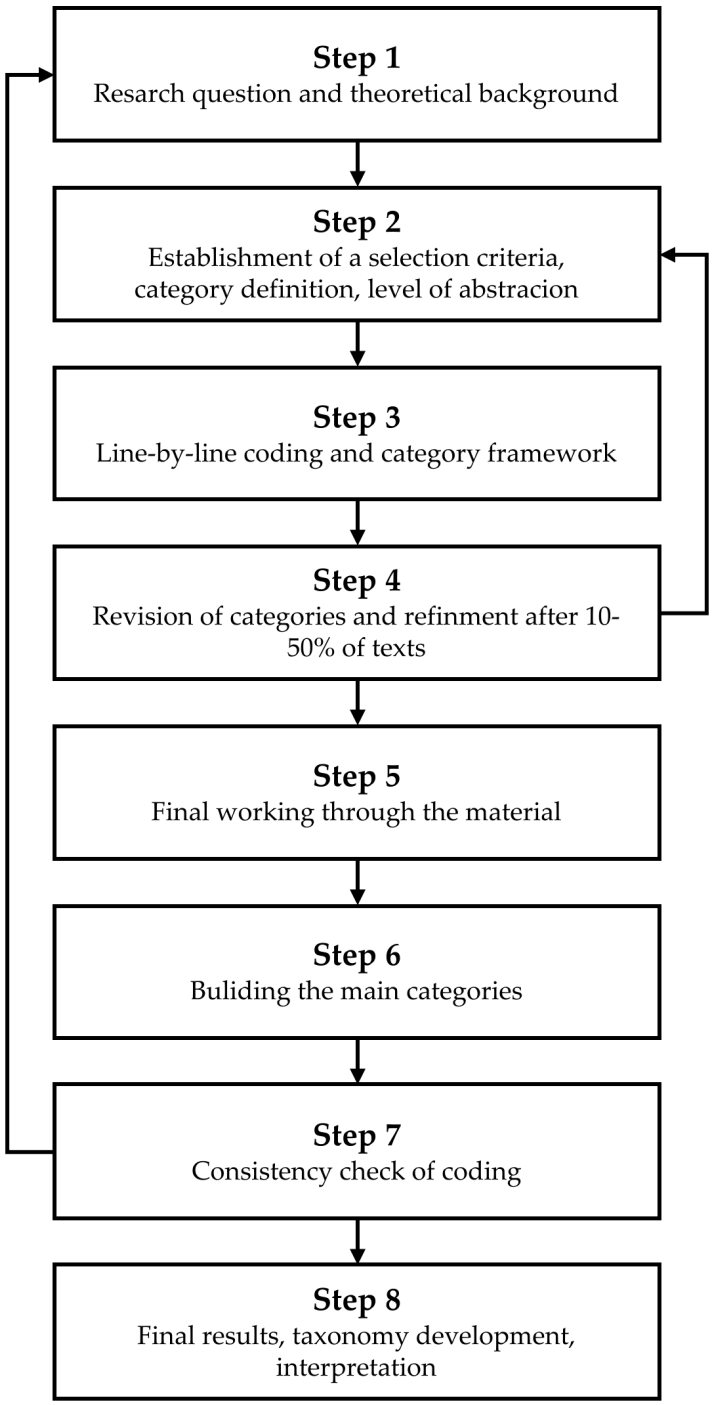

Figure 2. Steps of Inductive category development in Qualitative Content Analysis, adapted from [81].

Although all prompting techniques were initially applied to each use case, subsequent methodological refinement led to the selection of two tasks: analysis and synthesis. These tasks were retained due to their higher semantic variability and greater potential to produce differentiated outputs across prompting strategies.

The review task was excluded after preliminary testing revealed strong convergence of outputs across prompting techniques. Similarly, the general information retrieval task was excluded because it primarily involves factual recall and does not meaningfully benefit from

advanced prompting strategies. These exclusions reflect methodological considerations regarding the sensitivity of prompting effects across task types.

Table 2. Collection of GenAI Business Use Cases.

Business Use Case	Prompt (Raw)
Review/Text Enhancement	Can you please correct this email that I need to send to my boss? It needs to have a formal and professional tone of voice.
Analysis	Can you please tell me more about these financial KPIs?
Synthesis	Can you please summarize this market research article for me? I'd like to save time... give me the main context and outcomes.
General Information Retrieval	Please tell me what time is in Sao Paulo now.

3.2.2. Selection of Prompting Techniques

Prompting techniques were selected based on their frequency and relevance in the literature review on prompt engineering and Generative AI applications. To ensure methodological feasibility and avoid excessive cognitive load for participants, a subset of techniques was chosen. The study investigates:

- Zero-shot prompting;
- One-shot prompting;
- Few-shot prompting;
- Chain-of-thought prompting;
- One-shot prompting combined with Chain-of-Thought;
- Few-shot prompting combined with Chain-of-Thought.

Limiting the number of techniques was necessary to maintain survey quality and response reliability.

3.2.3. Prompt Construction and Output Generation

Raw prompts were derived from selected real-world tasks collected during the exploratory phase (Table 2), as described in Section 3.2.1. The analysis and synthesis prompts were then systematically enhanced by applying each prompting technique separately. To ensure comparability across outputs, each enhanced prompt was executed independently using Microsoft Copilot (free web version, accessed November–December 2025). Every prompt was submitted in a separate conversation to prevent contextual bias arising from conversation history. This procedure ensured that output variations were attributable to prompting manipulations rather than contextual carryover effects.

In total, 12 AI-generated outputs were produced: six for the analysis task and six for the synthesis task, each corresponding to a different prompting condition.

The complete prompt–output configurations evaluated in the experimental survey are reported in Appendix A to support transparency and reproducibility.

3.2.4. Data Collection and Sampling

The study used a homogeneous convenience sampling method, a non-probabilistic approach often employed in exploratory research where probability sampling is not possible [83,84]. This method limits participation to a specific subgroup, which enhances internal consistency but reduces generalizability.

The target population consisted of professionals engaged in business-related roles. Participants were recruited primarily through LinkedIn posts and direct email outreach. LinkedIn was selected due to its professional orientation and its suitability for reaching

individuals familiar with digital tools and Generative AI applications. Email recruitment complemented this strategy, enhancing response rates. Participation was voluntary and anonymous, and no financial incentives were provided. Participants were informed of the study's academic purpose prior to participation.

Within the survey, participants evaluated four prompt–output pairs: two from the analysis task and two from the synthesis task. The pairs were randomly selected from the pool of six prompting conditions per task. Randomization ensured balanced exposure across participants and reduced systematic bias.

3.2.5. Measurements

To evaluate Perceived Usefulness, the questionnaire included a structured assessment of both the AI-generated output and the prompt formulation that guided the system. The measurement framework was designed to distinguish between evaluations of the AI's response and evaluations of the prompt structure itself.

The evaluation of Perceived Output Usefulness was based on a factor-based Likert-scale instrument adapted from prior research on assessing Large Language Model (LLM) applications [85]. Participants rated each output on a five-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree") across five dimensions: relevance, informativeness, correctness, clarity, and absence of hallucinated or unsupported information.

In parallel, participants assessed the Perceived Prompt Quality. Drawing on the literature on prompt engineering, which emphasizes the central role of prompt formulation in shaping the effectiveness and usefulness of AI outputs [27], Perceived Prompt Quality was measured using a five-point Likert scale adapted from Abeysinghe & Circi (2024) [85]. Participants evaluated whether the prompt was relevant to the task, clearly formulated, error-free, and sufficiently informative to guide the AI system toward an appropriate response. The hallucination-related item was excluded from prompt evaluation, as it is conceptually applicable only to output assessment.

The Frequency of Generative AI Usage was measured to capture participants' level of familiarity and prior exposure to Generative AI tools. This variable was tested as a moderating factor in the relationship between prompting techniques and perceived prompt quality to examine whether evaluations differ by users' experience levels.

All constructs were measured using five-point Likert scales to ensure consistency across items. The reliability of the measurement instruments were subsequently assessed in Section 4 using Cronbach's alpha.

3.2.6. Conceptual Model

The conceptual framework builds on the findings of the systematic literature review, which identified recurring functional effects associated with prompt enhancement strategies. These effects primarily concern improvements in relevance, clarity, reasoning structure, and decision support. The empirical study examines whether such differences translate into variations in perceived usefulness within business-related tasks.

Perceived Usefulness is operationalized at two related levels: *Perceived Output Usefulness* and *Perceived Prompt Quality*. The former level reflects participants' evaluation of the AI-generated response, whereas the latter captures the perceived quality of the prompt formulation in guiding the AI system. This distinction allows the analysis to separate evaluations of outcomes from evaluations of prompt design.

In addition, the framework incorporates *Frequency of Generative AI Usage* as a moderating factor. Differences in prior exposure to AI tools may influence how users recognize and evaluate the contribution of structured prompting techniques. The moderating effect, therefore, tests whether familiarity with Generative AI conditions the relationship between

prompting techniques and perceived usefulness, such that differences in usage frequency may affect users' recognition and evaluation of the value of different prompting strategies.

Based on this framework, the following hypotheses are tested:

H1. *Prompting techniques significantly influence perceived usefulness.*

H2. *The Frequency of Generative AI Usage moderates the relationship between Prompting Techniques and Perceived Usefulness.*

Figure 3 presents the conceptual model guiding the empirical analysis.

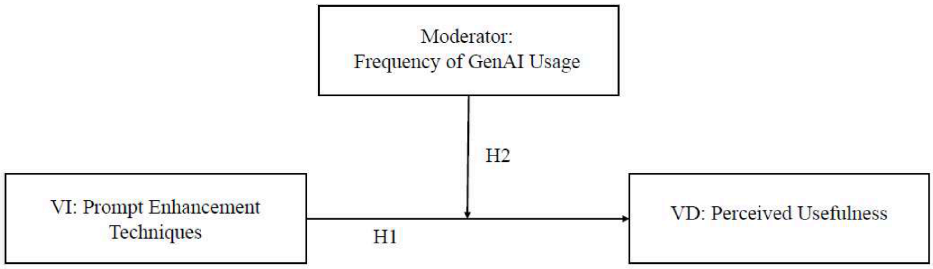

Figure 3. Conceptual Framework.

4. Results of Systematic Literature Review

Following the Structured Content Analysis, prompting techniques were organized into an outcome-oriented taxonomy reflecting their primary reported effects. Three macro-categories were identified: "Baseline Prompting Techniques", "Task Alignment Prompting Techniques" and "Output Transparency Prompting Techniques".

The taxonomy adopts a functional perspective, categorizing prompting techniques by their primary reported contribution to the resulting outputs, rather than assuming mutually exclusive effects. "Baseline Prompting Techniques" primarily support rapid output generation with minimal prompt specification and are commonly cited in the literature as reference points. "Task Alignment Prompting Techniques" focus on improving the relevance, structure and contextual appropriateness of outputs by better aligning model responses with task requirements and user expectations. "Output Transparency Prompting Techniques" aim to enhance the interpretability, robustness and verifiability of AI-generated outputs by making reasoning processes more explicit or by validating and stabilizing generated responses.

Each prompting technique is positioned within the taxonomy according to its primary reported effect. While several techniques contribute to multiple outcomes, they are classified acknowledging trade-offs among efficiency, interpretability, output quality and reliability. These trade-offs are examined in greater detail in the following sections. The taxonomy offers a comprehensive set of prompting techniques that professionals can apply depending on the desired outcome, organized into a two-level hierarchical structure comprising three categories and seven techniques, as illustrated in Figure 4.

4.1. Baseline Prompting Techniques

Baseline prompting methods are known for minimal prompt details and limited interaction, often serving as reference points in research. Instead of focusing on enhancing output quality, reasoning, or robustness, these methods emphasize speed and ease of use. Consequently, they are typically used for basic information retrieval for business queries or tasks where quick answers are needed, while designing prompts is kept inexpensive.

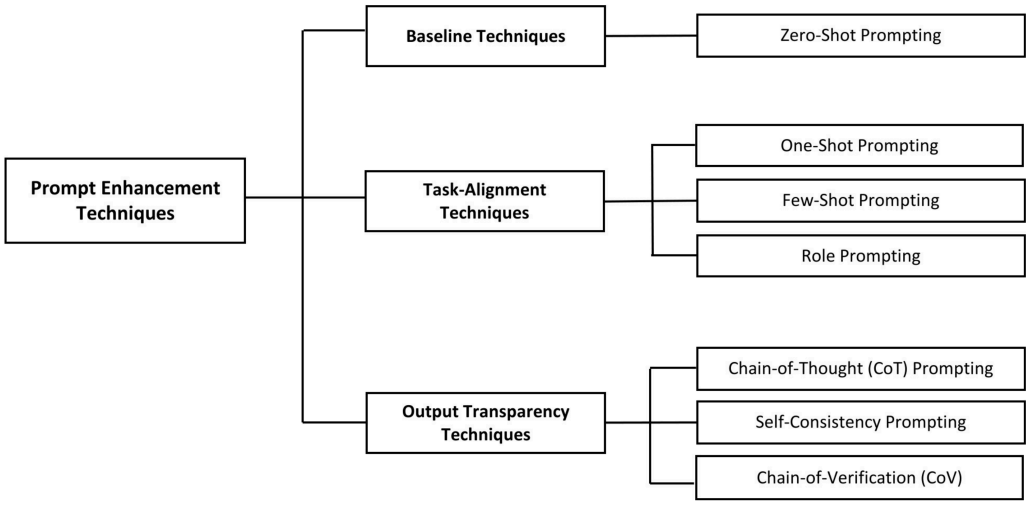

Figure 4. Taxonomy of Prompt Enhancement Techniques.

Zero-Shot Prompting

Across the studies reviewed, zero-shot prompting is primarily treated as a baseline or comparison point rather than as an optimized prompting technique. Several papers explicitly describe zero-shot settings as a minimum benchmark for evaluating more advanced methods, such as few-shot or chain-of-thought prompting. In terms of outcomes, zero-shot prompting often yields incomplete, generic or poorly reasoned responses, particularly in tasks involving prioritization, structured reasoning, or domain-specific judgment. For example, in construction cost estimation, zero-shot prompts generated incomplete answers with a low confidence score of 1.94 (64%), indicating a lack of focus and reduced reliability [30]. Similar issues are observed in peer-review assessments, where outputs are described as tentative and sentiment-based, with partial justifications and lower accuracy and recall than those from more structured prompts [31]. Several classification studies also report that zero-shot prompting yields variable performance, often with substantial variability across categories and models. In cybersecurity text classification of hacker forum posts, zero-shot prompting enabled basic categorization but yielded inconsistent results across categories and models, underscoring its role as a baseline rather than a robust method [32]. Similar findings are noted in traffic crash severity analysis, where zero-shot prompting achieved acceptable baseline accuracy but struggled with class imbalance and complex inference tasks, especially in fatal crash detection [33]. In medical and clinical contexts, zero-shot prompting is often linked to limited reasoning depth and reduced interpretability. For early sepsis diagnosis from clinical data, zero-shot prompting achieved acceptable recall but lower accuracy and F1 scores, with limited diagnostic reasoning compared to reasoning-based prompts [34]. In clinical question answering in nephrology, zero-shot prompts delivered concise but shallow responses, lacking explicit differential reasoning and justification, thereby limiting their usefulness for complex clinical decisions [35].

Many studies emphasize that zero-shot prompting depends heavily on the model and task. For instance, in multi-task harmful content detection, zero-shot enabled handling multiple tasks simultaneously but showed inconsistent performance across nuanced categories, heavily influenced by the model and output quality [36]. In multilingual scenarios, performance was sensitive to language and domain features, often serving only as a baseline comparison [37].

However, some studies report strong or acceptable zero-shot performance on narrowly defined or well-instructed tasks. For example, in drug–drug interaction classification, zero-shot prompting achieved stable, competitive F1 scores without training data, especially when prompts included clear domain instructions [38]. Similarly, in large-scale information

extraction and clustering, instruction-only zero-shot prompts effectively induced structure and improved clustering performance relative to traditional baselines [39]. Literature suggests that zero-shot prompting is rarely used as the main solution for complex or critical applications. Instead, it typically serves as an exploratory or baseline method, yielding quick yet shallow results that expose the limitations of minimal instructions and motivate the development of more structured prompting strategies.

4.2. Task Alignment Prompting Techniques

Task alignment prompting techniques enhance the relevance, structure, and context-appropriateness of outputs by aligning the model's behavior with specific task needs and user expectations. By using example-based conditioning, explicit instructions, or contextual framing, these techniques minimize ambiguity and steer the model toward generating more consistent, well-formatted, and suitable responses for professional or business decision-making environments. Their main role is to influence what the model produces, rather than how it reasons or validates its answers.

