LUISS

Degree Program in Marketing

Course of Business and Marketing Analytics

Generative AI Prompt Enhancement Techniques in Business: A Systematic Review and Empirical Evaluation

Prof. Andrea De Mauro

SUPERVISOR

Prof. POZHARLIEV RUMEN IVAYLOV

CO-SUPERVISOR

Alessia Cantini Matr. 780841

CANDIDATE

Academic Year 2024/2025

Table of Contents

List of Figures and Tables

Introduction

Artificial Intelligence plays an increasingly prominent role in modern organisations. Evidence indicates that AI is already widespread in workplaces: 58% of employees use AI at work, with 31% doing so daily (University of Melbourne & International, 2025). A study by the University of Melbourne shows that Generative AI tools are the most widely used solutions among employees, who note benefits such as higher efficiency, easier access to information and better work quality. However, widespread adoption also introduces significant risks. Over half of employees have made mistakes due to AI, primarily because of incorrect or hallucinated outputs and misinterpretations of AI-generated content (University of Melbourne & International, 2025). Notably, employees who receive AI training benefit more, achieving higher efficiency gains and greater involvement in revenue-generating tasks than untrained colleagues (University of Melbourne & International, 2025). Research describes the diffusion of AI as a structural shift in work, while simultaneously highlighting the gap between rapid adoption and organisational readiness (Mayer et al., n.d.), underscoring the importance of mechanisms that enable users to interact with generative AI systems more effectively. Generative AI is transforming work by automating and augmenting a substantial share of employees' tasks, with existing technologies capable of affecting 60-70% of current work activities (Chui et al., 2023). In this context, prompt engineering, defined as the craft of creating effective queries to improve AI responses, is crucial for maximising the effectiveness of generative AI systems (Sikha, 2023). 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 (Bozkurt, 2024).

The existing literature mainly investigates prompting techniques in domain-specific settings and highly technical tasks. For instance, Setyo Nugroho & Shaferi (2025) analyse how prompt engineering affects the quality of ChatGPT's stock recommendations in the Indonesian energy sector. J. Wang (2024) examines how prompting strategies, when combined with rhetorical analysis, can enhance the effectiveness of AI-generated business communication. Some studies also provide rigorous comparisons of prompting techniques, but such research is mostly limited to software engineering tasks, coding, structured data, and multimodal benchmarks (Elnashar et al., 2025; Mohanty et al., 2025; Santana et al., 2025; Tony et al., 2025). Therefore, there remains a gap in the literature for a systematic review of these techniques, focusing on their impact on output quality and practical usefulness across industries. Despite an expanding body of research on prompting methods, most studies focus on technical or domain-specific tasks, offering limited guidance for those applying Generative AI in everyday business activities. Additionally, Panneer Selvam Viswanathan (2025) notes that cross-domain applications present particularly promising opportunities for advancing prompt engineering research. Our study aims to develop an operational framework of prompting techniques to optimise routine real-world business queries. We evaluate the utility of a subset of prompt enhancement methods when applied to tasks within an enterprise setting. Supported by the University of Melbourne & International, evidence suggests that

users who receive AI-related training are more likely to benefit from AI applications. Building on this, our research investigates whether users' familiarity with AI, indicated by their frequency of use, influences their ability to recognise the value of different prompting strategies. Specifically, we explore whether usage frequency moderates the relationship between prompt enhancement techniques and their perceived usefulness in business-related tasks.

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?

The rest of this thesis is organised as follows. Chapter 1 introduces the theoretical framework, outlining key concepts related to Generative AI, Large Language Models and prompt engineering in business contexts. Chapter 2 describes the research design and methodology, followed by Chapter 3, which presents the results of the systematic literature review. Chapter 4 presents the empirical study's findings, while Chapter 5 examines the results and their managerial and theoretical implications. Lastly, the final chapter summarises the main contributions, limitations and suggestions for future research.

Chapter 1 Theoretical Framework

This chapter lays the framework for the study by presenting the essential concepts needed to grasp the role of Generative AI and prompt engineering in business settings. It begins with an overview of the technical principles behind Large Language Models, explores their practical uses within organisations and concludes with an introduction to prompt engineering as a key mechanism of human-AI interaction.

1.1 Generative AI and Large Language Models (LLMs)

Artificial intelligence (AI) is a broad term that refers to computational algorithms designed to replicate aspects of human intelligence, such as learning, reasoning, and decision-making. As illustrated in Figure 1.1, one of the main subfields of AI is Machine Learning (ML), which enables systems to improve their performance by learning patterns directly from data rather than relying on explicitly programmed rules. In machine learning, three primary learning paradigms are commonly distinguished: supervised, unsupervised and reinforcement learning. Supervised learning focuses on predicting outcomes from labelled datasets, whereas unsupervised learning aims to detect latent patterns and structures in unlabelled data. Reinforcement learning, in contrast, relies on trial-and-error interactions with an environment, allowing agents to learn behaviours that maximise a cumulative reward signal (Janiesch et al., n.d.). Most traditional machine learning approaches are discriminative, as they emphasise modelling the relationship between input features and output labels to analyse, classify, or predict properties of existing data. Recent advances in machine learning have been driven by the development of neural networks, complex computational architectures inspired by the structure of the human brain and composed of multiple interconnected layers. Building on these architectures, deep learning leverages deep neural networks with multiple hidden layers to analyse high-dimensional data and automatically uncover latent representations and correlations (Banh & Strobel, 2023). This architectural dimension of deep learning enables the development of deep generative models (DGMs), which are designed to learn the underlying probability distribution of the training data. By modelling this distribution, DGMs can generate new data samples that reflect the characteristics of the original data rather than merely assigning labels or predictions to existing observations(Banh & Strobel, 2023). Therefore, a fundamental distinction arises between traditional discriminative AI models, which primarily analyse and classify existing data, and deep generative models, which focus on the probabilistic generation of new data. Despite their powerful predictive and generative abilities, deep learning models are frequently considered black-box systems because their internal processes for converting inputs into outputs are complex and difficult to interpret, posing challenges for fully explaining or tracing specific decisions (Castelvecchi D., 2016).

Building on the mechanisms of deep generative modelling, Generative Artificial Intelligence refers to AI systems capable of producing new, meaningful content such as text, images or code. Within this domain, Large Language Models (LLMs) represent a prominent class of generative models that leverage deep learning architectures to generate natural language output and support interaction-based applications.

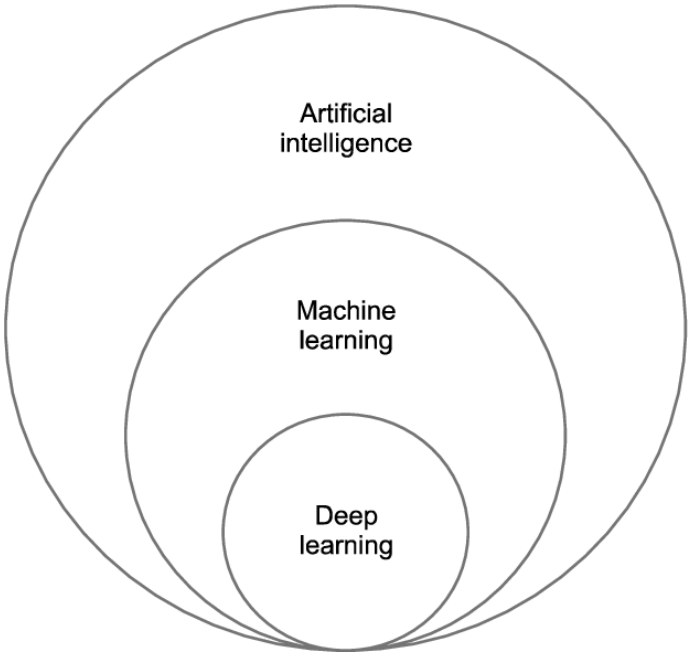

Figure 1.1 Diagram showing AI's subfields (Guney et al., 2021)

A core component of LLMs is the ability to convert human-language text into numerical representations that can be processed by neural architectures. This transformation is achieved through tokenisation and embedding mechanisms, which underpin modern LLM architectures. Textual input is first decomposed into tokens, which may correspond to words, subwords or characters depending on the tokenisation strategy employed (Naveed et al., 2025). Each token is then mapped to a dense vector representation, known as an embedding(Naveed et al., 2025). Embeddings represent tokens in a continuous vector space, in which semantic and syntactic relationships between linguistic units are captured by geometric proximity (Raiaan et al., 2024). Embeddings constitute a significant advance over earlier symbolic or discrete representations of language, as they enable models to encode semantic similarity and contextual relationships numerically. By operating in an embedding space, LLMs can generalise across linguistic contexts and capture relationships between words that are not explicitly encoded by rules (Raiaan et al., 2024). These embeddings are learned during training and are updated iteratively as the model optimises its language-modelling objective. Modern LLMs are predominantly based on transformer architectures that process embedded tokens via self-attention. Self-attention allows the model to assign different weights to tokens within an input sequence based on their mutual relevance, enabling the representation of both local and long-range dependencies in text. Unlike sequential architectures, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), transformers process input tokens in parallel, thereby improving scalability and enabling efficient training on large datasets (Naveed et al., 2025). LLMs are typically trained using self-supervised objectives, such as autoregressive next-token prediction or masked language modelling. Through these objectives, the model learns to predict missing or subsequent tokens from contextual information encoded in embeddings and attention layers. As model scale increases in terms of parameters, data and compute, LLMs exhibit improved generalisation and emergent

capabilities, including in-context learning and zero-shot task performance. AI applications rely on this embedding-based and attention-driven architecture to generate responses conditioned on user-provided input. User prompts are processed as text, converted into embeddings, and integrated into the model's context window, where they influence token-level predictions during generation. Because the internal representations and decision processes of LLMs are not directly observable, interaction with these systems occurs entirely through natural-language input and output (Raiaan et al., 2024).

1.2 GenAI in Business Contexts

The rapid spread of Generative Artificial Intelligence (GenAI) has led to its increased adoption within organisations, mainly to support knowledge-intensive and task-oriented activities. Unlike traditional automation tools, GenAI systems don't merely follow predefined procedures; they can generate new outputs that assist users with reasoning, content creation, and decision-making (Fui-Hoon Nah et al., 2023). From a business and information systems perspective, GenAI should be viewed as a socio-technical instrument integrated into organisational workflows, rather than merely as an independent analytical tool. Feuerriegel et al. (2024) describe GenAI at the model, system, and application levels, stressing that its business value arises from the interaction of technical features, system design, and organisational use cases. This view underscores that GenAI's success in organisations hinges not only on model accuracy but also on how effectively systems are integrated into existing processes and on how users interact with them in practice. Existing research on GenAI in business contexts has mainly concentrated on specific functional areas and narrowly defined tasks. Prior studies have investigated applications in fields such as marketing, customer service, education, healthcare, and software development, often assessing performance within narrowly defined use cases (Fui-Hoon Nah et al., 2023). While these task- and sector-focused approaches provide useful insights into localised applications, they tend to highlight outcomes specific to individual domains rather than cross-sector usage patterns. Consequently, the literature provides only a limited understanding of how GenAI systems are used across organisational functions and how similar interaction patterns can emerge independently of the application area. Feuerriegel et al. (2024) note that many studies focus on models or applications, making it difficult to compare findings across research and hindering the discovery of general mechanisms by which GenAI adds value within information systems. A key characteristic of GenAI in organisational settings is its role in AI-human collaboration. Rather than replacing human decision-makers, GenAI systems are generally seen as collaborative tools that enhance human capabilities by supporting ideation, problem-solving, and information synthesis. Fui-Hoon Nah et al. (2023) highlight that the success of GenAI applications depends on how humans and AI systems interact, with users actively guiding, validating, and contextualising AI outputs. This collaborative dynamic emphasises the need for interaction mechanisms that enable users to clearly communicate their intentions, constraints, and expectations to the system.

Despite their potential advantages, GenAI systems present notable challenges in business environments, including reliability, hallucinations, bias, data privacy, and governance. These risks are especially significant

in organisational settings, where inaccurate or misleading results can harm decision-making and operational performance (Fui-Hoon Nah et al., 2023). Therefore, organisations need to balance efficiency improvements with proper oversight and responsible management of GenAI. Overall, assessing GenAI's value in business contexts requires looking beyond technical performance metrics alone. Since these systems are interaction-driven and collaborative, their value should be evaluated by how well they support users in completing tasks. This viewpoint encourages a usage-focused analytical approach that moves beyond specific domains and focuses on how GenAI is used across organisations, offering more broadly applicable insights into its business benefits. Given that Generative AI systems in organisations operate through user interactions, it is crucial to understand how users convey tasks, constraints and expectations to these models. This communication mainly occurs via prompts, which will be discussed in the next section.

1.3 Prompt Engineering

After outlining the technical foundations and organisational importance of Generative AI, this section emphasises prompt engineering as the main way users engage with Large Language Model practice. In systems based on Large Language Models (LLMs), user interaction mainly occurs through prompts, which are natural language inputs that direct the model's behaviour and outputs. Instead of changing model parameters via retraining or fine-tuning, prompt engineering allows users to tailor pre-trained models for various tasks by carefully designing task instructions and contextual cues (Sahoo et al., 2025). Conceptually, prompt engineering involves the iterative process of creating, refining and assessing prompts to achieve the desired model responses. Schulhoff et al. (2025) stress that it is not a one-time task but a cyclical process that includes formulating prompts, evaluating outputs and adjusting, often repeatedly, until responses align with task-specific needs. This view portrays prompt engineering more as an interaction design than just a technical optimisation method. A core feature of prompt engineering is its independence from specific tasks. Unlike traditional machine learning pipelines that require training or fine-tuning models for each task, prompt engineering enables a single base model to handle diverse tasks by adjusting the prompts' structure and content. Sahoo et al. (2025) demonstrate that prompting techniques can be used across a wide range of applications, including question answering, summarisation, reasoning and code generation, all without modifying the model’s internal parameters. Simultaneously, current research emphasises that LLMs are highly sensitive to prompts, with small changes in wording, structure, or format leading to large variations in performance. Polo et al. (2024) show that LLM performance varies substantially across different prompt templates and suggest that assessing models with a single prompt can be misleading. This sensitivity highlights the importance of prompt engineering in guiding both model behaviour and evaluation results. Prompt engineering strategies vary widely in complexity and scope. For example, Gozzi & Di Maio (2024) compare single-task and multi-task prompts, demonstrating that no single approach is universally best: the success of a prompt depends on how prompt design, task features and model architecture interact. These insights highlight that prompt engineering results rely on context rather than fixed rules. Beyond experimental and benchmarking scenarios,

prompt engineering is vital in real-world productivity and knowledge tasks. Anam (2025) observes that users who employ more structured, contextual and specific prompts perceive greater usefulness and efficiency of LLMs in daily work and learning. This indicates that prompt engineering is not just a technical issue for researchers and developers but also an emerging skill for users within organisations. The literature considers prompt engineering as a key aspect of human-AI interaction in generative AI systems. Its success relies on prompt design, task presentation, and how users tweak prompts in response to model feedback. Nonetheless, although many prompting methods have been suggested, the field still suffers from inconsistent terminology and varying conceptual approaches (Sahoo et al., 2025; Schulhoff et al., 2025).

While prompt engineering describes the overall interaction process, the literature has identified recurring structural patterns that can be systematically analysed and compared. These patterns, referred to as prompt enhancement techniques, are introduced in the following section.

1.3.1 Prompt Enhancement Techniques

Prompting techniques are structured methods for creating prompts that steer large language models toward specific outputs. Although prompt engineering covers the overall process of designing and improving prompts, prompting techniques refer to common structural patterns seen across various tasks and scenarios (Geroimenko, n.d.). Their success relies on how well they communicate user intent and task details, since language models generate answers based on learned statistical patterns and cannot reliably infer unstated assumptions (Geroimenko, n.d.). In this context, some basic rules support effective prompt design: prompts should be clear and explicit, avoiding vague or underspecified instructions. Including relevant context can enhance output relevance, but too much or irrelevant information can impede clarity, underscoring the need for balance. Clearly defining output expectations, such as format or detail level, can also improve consistency and usefulness (Geroimenko, n.d.). Prompting techniques can also involve small changes in wording or structure that yield different results, reflecting the probabilistic nature of large language models. In practice, these strategies tend to be somewhat personal, leading to the development of structured taxonomies for systematically comparing and evaluating prompting methods across different contexts.

Chapter 2 Research Design and Methodology

This chapter explains the research design and methodological decisions made to answer the two research questions. It is organized into two main sections. The first part covers the systematic literature review conducted for RQ1, including paper selection, data extraction and qualitative content analysis used to develop the taxonomy of prompting techniques. The second part describes the empirical study for RQ2, detailing the experimental setup, prompt formulation, methods for measuring perceived usefulness, sampling approach, and statistical analysis.

2.1 Overall Research Design

This study employs a two-stage research approach. RQ1 is explored via a systematic literature review and a qualitative content analysis, focusing on identifying and categorising prompting techniques deemed useful in business settings. RQ2 is examined through an empirical study that assesses the effectiveness of different prompting techniques and their combinations in real-world business prompts.

2.2 Systematic Literature Review

To address RQ1, the proposed research conducted a systematic review in accordance with the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) Guidelines. A systematic review is a comprehensive method for collecting, synthesising, and analysing multiple findings in a consistent and unbiased manner. It is characterised by strict features that define the scope of research synthesis, beginning with clearly articulated research questions. These questions are essential because well-structured questions enhance research efficiency by optimising the time and resources needed to find relevant evidence. The scope of the question also influences the quality of conclusions: overly narrow or broad questions can limit the applicability of the results to different populations. After establishing the research foundation, data collection and assessment proceed according to a protocol, which ensures the process can be replicated. Several protocols are documented in the literature; organisations such as Cochrane have developed their own to guide researchers and to prevent selection bias (Pollock & Berge, 2018). The PRISMA guidelines, first published in 2009, promote transparency regarding research scope, process, and findings. Originally developed for health-related systematic reviews with meta-analyses, PRISMA's principles now apply more broadly across various interventions and reviews (Page et al., 2021). The 2020 update features a 27-item checklist, including recommendations for each item and its sub-items, an abstract checklist, and a flow diagram. This approach identifies four phases in the paper selection process: identification, screening, eligibility assessment, and inclusion in the synthesis. The data extracted from the final selection of papers were codified and studied through a Qualitative Content Analysis.

The resulting synthesis is a taxonomy of prompting techniques classified according to the primary reported effect. The objective of the review is not to compare performance across techniques, but to identify which prompting techniques are reported to be useful in business contexts.

2.2.1 Paper Identification and Screening

Literature retrieval was carried out using two academic databases, Scopus and Web of Science, selected for their broad coverage of peer-reviewed research in artificial intelligence and applied domains. An initial, exploratory search was performed to investigate how prompt engineering has been studied within business and marketing contexts. This first search employed the following query:

("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")

Although this query yielded a relatively small set of results, it proved valuable for this study. Specifically, 11 papers included in the final review were identified through this search, and these works provided important qualitative insights into how prompt engineering techniques are currently applied in business settings. Moreover, this exploratory phase enabled the identification of the most used prompting techniques and the types of effects typically reported, such as improvements in relevance, clarity, and decision support. The 11 studies identified through the exploratory search also met the final eligibility criteria and were therefore retained in the final sample. However, as the review progressed, it became evident that restricting the search exclusively to business-related applications would significantly limit the analysis. Many empirical studies investigating the same prompting techniques examine their effects in other domains, often with greater methodological depth or more systematic evaluation. Since the primary objective of this review is to analyse the effects, improvements, and limitations of prompting techniques themselves, rather than their performance within a single application domain, the search strategy was therefore broadened. Building on the insights gained from the exploratory phase, a second and final search query was formulated to capture empirical studies evaluating prompting techniques across a wider range of tasks and contexts, while maintaining a focus on output quality and effectiveness:

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")

This broader search strategy allowed for the inclusion of studies analysing prompting techniques across heterogeneous domains and experimental settings. Given the diversity of tasks, evaluation metrics and methodological approaches observed in the retrieved literature, the findings of this review were synthesised qualitatively to identify recurring patterns, trade-offs and contextual factors influencing the effectiveness of different prompting techniques.