4.2.1. One-Shot Prompting

One-shot prompting is a technique where the model is given a single example to demonstrate the task or output format. It is seen as an intermediate approach between zero-shot and few-shot prompting, offering some gains with a relatively simple prompt design. Studies show that one-shot prompting can improve task performance relative to zero-shot methods by reducing ambiguity and clarifying label boundaries. For instance, in cybersecurity text classification of dark web hacker forum posts, providing one example per class improves accuracy and stability by helping the model better understand class semantics [32]. Similarly, in medical ontology-based classification, one-shot prompting yields higher accuracy than zero-shot prompting, as even minimal context helps the model align with task requirements [40]. However, its effectiveness is limited in complex or highly specialized tasks. In multimodal medical diagnosis from retinal OCT images, one-shot prompting allows basic classification but only reaches baseline accuracy, especially for complex pathologies, indicating a single example can't capture all visual or condition-specific details [41]. In such cases, one-shot prompting mainly serves as a reference rather than a robust diagnostic method. Review articles emphasize that one-shot prompting mainly improves task alignment rather than reasoning or accuracy. In clinical decision support, providing one example helps constrain outputs and clarifies the task, but doesn't fully address limitations in reasoning depth or generalization [42]. This highlights that one-shot prompting is a lightweight enhancement over zero-shot, not a substitute for more detailed or example-rich approaches. Content analysis confirms that one-shot prompting gives modest but consistent improvements by reducing ambiguity and guiding task understanding. Its advantages are most apparent in well-defined classification tasks with representative examples. In complex, multimodal, or specialized domains, however, it remains insufficient and is mostly a baseline or intermediate strategy rather than the optimal prompting approach.

4.2.2. Few-Shot Prompting

Few-shot prompting involves supplying a small number of example cases within the prompt to steer the model's output and behavior. Across the reviewed studies, it consistently outperforms zero-shot prompting, particularly on tasks requiring domain adaptation, structured output, or fine-grained classification. Multiple empirical studies report notable performance improvements when using few-shot prompting for classification and information extraction. For instance, in knowledge graph construction for equipment operation and maintenance, few-shot prompting increased recall by 10%, indicating improved generaliza-

tion and task understanding [43]. Similarly, in peer-review evaluations, it boosted recall by 11.3% and F1 score by 10.7% compared to zero-shot approaches, though the explanations provided are often less comprehensive, covering only parts of the reviews [31].

Few-shot prompting proves highly effective in domain-specific and technical areas. It enhances code security and robustness compared to zero-shot methods, though outcomes are still influenced by the type of vulnerability and the chosen examples [12]. Similar gains are seen in harmful content detection, where few-shot prompting boosts classification accuracy and F1 scores in tasks like cyberbullying and sarcasm detection, though improvements vary greatly depending on the example selection [36]. Overall, these results underscore the importance of example quality as a key factor in performance. However, the evidence also shows significant variability and context dependence. In multilingual natural language understanding tasks, few-shot prompting yields mixed results: it can enhance performance compared to zero-shot prompting in certain situations, but improvements are inconsistent across different languages and tend to plateau quickly, especially in low-resource settings [37]. Similar challenges are observed in risk-of-bias evaluations for clinical trials, where using few-shot prompting with justification examples does not significantly outperform zero-shot prompting. This indicates that in-context examples alone may not be enough for tasks involving complex methodological reasoning [44]. In both multimodal and medical fields, few-shot prompting generally provides more reliable and stronger improvements, especially when examples are chosen carefully. For retinal disease classification with OCT images, using expert-selected reference images in few-shot prompting greatly enhances diagnostic accuracy across most conditions, with improvements of up to 64% in certain categories [41]. Similarly, in medical image classification tasks with descriptor-based prompts, selecting high-quality descriptors for few-shot prompts significantly boosts accuracy and consistency without needing to retrain the model [45]. These results show that few-shot prompting is most effective when examples highlight domain-relevant features rather than superficial patterns. Despite these benefits, multiple studies highlight diminishing returns and trade-offs with increasing example numbers. In cybersecurity text classification, shifting from two-shot to three-shot prompting does not reliably enhance performance and might cause confusion or bias due to excessive context [32]. Similarly, in traffic crash severity classification, few-shot prompting mainly boosts results for smaller models, while showing mixed effects across severity levels [33]. These findings imply that few-shot prompting does not scale linearly with the number of examples and could impair performance if overloaded with context.

In professional communication and evaluative tasks, few-shot prompting enhances output structure, relevance and perceived usefulness, but it does not fully solve issues related to reasoning transparency. In educational feedback, using a few example response–feedback pairs results in outputs that are seen as more useful and responsive than those written by humans [46]. Similarly, for automatic scoring of student explanations, few-shot prompting consistently boosts accuracy by shaping output structure and aligning model behavior with human scoring patterns [47]. Nonetheless, without explicit reasoning mechanisms, few-shot prompting alone remains limited in interpretability, especially in complex evaluative contexts [47].

The content analysis shows that few-shot prompting offers a strong balance between output quality and interaction effort. It consistently improves accuracy, relevance and task alignment relative to zero-shot prompting, particularly in domain-specific and multimodal tasks. However, its effectiveness heavily depends on example quality, task complexity, and context length, with improvements often plateauing quickly. Therefore, few-shot prompting is most suitable for situations in which high-quality, structured outputs are more important than efficiency and in which representative examples can be carefully selected.

4.2.3. Role Prompting

Role prompting assigns a specific professional identity or role to the language model to steer responses toward domain-relevant knowledge and perspectives [48]. By framing the model as a particular type of expert, this technique influences the tone, formality and depth of the generated output, often improving relevance and coherence [49]. While role prompting can enhance domain-specific reasoning and support decision-making tasks, it also introduces potential limitations. Inconsistent role adherence, superficial or performative displays of expertise and the introduction of role-induced biases or stereotypes may affect the objectivity of the output [48]. For this reason, careful validation remains necessary to ensure analytical reliability. Empirical findings further suggest that the effectiveness of role prompting is task-dependent. In business communication writing, role-based prompts are associated with positive but not statistically significant improvements in output quality, suggesting that their impact may vary with task complexity and evaluation criteria [9]. Conversely, in more evaluative settings such as user preference and recommendation tasks, role prompting has been shown to significantly improve output, suggesting that its benefits extend beyond stylistic alignment when prioritization or judgment is required [50]. Hence, from a functional perspective, role prompting primarily contributes to task alignment by constraining outputs to a specific perspective or professional frame, thereby shaping relevance and actionability rather than transparency of reasoning or output verification.

4.3. Output Transparency Prompting Techniques

Output transparency prompting techniques aim to improve the clarity, verifiability, and reliability of AI-generated outputs. Rather than just enhancing task alignment or superficial accuracy, these methods focus on making the reasoning process behind the output more explicit, consistent, or easier to verify. Increasing transparency helps users understand, validate, and trust the model's results, especially in complex, high-stakes or reasoning-heavy situations.

4.3.1. Chain-of-Thought (CoT) Prompting

Chain-of-Thought (CoT) prompting is a reasoning-focused technique that explicitly guides large language models through step-by-step reasoning before delivering a final answer. Evidence indicates that CoT significantly improves performance in tasks involving multi-step reasoning and logical consistency. In vulnerability detection, CoT-based prompting notably boosts both detection accuracy and interpretability, with a decline in F1 score observed when CoT instructions are removed [51]. Similarly, in automated program repair and code generation, CoT enhances correctness by enabling models to better understand control flow and constraints, resulting in higher pass rates and improved solutions [52,53]. In medical applications, CoT improves diagnostic accuracy, recall and interpretability in early sepsis detection, depression classification, and clinical question answering, often aligning model reasoning more closely with expert decisions [34,35,54].

However, some research indicates that traditional CoT may be less effective than simpler reasoning methods for fact-based or low-complexity questions, suggesting that deeper reasoning does not always yield higher accuracy [55]. Several studies highlight CoT's strength in long-form and complex generation tasks. In summarizing long documents, CoT improves factual accuracy, structural coherence, and content coverage by enforcing stepwise reasoning prior to summarization [56]. In instruction-following and length-controlled text generation, CoT reduces constraint violations and improves adherence to complex requirements, resulting in higher accuracy and greater control [57].

Despite these benefits, there are limitations and trade-offs. In cost estimation and peer review, CoT enhances output reliability and interpretability but also increases interac-

tion complexity and response times, lowering efficiency [30,31]. In therapeutic dialogues, it encourages more reflective responses but can reduce task effectiveness and protocol compliance, highlighting a trade-off between explainability and actionability [58]. Furthermore, CoT's benefits are not universal. In graph construction from short texts, CoT does not improve performance, suggesting its advantages depend on input length and task structure [43]. In multimodal and vision-language tasks, CoT improves reasoning transparency but may also lead to hallucinations or false positives, especially in safety-critical situations [59]. The effectiveness of CoT also depends on how reasoning is integrated. Comparing explicit CoT prompting with implicit learning or fine-tuning shows that embedding CoT during training can produce better reasoning coherence and interpretability than inference-only prompting [60,86]. Additionally, manually crafting CoT prompts can be costly and difficult to scale, thereby limiting their practical use [61].

The analysis suggests that CoT enhances the clarity and interpretability of reasoning and multi-step problem-solving. However, its effectiveness depends on the context and entails trade-offs among efficiency, scalability and reliability. Therefore, CoT is most effective when used selectively for tasks requiring explicit reasoning and explainability, rather than as a universal approach.

4.3.2. Self-Consistency Prompting

Self-consistency prompting is a technique in which the model produces multiple independent solutions to the same task and combines them to generate a result. By sampling different reasoning paths and selecting the most consistent answer, this method aims to reduce variability and enhance robustness, particularly in reasoning-intensive tasks. The available empirical evidence supports this view but also reveals notable limitations. In medical question answering, self-consistency does not consistently improve accuracy compared to a single prompt. Instead, it shows moderate effects, with the smallest gains for logic-based factual questions and larger improvements for causal judgment tasks [55]. This indicates that self-consistency primarily stabilizes reasoning rather than consistently increasing correctness across all task types. These findings highlight a clear trade-off between robustness and efficiency. While multiple reasoning paths can reduce response variability and increase reliability in certain scenarios, they require multiple generations per task, thereby increasing computational and token costs [48]. Conversely, strong off-the-shelf models using simpler prompts can often achieve similar or better accuracy at lower costs, suggesting that model capability can sometimes outweigh the marginal gains from self-consistency [48,55].

It is a robustness-focused technique with benefits that depend on the task, particularly in ambiguous reasoning or causal inference situations. Therefore, self-consistency prompting is most suitable for applications where stability and reliability are more important than scalability and efficiency. Within the proposed taxonomy, self-consistency contributes to output transparency by stabilizing reasoning outcomes rather than by increasing interpretability or task alignment.

4.3.3. Chain-of-Verification (CoV) Prompting

Chain-of-Verification (CoV) prompting is designed to mitigate hallucinations by introducing an explicit verification stage into the generation process. The model first produces an initial response, which is then evaluated through a set of targeted verification questions. These questions are addressed independently to reduce confirmation bias, after which the model synthesizes a final, verified output. By explicitly separating generation and validation, CoV reduces ambiguity and increases confidence in the resulting responses [62]. In business idea generation and advisory contexts, CoV is particularly valuable for supporting

accurate and informed decision-making, especially when validating intermediate reasoning steps is critical. By forcing the model to reassess its outputs through structured verification prompts, the technique mirrors an iterative self-audit or risk-assessment process in which each assumption is evaluated before proceeding [62].

The evidence suggests that Chain-of-Verification prompting enhances output reliability by reducing unchecked or weakly justified conclusions. However, this benefit comes with more interaction and processing steps than simpler prompting strategies. As a result, CoV is best suited to business tasks where correctness, traceability and validation outweigh efficiency considerations.

5. Results of Empirical Evaluation

The analysis is organized around testing the research hypotheses: initially, the direct effects of Prompting Techniques on Perceived Output Usefulness are evaluated (H1), followed by the examination of the moderating effect of Frequency of GenAI Usage on the relation between Prompting Techniques and Perceived Prompt Quality (H2).

5.1. Sample Characteristics

A total of 127 participants completed the survey. After removing incomplete responses, the final valid sample included 105 participants. Geographically, most respondents were from Italy (85.25%), followed by the United Kingdom (7.38%), North America (3.28%), and other regions (4.10%). Figure 5 presents the geographical distribution of the sample.

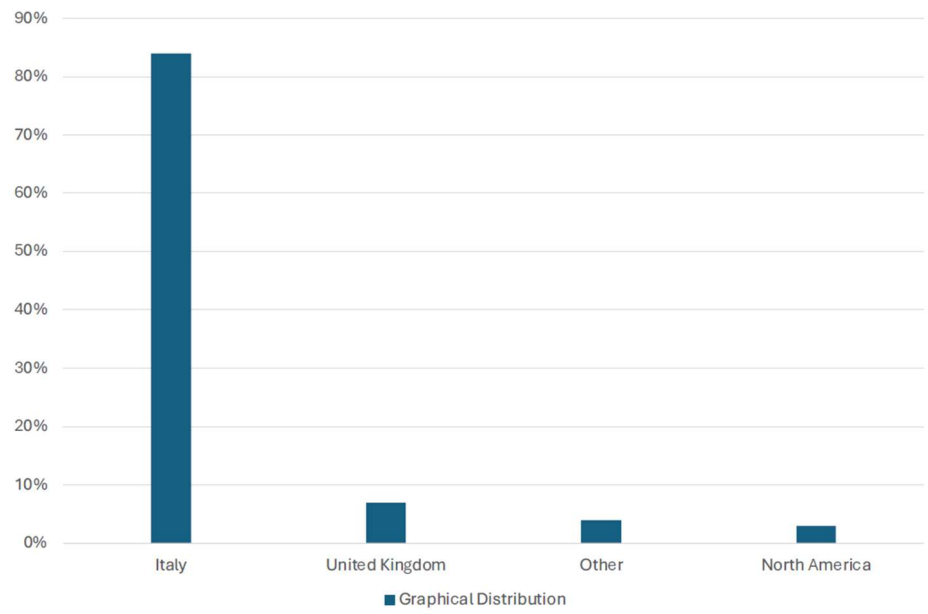

Figure 5. Geographical distribution of participants.

Participants had diverse professional backgrounds, primarily in business functions. The largest group worked in Marketing (28.69%), followed by Sales (12.30%), Staff roles (7.38%), and Advisory roles (4.92%), and nearly half specified "Other" (46.72%), indicating varied professional profiles. Figure 6 illustrates the professional backgrounds of respondents, indicating a heterogeneous sample composed primarily of business professionals.

Regarding AI usage, respondents reported different frequencies, with most using AI frequently, mainly 5–6 times per week or daily, showing significant exposure to Generative AI tools. Figure 7 reports the frequency of Generative AI usage among participants, highlighting a high level of exposure to GenAI tools within the sample.

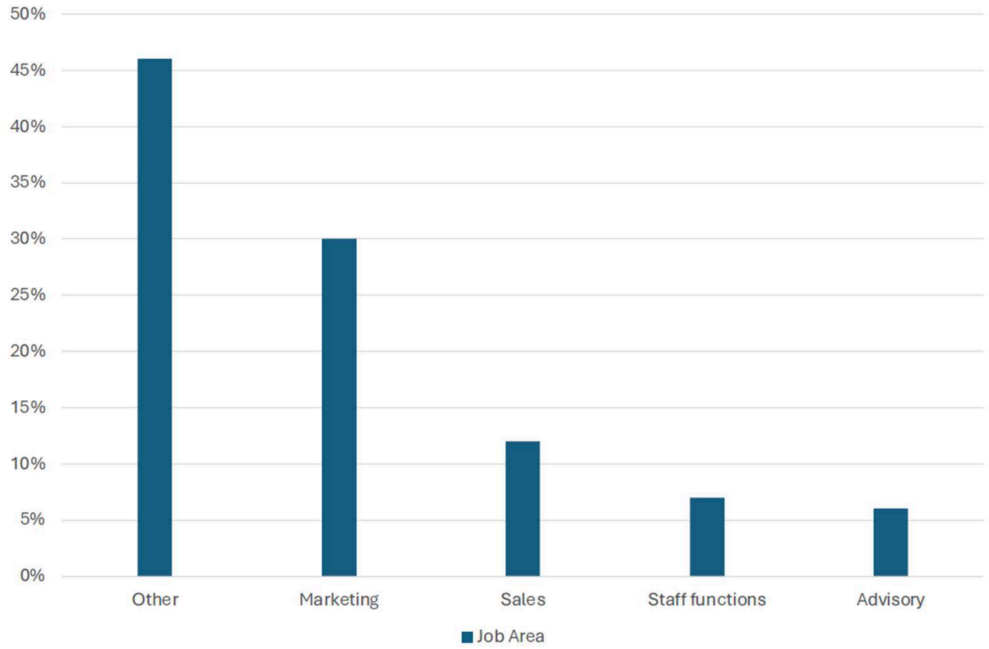

Figure 6. Professional background of participants.