Based on the selected protocol, the eligibility criteria were established 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 conceptualise prompting as a methodological element, by defining, analysing, 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 optimisation without substantive analysis of prompting techniques.

Figure 2.1 presents the PRISMA flow diagram detailing the study selection process (Haddaway, 2022) . A total of 317 records were initially identified across the selected databases. Of these, 130 papers were excluded prior to screening due to evident irrelevance to the scope of the review. In particular, a substantial number of studies used the term “zero-shot” to describe scenarios in which newly trained models are evaluated on unseen tasks or classes without task-specific supervision. This interpretation differs from zero-shot prompting as adopted in this review, where a pre-trained language model is queried at inference time without providing examples within the prompt. Since the focus of this review is on prompt design rather than model training or generalization, such studies were excluded.

After removing duplicates, the remaining records were screened based on title and abstract, resulting in the exclusion of 115 additional studies that did not meet the inclusion criteria. Many of these studies employed large language models through prompting but treated prompts merely as an input mechanism, without analysing prompt design, structure, comparison or evaluation. In several cases, the primary contribution concerned model fine-tuning or model architectural modifications, with prompting playing only a secondary role. However, studies proposing automated or adapted prompting approaches were included when the original

authors explicitly isolated the underlying baseline prompting techniques and documented their evaluation separately, allowing the prompting strategy to be analysed independently from other system components, without synthesising or quantitatively comparing the reported outcomes. Screened papers were then read once full text was available, resulting in the exclusion of 5 papers. For eligibility, only findings explicitly addressing the effects of prompting techniques were included. Importantly, studies were evaluated primarily on the effects of prompting techniques on the task being performed, rather than on the specific application domain in which they were applied. As a result, studies were considered eligible if they assessed the effects of prompting on tasks such as classification, risk assessment, decision support, reasoning, or information synthesis, even when conducted in non-business domains. Conversely, studies were excluded when the reported effects were highly context-specific and could not be meaningfully disentangled from domain-dependent constraints.

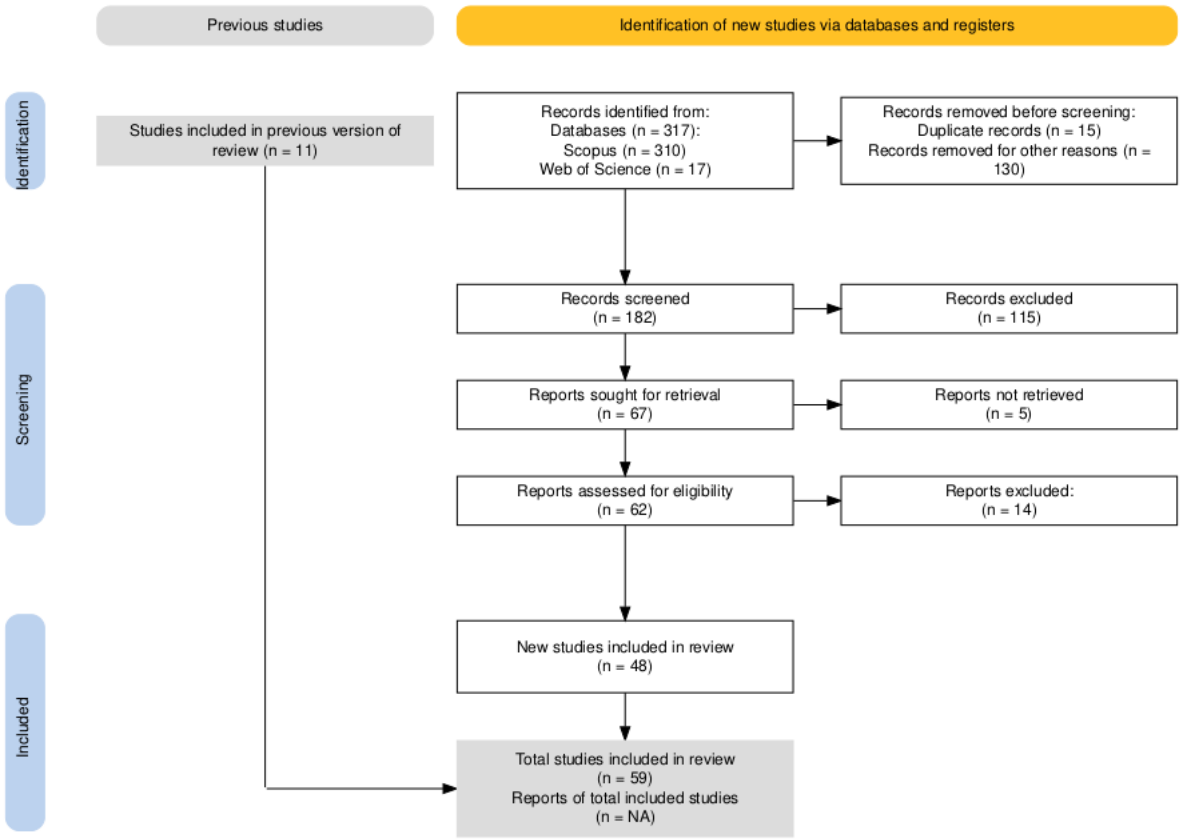

Figure 2.1 Prisma Flow Diagram (Haddaway, 2022)

2.2.2 Data Extraction

Following the eligibility assessment, data were systematically extracted from all included studies using a structured table that captured the main relevant information. The complete table is provided in the Appendix to ensure process transparency. The extraction focused on the core analytical dimensions needed to address RQ1. Specifically, the table includes fields documenting:

- The prompting technique studied

- The task or application context in which the prompting technique was evaluated
- The type of evaluation and metrics reported in the study
- The effects and key insights attributed by the authors to the prompting technique.

When a single study examined multiple prompting techniques, each was recorded as a separate observation. Consequently, a study could contribute multiple rows to the extraction table. This method enabled a detailed analysis of prompting techniques across diverse tasks and study designs, while maintaining the connection between techniques and their reported effects. Furthermore, quantitative metrics were collected to identify directional effects, when available; however, no statistical aggregation was performed due to variability across tasks and evaluation settings.

The completed extraction table, reported in Appendix A, provided the empirical foundation for the subsequent structured content analysis and for the creation of the taxonomy of prompting techniques.

2.2.3 Qualitative Content Analysis

Qualitative content analysis was used to examine and interpret textual data by systematically describing and synthesising phenomena reported in the literature (Elo et al., 2014). This method is particularly suitable when empirical evidence is fragmented or heterogeneous, as it allows categories to emerge naturally from the data rather than being predefined (Hsieh & Shannon, 2005). Therefore, this study employs an inductive approach to content analysis, aligned with established qualitative methods. Figure 2.2 shows the steps of analysis based on Mayring's framework of Inductive Content Analysis (Mayring, n.d.).

Consistent with Mayring's inductive approach, the research questions for the analysis are exploratory, focusing on discovering patterns, effects and trade-offs associated with prompt enhancement techniques rather than on testing specific hypotheses. The goal is to synthesise the use and assessment of prompting techniques across various tasks and domains, with particular attention to their reported effects on efficiency, clarity, output quality, decision support, and reliability. Before coding, we established a set of category definitions and an abstraction level to guide material selection and maintain analysis consistency. These definitions clarified that only text segments explicitly discussing prompting techniques, their contexts and reported effects were relevant. The chosen level of abstraction was designed to capture the functional effects of prompting techniques, enabling comparisons across diverse studies without fragmenting into overly specific task- or domain-focused categories. The coding process involved systematically reviewing each eligible study line by line. Relevant passages were coded and grouped into categories that closely reflected the original wording and meaning. When new passages matched existing categories, they were added accordingly; otherwise, new categories were created. This iterative approach ensured that categories were firmly based on the empirical evidence from the reviewed studies.

After coding an initial subset of the material, the emerging category system was reviewed during a pilot phase, in accordance with Mayring's recommendation to revise categories once they had reached a preliminary level of stability. In this phase, categories were assessed for conceptual coherence, alignment with research

questions and a suitable level of abstraction. When needed, category definitions were refined and overlapping or overly specific categories were merged to enhance clarity. After this revision, the entire set of eligible studies was re-examined using the finalised category system and uniform coding rules. Notably, the unit of analysis was the prompting technique itself, not the individual study. Consequently, a single paper could yield multiple observations if it assessed more than one prompting strategy. This method allowed for a more detailed comparison of techniques and helped identify recurring patterns and trade-offs across various contexts. At the conclusion of the coding process, related categories were grouped into higher-order categories where this was deemed useful for addressing the research questions. This step led to the development of a set of outcome-oriented categories that formed the basis for the taxonomy of prompt enhancement techniques presented in the results section. In line with Mayring's quality criteria, the outcome of the inductive content analysis consists of a structured category system and its interpretation with respect to the research questions.

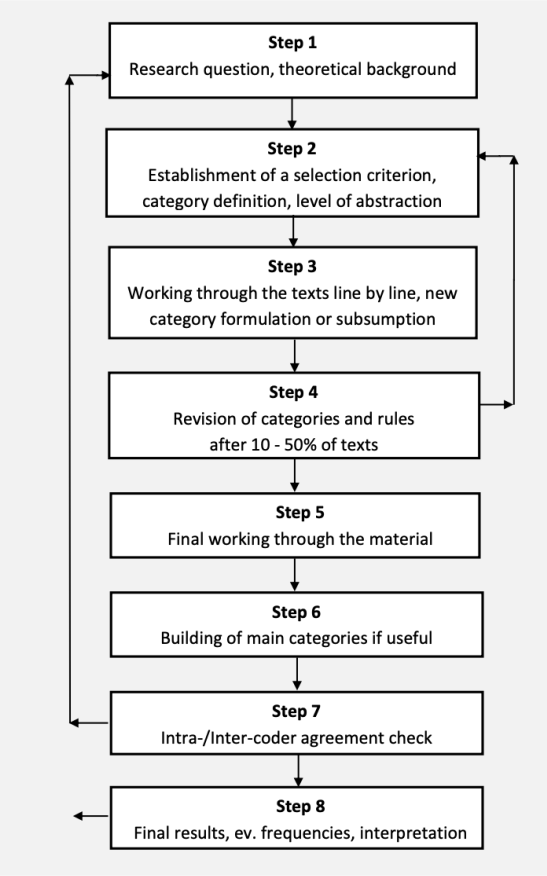

Figure 2.2 Steps of Inductive category development in Qualitative Content Analysis (Mayring, n.d.)