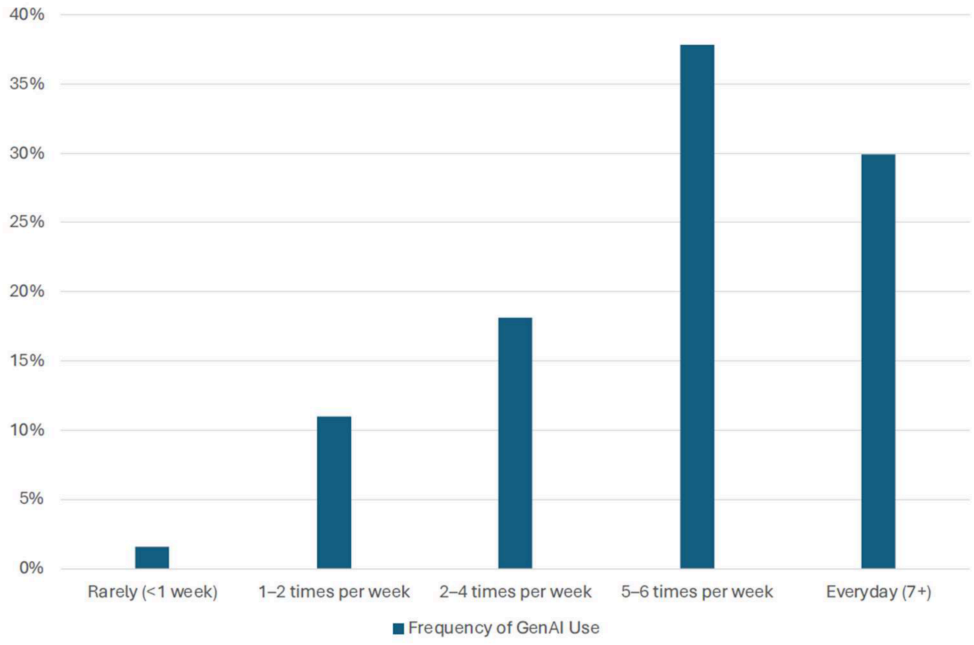

Figure 7. Frequency of Generative AI Use.

5.2. Measurements Reliability and Data Preparation

Before hypothesis testing, the survey data were prepared and structured to support empirical analyses. The questionnaire included multiple Likert-scale items designed to capture participants' evaluations of AI-generated outputs and the prompts used to generate them [85]. All evaluative items were measured on five-point Likert scales from 1 (strongly disagree) to 5 (strongly agree).

Perceived Output Usefulness was assessed with five items related to relevance, informativeness, correctness, clarity and hallucination absence. *Perceived Prompt Quality* was measured with four items on relevance, informativeness, correctness, and clarity. Internal consistency was evaluated with Cronbach's alpha, showing excellent reliability

(α = 0.947 for output usefulness and α = 0.994 for prompt quality), indicating high internal consistency. The details of the scales are reported in Appendix C.

Because each participant evaluated multiple prompts and outputs, the dataset was restructured from wide to long format for observation-level analysis. Each respondent evaluated four prompts and outputs, yielding 420 observations for both analyses of output usefulness and prompt quality. This structure captured within-participant variation and accounted for the non-independence of repeated evaluations. To link evaluations to experimental conditions, prompt–output pairs were matched to their respective prompting technique and task type using a mapping table. Prompting techniques were treated as categorical independent variables, including zero-shot, one-shot, few-shot, chain-of-thought (CoT), and their combinations. The *Frequency of GenAI Use*, measured via a self-report item about how often participants used GenAI tools, served as a continuous moderator.

Given the hierarchical data structure, multiple evaluations nested within respondents, linear mixed-effects models (LMEMs) were used. Random intercepts for respondents were included to account for individual-level heterogeneity arising from repeated evaluations and the fact that each participant assessed only a subset of the available prompt–output pairs. This approach provided robust estimates of fixed effects while managing repeated measures.

Two separate analytical models were estimated to test the two research hypotheses. In the first analysis, perceived output usefulness was modelled as a function of prompting techniques and task type to test the main effect of prompting techniques (H1). Post hoc comparisons with estimated marginal means examined differences between techniques. In the second analysis, Perceived Prompt Quality was modelled as a function of prompting techniques and the Frequency of GenAI Usage. An interaction between prompting techniques and usage frequency was tested to determine whether Frequency of GenAI Usage moderated the relationship between prompting techniques and Perceived Prompt Quality (H2). Likelihood-ratio tests compared models with and without the interaction to assess improvements in fit.

5.3. Descriptive Statistics of Key Variables

This section presents descriptive statistics for the key variables included in the empirical analyses. All evaluative measures were collected using a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Descriptive statistics are reported at the observation level to reflect the repeated-evaluation structure of the data, whereby each participant provided multiple ratings across different prompt–output pairs.

Overall, *Perceived Output Usefulness* was relatively high across the full set of evaluations (N = 420), with a mean score of 4.18 (SD = 0.67), indicating generally positive perceptions of AI-generated outputs. Observed usefulness ratings ranged from 2 to 5 on the Likert scale, suggesting sufficient variability in responses to support subsequent inferential analyses.

When Perceived Output Usefulness was examined across prompting techniques, systematic differences emerged between conditions. Zero-shot prompting yielded the lowest average usefulness rating (M = 3.26, SD = 0.55), whereas prompting techniques that incorporated additional structure or guidance were associated with higher ratings. Combined techniques achieved the highest mean usefulness scores: few-shot with chain-of-thought (M = 4.79, SD = 0.52) and one-shot with chain-of-thought (M = 4.82, SD = 0.40).

Intermediate techniques, including chain-of-thought alone (M = 4.11, SD = 0.33), few-shot prompting (M = 4.09, SD = 0.25), and one-shot prompting (M = 4.01, SD = 0.39), showed comparable average levels of perceived usefulness.

A similar descriptive pattern was observed for *Perceived Prompt Quality* evaluations. Prompts that combined example-based prompting with explicit reasoning guidance re-

ceived higher quality ratings, whereas zero-shot prompts were consistently evaluated as lower in quality. Taken together, these descriptive results suggest that the structure and composition of prompting techniques are associated with meaningful differences in user evaluations, a pattern that is formally tested in the following hypothesis-testing sections.

5.4. Hypothesis Testing

To evaluate the proposed hypotheses, linear mixed-effects models were used with observation-level data. This approach was chosen to address the dataset's hierarchical structure, where multiple evaluations are nested within respondents. All models included random intercepts for respondents to account for individual differences in baseline rating tendencies. Two analyses were performed: first, the effect of prompting techniques on Perceived Output Usefulness was examined (H1); second, the moderating role of Frequency of GenAI Usage on the relationship between Prompting Techniques and Perceived Prompt Quality was tested using an interaction model (H2).

5.4.1. Effects of Prompting Techniques on Perceived Output Usefulness (H1)

To test Hypothesis 1, a linear mixed-effects model was estimated to examine the effect of prompting techniques on perceived output usefulness. The model was specified as:

$$\text{Usefulness} \sim \text{Technique} + \text{Task Type} + (1 \mid \text{ResponseId})$$

Perceived output usefulness represents the dependent variable measured on a five-point Likert scale. Prompting technique is included as a categorical fixed effect, with zero-shot prompting as the reference category, allowing all estimated coefficients to be interpreted as deviations from the zero-shot baseline. Task type (analysis vs. synthesis) was included as a control variable to account for potential differences in task complexity. A random intercept for each respondent (ResponseId) was added to account for repeated evaluations nested within participants and to capture stable individual differences in rating tendencies.

The model intercept ($\beta = 3.28$, $p < 0.001$) represents the expected usefulness rating for zero-shot prompts in the reference task condition. The coefficients associated with the remaining prompting techniques indicate how much perceived usefulness differs from the zero-shot prompting condition. Positive coefficients indicate higher perceived usefulness compared with zero-shot prompting, whereas negative coefficients would indicate lower perceived usefulness. Figure 8 visualizes these estimated differences relative to the zero-shot reference level (0).

Results show a strong and statistically significant main effect of prompting technique ($p < 0.001$). Compared to zero-shot prompts, all enhanced prompting techniques significantly increased perceived output usefulness. Specifically, chain-of-thought prompting increased usefulness by 0.84 points ($\beta = 0.84$, $p < 0.001$), few-shot prompting by 0.82 points ($\beta = 0.82$, $p < 0.001$), and one-shot prompting by 0.74 points ($\beta = 0.74$, $p < 0.001$). Even larger effects emerged for combined techniques: few-shot with chain-of-thought increased usefulness by 1.52 points ($\beta = 1.52$, $p < 0.001$), while one-shot with chain-of-thought produced the strongest improvement, increasing usefulness by 1.55 points relative to zero-shot prompting ($\beta = 1.55$, $p < 0.001$). Finally, task type did not have a significant effect on perceived usefulness ($\beta = -0.03$, $p = 0.369$), indicating that the observed differences are primarily attributable to prompting techniques rather than to task characteristics (Figure 8).

5.4.2. Estimated Marginal Means

Estimated marginal means are used to compare perceived output usefulness across prompting techniques while accounting for the experimental design and task-level variabil-

ity. When averaged across task types, these estimates reveal a clear hierarchy of prompting methods by perceived usefulness (see Figure 8).

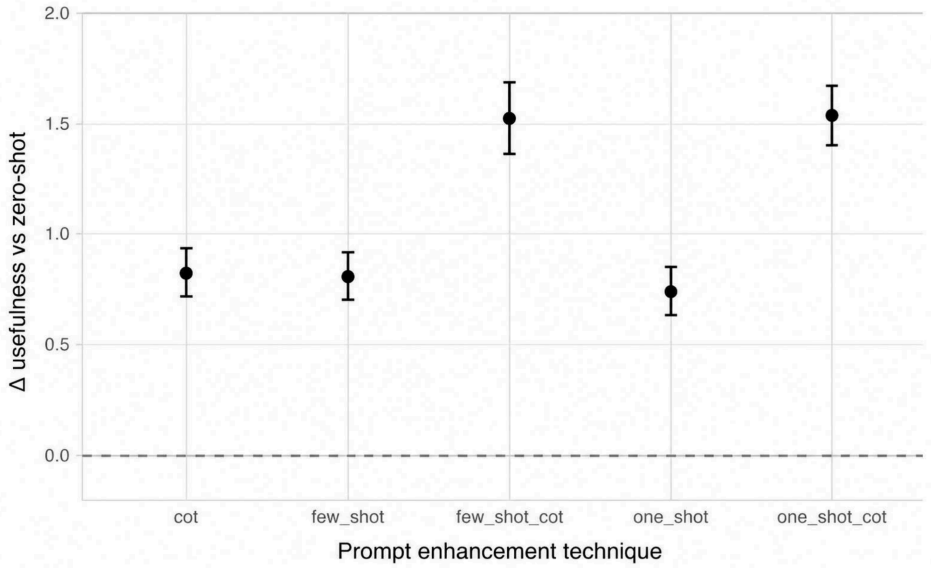

Figure 8. Estimated effects of prompting techniques relative to the zero-shot baseline (95% CI).

Zero-shot prompting yielded the lowest usefulness ratings (EMM = 3.27, 95% CI [3.17, 3.37]), establishing a clear baseline. All other techniques produced substantially higher evaluations. Techniques relying on minimal or moderate structures, chain-of-thought alone (EMM = 4.11), few-shot (EMM = 4.09), and one-shot (EMM = 4.01), clustered around a similar level of perceived usefulness. Their confidence intervals largely overlapped, indicating comparable performance. The highest usefulness scores were observed for combined techniques, which integrate example-based prompting with explicit reasoning instructions. Few-shot with chain-of-thought achieved an estimated mean of 4.79 (95% CI [4.69, 4.89]), while one-shot with chain-of-thought reached an even slightly higher mean of 4.82 (95% CI [4.72, 4.92]). These techniques consistently outperformed all other conditions.

Figure 9 displays the estimated marginal means and 95% confidence intervals for perceived output usefulness across different prompting techniques. The visualization emphasizes a clear stratification among these techniques. Zero-shot prompting sits at a distinctly lower position, with confidence intervals that do not overlap with any of the enhanced techniques. Conversely, chain-of-thought, few-shot, and one-shot prompting exhibit significant overlap in their confidence intervals, suggesting similar levels of perceived usefulness. The highest estimated means are observed for the combined techniques, whose confidence intervals are clearly separate from those of single-component approaches. Notably, the confidence intervals for few-shot with chain-of-thought and one-shot with chain-of-thought largely overlap, indicating no meaningful difference between the two once explicit reasoning is incorporated.

5.4.3. Pairwise Comparisons

Pairwise comparisons are conducted to identify which specific prompting techniques differ significantly in perceived output usefulness. Zero-shot prompting was rated significantly lower than every other technique (all $p < 0.001$), confirming that the absence of examples or reasoning guidance substantially reduces perceived output usefulness. Among non-zero-shot techniques, no significant differences emerged between chain-of-thought, few-shot, and one-shot prompting when used in isolation (all $p > 0.65$). This suggests that adding either examples or reasoning alone yields similar gains over zero-shot prompting

but does not meaningfully differentiate usefulness among these approaches. In contrast, both combined techniques, few-shot with chain-of-thought and one-shot with chain-of-thought, were rated significantly higher than all single-component techniques (all $p < 0.001$). Notably, no significant difference was found between the two combined techniques themselves ($p = 0.998$), indicating that once both examples and reasoning are present, the number of examples (one vs. few) does not further increase perceived usefulness.

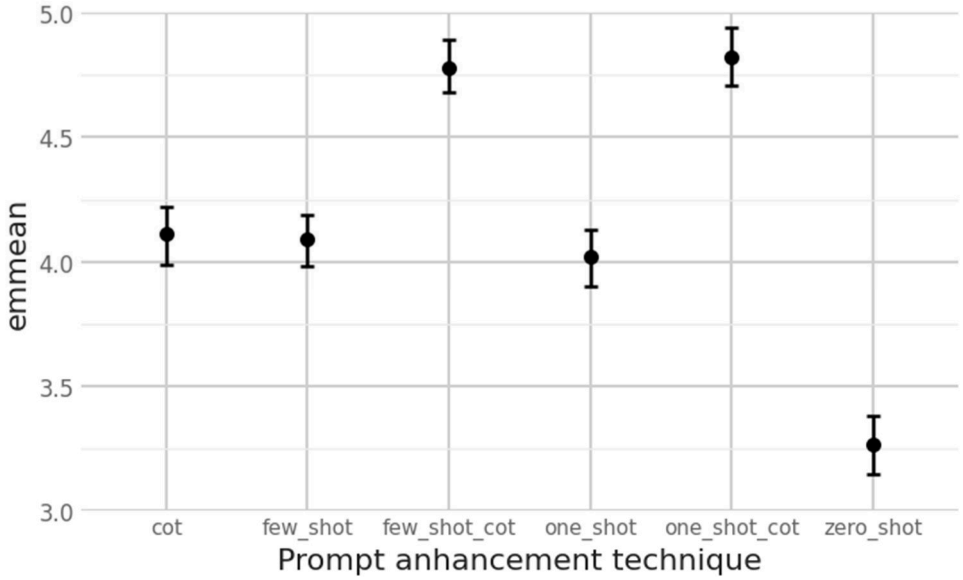

Figure 9. Estimated marginal means of perceived output usefulness by prompting technique.

Detailed pairwise comparison results and full model outputs for H1 are reported in Appendix C.1. These analyses collectively support the evaluation of H1, which examines differences in perceived usefulness across prompting techniques.

5.5. Moderating Role of Generative AI Usage (H2)

Unlike H1, which focuses on Perceived Output Usefulness, H2 examines whether users' Frequency of Generative AI Usage moderates the relationship between Prompting Techniques and Perceived Prompt Quality. As a first step, a baseline mixed-effects model including only main effects was estimated:

$$\text{Prompt quality} \sim \text{Technique} + \text{Frequency of use} + (1|\text{ResponseId})$$

In this specification, prompt quality is the dependent variable, measured on a five-point Likert scale. Prompting technique is included as a categorical fixed effect, with zero-shot prompting as the reference category, allowing coefficients to be interpreted as deviations from the zero-shot baseline. Frequency of use captures respondents' self-reported frequency of Generative AI usage and is modelled as a continuous covariate. A random intercept for each respondent accounts for repeated evaluations nested within individuals. Results from this baseline model indicate a strong and statistically significant main effect of prompting technique on perceived prompt quality (all $p < 0.001$). Relative to zero-shot prompting, all enhanced techniques were associated with substantially higher quality ratings, with the largest coefficients observed for combined techniques.

In contrast, the frequency of Generative AI use did not have a significant main effect on prompt quality ($\beta = -0.04$, $p = 0.169$), suggesting that, on average, more frequent users did not systematically rate prompts as higher or lower quality after controlling for the prompting technique. Results for perceived prompt quality were broadly consistent with

those observed for perceived output usefulness, with structured prompting techniques systematically receiving higher evaluations than zero-shot prompting. Given the substantial overlap in the observed patterns across outcomes, detailed graphical representations of prompt quality results are omitted for brevity and to improve readability.

Interaction Model: Testing Moderation by Usage Frequency

To formally assess whether the effect of the prompting technique varies as a function of Frequency of GenAI Usage, a second model including the interaction term (*) was estimated:

$$\text{Prompt quality} \sim \text{Technique} * \text{Frequency of use} + (1|\text{ResponseId})$$

In the model specification, an interaction term (*) is included between the prompting technique and Frequency of GenAI Usage. Following standard R model-formula notation, the asterisk indicates that the model estimates both the main effects of the two variables and their interaction (i.e., Technique + Frequency of GenAI Usage + Technique:Frequency of GenAI Usage).This specification allows the effect of a given prompting technique on prompt quality to vary across different levels of AI usage frequency. Model comparison between the baseline and interaction models, conducted via likelihood-ratio testing, did not reveal a significant improvement in model fit ($\chi^2(5) = 3.24$, $p = 0.662$). None of the interaction terms between the prompting technique and frequency of use was statistically significant (all $p > 0.33$). These results indicate that the relationship between prompting technique and perceived prompt quality is stable across different levels of Generative AI usage, providing no empirical support for a moderating role of usage frequency.

Complete model estimates and additional diagnostic statistics for the moderation analysis are reported in Appendix C.2.

5.6. Summary of Hypothesis Testing

The empirical analyses provide strong support for Hypothesis 1. The linear mixed-effects model revealed a statistically significant main effect of prompting technique on perceived output usefulness ($p < 0.001$). Relative to zero-shot prompting (intercept $\beta = 3.28$, $p < 0.001$), all enhanced techniques produced significant positive effects, including chain-of-thought ($\beta = 0.84$, $p < 0.001$), few-shot ($\beta = 0.82$, $p < 0.001$), and one-shot prompting ($\beta = 0.74$, $p < 0.001$). The largest effects were observed for combined techniques: few-shot with chain-of-thought ($\beta = 1.52$, $p < 0.001$) and one-shot with chain-of-thought ($\beta = 1.55$, $p < 0.001$).

Estimated marginal means further illustrate this hierarchy. Zero-shot prompting yielded the lowest usefulness ratings (EMM = 3.27, 95% CI [3.17, 3.37]), whereas combined techniques achieved the highest evaluations, with few-shot plus chain-of-thought (EMM = 4.79) and one-shot plus chain-of-thought (EMM = 4.82). Pairwise comparisons confirmed that both combined techniques significantly outperformed single-component approaches (all $p < 0.001$), while no significant differences emerged among chain-of-thought, few-shot, and one-shot prompting when used independently (all $p > 0.65$). These findings indicate that integrating example-based prompting with explicit reasoning instructions substantially enhances perceived output usefulness.

Hypothesis 2 was not supported. Although prompting technique significantly influenced perceived prompt quality (all $p < 0.001$), the interaction between prompting technique and GenAI usage frequency was not statistically significant ($\chi^2(5) = 3.24$, $p = 0.662$), and none of the interaction coefficients reached significance (all $p > 0.33$). Frequency of GenAI Usage also did not exert a significant main effect on perceived prompt quality ($\beta = -0.04$, $p = 0.169$). These results indicate that the effect of prompting techniques on perceived prompt quality remains stable across different levels of Generative AI usage.

The findings demonstrate that prompt structure plays a decisive role in shaping perceived usefulness, whereas user familiarity with Generative AI does not significantly condition these effects.

6. Discussion

This study combined a systematic literature review with an empirical investigation to examine how prompt enhancement techniques influence perceived usefulness in business-related tasks. The findings provide converging evidence that prompt structure plays a decisive role in shaping user evaluations of AI-generated outputs.

The results strongly support Hypothesis 1. Prompting techniques significantly influenced Perceived Output Usefulness ($p < 0.001$), with all enhanced strategies outperforming zero-shot prompting. Techniques that incorporated explicit reasoning (chain-of-thought) or example-based guidance (one-shot and few-shot) yielded substantial improvements relative to the zero-shot baseline (β ranging from 0.74 to 0.84, all $p < 0.001$). The most pronounced effects emerged when these approaches were combined. Few-shot with chain-of-thought ($\beta = 1.52$, $p < 0.001$) and one-shot with chain-of-thought ($\beta = 1.55$, $p < 0.001$) produced the highest increases in perceived output usefulness, with estimated marginal means approaching the upper bound of the Likert scale (EMM ≈ 4.8). These findings indicate that structured prompt design, particularly the integration of examples with explicit reasoning instructions, substantially enhances users' perceptions of output usefulness.

Importantly, no meaningful differences were observed among chain-of-thought, few-shot, and one-shot prompting when used independently (all $p > 0.65$), suggesting that isolated enhancements may yield comparable gains, whereas their combination generates a cumulative effect. This pattern aligns with the systematic review, which identified improvements in relevance, clarity, and reasoning coherence as recurrent functional outcomes of structured prompting techniques. The empirical evidence, therefore, reinforces the taxonomy developed in the review phase and confirms that these functional mechanisms translate into higher perceived usefulness in applied business contexts.

In contrast, Hypothesis 2 was not supported. Although prompting techniques significantly influenced Perceived Prompt Quality (all $p < 0.001$), the interaction between prompting technique and Frequency of Generative AI Usage was not statistically significant ($\chi^2(5) = 3.24$, $p = 0.662$), and frequency of use did not exert a significant main effect ($\beta = -0.04$, $p = 0.169$). These results indicate that the impact of prompting techniques on perceived prompt quality remains stable across different levels of GenAI usage frequency.

The absence of a moderating effect offers an important theoretical insight. While prior exposure to AI tools might be expected to enhance users' sensitivity to prompt design differences, the findings suggest that structured prompting strategies are perceived as beneficial regardless of usage frequency. This implies that the effectiveness of prompt enhancement techniques may not depend primarily on accumulated experience, but rather on the intrinsic clarity and guidance embedded in the prompt formulation itself. In other words, well-designed prompts appear to function as universally effective interaction mechanisms rather than expertise-dependent tools.

The alignment between the systematic review and the empirical findings strengthens the conclusion that prompt engineering constitutes a central mechanism in improving the perceived usefulness of Generative AI in business settings. At the same time, the absence of moderation underscores the need to move beyond usage frequency as a proxy for AI expertise and to develop more precise measures of prompt literacy and prompt-engineering competence in future research.

7. Managerial Contributions

The study offers practical guidance for organizations integrating Generative AI into business processes by combining insights from a structured taxonomy of prompt enhancement techniques with empirical evidence on user evaluations.

The taxonomy developed through the systematic review clarifies how different prompting strategies contribute to functional improvements such as task alignment, reasoning transparency, and output coherence. By organizing prompting techniques according to their primary effects rather than solely their structural format, the framework provides practitioners with a structured basis for selecting and combining strategies appropriate to specific business tasks.

The empirical findings reinforce this framework. The substantial gap between zero-shot prompting and structured combinations ($\beta > 1.5$ relative to baseline, $p < 0.001$) demonstrates that prompt design materially influences Perceived Output Usefulness. Combining example-based guidance with explicit reasoning instructions consistently produced the highest evaluations. This suggests that organizations can enhance AI-supported workflows by formalizing prompt templates that embed these elements. The absence of a moderating effect of AI usage frequency further indicates that effective prompting does not depend solely on accumulated experience. Increasing exposure to AI tools does not automatically improve the recognition or evaluation of well-designed prompts. As a result, structured prompt design training and organizational standards may be more impactful than relying on usage-based learning.

Together, the taxonomy and empirical findings position prompt engineering as an organizational capability that can be systematized, governed, and embedded into AI deployment strategies.

8. Theoretical Implications

This study contributes to the literature on Generative AI and human–AI interaction by integrating prompt engineering research with a business-oriented and user-centered perspective. Prior studies have shown that integrating AI into organizational processes reshapes work design and employee task execution [1], while empirical evidence demonstrates that Generative AI can enhance productivity, output quality, and learning outcomes in knowledge work [2]. At the same time, research on human–AI interaction highlights risks related to misinterpretation, overreliance, and workflow disruption when AI-generated outputs are not effectively managed [3]. Within this context, prompt engineering has been increasingly recognized as a key mechanism for improving the reliability and effectiveness of generative models [4,5].

Building on this stream of research, the systematic literature review conducted in this study consolidates fragmented evidence on prompt enhancement techniques into an outcome-oriented taxonomy. By classifying prompting strategies according to their primary functional contribution, the study responds to calls for greater conceptual clarity in prompt engineering research [23,24]. The taxonomy clarifies how different prompting strategies influence output relevance, reasoning coherence, and robustness across heterogeneous contexts, moving beyond isolated or purely technical descriptions of prompting methods.

The empirical findings extend this conceptual contribution by demonstrating that these functional distinctions translate into systematic differences in perceived usefulness in cross-domain business tasks. The results reveal a hierarchical pattern: zero-shot prompting yields significantly lower evaluations; single-component techniques such as one-shot, few-shot, and chain-of-thought prompting produce comparable intermediate improvements; and combinations that integrate example-based conditioning with explicit reasoning scaffolding generate the highest perceived usefulness. This pattern suggests that task-

alignment mechanisms and reasoning-transparency mechanisms operate as complementary dimensions, whose joint application produces cumulative effects in cognitively demanding business contexts.

Furthermore, this research extends the business-oriented Generative AI literature, which has largely focused on model capabilities, system architectures, or domain-specific applications [21,22], by shifting attention to how users formulate prompts in everyday work tasks. Rather than treating prompting as a purely technical adjustment, the study shows that the structure of instructions directly shapes how AI systems are perceived and evaluated in business contexts. In line with research emphasizing that the effectiveness of AI depends not only on technological capability but also on how employees interact with it [15], the findings suggest that prompt structure plays a central role in translating model potential into perceived usefulness.

The empirical results further refine theoretical assumptions regarding user experience and AI familiarity. Although prior research emphasizes the importance of employee capabilities and organizational conditions in realizing AI-driven performance improvements [15], the present findings indicate that the effectiveness of structured prompting techniques remains stable across different levels of Generative AI usage frequency. The absence of a moderating effect suggests that well-designed prompts can enhance perceived usefulness independently of accumulated usage experience. This finding nuances capability-based views of AI value creation and highlights the need for more precise conceptualizations of prompt-related competencies in future research.

9. Limitations and Future Research

While this study makes valuable contributions, it also has limitations that merit recognition and suggest directions for future research.

The reviewed studies exhibit substantial heterogeneity in terms of application domains, task types, evaluation metrics, and methodological designs. This diversity reflects the cross-domain nature of prompt engineering research but limits the direct comparability of findings across studies. As a result, the review adopts a qualitative content analysis rather than a quantitative meta-analytic approach. Future research could address this limitation by focusing on narrower task categories or by developing standardized evaluation frameworks that enable meta-analyses and more precise cross-study comparisons.

Furthermore, the literature on prompt engineering is characterised by inconsistent terminology and overlapping conceptual definitions [24], which complicates synthesis and categorization. Although this review addresses this issue by organizing techniques based on their primary reported effects, future research would benefit from greater conceptual standardization and shared taxonomies. Longitudinal reviews could also examine how prompting techniques evolve as models and user practices mature.

The empirical analysis relies on a small-scale study with limited resources. Although the sample size and experimental design are suitable for initial exploration, the results should be viewed cautiously when considering their generalizability. Future work could build on this by conducting larger studies with more diverse and representative samples, employing stronger probabilistic sampling methods, and exploring a wider variety of prompting techniques beyond those tested here. Increasing the number of tasks and expanding the range of prompting strategies would enable more thorough comparisons and detailed insights into how different methods perform in various business contexts.

The study also found that how often Generative AI is used does not significantly affect the results. This indicates that greater usage does not necessarily mean a better understanding of prompt enhancement techniques. It highlights a limitation of using frequency alone as a measure of AI skill. Future research should focus on more detailed, task-specific

assessments of user ability, such as Gibreel & Arpaci's Prompt Engineering Competence Scale (PECS) [87]. This five-point Likert-type scale ranges from "strongly disagree" (1) to "strongly agree" (5), with higher scores indicating greater prompt engineering competence. The scale includes 9 items, such as adaptability to different AI models, efficiency in prompt optimization, use of contextual constraints and diverse formats, handling AI limitations and biases, improving response relevance with examples and scenarios, and critically interpreting and refining AI responses [87].

Additionally, the study relies on self-reported evaluations of perceived usefulness, which, while appropriate for capturing user-centered judgments, may be influenced by subjective biases or individual response tendencies. Future research could complement perceptual measures with objective task-based performance indicators, such as decision accuracy, time savings or error reduction, to triangulate findings and strengthen causal inference.

Finally, the empirical evaluation focuses on a limited set of business tasks that, while representative of common knowledge work, do not cover the full range of organizational applications of Generative AI. Future studies could explore prompting techniques in more complex, high-stakes, or domain-specific business scenarios, as well as longitudinal designs to assess how prompt literacy and perceived usefulness evolve over time with continued AI use and training.

10. Conclusions

This study was motivated by two central research questions concerning the identification and evaluation of prompt enhancement techniques in business-related applications of Generative AI.

First, the paper examined which prompt enhancement techniques reported in the literature improve output usefulness in business settings and how they contribute to general business enquiries. To address RQ1, the systematic review organized prompting techniques into baseline approaches, task-alignment techniques, and output-transparency techniques. Baseline prompting, such as zero-shot, primarily enables rapid response generation but is frequently associated with limited depth and robustness. In contrast, example-based techniques (one-shot and few-shot) enhance contextual alignment and structural clarity, while reasoning-oriented strategies such as chain-of-thought improve interpretability and coherence. By consolidating fragmented evidence into an outcome-oriented taxonomy, this study confirms prior research that positions prompt engineering as a key mechanism for improving reliability and reasoning quality in Generative AI systems [4,5], while extending the existing literature that has predominantly focused on technical benchmarks and domain-specific contexts to cross-domain business tasks [23,24].

The second research question (RQ2) investigated the extent to which different prompting techniques influence perceived usefulness in business-related tasks and whether this relationship is moderated by users' frequency of Generative AI usage. The empirical findings demonstrate that structured prompting strategies significantly increase perceived usefulness compared to zero-shot prompting. While one-shot, few-shot, and chain-of-thought techniques independently produce intermediate improvements, their combination yields the highest usefulness evaluations, indicating a cumulative effect between task-alignment and reasoning-transparency mechanisms.

This pattern is consistent with prior empirical evidence showing that structured prompt training enhances task performance compared to unstructured AI use [88], that structured prompting frameworks increase perceived response quality and user satisfaction [89], and that hybrid prompting strategies combining explicit instructions and reasoning scaffolds outperform simpler approaches in complex analytical contexts [90].

The present findings therefore confirm that more structured and integrated prompting designs generally improve output quality and user evaluations across domains.

However, contrary to perspectives that emphasize the decisive role of user experience in shaping AI-related performance gains [15], the analysis did not reveal a statistically significant moderating effect of AI usage frequency. This suggests that the prompt structure itself is the primary determinant of perceived usefulness in business contexts, independent of accumulated exposure to Generative AI tools.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/bdcc10070224/s1, PRISMA 2023 Checklist.

Author Contributions: Conceptualization, A.C. and A.D.M.; methodology, A.C. and A.D.M.; formal analysis, A.C.; investigation, A.C.; data curation, A.C.; writing—original draft preparation, A.C.; writing—review and editing, A.C. and A.D.M.; visualization, A.C.; supervision, A.D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: This study was conducted in accordance with the ethical principles of the Declaration of Helsinki. Ethical review and approval by an Ethics Committee were not required for this study under the applicable Italian legislation. In Italy, the competence of the Territorial Ethics Committees (Comitati Etici Territoriali) concerns clinical trials on medicinal products for human use, clinical investigations of medical devices, and pharmacological observational studies, pursuant to Article 2, paragraph 10, of Law No. 3 of 11 January 2018 and Article 1, paragraph 1, of the Ministerial Decree of 30 January 2023 ('Definizione dei criteri per la composizione e il funzionamento dei comitati etici territoriali'), read in conjunction with Regulation (EU) No 536/2014. The present study falls within none of these categories: it was non-clinical and non-interventional and did not concern medicinal products, medical devices, or any pharmacological or health-related matter. It did not involve patients, vulnerable subjects, clinical procedures, drugs, medical devices, biological samples, or any health-related intervention. Data were collected through a fully anonymous questionnaire; no directly identifiable personal data and no special categories of personal data were collected. Participants were informed in advance of the purpose of the study, the voluntary nature of their participation, the anonymous handling of their responses, the intended use of the data, and the absence of any foreseeable risk. Participation was voluntary, and informed consent to participate was obtained from all participants.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data is contained within the article or Supplementary Materials. The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding authors.