2.3 Experimental Design

Although the systematic literature review concentrates on isolated prompting techniques for comparability, real-world business use of generative AI often involves combining multiple strategies. Most prior research has evaluated prompting methods in controlled, isolated settings, creating a gap in understanding their interactions in actual organisational environments. To fill this gap, an empirical study was carried out. It explores how individual prompting strategies and their combinations affect the perceived usefulness of GenerativeAI

prompts and outputs for common business tasks. The study specifically employs human evaluation to assess perceived usefulness and decision-making support in practical organisational contexts.

2.3.1 Selection of Prompting Techniques

A systematic literature review identified and classified the most common prompting techniques in business based on their impact on output quality. However, to measure their influence on perceived usefulness and allow for meaningful comparisons, an empirical study was necessary. The literature also reveals a gap in applying prompting techniques to real-world business tasks, specifically, tasks routinely performed by knowledge workers to support daily activities that are not industry specific. To fill this gap, preliminary data were collected via a Microsoft Forms questionnaire within a global organisation. Respondents reported common business tasks performed with generative AI tools such as ChatGPT and Copilot, along with the prompts they used. This process identified four main business use cases, listed in Table 1.

Table 1 Collection of GenAI Business Use Cases

Business Use Case	**Prompt (raw)**
Review	*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 San Paulo now.*

Initially, all prompting techniques were applied to each use case. However, to ensure robust and meaningful experimental results, the study ultimately focused on the analysis and synthesis tasks. These tasks were selected due to their higher capacity to generate diverse and informative outputs. The text-enhancement (review) task was initially included as a realistic business use case. However, preliminary testing revealed that large language models tend to converge toward highly similar outputs when asked to “improve” or “polish” a text. This behaviour is consistent with the low-entropy nature of enhancement tasks, in which the model relies on strong stylistic priors such as clarity, conciseness, and professional tone. As a result, this task produced no meaningful variation across prompting techniques and was excluded from the experimental phase. This finding

is itself relevant, as it suggests that prompting strategies have a limited impact in tasks where genre conventions and optimisation patterns strongly constrain the model's objective. Similarly, the general information retrieval task was excluded, as it primarily involves factual recall and does not meaningfully benefit from advanced prompting strategies. To ensure a feasible and methodologically sound survey design, a subset of prompting techniques was selected based on their frequency of occurrence in the literature. Limiting the number of techniques and tasks was necessary to avoid excessive cognitive load for participants and to ensure sufficient response quality. The prompting techniques investigated in this study are:

- Zero-shot prompting
- One-shot prompting
- Few-shot prompting
- Chain-of-thought prompting
- One-shot prompting with Chain-of-thought prompting
- Few-shot prompting with Chain-of-thought prompting

2.3.2 Business Prompts Collection and Prompt Construction

After defining the prompting techniques and business use cases, raw prompts were created from real-world examples. Each prompt was then systematically improved by applying each technique separately and in selected combinations. These enhanced prompts were run independently with Microsoft Copilot, with each prompt in a separate conversation to prevent biases from conversation history. This approach ensured comparable conditions for each output. In total, the process yielded 12 AI-generated outputs, six for each task (analysis and synthesis), covering different prompting techniques and their combinations.

2.3.3 Conceptual Framework and Procedure

This study's conceptual framework is illustrated in Figure 2.3. The framework explores how prompting enhancement techniques influence perceived usefulness, considering how often participants use generative AI as a moderating factor. Prompting techniques are defined as structured design choices applied to prompts to guide AI behaviour and enhance output quality. Perceived usefulness reflects users' judgment of the effectiveness and value of AI-assisted interactions. The frequency of AI use is included as a moderator to account for variations in users' exposure to and familiarity with AI systems.

The variables included in the model are defined as follows:

• Independent Variable (IV): Prompting Enhancement Techniques

• Dependent Variable (DV): Perceived usefulness

• Moderator: Frequency of Generative AI Usage

In the empirical analyses, perceived usefulness is defined in two related but distinct ways, representing different levels of user evaluation. First, perceived output usefulness refers to participants' judgments of the

quality and usefulness of the AI-generated outputs, including relevance, clarity, correctness, and the absence of hallucinations. Second, perceived prompt quality pertains to participants' assessments of the prompt itself, highlighting how effectively, informatively, and appropriately the structure and wording of the prompt guide the AI. This dual approach enables the study to distinguish between assessments of outcomes and evaluations of the prompting strategies that produce them. Specifically, while output usefulness measures the final result of human-AI interaction, prompt quality more directly reflects users' recognition of the value within various prompting techniques.

The study tests the following hypotheses:

• H1: Prompting techniques significantly influence the perceived usefulness of AI-generated outputs.

• H2: The frequency of generative AI usage moderates 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 techniques.

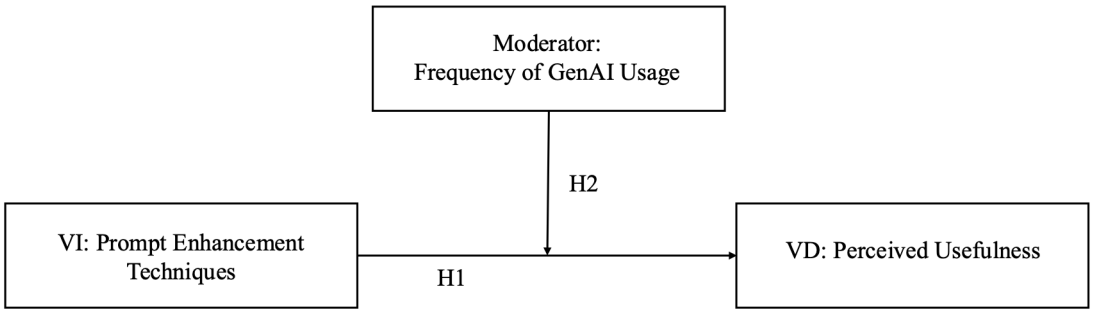

Figure 2.3 Conceptual Framework

The study employs an experimental design, with the independent variable manipulated by systematically applying different prompting techniques to the same business tasks. A human-evaluation approach was employed, using a survey to assess AI-generated outputs. A mixed-subjects design was adopted, in which each participant evaluated multiple AI-generated outputs, but not the full set. This design choice was motivated by several methodological considerations. First, it reduces participant fatigue and cognitive overload, which are particularly relevant when evaluating complex textual outputs. Second, it limits potential learning and anchoring effects that may arise if participants are exposed to all prompting techniques sequentially. Participants therefore evaluated multiple outputs, but not all available outputs, ensuring a balance between experimental control and ecological validity. Therefore, the mixed-subjects approach allows for sufficient variability in evaluations while maintaining a manageable survey length, thereby improving response quality and reliability. Participants evaluated multiple outputs, but not all available outputs, to reduce fatigue.

Figure 2.4 illustrates the survey flow, created in Qualtrics XM, which includes an introduction, demographic questions (e.g., the frequency of generative AI use), some basic instructions, and two randomised evaluation blocks, one for each task. In the first randomised block, participants were presented with two prompts, each

enhanced with a different prompting technique, along with their corresponding AI-generated outputs. These prompts and outputs were randomly selected from the six pairs generated for the analysis task. In the second randomised block, participants evaluated two additional prompt-output pairs, randomly selected from the sets generated for the synthesis task and enhanced using the same set of prompting techniques. Randomisation was uniformly distributed across participants. Overall, each participant evaluated four prompt-output pairs.

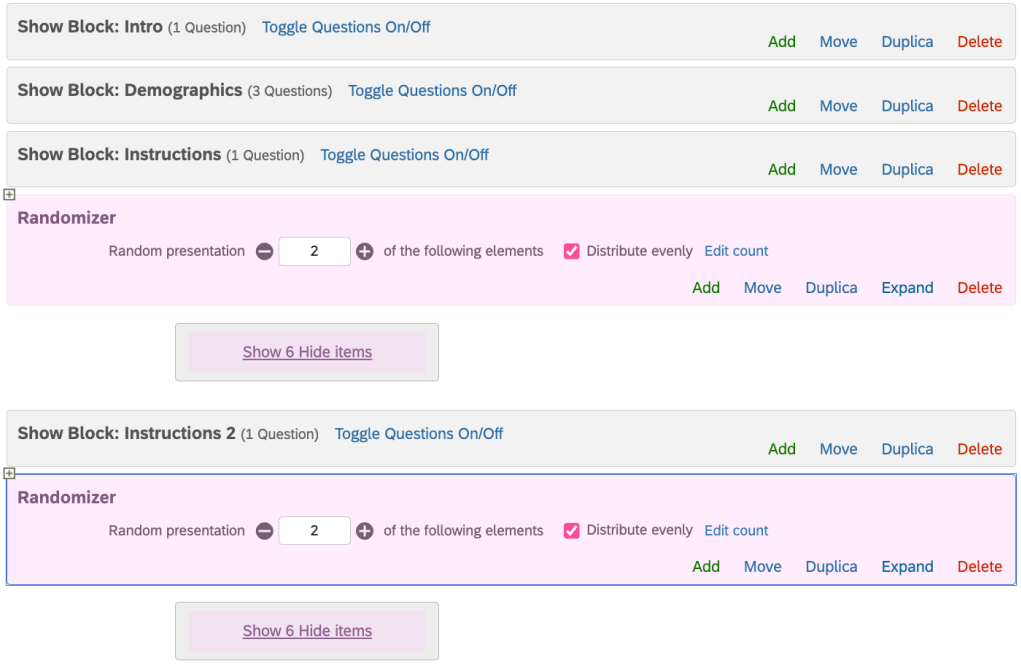

Figure 2.4 Survey Flow and experimental procedure

Below an example of a prompt-output pair showing how a prompt and its AI-generated output were displayed. In this example, the prompt requested the AI to summarise a brief text about the European Union's energy transition, followed by the original input and the resulting AI-generated summary. This example was shown to participants before the evaluation items, enabling them to base their judgments on a realistic, representative human-AI interaction scenario.