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

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DGM	Deep Generative Model
GenAI	Generative Artificial Intelligence
LLM	Large Language Model
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SLR	Systematic Literature Review
KPI	Key Performance Indicator
CoT	Chain-of-Thought
CoV	Chain-of-Verification

IV	Independent Variable
DV	Dependent Variable
LMEM	Linear Mixed-Effects Model
EMM	Estimated Marginal Means
PECS	Prompt Engineering Competence Scale
OS	One-Shot Prompting
FS	Few-Shot Prompting

Appendix A. Prompting Configurations Used in the Experimental Study

The appendix reports the exact prompt–output configurations evaluated by participants in the experimental survey.

Appendix A.1. Analysis Task

Table A1. Prompting configurations and generated outputs for the analysis task.

Technique	Prompt	Output
Zero-shot	Explain the following marketing KPIs in a clear and professional way and classify them into meaningful categories for a quarterly performance review: Churn Rate, Conversion Rate, Customer Acquisition Cost (CAC), Customer Lifetime Value (CLV), Return on Investment (ROI), Click-Through Rate (CTR).	Churn Rate measures how many customers stop using a product or service. Conversion Rate shows the percentage of users who complete a desired action. CAC indicates how much it costs to acquire a new customer. CLV estimates the total revenue a customer is expected to generate over time. ROI compares the financial return of an initiative to its cost. CTR measures how many users click on an ad relative to the number of impressions. Categories: - Acquisition: CAC, - Conversion Rate, CTR - Retention: Churn Rate - Value: CLV - Efficiency: ROI
One-shot	Explain the following marketing KPIs and classify them into meaningful categories for a quarterly performance review. Example of expected style: “This metric helps leadership understand whether our current initiatives are creating meaningful business impact and where strategic adjustments may be required.” KPIs: Churn Rate, Conversion Rate, Customer Acquisition Cost (CAC), Customer Lifetime Value (CLV), Return on Investment (ROI), Click-Through Rate (CTR).	Churn Rate indicates how many customers discontinue their relationship with the company, highlighting potential retention issues. Conversion Rate reflects how effectively users move from interest to action, informing where strategic adjustments may be needed. CAC shows the investment required to acquire each new customer, helping evaluate whether acquisition efforts are sustainable. CLV estimates the long-term value of each customer, supporting decisions on where to prioritize resources. ROI assesses whether an initiative generates meaningful business impact relative to its cost. CTR reveals how effectively our ads capture user interest. Categories: - Acquisition Metrics: CAC, Conversion Rate, CTR - Retention Metrics: Churn Rate - Value Metrics: CLV - Financial Performance: ROI

Table A1. *Cont.*

Technique	Prompt	Output
Few-shot	Explain the following marketing KPIs and classify them into meaningful categories for a quarterly performance review. Follow the examples below: # Example 1: "This metric highlights whether our current strategy is generating sustainable growth and indicates where resource allocation may need to shift in the next quarter." # Example 2 "This metric shows how efficiently our campaigns convert attention into action, helping identify which channels require optimisation to improve performance." # Example 3 "A sudden change in this metric often signals emerging risks in customer behaviour and prompts a deeper review of retention and engagement patterns." KPIs: Churn Rate, Conversion Rate, Customer Acquisition Cost (CAC), Customer Lifetime Value (CLV), Return on Investment (ROI), Click-Through Rate (CTR).	Churn Rate highlights potential risks in customer retention and signals when engagement patterns may require deeper investigation. Conversion Rate shows how efficiently we convert interest into action, helping identify which parts of the funnel need optimisation. CAC indicates whether our acquisition strategy is sustainable and where spending may need to be adjusted. CLV provides insight into long-term revenue potential and helps determine where to focus investment. ROI shows whether our initiatives are generating sufficient impact to justify continued resource allocation. CTR reflects how effectively our messaging captures attention and whether creative adjustments are needed. Categories: - Top-of-Funnel Efficiency: CTR - Acquisition Performance: Conversion Rate, CAC - Retention Health: Churn Rate - Customer Value: CLV - Financial Impact: ROI
Chain-of-Thought	Explain the following marketing KPIs and classify them into meaningful categories for a quarterly performance review. Before answering, think through the steps: what each KPI measures, which part of the funnel it belongs to and why it matters for quarterly evaluation. Provide only the final explanation. KPIs: Churn Rate, Conversion Rate, Customer Acquisition Cost (CAC), Customer Lifetime Value (CLV), Return on Investment (ROI), Click-Through Rate (CTR).	Churn Rate measures the percentage of customers who discontinue their relationship with the company and helps assess retention performance. Conversion Rate shows how effectively users complete a desired action, indicating funnel efficiency. CAC reflects the average cost of acquiring a new customer and is essential for evaluating acquisition scalability. CLV estimates the long-term value of each customer, supporting investment decisions. ROI evaluates whether an initiative generates more value than it costs. CTR measures how many users engage with an ad relative to impressions, indicating the relevance of campaign messaging. Categories: - Acquisition: CAC, Conversion Rate, CTR - Retention: Churn Rate - Customer Value: CLV - Financial Efficiency: ROI

Table A1. *Cont.*

Technique	Prompt	Output
One-shot + Chain-of-Thought	Explain the following KPIs and classify them into meaningful categories for a quarterly performance review. Example of expected style: "This metric helps leadership understand whether our current initiatives are creating meaningful business impact and where strategic adjustments may be required." Think step by step about what each KPI measures and where it fits in the marketing funnel, but provide only the final explanation. KPIs: Churn Rate, Conversion Rate, Customer Acquisition Cost (CAC), Customer Lifetime Value (CLV), Return on Investment (ROI), Click-Through Rate (CTR).	Churn Rate shows the percentage of customers who stop using our service and helps identify retention challenges. Conversion Rate measures how effectively users complete a targeted action, indicating funnel performance. CAC reflects the average cost of acquiring a new customer and helps assess acquisition efficiency. CLV estimates the long-term value of each customer and supports strategic investment decisions. ROI compares the net value created by an initiative to its cost. CTR measures how many users click on an ad relative to impressions, indicating the relevance of our messaging. Categories: - Acquisition Metrics: CAC, Conversion Rate, CTR - Retention Metrics: Churn Rate - Value Metrics: CLV - Financial Metrics: ROI
Few-shot + Chain-of-Thought	Explain the following KPIs and classify them into meaningful categories for a quarterly performance review. Follow the examples below. # Example 1: "This metric highlights whether our current strategy is generating sustainable growth and indicates where resource allocation may need to shift in the next quarter." # Example 2: "This metric shows how efficiently our campaigns convert attention into action, helping identify which channels require optimisation to improve performance." # Example 3: "A sudden change in this metric often signals emerging risks in customer behaviour and prompts a deeper review of retention and engagement patterns." Think step by step about measurement, funnel stage, and strategic relevance, but provide only the final explanation. KPIs: Churn Rate, Conversion Rate, Customer Acquisition Cost (CAC), Customer Lifetime Value (CLV), Return on Investment (ROI), Click-Through Rate (CTR).	Churn Rate indicates how many customers we lose within a given period and highlights emerging retention risks. Conversion Rate shows how effectively we turn interest into action, helping identify which parts of the funnel require optimisation. CAC measures the total cost of acquiring each new customer and is essential for evaluating acquisition sustainability. CLV estimates the long-term revenue potential of each customer and supports decisions on where to focus investment. ROI evaluates whether an initiative generates sufficient financial return relative to its cost. CTR reflects how engaging and relevant our ads are to the target audience. Categories: - Acquisition Performance: CAC, Conversion Rate, CTR - Retention and Loyalty: Churn Rate - Customer Value: CLV - Financial Impact: ROI

Appendix A.2. Synthesis Task

Source text: In recent years, the European Union has increased investments in the energy transition, aiming to accelerate the deployment of renewable sources and reduce dependence on fossil fuels. However, recent analyses indicate that the current pace is not sufficient to meet the 2030 climate targets. Many Member States are facing delays in permitting processes, infrastructure gaps, and difficulties in attracting private capital. At the same time, rising electricity demand requires significant grid expansion and greater

integration of digital technologies. Experts emphasize the need for more coordinated investment strategies, targeted incentives, and a stable regulatory framework to support innovation. Without a decisive acceleration, the EU risks slowing its competitiveness and failing to fully capture the economic potential of the energy transition.

Table A2. Prompting configurations and generated outputs for the synthesis task.

Technique	Prompt	Output
Zero-shot	Please summarise the source text.	The EU has increased investments in the energy transition, but progress remains too slow to meet 2030 climate goals. Member States face delays in permitting, infrastructure limitations, and challenges in attracting private capital. Rising electricity demand also requires major grid expansion and digital integration. Experts argue that coordinated strategies, targeted incentives, and stable regulations are needed to support innovation and competitiveness.
One-shot	Summarize the source text in a clear and concise way. Example of desired style: "Provide a short summary that highlights the main goals, challenges and implications."	Goals: Accelerate renewable energy deployment and reduce fossil-fuel dependence. Challenges: Slow permitting, infrastructure gaps, limited private investment, and rising electricity demand. Implications: Without coordinated strategies, incentives, and regulatory stability, the EU risks missing 2030 climate targets and weakening its economic competitiveness.
Few-shot	Summarise source text following the style shown in the examples. # Example 1: "Provide a structured summary that identifies the main objectives and the key obstacles." # Example 2: "Include a brief explanation of why the topic is relevant for policy or strategic decision-making	The EU is increasing investments in renewable energy, yet current efforts fall short of what is needed to meet 2030 climate objectives. Key obstacles include slow permitting procedures, insufficient infrastructure, and difficulties in mobilizing private capital. Growing electricity demand further intensifies the need for grid expansion and digital integration. For policymakers, the findings highlight the importance of coordinated investment strategies, targeted incentives, and a stable regulatory environment to sustain innovation and maintain competitiveness.
Chain-of-Thought	First identify the key ideas in the text. Then summarise the source text in a clear and concise way based on those ideas.	Key ideas: - EU investment in energy transition is rising. - Current pace is insufficient for 2030 goals. - Barriers: permitting delays, infrastructure gaps, limited private capital. - Electricity demand is increasing. - Need for coordinated strategies, incentives, and regulatory stability. Summary: Although the EU is investing more in renewable energy, progress is too slow to meet 2030 climate targets. Structural barriers such as slow permitting, inadequate infrastructure, and limited private investment hinder advancement. With electricity demand rising, experts call for coordinated strategies, targeted incentives, and stable regulations to support innovation and competitiveness.

Table A2. *Cont.*

Technique	Prompt	Output
One-shot + Chain-of-Thought	Summarise the source text in a clear and concise way. Example of desired style: "Provide a short summary that highlights the main goals, challenges, and implications.". Before summarising, identify the key ideas in the text	Key ideas: - The EU is increasing investments in the energy transition. - Current progress is insufficient for 2030 climate targets. - Barriers include slow permitting, infrastructure gaps, and limited private capital. - Rising electricity demand requires grid expansion and digital integration. - Experts call for coordinated strategies, incentives, and regulatory stability. Summary: - Goals: Accelerate renewable deployment and reduce fossil-fuel dependence. - Challenges: Slow permitting, infrastructure limitations, weak private-capital mobilization, and rising electricity demand. - Implications: Without coordinated strategies and stable regulations, the EU risks missing 2030 climate targets and weakening its competitiveness.
Few-shot + Chain-of-Thought	Summarise the source text following the style shown in the examples. # Example 1: "Provide a structured summary that identifies the main objectives and the key obstacles." # Example 2: "Include a brief explanation of why the topic is relevant for policy or strategic decision-making." Before summarising, identify the key ideas in the text.	Key ideas: - EU investment in the energy transition is rising. - Current pace is insufficient for 2030 goals. - Obstacles: permitting delays, infrastructure gaps, limited private investment. - Electricity demand is increasing. - Need for coordinated strategies, incentives, and regulatory stability. Summary: - Objectives: Accelerate renewable-energy deployment and reduce fossil-fuel dependence. - Obstacles: Slow permitting, inadequate infrastructure, and challenges in attracting private capital, combined with rising electricity demand. - Policy relevance: These issues highlight the need for coordinated investment strategies, targeted incentives, and a stable regulatory framework to sustain innovation and maintain the EU's competitiveness.

Appendix B. Data Extraction Tables (SLR)

Appendix B.1. Baseline Prompting Techniques

Table A3. Data Extraction Table of Baseline Prompting Techniques.

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Ghimire, P. et al. [30]	Zero-shot	Cost estimation in construction	Human evaluation	Incomplete responses with an average confidence score of 1.91 (64%). Lack of predefined priority logic, reducing reliability and indicating necessity of structured prompting strategies
Bharti, P.K et al. [31]	Zero-shot	Peer reviews assessment	Human evaluation, recall, precision, f3	The model's classification was tentative and sentiment-biased. justification was partial and lacked confidence, leading to moderate recall and lower precision.
Tony et al. [12]	Zero-shot	Code generation/software security	Vulnerability count; security rule violations	Zero-shot prompting produces insecure code with a high number of vulnerabilities, offering limited control over security properties.
Razumovskaia et al. [37]	Zero-shot	Multilingual NLU (Intent Detection, NER, NLI)	F1, Accuracy	Zero-shot in-context learning shows very low performance across multilingual tasks, especially for low-resource languages, often close to baseline.
Indira Kumar et al. [36]	Zero-shot	Harmful meme text classification (cyberbullying, sarcasm, sentiment, emotion, harmfulness)	Accuracy, Precision, Recall, F1	Zero-shot prompting enables multi-task classification without labeled data, but performance varies widely across tasks and models, especially for nuanced categories. It is efficient in handling multiple tasks simultaneously, but its reliability is influence by model and quality of output
Giannilias et al. [32]	Zero-shot	Cybersecurity text classification (dark web hacker forum posts)	Accuracy, Precision, Recall, F1	Zero-shot prompting enables classification of unstructured hacker forum posts but shows uneven performance across categories and high variance between models.
Jaradat et al. [63]	Zero-shot	Traffic crash analysis (classification & IE)	Accuracy, F1, Jaccard	Zero-shot enables immediate inference but underperforms compared to few-shot and fine-tuning in complex multimodal tasks.
Zhang et al. [34]	Zero-shot	Early sepsis diagnosis from structured clinical data	Accuracy, Precision, Recall, F1	Standard prompts yield acceptable recall but lower accuracy and F1, with limited interpretability and less structured diagnostic reasoning.
Yin & Guo [64]	Zero-shot	Financial forecasting (earnings direction prediction from numerical financial statements)	Accuracy, Precision, Recall, F1	Direct-answer prompting yields stable but lower predictive performance and provides no transparent reasoning for financial decision-making.

Table A3. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Shafikuzzaman et al. [65]	Zero-shot	Software requirements classification (FR vs. NFR, NFR subclasses, security)	Precision, Recall, F1 (macro)	Zero-shot LLMs (GPT-4o) achieve strong performance without training data, sometimes approaching FS models.
Byra et al. [45]	Zero-shot	Medical image classification (chest X-rays: pneumonia vs. normal; breast US: benign vs. malignant)	Accuracy, AUC, ICC	Zero-shot classification using GPT-4-generated shape and texture descriptors achieves reasonable AUC (≈0.72–0.76) but shows high variability depending on descriptor quality.
Šuster et al. [44]	Zero-shot	Risk-of-bias assessment in clinical trials (RoB2, systematic reviews)	Macro F1, Accuracy	Zero-shot prompting fails to capture complex methodological reasoning required for RoB2 assessment, yielding F1 scores close to random or majority baselines.
Lee, A.V.Y. et al. [66]	Zero-shot	Education; commonsense reasoning; knowledge creation discourse	Human qualitative comparison (length, depth, idea quality)	Zero-shot prompting produces reasonable answers but does not sufficiently support idea elaboration or improvement in student discourse.
Zhen et al. [33]	Zero-shot	Traffic crash severity classification (road safety)	Macro F1, Macro Accuracy	Plain zero-shot prompting struggles with class imbalance and fails to reliably detect fatal crashes, especially for smaller models.
Shah et al. [42]	Zero-shot	Clinical QA, summarization, decision support	Not applicable (review)	Zero-shot prompting can be effective for simple clinical tasks but is prone to hallucinations and misapplication of knowledge in complex medical contexts.
Meshkin et al. [38]	Zero-shot	Regulatory NLP; PK drug–drug interaction (DDI) sentence classification	Precision, Recall, F1-score, Specificity	Several open-source LLMs (notably Flan-T5) achieve F1 ≈0.88–0.89 in zero-shot settings, matching or exceeding a BioBERT model trained on >20 k sentences.
Meshkin et al. [38]	Zero-shot	Regulatory NLP; identification of intrinsic factors affecting drug exposure	Precision (≈78.5%)	Zero-shot prompting enables large-scale extraction of clinically relevant intrinsic factors from >700 k FDA label sentences without fine-tuning.
Ono et al. [67]	Zero-shot	Histopathology image classification (tau lesions: AP, NP, TA)	Accuracy, Sensitivity, Specificity	Zero-shot GPT-4V accurately recognises staining and tissue type but fails to reliably identify specific tau lesions, achieving only ~40% accuracy.
Viswanathan et al. [39]	Zero-shot (instruction-only)	Text clustering (entity canonicalization, intent clustering, tweet clustering)	Macro F1, Micro F1, Pairwise F1, Accuracy, NMI	Even instruction-only prompts (no demonstrations) enable LLMs to inject task-specific structure into text representations, improving clustering quality.

Table A3. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Lee et al. [47]	Zero-shot	Automatic scoring of student-written science explanations	Accuracy, Precision, Recall, F1, QWK	Zero-shot prompting provides a feasible but limited baseline for automatic scoring, particularly struggling with minority proficiency categories.
Choi, J. [68]	Zero-shot	Relevance evaluation in information retrieval	Cohen's kappa	Zero-shot prompts consistently outperform few-shot prompts, suggesting that GPT models leverage their pretrained knowledge more effectively without in-context examples.
Hebenstreit, K. et al. [69]	Zero-shot	Multi-domain multiple-choice question answering	Krippendorff's alpha; accuracy	Direct prompting without explicit reasoning instructions consistently underperforms reasoning-based prompts across all evaluated models and datasets.
Miao, J. et al. [35]	Zero-shot	Medical QA and general reasoning	Accuracy (reported in prior literature), qualitative review	Zero-shot prompting enables broad task generalization but often lacks the specificity and accuracy required for nuanced clinical decision-making in nephrology.
Filienko, D. et al. [58]	Zero-shot	Problem-Solving Therapy dialogue (symptom assessment, goal setting)	Human Likert ratings; empathy metrics (ER, IP, EX)	Zero-shot prompting with explicit task instructions improves structure but fails to reliably capture implicit therapeutic skills required for PST, leading to lower overall quality than other techniques.
Thanasi-Boçe, M & Hoxa [62]	Zero-shot	entrepreneurship-learning,	Empirical (mixed-methods)	Zero-shot prompting is a technique where the prompt does not provide any prior information about the task that the LLM is supposed to perform.
Ayad, S; Alsayoud, F [70]	Zero-shot	Business process modeling	Empirical (experimental/applied)	Prompt without examples used to generate BPM elements from textual input
Carlson, NA; Burbano, V [48]	Zero-shot	Data annotation/text classification	Empirical (mixed-methods)	Direct task instruction without examples
Abou El Karam et al. [49]	Zero-shot	Change management/employee attitude assessment	Empirical (experimental)	Classification of open-ended employee responses without examples
Wang, J. [9]	Zero-shot	Business communication writing (negative message)	Empirical (experimental)	Direct prompt without additional guidance or interaction
Jovic, M et al. [71]	Zero-shot	Written feedback generation	Empirical (experimental)	Minimal instruction without contextual or iterative guidance
Jovic, M; et al. [71]	Zero-shot	Written feedback generation	Empirical (experimental)	Baseline AI feedback without structured context

Table A3. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Tocev, T.; Atanasovski, A. [72]	Zero-shot	IFRS advisory/financial reporting	Empirical (experimental)	Single-pass query without examples or stepwise decomposition
Tocev, T.; Atanasovski, A. [72]	Zero-shot	IFRS advisory/financial reporting	Empirical (experimental)	Direct prompt describing the case once

Appendix B.2. Task Alignment Prompting Techniques

Table A4. Data Extraction Table of Task Alignment Prompting Techniques.

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Golnari et al. [40]	One-shot	Medical		One-shot offers an improvement in respect to zero-shot
Giannilias et al. [32]	One-shot	Cybersecurity text classification (dark web hacker forum posts)	Accuracy, Precision, Recall, F1	Providing a single representative example per class improves classification accuracy by reducing ambiguity and clarifying label boundaries.
Agbareia et al. [41]	One-shot	Retinal disease classification from OCT images (AMD, DR, CSR, MH, Normal)	Accuracy (%), CI	Single-shot prompting allows basic multimodal diagnosis but shows limited accuracy, especially for complex retinal pathologies.
Shah et al. [42]	One-shot	Clinical decision support; diagnostic illustration	Not applicable (review)	Providing a single exemplar helps constrain model outputs and improves task understanding in narrowly defined clinical tasks.
Razumovskaia et al. [37]	Few-shot	Multilingual NLU (Intent Detection, NER, NLI)	F1, Accuracy	Few-shot prompting improves performance over zero-shot but remains significantly weaker than supervised approaches, particularly in cross-lingual settings.
Byra et al. [45]	Few-shot	Medical image classification (chest X-rays; breast US)	Accuracy, AUC	Few-shot descriptor selection ($n \geq 10$) substantially improves accuracy (up to 0.81–0.83) by removing poorly performing text descriptors, without retraining the VLM.
Viswanathan et al. [39]	Few-shot	Text clustering	Macro F1, Micro F1, Pairwise F1, Accuracy, NMI	Few-shot prompting for key phrase expansion is the most effective LLM integration strategy, outperforming classical and neural baselines on most datasets.
Lee et al. [47]	Few-shot	Automatic scoring in education	Accuracy, Precision, Recall, F1, QWK	Few-shot prompting improves scoring accuracy by constraining output structure and aligning model behavior with human scoring patterns.

Table A4. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Jiang, Y. et al. [73]	Few-shot	Out-of-scope intent classification in task-oriented dialogue	AU-IOC, in-scope accuracy, OOS recall	Prompt-based learning that incorporates natural-language intent descriptions achieves higher AU-IOC scores across all datasets, especially in extremely low-data regimes (1–5 shots).
Qi, X. et al. [43]	Few-shot	Construction of Knowledge Graphs in the domain of equipment O&M	Precision, Recall, F1	10% increase in Recall, Few-shot learning helps LLMs better understand and adapt to task requirements, thereby substantially enhancing their generalization capabilities.
Bharti, P.K. et al. [31]	Few-shot	Peer reviews assessment	human evaluation, recall, precision, f2	By leveraging example-driven analogical reasoning, this approach improved classification accuracy. It correctly identified the trivial review's superficiality and the exhaustive review's depth. However, the generated explanations were less detailed, covering partial aspects of the reviews.
Tony et al. [12]	Few-shot	Code generation/software security	Vulnerability count; security rule violations	Few-shot prompting reduces vulnerabilities compared to zero-shot, but effectiveness depends on vulnerability type and examples provided.
Golnari et al. [40]	Few-shot	Medical	-	The performance of the baseline prompt improves by 13.3% for detecting paroxysmal events using a two-shot approach as compared to zero-shot.
Du et al. [74]	Few-shot	power equipment defect grading	Accuracy	Few-shot examples improve performance over zero-shot but show limited gains in domain-specific reasoning.
Indira Kumar et al. [36]	Few-shot	Harmful meme text classification (cyberbullying, sarcasm, sentiment, emotion, harmfulness)	Accuracy, Precision, Recall, F1	Few-shot prompting improves F1 and accuracy for several tasks when compared to zero-shot, particularly cyberbullying detection, though gains are uneven and sensitive to example choice.
Jaradat et al. [63]	Few-shot	Traffic crash analysis (severity, driver fault, actions)	Accuracy, F1, Jaccard	Few-shot prompting with GPT-4.5 achieves exceptional performance in classifications with accuracy reaching 98.9% and precision as high as 100%.
Shafikuzzaman et al. [65]	Few-shot	Software requirements classification (FR/NFR, NFR subclasses, security)	Precision, Recall, F1 (macro)	Few-shot prompting reliably improves classification performance across all tasks.

Table A4. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Agbareia et al. [41]	Few-shot	Retinal disease classification from OCT images	Accuracy (%), CI, *p*-values	Few-shot prompting with reference images substantially improves diagnostic accuracy across most retinal conditions, with gains up to +64% in some classes.
Šuster et al. [44]	Few-shot	Risk-of-bias assessment in clinical trials	Macro F1, Accuracy	Few-shot prompting with justification exemplars does not substantially improve performance, suggesting that in-context examples are insufficient for complex evidence appraisal tasks.
Lee, A.V.Y. et al. [66]	Few-shot	Education; commonsense reasoning; knowledge creation discourse	Human qualitative comparison	Few-shot prompting improves response relevance and structure by providing exemplars aligned with human expectations.
Zhen et al. [33]	Few-shot	Traffic crash severity classification	Macro F1, Macro Accuracy	Few-shot prompting improves performance mainly for smaller models (LLaMA3-8B), but introduces trade-offs across severity categories.
Shah et al. [42]	Few-shot	Clinical QA, medical education, classification	Not applicable (review)	Few-shot prompting improves performance by clarifying task structure and reducing ambiguity, particularly in domain-specific medical queries.
Meshkin et al. [38]	Few-shot	Regulatory NLP; PK-DDI classification	Precision, Recall, F1-score	Few-shot prompting provides modest gains for some models but does not consistently outperform zero-shot learning in imbalanced regulatory datasets.
Ono, Dickson & Koga [67]	Few-shot	Histopathology image classification	Accuracy, Sensitivity, Specificity	Few-shot prompting with a small number of reference images improves diagnostic accuracy to ~50–60%, but performance remains unstable.
Viswanathan et al. [39]	Few-shot	Semi-supervised text clustering	Macro F1, Pairwise F1	LLMs can act as effective pseudo-oracles for pairwise constraints, approaching human-supervised clustering performance at lower cost.
Gu, X. et al. [61]	Few-shot	Sentiment classification	Accuracy, F1	Few-shot AGCVT prompting achieves the highest accuracy while preserving interpretability and low data requirements.
Choi, J. [68]	Few-shot	Relevance evaluation in information retrieval	Cohen's kappa	Few-shot prompting introduces variability and bias that can reduce alignment with human relevance judgments, even when increasing the number of examples.
Wan, T.; Chen, Z. [46]	Few-shot	Feedback generation for physics conceptual questions	Usefulness ratings	Providing a small number of example response–feedback pairs enables GPT to generate feedback that students perceive as more useful and more responsive to their reasoning.

Table A4. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Miao, J. et al. [35]	Few-shot	Medical QA and clinical reasoning	Qualitative synthesis of prior studies	Few-shot prompting improves task alignment and reduces hallucinations, but performance gains tend to plateau after a small number of examples.
Filienko, D. et al. [58]	Few-shot	Problem-Solving Therapy dialogue	Human Likert ratings; empathy metrics	Few-shot prompting with clinician-curated examples enables the model to internalize implicit therapeutic norms, resulting in more coherent, personalized, and protocol-adherent PST dialogues.
Ayad, S; Alsayoud, F [70]	Few-shot	Business process modeling	Empirical (experimental/applied)	Prompt includes examples to guide BPM model generation
Carlson, NA; Burbano, V [48]	Few-shot	Data annotation/text classification	Empirical (mixed-methods)	Including examples in prompt
Wang, J. [9]	Few-shot	Business communication writing	Empirical (experimental)	Providing genre-based outlines or examples
Tocev, T.; Atanasovski, A. [72]	Few-shot	IFRS advisory/financial reporting	Empirical (experimental)	Prompt includes worked IFRS examples before the target case
Carlson, NA; Burbano, V [48]	Role	Data annotation/text classification	Empirical (mixed-methods)	Assigning an identity to focus responses on relevant domain knowledge
Karam et al. [49]	Role	Change management/decision support	Empirical (experimental)	Implicit expert framing via allies strategy categories
Wang, J. [9]	Role	Business communication writing	Empirical (experimental)	Assigning ChatGPT an expert professional role
Tocev, T.; Atanasovski, A. [72]	Role	IFRS advisory/professional decision support	Empirical (experimental)	Assigning GPT-4 the role of an IFRS expert
Mao, W. et al. [50]	Role	Recommendation systems	Empirical (experimental)	Assigning expert or recommender roles to the LLM
Cheung, K.S. [75]	Role	Professional decision support	Conceptual (viewpoint)	Framing the LLM as a professional valuer

Appendix B.3. Output Transparency Prompting Techniques

Table A5. Data Extraction Table of Output Transparency Prompting Techniques.

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Chen et al. [51]	Chain-of-Thought	Vulnerability detection	F1, Accuracy, Recall	CoT-based prompting significantly improves both vulnerability detection performance and interpretability; removing CoT leads to a clear drop in F1.
Qi et al. [43]	Chain-of-Thought	Construction of Knowledge Graphs in the domain of equipment O&M	Precision, Recall, F1	slight 2% increase in metrics. The dataset primarily consists of short texts, where the advantages of CoT—typically more pronounced with longer texts—are not fully realized.
Ghimire et al. [30]	Chain-of-Thought	Cost estimation in construction	human evaluation	2.52 (84%) av confidence score marks 20% improvement from ZS. This improvement confirms that structured, step-by-step reasoning enhances both response reliability and alignment with estimation logic, highlighting improvements in response quality and module adherence.
Bharti et al. [31]	Chain-of-Thought	Peer reviews assessment	human evaluation, recall, precision, f1	Most detailed and interpretable outputs and provided well structured, evidence based justifications
Köksal, A.; Alatan, A.A. [60]	Chain-of-Thought	small scale remote sensing	recall, precision, f1, param, rs	Model with CoT reasoning in its training captions achieved more than 1.5× military-related answers, while the precision drops by 3%, due to increasing false positives in C2. The intermediate reasoning in the CoT training likely taught the model what clues to look for. This aligns with observations that CoT can make models better at justifying and thereby correctly executing a task. Thus, incorporating reasoning-focused data is beneficial for fine-tuning multimodal models in this context.
Darwiyanto et al. [52]	Chain-of-Thought	automated program repair	plausible patches	The designed CoT prompting structure has been generally shown to improve the ability of LLMs to generate solutions for APR tasks.
Qiao et al. [76]	Chain-of-Thought	Medical	Accuracy; Recall; BLEU-1; reasoning-quality metrics	Embedding structured Chain-of-Thought during training improves reasoning coherence and interpretability compared to unstructured CoT.
Tony et al. [12]	Chain-of-Thought (CoT)	Code generation/software security	Vulnerability count; human evaluation	Chain-of-Thought improves reasoning transparency and helps the model avoid obvious insecure patterns during code generation.
Jeon at al [55]	Chain-of-Thought	Medical	Cohen's d effect sizes calculated against this Control condition (baseline prompt)	Most consistent performance (M = 65.61%, SD = 17.17, 95% CI [60.68, 70.54]). stable performance compared to Control (d = −0.301 to 0.308, lowest in logic and facts tasks and highest in Chinese tasks)

Table A5. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Du et al. [74]	Chain-of-Thought	power equipment defect grading	human eval., accuracy, ecs	Improves raw grading accuracy but also generates explanations that are both practically useful and theoretically grounded. The high Trustworthiness scores underscore the value of grounding model inferences in curated domain knowledge, while the expert suggestions point the way toward future enhancements that integrate richer contextual features and uncertainty quantification into the reasoning pipeline.
Teng et al. [54]	Chain-of-Thought	Depression detection (clinical text analysis)	CCC, MAE, Accuracy	Chain-of-Thought prompting improves both severity estimation and transparency of clinical reasoning.
Chen et al. [56]	Chain-of-Thought	Long-document abstractive summarization (scientific, biomedical, legal, governmental texts)	ROUGE-1/2/L (F1), BLEU, BERTScore, FactCC, human evaluation	Chain-of-Thought prompting improves factual consistency, structural coherence, and content coverage in long-document summarization by enforcing step-by-step reasoning before summary generation.
Chen & Li [57]	Chain-of-Thought	Instruction following; length-controlled text generation	Acc%, Vlt%, Avg. response length, ROUGE, BLEU	Chain-of-Thought prompting significantly reduces length violations and improves reasoning coherence, enabling better compliance with strict length constraints. By introducing and analyzing COT, we gain deeper insights into model reasoning, enhancing output accuracy and enabling finer control over output length.
Zhang et al. [34]	Chain-of-Thought	Early sepsis diagnosis from structured clinical data	Accuracy, Precision, Recall, F1	Chain-of-Thought prompting significantly improves F1 score (≈+7–8 pp vs. ML baselines) and enhances clinical interpretability by simulating expert reasoning without task-specific training.
Zhang et al. [77]	Chain-of-Thought	Reasoning tasks across NLP, multimodal tasks, and language agents	Not applicable (survey)	Across a wide range of studies, Chain-of-Thought prompting consistently enhances multi-step reasoning, interpretability, and task generalization, especially in large-scale LLMs (>10B parameters).
Yin & Guo [64]	Chain-of-Thought	Financial forecasting and portfolio optimization (quantitative finance)	Accuracy, Precision, Recall, F1; Sharpe Ratio; Alpha; Max Drawdown	Chain-of-Thought prompting enables step-by-step numerical reasoning that outperforms human analysts in earnings direction prediction and supports profitable, risk-controlled investment strategies. CoT enhanced model achieved superior accuracy, demonstrating LLM's capacity to automate financial reasoning and reducing reliance on specialized expertise
Hlaing et al. [78]	Chain-of-Thought	Dental licensing exam QA (prosthodontics, Korean language)	Accuracy (%), confidence intervals, reliability (Cronbach's α)	Models with native or elicited Chain-of-Thought reasoning achieve passing-level accuracy comparable to human averages, outperforming language-optimized models in a non-Indo-European language.
Park et al. [79]	Chain-of-Thought	Radiology report classification (TRC)	AUROC, Accuracy, Precision, Recall, F1	CoT improves interpretability and reasoning structure but underperforms ICL in difficult categories when used alone.