#Request to the AI:
Please summarise the text below.

#Input text included in the request:
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.

#AI-generated response:
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.

2.3.4 Measurement of Perceived Usefulness

To assess the quality of human-AI interaction, the questionnaire included a dedicated section evaluating both the prompt provided to the AI system and the textual output generated in response. The evaluation of the AI-generated output was based on a factor-based Likert-scale instrument adapted from prior research on the assessment of Large Language Model (LLM) applications (Abeysinghe & Circi, 2024a). Specifically, participants rated the output on a 5-point Likert scale ranging from 1 ("strongly disagree") to 5 ("strongly agree") with five items: relevance, informativeness, correctness, clarity and the absence of hallucinated or unsupported information. This multidimensional approach allows for a more nuanced evaluation of AI-generated content compared to a single overall quality rating.

The same measurement framework was adapted to evaluate the prompt itself. Drawing on the literature on prompt engineering, which highlights the central role of prompt formulation in shaping the accuracy, effectiveness, and usefulness of AI outputs (Anam, 2025), participants assessed whether the prompt was relevant to the task, clearly formulated, free of errors, and sufficiently informative for the AI system to generate an appropriate response. For this purpose, the hallucination-related item was excluded, as it is not applicable to prompt evaluation. All prompt-related items were rated using the same 5-point Likert scale as the output evaluation.

2.3.5 Sampling Method and Survey Distribution

The survey used a homogeneous convenience sampling method, a non-probabilistic approach in which researchers select participants based on proximity, willingness, or network connections (X. Wang, 2024). This technique is considered easy and cost-effective, especially in resource-limited settings and exploratory research (Stratton, 2021). However, it carries risks, including motivation bias, low participation, poor sample representation, and limited generalizability (Stratton, 2021). Jager et al. (2017) differentiate between conventional (heterogeneous) and homogenous convenience sampling. The former involves selecting participants based on accessibility without restricting key characteristics, leading to a more diverse sample but possibly affecting clarity in generalisation. Homogeneous convenience sampling, on the other hand, restricts

the sample to a specific subgroup, resulting in a more uniform population. While it may limit broader generalisation, this approach provides more representative samples and valid estimates, making it a preferable method when probability sampling isn't possible (Jager et al., 2017). Since our survey targeted professionals in business-related roles, we focused our recruitment strategy on reaching this group. Hence, participants were recruited through LinkedIn posts and direct emails sent to professional contacts, mostly colleagues in business settings. LinkedIn was selected as the primary channel for its professional focus and its ability to connect with individuals familiar with digital tools and generative AI. Email outreach was used to supplement online recruitment and increase response rates among professionals. Participation was voluntary and anonymous, with no financial or material incentives offered. Participants were informed about the study's academic purpose before participating.

Chapter 3 Results of Systematic Literature Review

This chapter presents the findings from the systematic literature review conducted to answer RQ1. It introduces an outcome-oriented taxonomy of prompt enhancement techniques, developed through structured content analysis. The chapter is structured into macro-categories that reflect the main contribution of each technique, with each section detailing the empirical evidence, strengths, limitations and trade-offs of each prompting approach.

3.1 Results of Prompt Enhancement Techniques in Business

Following the Structured Content Analysis, prompting techniques were organised 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 by categorising prompting techniques according to 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 used as reference points in the literature. 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 stabilising 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, organised into a two-level hierarchical structure comprising three categories and seven techniques, as illustrated in 3.1. The complete list of reviewed studies used to derive the taxonomy is reported in Appendix A.

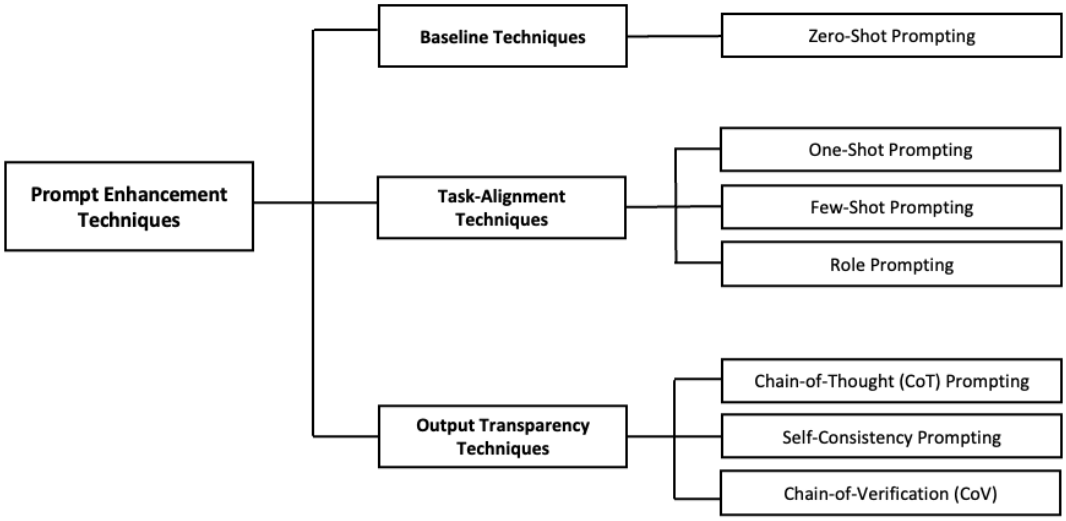

Figure 3.1 Taxonomy of Prompt Enhancement Techniques

3.2 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 emphasise speed and ease of use. Consequently, they are typically used for simple business queries or tasks where quick answers are needed and designing prompts is kept inexpensive.

3.2.1 Zero-Shot Prompting

Across the studies reviewed, zero-shot prompting is primarily treated as a baseline or comparison point rather than as an optimised 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 prioritisation, 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 (Ghimire et al., 2026). 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 (Bharti et al., 2026). 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 categorisation but yielded inconsistent results across categories and models, underscoring its role as a baseline rather than a robust method (Giannilias et al., 2025) 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 (Zhen et al., 2024) 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 provided acceptable recall but lower accuracy and F1 scores, with limited diagnostic reasoning compared to reasoning-based prompts (Zhang et al., 2025) . 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 (Miao et al., 2024).

Many studies emphasise 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 (Indira Kumar et al., 2025). In multilingual scenarios, performance was sensitive to language and domain features, often serving only as a baseline comparison (Razumovskaia et al., 2025).

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 (Meshkin et al., 2024). Similarly, in large-scale information extraction and clustering, instruction-only zero-shot prompts effectively induced structure and improved clustering performance relative to traditional baselines (Viswanathan et al., 2023). The evidence 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.

3.3 Task Alignment Prompting Techniques

Task alignment prompting techniques enhance the relevance, structure, and context-appropriateness of outputs by aligning the model's behaviour with specific task needs and user expectations. Using methods like example-based conditioning, explicit instructions, or contextual framing, these techniques minimise 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.

3.3.1 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 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 (Giannilias et al., 2025). Similarly, in medical classification using ontologies, one-shot prompting yields higher accuracy than zero-

shot prompting, as even minimal context helps the model align with task requirements (Golnari et al., 2025). However, its effectiveness is limited in complex or highly specialised 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 (Agbareia et al., 2025). In such cases, one-shot prompting mainly serves as a reference rather than a robust diagnostic method. Review articles emphasise 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 generalisation (Shah et al., 2024). 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 specialised domains, however, it remains insufficient and is mostly a baseline or intermediate strategy rather than the optimal prompting approach.

3.3.2 Few-Shot Prompting

Few-shot prompting involves supplying a small number of example cases within the prompt to steer the model's output and behaviour. Across the reviewed studies, it consistently outperforms zero-shot prompting, particularly in tasks that require domain adaptation, structured output, or fine-grained classification. Multiple empirical research reports 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 generalisation and understanding of the task (Qi et al., 2026). 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 (Bharti et al., 2026).

Few-shot prompting proves highly effective in domain-specific and technical areas. It enhances code security and robustness over zero-shot methods, though outcomes are still influenced by the type of vulnerability and chosen examples (Tony et al., 2025). 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 (Indira Kumar et al., 2025). 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 (Razumovskaia et al., 2025). 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 (Šuster et al., 2024). 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 (Agbareia et al., 2025). 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 (Byra et al., 2025). 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 (Giannilias et al., 2025). Similarly, in traffic crash severity classification, few-shot prompting mainly boosts results for smaller models, while showing mixed effects across severity levels (Zhen et al., 2024). 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 (Wan & Chen, 2024). 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. Nonetheless, without explicit reasoning mechanisms, few-shot prompting alone remains limited in interpretability, especially in complex evaluative contexts (Lee et al., 2024).

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.

3.3.3 Role Prompting

Role prompting assigns a specific professional identity or role to the language model to steer responses toward domain-relevant knowledge and perspectives (Carlson & Burbano, 2025). 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 (Abou et al., 2025; Anam, 2025). 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 (Carlson & Burbano, 2025). 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 (J. Wang, 2024). 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 prioritisation or judgment is required (Mao et al., 2025). 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.

3.4 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.

3.4.1 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 (Chen et al., 2026). 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 (Darwiyanto et al., 2025; Yang et al., 2024). In classification and decision-making contexts, CoT consistently improves reasoning quality and aligns more closely with context, although the extent of improvement varies across domains. 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 (Miao et al., 2024; Teng et al., 2025; Zhang et al., 2025).

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 (Jeon & Kim, 2025). Several studies highlight CoT's strength in long-form and complex generation tasks. In summarising long documents, CoT improves factual accuracy, structural coherence, and content coverage by enforcing stepwise reasoning prior to summarisation (X. Chen et al., 2025). 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 (P. Chen & Li, 2025).