Table A5. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Bukhary et al. [59]	Chain-of-Thought	Visual defect detection in AV software requirements	Precision, Recall, F1-score, Correctness score	Chain-of-Thought prompting improves reasoning quality and explanation completeness but may introduce over-analysis and hallucinated defects, especially in safety-critical visuals.
Liao et al. [80]	Chain-of-Thought	Traffic scene understanding & motion forecasting (autonomous driving)	BERTScore (Precision, Recall, F1)	CoT prompting enables LLMs to generate structured, human-like reasoning over traffic scenes, significantly enhancing semantic understanding without updating model weights.
Zhen et al. [33]	Chain-of-Thought	Traffic crash severity analysis & inference	Macro F1, Macro Accuracy	CoT prompting enables step-by-step reasoning over crash causes, improving overall inference quality and interpretability.
Shah et al. [42]	Chain-of-Thought	Clinical reasoning, decision support, education	Not applicable (review)	Chain-of-Thought prompting enhances multi-step clinical reasoning and interpretability, making LLM outputs more aligned with clinician expectations.
Lee et al. [47]	Chain-of-Thought	Automatic scoring	Accuracy, F1	Zero-shot CoT alone does not substantially improve scoring accuracy, indicating that generic reasoning is insufficient for rubric-based grading.
Gu et al. [61]	Chain-of-Thought	Sentiment classification	Accuracy, F1	Chain-of-thought improves reasoning transparency but manual construction is costly and non-scalable.
Yang, G. et al. [53]	Chain-of-Thought	Code generation	Pass@1, CoT-Pass@1	Providing explicit reasoning steps enables lightweight language models to better understand control flow and constraints, resulting in significantly higher code correctness.
Hebenstreit et al. [69]	Chain-of-Thought	Multi-domain multiple-choice question answering	Krippendorff's alpha; accuracy	The standard zero-shot CoT trigger "Let's think step by step" improves performance consistently across models and datasets, demonstrating strong generalizability.
Miao et al. [35]	Chain-of-Thought	Clinical diagnosis and treatment planning in nephrology	Diagnostic accuracy, qualitative clinical evaluation	Chain-of-thought prompting enables step-by-step clinical reasoning that mirrors physician decision-making, leading to more accurate diagnoses and clearer justification of medical decisions.
Miao et al. [35]	Chain-of-Thought	Clinical decision support and auditing	Explainability analysis	By externalizing reasoning steps, chain-of-thought prompting enhances explainability, enabling error tracing, auditing, and greater trust in high-stakes medical decisions.
Filienko et al. [58]	Chain-of-Thought	Problem-Solving Therapy dialogue	Human Likert ratings; empathy metrics	Zero-shot chain-of-thought prompting enhances exploratory and reflective responses but can reduce actionability and protocol adherence in dialogue-based therapeutic tasks.
Thanasi-Boçe, M; Hoxha, J [62]	Chain-of-Thought	entrepreneurship-learning, critical thinking, problem solving	Empirical (mixed-methods)	It involves providing a prompt that presents a sequence of intermediated reasoning instructions or steps to complete a task
Ayad, S; Alsayoud, F [70]	Chain-of-Thought	Business process modeling	Empirical (experimental/applied)	Prompt requests step-by-step reasoning before generating BPM models

Table A5. *Cont.*

Authors	Technique	Task/Domain	Evidence Type	Reported Effects
Carlson, NA; Burbano, V [48]	Chain-of-Thought	Data annotation/text classification	Empirical (mixed-methods)	Explicit step-by-step reasoning process
Karam et al. [49]	Chain-of-Thought	Organizational analysis	Empirical (experimental)	Stepwise reasoning via explicit class descriptions
Tocev, T.; Atanasovski, A. [72]	Chain-of-Thought	IFRS advisory/financial reporting	Empirical (experimental)	Sequential decomposition into subquestions with stepwise reasoning
Mao, W. et al. [50]	Chain-of-Thought	Personalized recommendation	Empirical (experimental)	Step-by-step reasoning to infer user preferences
Cheung, K.S. [75]	Chain-of-Thought	Property valuation reporting	Conceptual (viewpoint)	Structured step-by-step prompt aligned with RICS "Red Book" valuation standards
Jeon at al. [55]	Self-Consistency	Medical	Empirical	Moderate variability (d = −0.623 to 0.362), lowest in logic global facts and highest in causal judgement
Thanasi-Boçe, M; Hoxha, J [62]	Chain of verification	Transactional decision-making	Empirical (mixed-methods)	CoV will first generate an initial response, and after, it will create verification questions to fact-check its response.
Carlson, NA; Burbano, V [48]	Self-Consistency	Data annotation/text classification	Empirical (mixed-methods)	Multiple independent attempts at same task

Appendix C. Hypothesis Testing

Appendix C.1. H1 Testing (Perceived Output Usefulness)

Table A6. Reliability Analysis of Perceived Output Usefulness.

Construct	Items	Cronbach's α
Perceived Output Usefulness	5	0.947

Table A7. Linear Mixed-Effects Model—Effect of Prompting Technique on Output Usefulness (H1).

Predictor	Estimate	SE	df	t	*p*
Zero-Shot (Intercept)	3.285	0.053	691	61.45	<0.001 ***
CoT	0.843	0.067	1249	12.65	<0.001 ***
Few-shot	0.823	0.067	398	12.27	<0.001 ***
Few-shot + CoT	1.522	0.070	398	22.74	<0.001 ***
One-shot	0.741	0.068	399	11.03	<0.001 ***
One-shot + CoT	1.553	0.067	383	22.88	<0.001 ***
Task Type (Synthesis)	−0.033	0.037	311	−0.89	0.36

Note: *** $p < 0.001$.

Table A8. ANOVA (Type III)—Fixed Effects.

Effect	Num df	Den df	F	*p*
Technique	5	304	144.64	<0.001 ***
Task Type	1	304	0.81	0.370
Interaction	5	304	0.99	0.427

Note: *** $p < 0.001$.

Table A9. Estimated Marginal Means (Post Hoc).

Technique	Mean	SE	95% CI Lower	95% CI Upper
Zero-shot	3.27	0.051	3.17	3.37
CoT	4.11	0.051	4.01	4.21
Few-shot	4.09	0.051	3.99	4.19
Few-shot + CoT	4.79	0.050	4.69	4.89
One-shot	4.01	0.051	3.91	4.11
One-shot + CoT	4.82	0.051	4.72	4.92

Appendix C.2. H2 Testing (Perceived Prompt Usefulness)

Table A10. Reliability Analysis of Perceived Prompt Usefulness.

Construct	Items	Cronbach's α
Perceived Prompt Usefulness	4	0.994

Table A11. Linear Mixed-Effects Model—Main Effects (H2).

Predictor	Estimate	SE	df	t	*p*
Zero-Shot (Intercept)	3.078	0.111	135	27.65	<0.001 ***
CoT	1.051	0.066	376	15.83	<0.001 ***
Few-shot	1.089	0.067	377	16.15	<0.001 ***
Few-shot + CoT	1.778	0.069	377	26.61	<0.001 ***
One-shot	0.948	0.069	377	14.18	<0.001 ***
One-shot + CoT	1.876	0.068	380	27.71	<0.001 ***
Frequency of Use	−0.037	0.026	103	−1.39	0.169

Note: *** $p < 0.001$.

Table A12. Random Effect for Main-Effects Model (H2).

Random Effect	Variance	SD
Respondent (Intercept)	0.038	0.196
Residual	0.141	0.376

Table A13. Linear Mixed-Effects Model—Interaction Model (Moderation Test).

Predictor	Estimate	SE	df	t	*p*
(Intercept)	3.154	0.190	379.80	16.63	<0.001 ***
CoT	1.053	0.263	358.90	4.01	<0.001 ***
Few-shot	0.852	0.246	364.00	3.46	0.001 ***
Few-shot + CoT	1.905	0.257	380.50	7.42	<0.001 ***
One-shot	0.816	0.259	349.70	3.15	0.002 **
One-shot + CoT	1.667	0.262	373.60	6.36	<0.001 ***
Frequency of Use	−0.057	0.049	379.90	−1.16	0.245
CoT × Frequency	0.000	0.007	361.10	0.00	0.999
Few-shot × Frequency	0.062	0.064	366.40	0.97	0.331
Few-shot + CoT × Frequency	−0.036	0.068	380.90	−0.54	0.593
One-shot × Frequency	0.035	0.066	354.80	0.53	0.596
One-shot + CoT × Frequency	0.056	0.067	372.90	0.82	0.421

Note: ** $p < 0.01$; *** $p < 0.001$.

Table A14. Random Effect for Interaction Model (Moderation test).

Random Effect	Variance	SD
Respondent (Intercept)	0.038	0.194
Residual	0.142	0.377

Table A15. Likelihood Ratio Test (Moderation).

Model Comparison	$\Delta\chi^2$	df	p
Main effects vs. Interaction	3.245	5	0.662

References

1. Bankins, S.; Ocampo, A.C.; Marrone, M.; Restubog, S.L.D.; Woo, S.E. A multilevel review of artificial intelligence in organizations: Implications for organizational behavior research and practice. *J. Organ. Behav.* **2024**, *45*, 159–182. [CrossRef]
2. Brynjolfsson, E.; Li, D.; Raymond, L.R. Generative AI at Work. *Q. J. Econ.* **2025**, *140*, 889–942. [CrossRef]
3. Simkute, A.; Tankelevitch, L.; Kewenig, V.; Scott, A.E.; Sellen, A.; Rintel, S. Ironies of generative AI: Understanding and mitigating productivity loss in human-AI interaction. *Int. J. Hum. Comput. Interact.* **2025**, *41*, 2898–2919. [CrossRef]
4. Sikha, V.K.; Siramgari, D.; Korada, L. Mastering Prompt Engineering: Optimizing Interaction with Generative AI Agents. *J. Eng. Appl. Sci. Technol.* **2023**, *5*, 1–8. [CrossRef]
5. Bozkurt, A. Tell Me Your Prompts and I Will Make Them True: The Alchemy of Prompt Engineering and Generative AI. *Open Prax.* **2024**, *16*, 111–118. [CrossRef]
6. Alostad, H. Large Language Models as Kuwaiti Annotators. *Big Data Cogn. Comput.* **2025**, *9*, 33. [CrossRef]
7. Wardle, G.; Sušnjak, T. Image First or Text First? Optimising the Sequencing of Modalities in Large Language Model Prompting and Reasoning Tasks. *Big Data Cogn. Comput.* **2025**, *9*, 149. [CrossRef]
8. Nugroho, H.S.; Shaferi, I. Analysis of Prompt Engineering Effectiveness in Stock Recommendation by ChatGPT: An Experimental Study in the Indonesian Market. *Int. Conf. Sustain. Econ. Manag. Account. Proc. (ICSEMA)* **2025**, *1*, 1755–1763. [CrossRef]
9. Wang, J. Improving ChatGPT's Competency in Generating Effective Business Communication Messages: Integrating Rhetorical Genre Analysis into Prompting Techniques. *J. Tech. Writ. Commun.* **2024**, *54*, 369–395. [CrossRef]
10. Elnashar, A.; White, J.; Schmidt, D.C. Enhancing structured data generation with GPT-4o evaluating prompt efficiency across prompt styles. *Front. Artif. Intell.* **2025**, *8*, 1558938. [CrossRef] [PubMed]
11. Bae, J.; Kwon, S.; Myeong, S. Enhancing software code vulnerability detection using GPT-4o and Claude-3.5 Sonnet: A study on prompt engineering techniques. *Electronics* **2024**, *13*, 2657. [CrossRef]
12. Tony, C.; Díaz Ferreyra, N.E.; Mutas, M.; Dhif, S.; Scandariato, R. Prompting techniques for secure code generation: A systematic investigation. *ACM Trans. Softw. Eng. Methodol.* **2025**, *34*, 1–53. [CrossRef]
13. Son, M.; Lee, S. Advancing multimodal large language models: Optimizing prompt engineering strategies for enhanced performance. *Appl. Sci.* **2025**, *15*, 3992. [CrossRef]
14. Panneer, S.V. Prompt Engineering for Conversational AI Systems: A Systematic Review of Techniques and Applications. *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.* **2025**, *11*, 733–741. [CrossRef]
15. Chowdhury, S.; Budhwar, P.; Dey, P.K.; Joel-Edgar, S.; Abadie, A. AI-employee collaboration and business performance: Integrating knowledge-based view, socio-technical systems and organisational socialisation framework. *J. Bus. Res.* **2022**, *144*, 31–49. [CrossRef]
16. Janiesch, C.; Zschech, P.; Heinrich, K. Machine learning and deep learning. *Electron. Mark.* **2021**, *31*, 685–695. [CrossRef]
17. Banh, L.; Strobel, G. Generative artificial intelligence. *Electron. Mark.* **2023**, *33*, 63. [CrossRef]
18. Naveed, H.; Khan, A.U.; Qiu, S.; Saqib, M.; Anwar, S.; Usman, M.; Akhtar, N.; Barnes, N.; Mian, A. A Comprehensive Overview of Large Language Models. *ACM Trans. Intell. Syst. Technol.* **2025**, *16*, 1–72. [CrossRef]
19. Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access* **2024**, *12*, 26839–26874. [CrossRef]
20. Castelvecchi, D. Can we open the black box of AI? *Nature* **2016**, *538*, 20–23. [CrossRef] [PubMed]
21. Fui-Hoon Nah, F.; Zheng, R.; Cai, J.; Siau, K.; Chen, L. Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *J. Inf. Technol. Case Appl. Res.* **2023**, *25*, 277–304. [CrossRef]
22. Feuerriegel, S.; Hartmann, J.; Janiesch, C.; Zschech, P. Generative AI. *Bus. Inf. Syst. Eng.* **2024**, *66*, 111–126. [CrossRef]
23. Liu, Y.Y.; Zheng, Z.; Zhang, F.; Feng, J.C.; Fu, Y.Y.; Zhai, J.D.; He, B.-S.; Zhang, X.; Du, X.Y. A comprehensive taxonomy of prompt engineering techniques for large language models. *Front. Comput. Sci.* **2026**, *20*, 2003601. [CrossRef]