Despite these benefits, there are limitations and trade-offs. In cost estimation and peer review, CoT enhances output reliability and interpretability but also increases interaction complexity and response times, lowering efficiency (Bharti et al., 2026; Ghimire et al., 2026). In therapeutic dialogues, it encourages more reflective responses but can reduce task effectiveness and protocol compliance, highlighting a trade-off between explainability and actionability (Filienko et al., 2024). 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 (Qi et al., 2026). In multimodal and vision-language tasks, CoT improves reasoning transparency but may also lead to hallucinations or false positives, especially in safety-critical situations (Bukhary et al., 2025). 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 (Huang et al., 2026; Köksal & Alatan, 2026). Additionally, manually crafting CoT prompts can be costly and difficult to scale, thereby limiting their practical use (Gu et al., 2024).

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.

3.4.2 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 (Jeon & Kim, 2025) . This indicates that self-consistency primarily stabilises 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 (Carlson & Burbano, 2025). 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 (Carlson & Burbano, 2025; Jeon & Kim, 2025)

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 stabilising reasoning outcomes rather than by increasing interpretability or task alignment.

3.4.3 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 synthesises a final, verified output. By explicitly separating generation and validation, CoV reduces ambiguity and increases confidence in the resulting responses (Thanasi-Boçe & Hoxha, 2024). 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 (Thanasi-Boçe & Hoxha, 2024)

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.

Chapter 4 Empirical Study Results

This chapter details the results of the empirical study addressing RQ2. The analysis is organised around testing the research hypotheses: initially, the direct effects of prompting techniques are evaluated (H1), followed by the examination of the moderating effect of Generative AI usage frequency (H2).

4.1 Sample Characteristics and Descriptive Statistics

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 4.1 presents the geographical distribution of the sample.

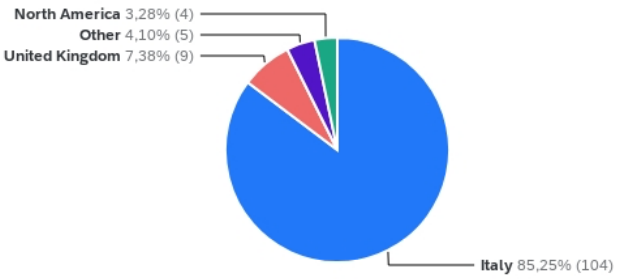

Figure 4.1 Geographical distribution of participants

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%), Advisory roles (4.92%), with nearly half identifying as "Other" (46.72%), indicating varied professional profiles. Figure 4.2 illustrates the professional backgrounds of respondents, indicating a heterogeneous sample composed primarily of business professionals.

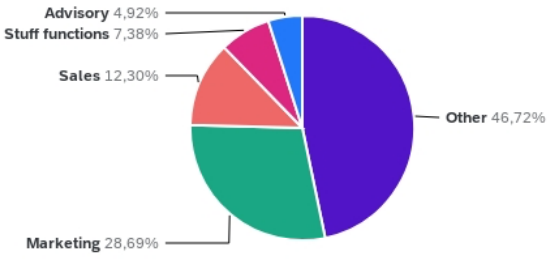

Figure 4.2 Professional background of participants

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 4.3 reports the frequency of generative AI usage among participants, highlighting a high level of exposure to AI tools within the sample.

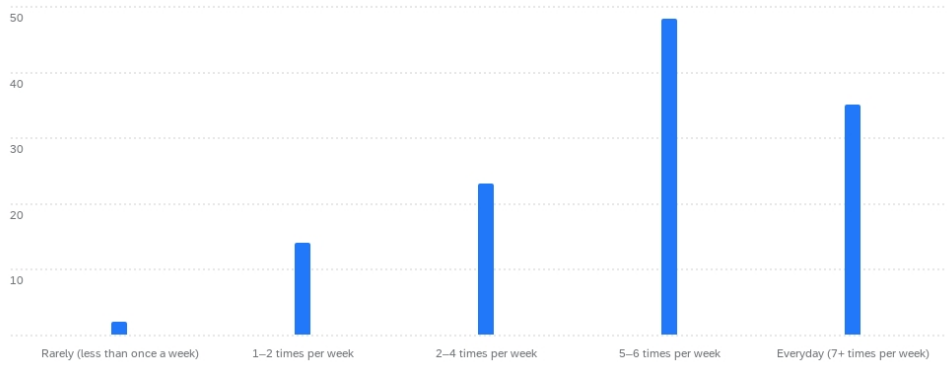

Figure 4.3 Generative AI Use Frequency

4.2 Scale Reliability Measurement 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 (Abeysinghe & Circi, 2024b). All evaluative items were measured on five-point Likert scales from 1 (strongly disagree) to 5 (strongly agree). Two composite constructs were created: perceived output usefulness and prompt quality. Perceived output usefulness was assessed with five items related to relevance, informativeness, correctness, clarity and hallucination absence. Prompt quality was measured with four items on relevance, informativeness, correctness, and clarity. Internal consistency was evaluated with Cronbach's alpha, showing excellent reliability ($\alpha = 0.947$ for output usefulness and $\alpha = 0.994$ for prompt quality), indicating high internal consistency. The scale details are reported in Appendix B.1

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 AI use, measured via a self-report item about how often participants used AI tools, served as a continuous moderator. Given the hierarchical data structure, multiple evaluations nested within respondents, linear mixed-effects models (LMEM) 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, prompt quality was modelled as a function of prompting techniques and AI usage frequency. An interaction between prompting techniques and usage frequency was tested to determine whether AI use 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.

4.2.1 Linear Mixed-Effects Models (LMEM)

To motivate the analytical approach adopted in the results section, linear mixed-effects models (LMMs) were used to analyse the survey data because they offer a robust and flexible approach to handling hierarchically structured, non-independent observations, which are common in repeated-evaluation setups. LMMs extend basic linear regression by estimating fixed effects that capture average relationships across the population and random effects that account for variability at higher levels, such as individual respondents or items (Bates et al., 2014; Brown, 2021). This structure allows LMMs to explicitly account for correlations among repeated measurements within the same participant, thereby preventing violations of the independence assumption that could bias traditional models (Meteyard & Davies, 2020). Unlike repeated-measures ANOVA, which requires balanced data and often involves aggregating responses, LMMs analyse raw, unaggregated data and manage unbalanced designs, missing data and unequal evaluation counts without discarding cases (Brown, 2021). This flexibility is particularly valuable in our empirical experiment, where participants evaluate multiple stimuli, but only certain prompt-output pairs combinations. By including random intercepts for respondents, LMMs account for individual differences in baseline ratings (e.g., stricter or more lenient evaluations), while fixed effects estimate overall impacts of prompting methods, task types and usage frequency (Meteyard & Davies, 2020). The framework also allows for the inclusion of continuous moderators and interaction effects, facilitating the testing of moderation hypotheses within a single model (Bates et al., 2014). Evidence shows that LMMs provide stable estimates of fixed effects even with moderate violations of residual or random-effect assumptions, thereby improving reliability for complex behavioural and survey data (Schielzeth et al., 2020). These features justify the use of linear mixed-effects models to examine perceived output usefulness and prompt quality when multiple evaluations are nested within respondents.

4.3 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 prompt quality evaluations. Prompts that combined example-based prompting with explicit reasoning guidance received 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.

4.4 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 Generative AI usage frequency on prompt quality evaluations was tested through an interaction model (H2).

4.5 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:

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

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 (β = 3.28, $p < .001$) represents the expected usefulness rating for zero-shot prompts in the reference task condition. All coefficients associated with prompting techniques, therefore, indicate how much perceived usefulness increases (or decreases) relative to zero-shot prompting (Figure 4.4).

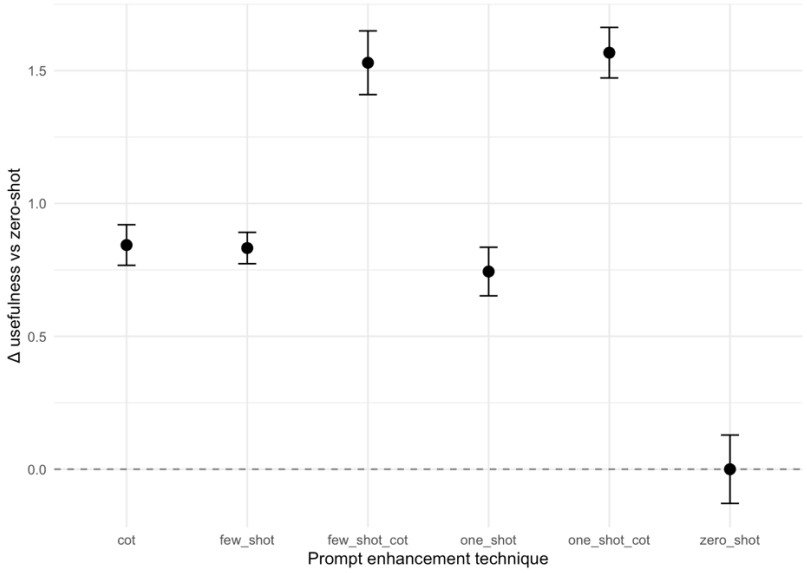

Figure 4.4 Δ usefulness vs zero-shot

Results show a strong and statistically significant main effect of prompting technique ($p < .001$). 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 (β = 0.84, $p < .001$), few-shot prompting by 0.82 points (β = 0.82, $p < .001$), and one-shot prompting by 0.74 points (β = 0.74, $p < .001$). Even larger effects emerged for combined techniques: few-shot with chain-of-thought increased usefulness by 1.52 points (β = 1.52, $p < .001$), while one-shot with chain-of-thought produced the strongest improvement, increasing usefulness by 1.55 points relative to zero-shot prompting (β = 1.55, $p < .001$).

Finally, task type did not have a significant effect on perceived usefulness (β = −0.03, $p = .369$), indicating that the observed differences are primarily attributable to prompting techniques rather than to task characteristics.