24. Liu, X.; Wang, J.; Yuan, X.; Sun, J.; Dong, G.; Di, P.; Wang, W.; Wang, D. Prompting frameworks for large language models: A survey. *ACM Comput. Surv.* **2026**, *58*, 1–38. [CrossRef]
25. Polo, F.M.; Xu, R.; Weber, L.; Silva, M.; Bhardwaj, O.; Choshen, L.; de Oliveira, A.F.M.; Sun, Y.; Yurochkin, M. Efficient multi-prompt evaluation of LLMs. In *Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS '24)*; Curran Associates Inc.: Red Hook, NY, USA, 2024; pp. 22483–22512. [CrossRef]
26. Gozzi, M.; Di Maio, F. Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts. *Electronics* **2024**, *13*, 4712. [CrossRef]
27. Vaira, L.A.; Lechien, J.R.; Abbate, V.; Gabriele, G.; Frosolini, A.; De Vito, A.; Maniaci, A.; Mayo-Yáñez, M.; Boscolo-Rizzo, P.; Saibene, A.M.; et al. Enhancing AI chatbot responses in health care: The SMART prompt structure in head and neck surgery. *OTO Open* **2025**, *9*, e70075. [CrossRef] [PubMed]
28. Geroimenko, V. *The Essential Guide to Prompt Engineering: Key Principles, Techniques, Challenges, and Security Risks*; SpringerBriefs in Computer Science; Springer Nature: Cham, Switzerland, 2025. [CrossRef]
29. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *J. Clin. Epidemiol.* **2021**, *134*, 178–189. [CrossRef] [PubMed]
30. Ghimire, P.; Kim, K.; Stentz, T.; Roy, T. Modular Chain-of-Thought (CoT) for LLM-Based Conceptual Construction Cost Estimation. *Buildings* **2026**, *16*, 396. [CrossRef]
31. Bharti, P.K.; Panchal, M.; Dalal, V.; Agarwal, M.; Ekbal, A. Not all peer reviews are significant: A dataset of exhaustive vs. trivial scientific peer reviews leveraging chain-of-thought reasoning. *Scientometrics* **2026**, *131*, 2261–2301. [CrossRef]
32. Giannilias, T.; Papadakis, A.; Nikolaou, N.; Zahariadis, T. Classification of Hacker's Posts Based on Zero-Shot, Few-Shot, and Fine-Tuned LLMs in Environments with Constrained Resources. *Future Internet* **2025**, *17*, 207. [CrossRef]
33. Zhen, H.; Shi, Y.; Huang, Y.; Yang, J.J.; Liu, N. Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference. *Computers* **2024**, *13*, 232. [CrossRef]
34. Zhang, W.; Wu, M.; Zhou, L.; Shao, M.; Wang, C.; Wang, Y. A sepsis diagnosis method based on Chain-of-Thought reasoning using Large Language Models. *Biocybern. Biomed. Eng.* **2025**, *45*, 269–277. [CrossRef]
35. Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. Chain of Thought Utilization in Large Language Models and Application in Nephrology. *Medicina* **2024**, *60*, 148. [CrossRef] [PubMed]
36. Indira Kumar, A.K.; Sthanusubramoniani, G.; Gupta, D.; Nair, A.R.; Alotaibi, Y.A.; Zakariah, M. Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches. *Egypt. Inform. J.* **2025**, *30*, 100683. [CrossRef]
37. Razumovskaia, E.; Vulić, I.; Korhonen, A. Analyzing and Adapting Large Language Models for Few-Shot Multilingual NLU: Are We There Yet? *Trans. Assoc. Comput. Linguist.* **2025**, *13*, 1096–1120. [CrossRef]
38. Meshkin, H.; Zirkle, J.; Arabidarrehdor, G.; Chaturbedi, A.; Chakravartula, S.; Mann, J.; Thrasher, B.; Li, Z. Harnessing large language models' zero-shot and few-shot learning capabilities for regulatory research. *Brief. Bioinform.* **2024**, *25*, bbae354. [CrossRef] [PubMed]
39. Viswanathan, V.; Gashteovski, K.; Lawrence, C.; Wu, T.; Neubig, G. Large language models enable few-shot clustering. *Trans. Assoc. Comput. Linguist.* **2024**, *12*, 321–333. [CrossRef]
40. Golnari, P.; Prantzalos, K.; Hood, V.; Meskis, M.A.; Isom, L.L.; Wilcox, K.; Parent, J.M.; Lal, D.; Lhatoo, S.D.; Goodkin, H.P.; et al. Ontology accelerates few-shot learning capability of large language model: A study in extraction of drug efficacy in a rare pediatric epilepsy. *Int. J. Med. Inf.* **2025**, *201*, 105942. [CrossRef] [PubMed]
41. Agbareia, R.; Omar, M.; Zloto, O.; Glicksberg, B.S.; Nadkarni, G.N.; Klang, E. Multimodal LLMs for retinal disease diagnosis via OCT: Few-shot versus single-shot learning. *Ther. Adv. Ophthalmol.* **2025**, *17*, 25158414251340569. [CrossRef] [PubMed]
42. Shah, K.; Xu, A.Y.; Sharma, Y.; Daher, M.; McDonald, C.; Diebo, B.G.; Daniels, A.H. Large Language Model Prompting Techniques for Advancement in Clinical Medicine. *J. Clin. Med.* **2024**, *13*, 5101. [CrossRef] [PubMed]
43. Qi, X.; Yang, B.; Wang, S.; Zhang, Z.; Zhang, Y.; Du, K. Few-Shot and Chain-of-Thought Prompting for Equipment Maintenance Knowledge Graph Construction Via Large Language Models. *SSRN* **2025**, *335*, 115266. [CrossRef]
44. Šuster, S.; Baldwin, T.; Verspoor, K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Res. Synth. Methods* **2024**, *15*, 988–1000. [CrossRef] [PubMed]
45. Byra, M.; Rachmadi, M.F.; Skibbe, H. Few-shot medical image classification with simple shape and texture text descriptors using vision-language models. *Bull. Pol. Acad. Sci. Tech. Sci.* **2025**, *73*, 153838. [CrossRef]
46. Wan, T.; Chen, Z. Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Phys. Rev. Phys. Educ. Res.* **2024**, *20*, 010152. [CrossRef]
47. Lee, G.-G.; Latif, E.; Wu, X.; Liu, N.; Zhai, X. Applying large language models and chain-of-thought for automatic scoring. *Comput. Educ. Artif. Intell.* **2024**, *6*, 100213. [CrossRef]

48. Carlson, N.A.; Burbano, V. The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strateg. Manag. J.* **2026**, *47*, 699–725. [CrossRef]
49. Karam, B.A.E.; Fissaa, T.; Marghoubi, R. AI-Powered Assessment of Resistance to Change in the Context of Digital Transformation. *Int. J. Adv. Comput. Sci. Appl.* **2025**, *16*, 546. [CrossRef]
50. Mao, W.; Wu, J.; Chen, W.; Gao, C.; Wang, X.; He, X. Reinforced prompt personalization for recommendation with large language models. *ACM Trans. Inf. Syst.* **2025**, *43*, 1–27. [CrossRef]
51. Chen, Y.; Huang, Y.; Chen, X.; Shen, P.; Yun, L. GPTVD: Vulnerability detection and analysis method based on LLM's chain of thoughts. *Autom. Softw. Eng.* **2026**, *33*, 3. [CrossRef]
52. Darwiyanto, E.; Gusnaen, R.A.; Nurtantyana, R. A comparative study of large language models with chain-of-thought prompting for automated program repair. *IAES Int. J. Artif. Intell.* **2025**, *14*, 4579–4589. [CrossRef]
53. Yang, G.; Zhou, Y.; Chen, X.; Zhang, X.; Zhuo, T.Y.; Chen, T. Chain-of-thought in neural code generation: From and for lightweight language models. *IEEE Trans. Softw. Eng.* **2024**, *50*, 2437–2457. [CrossRef]
54. Teng, S.; Liu, J.; Jain, R.K.; Chai, S.; Hou, R.; Tateyama, T.; Lin, L.; Chen, Y.W. Enhancing depression detection with chain-of-thought prompting: From emotion to reasoning using large language models. In Proceedings of the 2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Copenhagen, Denmark, 14–18 July 2025; pp. 1–6. [CrossRef] [PubMed]
55. Jeon, S.; Kim, H.-G. A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Comput. Biol. Med.* **2025**, *196*, 110614. [CrossRef] [PubMed]
56. Chen, X.; Chen, Z.; Cheng, S. CoTHSSum: Structured long-document summarization via chain-of-thought reasoning and hierarchical segmentation. *J. King Saud. Univ. Comput. Inf. Sci.* **2025**, *37*, 40. [CrossRef]
57. Chen, P.; Li, Z. Length Instruction Fine-Tuning with Chain-of-Thought (LIFT-COT): Enhancing Length Control and Reasoning in Edge-Deployed Large Language Models. *Electronics* **2025**, *14*, 1662. [CrossRef]
58. Filienko, D.; Wang, Y.; Jazmi, C.E.; Xie, S.; Cohen, T.; De Cock, M.; Yuwen, W. Toward large language models as a therapeutic tool: Comparing prompting techniques to improve GPT-delivered problem-solving therapy. *AMIA Annu. Symp. Proc.* **2025**, *2024*, 417–426. [PubMed]
59. Bukhary, N.; Ahmad, M.; Rashad, K.; Rai, S.; Shapsough, S.; Kaddoura, Y.; Dghaym, D.; Zualkernan, I. Few-Shot Evaluation of Vision Language Models for Detecting Visual Defects in Autonomous Vehicle Software Requirement Specifications. *IEEE Access* **2025**, *13*, 117914–117942. [CrossRef]
60. Köksal, A.; Alatan, A.A. SAMChat: Introducing Chain-of-Thought Reasoning and GRPO to a Multimodal Small Language Model for Small-Scale Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2026**, *19*, 795–804. [CrossRef]
61. Gu, X.; Chen, X.; Lu, P.; Li, Z.; Du, Y.; Li, X. AGCVT-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning. *Eng. Appl. Artif. Intell.* **2024**, *132*, 107907. [CrossRef]
62. Thanasi-Boçe, M.; Hoxha, J. From ideas to ventures: Building entrepreneurship knowledge with LLM, prompt engineering, and conversational agents. *Educ. Inf. Technol.* **2024**, *29*, 24309–24365. [CrossRef]
63. Jaradat, S.; Elhenawy, M.; Nayak, R.; Paz, A.; Ashqar, H.I.; Glaser, S. Multimodal Data Fusion for Tabular and Textual Data: Zero-Shot, Few-Shot, and Fine-Tuning of Generative Pre-Trained Transformer Models. *AI* **2025**, *6*, 72. [CrossRef]
64. Yin, M.; Guo, M. Complex forecasting and investment strategy optimization via chain-of-thought of large language models. *Expert Syst. Appl.* **2026**, *298*, 129913. [CrossRef]
65. Shafikuzzaman, M.; Islam, M.R.; Zaman, S.; Ma, A.; Sifat, A.I. On the Effectiveness of Zero-Shot and Few-Shot Pretrained Language Models for Software Requirement Classification. *IEEE Access* **2025**, *13*, 159439–159453. [CrossRef]
66. Lee, A.V.Y.; Teo, C.L.; Tan, S.C. Prompt Engineering for Knowledge Creation: Using Chain-of-Thought to Support Students' Improvable Ideas. *AI* **2024**, *5*, 1446–1461. [CrossRef]
67. Ono, D.; Dickson, D.W.; Koga, S. Evaluating the efficacy of few-shot learning for GPT-4Vision in neurodegenerative disease histopathology: A comparative analysis with convolutional neural network model. *Neuropathol. Appl. Neurobiol.* **2024**, *50*, e12997. [CrossRef] [PubMed]
68. Choi, J. Binary or Graded, Few-Shot or Zero-Shot: Prompt Design for GPTs in Relevance Evaluation. *Adv. Artif. Intell. Mach. Learn.* **2024**, *4*, 2687–2702. [CrossRef]
69. Hebenstreit, K.; Praas, R.; Kiesewetter, L.P.; Samwald, M. A comparison of chain-of-thought reasoning strategies across datasets and models. *PeerJ Comput. Sci.* **2024**, *10*, e1999. [CrossRef] [PubMed]
70. Ayad, S.; Alsayoud, F. Prompt engineering techniques for semantic enhancement in business process models. *Bus. Process Manag. J.* **2024**, *30*, 2611–2641. [CrossRef]
71. Jovic, M.; Papakonstantinidis, S.; Kirkpatrick, R. From red ink to algorithms: Investigating the use of large language models in academic writing feedback. *Lang. Test. Asia* **2025**, *15*, 59. [CrossRef]
72. Tocev, T.; Atanasovski, A. AI Chatbot as IFRS Advisory Tool: GPT-4 Experimental Design. *Intell. Syst. Account. Financ. Manag.* **2026**, *33*, e70031. [CrossRef]

73. Jiang, Y.; De Raedt, M.; Deleu, J.; Demeester, T.; Develder, C. Few-shot out-of-scope intent classification: Analyzing the robustness of prompt-based learning. *Appl. Intell.* **2024**, *54*, 1474–1496. [CrossRef]
74. Du, J.; Li, B.; Chen, Z.; Shen, L.; Liu, P.; Ran, Z. Knowledge-Augmented Zero-Shot Method for Power Equipment Defect Grading with Chain-of-Thought LLMs. *Electronics* **2025**, *14*, 3101. [CrossRef]
75. Cheung, K.S. Real Estate Insights Unleashing the potential of ChatGPT in property valuation reports: The 'Red Book' compliance Chain-of-thought (CoT) prompt engineering. *J. Prop. Investig. Financ.* **2024**, *42*, 200–206. [CrossRef]
76. Qiao, J.; Li, S.; Liu, J.; Yu, H.; Xiao, Y.; Yu, H.; Zheng, Y. Med-SCoT: Structured chain-of-thought reasoning and evaluation for enhancing interpretability in medical visual question answering. *Comput. Med. Imaging Graph.* **2025**, *126*, 102659. [CrossRef] [PubMed]
77. Zhang, Z.; Yao, Y.; Zhang, A.; Tang, X.; Ma, X.; He, Z.; Wang, Y.; Gerstein, M.; Wang, R.; Zhao, H.; et al. Igniting Language Intelligence: The Hitchhiker's Guide from Chain-of-Thought Reasoning to Language Agents. *ACM Comput. Surv.* **2025**, *57*, 1–39. [CrossRef]
78. Hlaing, N.H.M.M.; Park, K.; Hahn, S.; Lee, S.Y.; Yeo, I.-S.L.; Lee, J.-H. Chain-of-Thought reasoning versus linguistic optimization for artificial intelligence models on the prosthodontics section of a dental licensing examination. *J. Prosthet. Dent.* **2026**, *135*, 394.e1–394.e8. [CrossRef] [PubMed]
79. Park, J.; Sim, W.S.; Yu, J.Y.; Park, Y.R.; Lee, Y.H. Evaluation of Context-Aware Prompting Techniques for Classification of Tumor Response Categories in Radiology Reports Using Large Language Model. *J. Imaging Inform. Med.* **2025**, *39*, 2782–2793. [CrossRef] [PubMed]
80. Liao, H.; Kong, H.; Wang, B.; Wang, C.; Wang, K.Y.; He, Z.; Xu, C.; Li, Z. CoT-Drive: Efficient motion forecasting for autonomous driving with LLMs and chain-of-thought prompting. *IEEE Trans. Artif. Intell.* **2026**, *7*, 625–641. [CrossRef]
81. Mayring, P. Qualitative Content Analysis: Theoretical Background and Procedures. In *Advances in Mathematics Education*; Bikner-Ahsbahs, A., Knipping, C., Presmeg, N., Eds.; Springer: Dordrecht, The Netherlands, 2015; pp. 365–380. [CrossRef]
82. Hsieh, H.-F.; Shannon, S.E. Three Approaches to Qualitative Content Analysis. *Qual. Health Res.* **2005**, *15*, 1277–1288. [CrossRef] [PubMed]
83. Stratton, S.J. Population Research: Convenience Sampling Strategies. *Prehospital Disaster Med.* **2021**, *36*, 373–374. [CrossRef] [PubMed]
84. Jager, J.; Putnick, D.L.; Bornstein, M.H., II. More Than Just Convenient: The Scientific Merits of Homogeneous Convenience Samples. *Monogr. Soc. Res. Child Dev.* **2017**, *82*, 13–30. [CrossRef] [PubMed]
85. Abeysinghe, B.; Circi, R. The challenges of evaluating LLM applications: An analysis of automated, human, and LLM-based approaches. In Proceedings of the First Workshop Large Language Models for Evaluation in Information Retrieval, Washington, DC, USA, 18 July 2024. Available online: https://ceur-ws.org/Vol-3752/paper1.pdf (accessed on 17 June 2026).
86. Huang, Y.; Ma, T.; Yang, K.; Zhang, Z. FinSent-DistillQ: A distilled large language model with chain-of-thought fine-tuning for financial sentiment analysis. *J. Intell. Inf. Syst.* **2026**, *64*, 735–771. [CrossRef]
87. Gibreel, O.; Arpaci, I. Development and validation of the prompt engineering competence scale (PECS). *Inf. Dev.* **2025**, 02666669251336455. [CrossRef]
88. Garg, A.; Nisumba Soodhani, K.; Rajendran, R. Enhancing data analysis and programming skills through structured prompt training: The impact of generative AI in engineering education. *Comput. Educ. Artif. Intell.* **2025**, *8*, 100380. [CrossRef]
89. Tassoti, S. Assessment of Students Use of Generative Artificial Intelligence: Prompting Strategies and Prompt Engineering in Chemistry Education. *J. Chem. Educ.* **2024**, *101*, 2475–2482. [CrossRef]
90. Vilakati, S. Prompt engineering for accurate statistical reasoning with large language models in medical research. *Front. Artif. Intell.* **2025**, *8*, 1658316. [CrossRef] [PubMed]