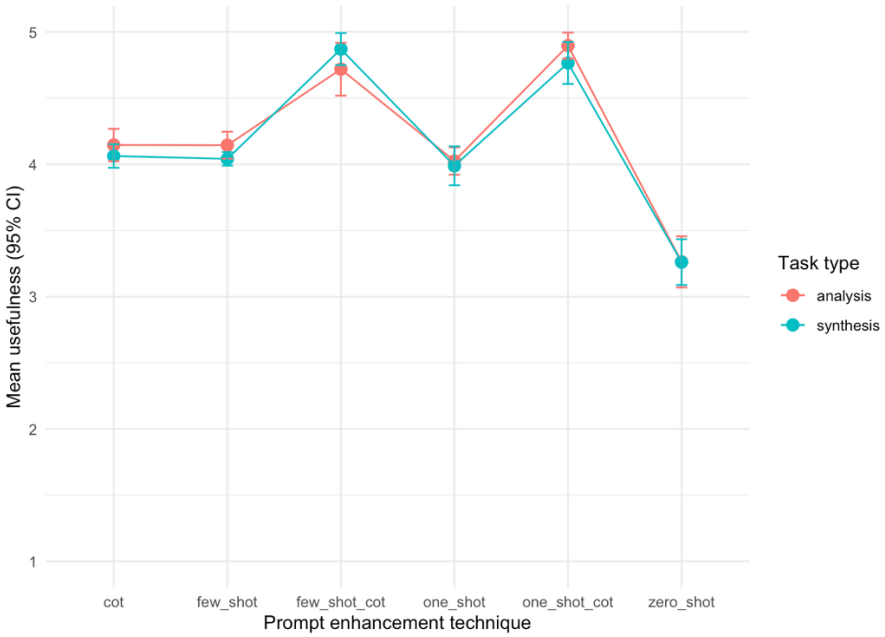

Figure 4.5 Interaction plot, technique and task type

4.5.1 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 variability. When averaged across task types, these estimates reveal a clear hierarchy of prompting methods by perceived usefulness (see Figure 4.6).

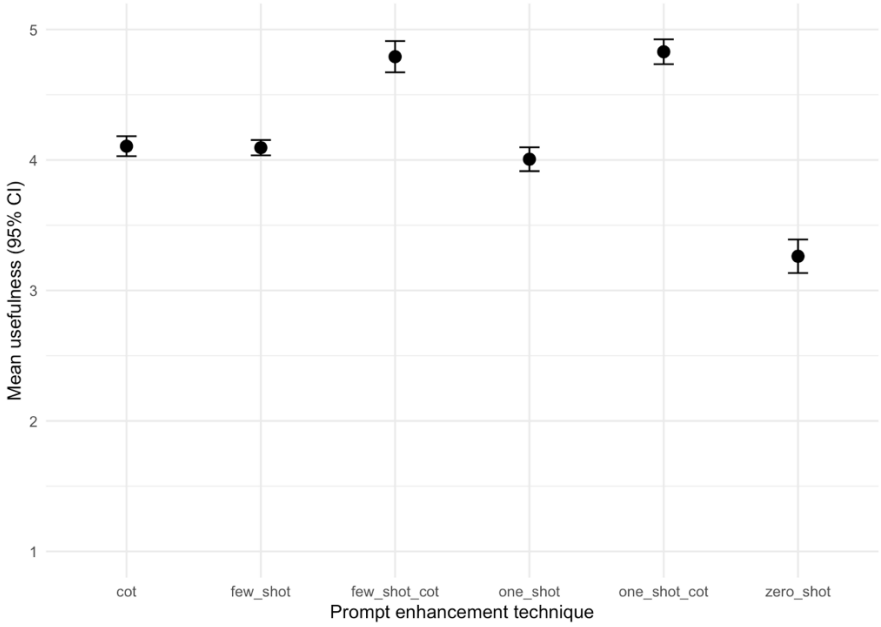

Figure 4.6 Mean usefulness by technique

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 structure, 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 4.7 displays the estimated marginal means and 95% confidence intervals for perceived output usefulness across different prompting techniques. The visualisation emphasises 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.

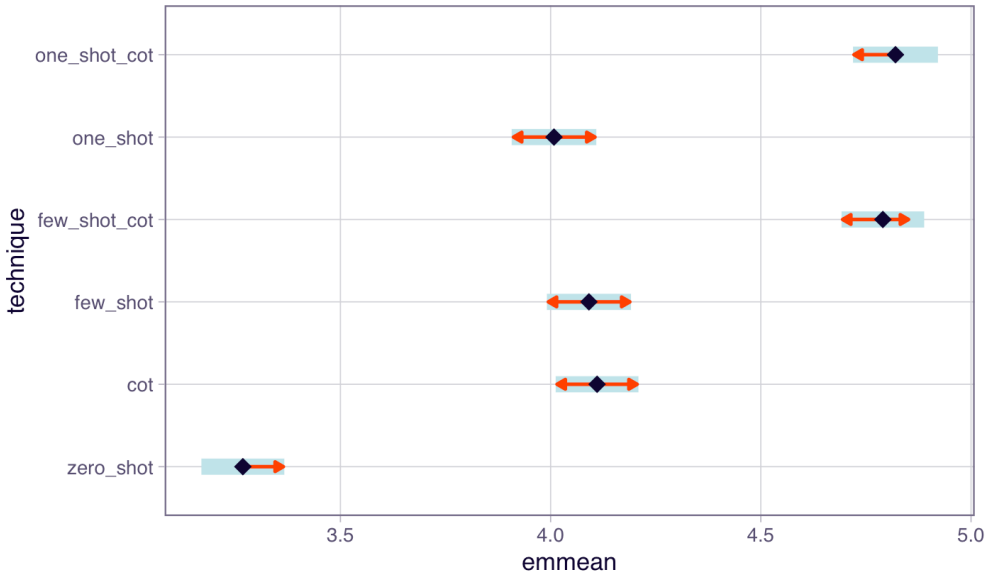

Figure 4.7 Estimated marginal means of perceived output usefulness by prompting technique

4.5.2 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 $p < .001$), 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 $p > .65$). 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 $p < .001$). Notably, no significant difference was found between the two combined techniques

themselves (p = .998), indicating that once both examples and reasoning are present, the number of examples (one vs. few) does not further increase perceived usefulness.

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

4.6 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:

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

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 < .001). Relative to zero-shot prompting, all enhanced techniques were associated with substantially higher quality ratings, with the largest coefficients observed for combined techniques (Figure 4.8).

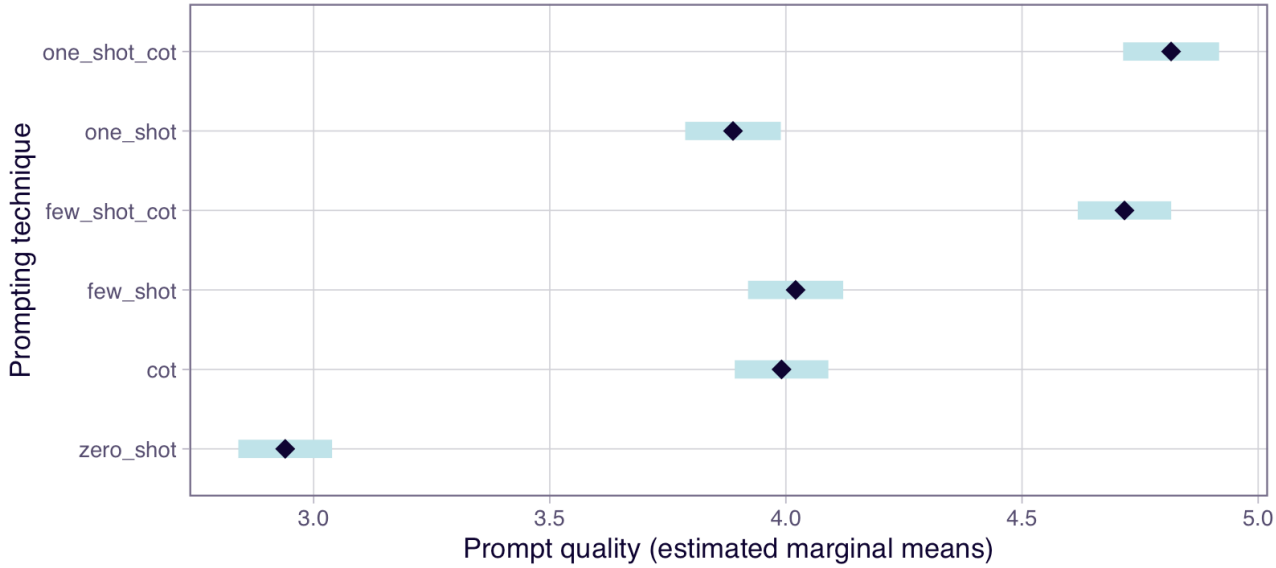

Figure 4.8 Perceived prompt quality by prompting technique

In contrast, the frequency of Generative AI use did not have a significant main effect on prompt quality ($\beta = -0.04$, $p = .169$), suggesting that, on average, more frequent users did not systematically rate prompts as higher or lower quality after controlling for the prompting technique.

While this section introduces the conceptual moderating relationship, the following subsection presents the regression specification used to estimate this effect, including main effects, interaction terms and control variables.

4.6.1 Interaction Model: testing moderation by usage frequency

To formally assess whether the effect of the prompting technique varies as a function of Generative AI usage frequency, a second model including the interaction term (*) was estimated:

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

In the model specification, an interaction term (*) *is* included between the prompting technique and Generative AI usage frequency. The asterisk indicates that the model estimates both the main effects of the two variables and their interaction, allowing 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 ($\chi^2(5) = 3.24$, $p = .662$). None of the interaction terms between the prompting technique and frequency of use was statistically significant (all $p > .33$). 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 (Figure 4.9).

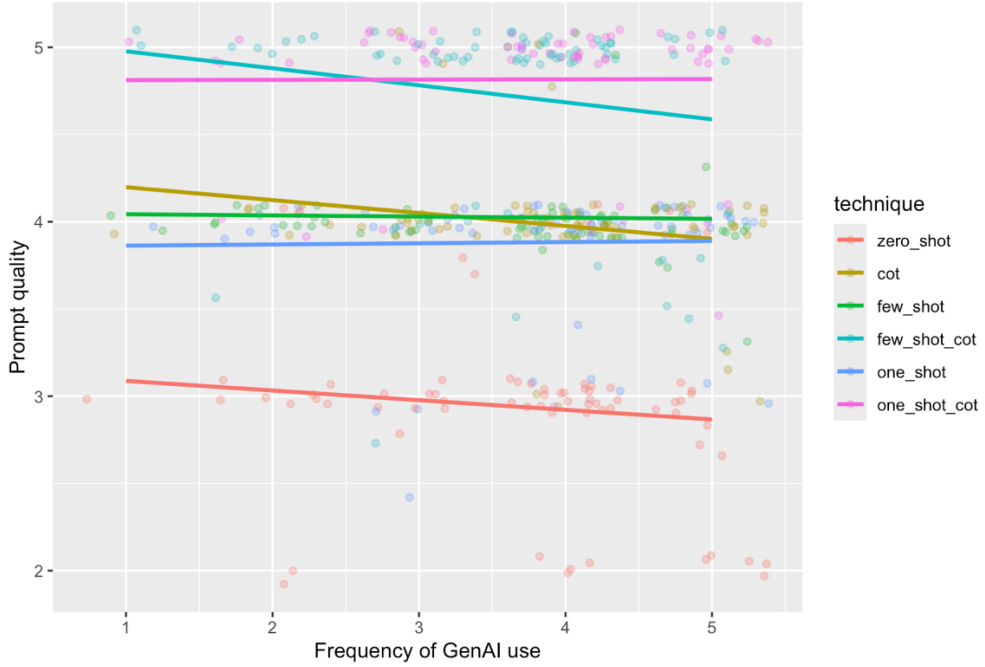

Figure 4.9 Prompt quality by prompting technique and frequency of use

A visual inspection of predicted values across low, medium, and high usage frequency levels further supports this conclusion. As shown in Figure 4.9, the relative ordering of prompting techniques remains largely unchanged across frequency levels, and the slopes associated with frequency of use are shallow and highly similar across techniques. Complete model estimates and additional diagnostic statistics for the moderation analysis are reported in Appendix B.2.

4.7 Summary of Hypotheses Testing

This study explored how prompting techniques influence users' assessments of AI-generated content and whether how often users use Generative AI modifies these effects. The findings strongly support H1, showing that prompting methods significantly impact how useful users find AI outputs. In both task types, improved prompting strategies consistently outperformed zero-shot prompts. Notably, combining example-based prompting with explicit reasoning instructions yielded the highest usefulness ratings. These results highlight the critical role of prompt structure in perceived output quality and confirm that structured prompt design strategies are effective. However, the second hypothesis was not supported: while prompting techniques affected perceptions of prompt quality, AI usage frequency did not moderate this relationship. Users with higher AI usage did not evaluate prompting techniques differently nor did they distinguish more between simple and advanced prompts.

Chapter 5 Discussion

This chapter presents the results of both the systematic literature review and the empirical study. These findings are analysed within the context of existing research and then explored for their managerial and theoretical implications.

5.1 Interpretation of Findings

This study combined a systematic literature review with an empirical investigation to examine how prompt enhancement techniques influence the perceived usefulness of AI-generated outputs in business tasks. The systematic review identified prompt engineering as a central mechanism for improving the quality, reliability, and practical usefulness of generative AI outputs across domains. In particular, the reviewed literature consistently links techniques such as few-shot prompting and chain-of-thought prompting to improved task alignment, more coherent outputs, and enhanced performance in complex or reasoning-intensive tasks. At the same time, the review revealed a clear imbalance in existing research. While prompt enhancement techniques are frequently evaluated using technical benchmarks or task-specific accuracy metrics, user-centred evaluations in real business contexts remain limited. As a result, the literature provides only partial insight into how these techniques are perceived and valued by end users in applied organisational settings. The empirical findings of this study directly respond to this gap by assessing prompting strategies through human evaluations of perceived usefulness in realistic business tasks. Hypothesis 1 was strongly supported, showing that prompting techniques significantly influence perceived usefulness. All enhanced methods outperformed zero-shot prompting, especially when combining example-based prompts with explicit reasoning instructions. However, Hypothesis 2 was not supported. The impact of prompting techniques on perceived usefulness and prompt quality remained consistent across different levels of AI usage, suggesting that users generally accept these techniques regardless of how often they use AI tools. Rather than reflecting varying expertise or familiarity, perceived usefulness is mainly determined by the intrinsic quality of prompt design. These findings emphasise a key difference between how often AI is utilised and users' proficiency in prompt engineering. This might indicate that, in some cases, prompt improvement methods can serve as universally effective design tools rather than skills dependent solely on experience. Finally, the alignment between the systematic review and empirical data reinforces the conclusion that prompt engineering significantly enhances the perceived usefulness of generative AI in business. Nonetheless, the absence of a moderation effect highlights the need for future research to go beyond using usage frequency as a proxy for expertise and to develop more accurate measures of prompt-engineering ability and AI literacy.

5.2 Managerial Contributions

This study provides several managerial insights by translating findings from the systematic review and empirical analysis into practical guidance for organisations implementing Generative AI in business settings. The taxonomy of prompt enhancement techniques, developed through the review, offers managers a clear,

organised framework for understanding how different prompting methods work and how they can be combined. Instead of viewing prompt engineering as an improvised or purely technical task, the taxonomy elucidates the roles of key techniques like zero-shot, one-shot, few-shot, and chain-of-thought prompting, along with their combinations. This classification enables managers and practitioners to go beyond trial-and-error and make more informed choices when designing prompts for tasks such as analysis, reporting, or decision-making.

Second, findings show that prompt design significantly and systematically influences perceived usefulness, regardless of users' experience levels. For managers, this means that enhancing AI performance doesn't require advanced technical skills or highly experienced users; instead, adopting well-structured prompting strategies can suffice. According to our study, combining examples with clear reasoning instructions consistently results in higher evaluations, indicating that organisations can improve AI effectiveness by standardising prompt templates that include these elements. Additionally, the lack of a moderating effect from how often AI is used has important implications for training and adoption. It suggests that merely increasing usage frequency doesn't improve recognition or evaluation of effective prompt techniques. Therefore, targeted training on prompt design is essential, rather than relying solely on frequent exposure. Lastly, the difference between prompt quality and output usefulness emphasises that organisations should assess AI systems not only by final results but also by the quality of the interaction design leading to those results. Integrating prompt engineering guidelines into workflows, documentation, or AI governance practices can help ensure more consistent and reliable outcomes across teams and use cases.

Overall, this research suggests that prompt engineering is best viewed as a managerial skill rather than merely a technical one. Using a clear taxonomy combined with empirical data equips managers to systematically integrate prompt enhancement methods into business processes, fostering more effective, scalable and inclusive use of Generative AI across organisations.

5.3 Theoretical Implications

This study contributes to the literature on Generative AI and human-AI interaction by integrating insights from prompt engineering research with a business-oriented and user-centred perspective. Prior studies have widely acknowledged that generative AI systems can significantly augment knowledge work, while also introducing risks related to reliability, hallucinations and misinterpretation of outputs (Chui et al., 2023; University of Melbourne & International, 2025). In this context, prompt engineering has been increasingly recognised as a critical mechanism for improving the effectiveness of human-AI interaction (Sikha, 2023; Bozkurt, 2024). Building on this stream of research, the systematic literature review conducted in this study consolidates and organises fragmented evidence on prompt enhancement techniques into an outcome-oriented taxonomy. By classifying prompting strategies according to their primary functional contribution, baseline efficiency, task alignment and output transparency, the study addresses the conceptual fragmentation noted in prior work on prompt engineering (Sahoo et al., 2025; Schulhoff et al., 2025). This

taxonomy provides a structured theoretical framework that clarifies how different prompting techniques contribute to output quality, interpretability and reliability, rather than treating them as isolated or purely technical design choices. Furthermore, this research extends existing business-oriented GenAI literature, which has predominantly focused on model capabilities or domain-specific applications (Fui-Hoon Nah et al., 2023; Feuerriegel et al., 2024), by explicitly conceptualising prompting techniques as mechanisms of value creation in human-AI collaboration. While prior studies highlight the importance of interaction design for extracting business value from generative AI, empirical evidence on how users perceive and evaluate prompting strategies has remained limited. By linking prompt enhancement techniques to perceived usefulness, this study advances a usage-focused theoretical perspective that complements performance-based evaluations and aligns with calls for more user-centred research in organisational AI adoption (Feuerriegel et al., 2024).

The study contributes to theory by bridging prompt engineering research and information systems literature, demonstrating that prompting techniques should be understood not only as optimisation tools for model outputs, but as central components of effective human–AI interaction in business contexts.

5.4 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 standardised 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 (Schulhoff et al., 2025), which complicates synthesis and categorisation. Although this review addresses this issue by organising techniques based on their primary reported effects, future research would benefit from greater conceptual standardisation 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). 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 optimisation, 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(Gibreel & Arpaci, 2025).

Additionally, the study relies on self-reported evaluations of perceived usefulness, which, while appropriate for capturing user-centred 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 organisational 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.

Conclusion

This thesis examines how prompt enhancement techniques influence the perceived usefulness of generative AI in business-related tasks by combining a systematic literature review with an empirical, user-centred investigation. This mixed method approach addresses a limitation in existing research, which has predominantly focused on technical benchmarks and task-specific performance, offering limited insight into how users evaluate prompting strategies in applied organisational contexts. The systematic literature review synthesised fragmented research on prompt engineering and organised prompt enhancement techniques into an outcome-oriented taxonomy, distinguishing between baseline prompting, task alignment techniques, and output transparency techniques. This framework contributes to theory by clarifying the functional role of prompting strategies and the trade-offs between efficiency, relevance, interpretability and reliability. In doing so, the study positions prompt engineering as a key mechanism of human-AI interaction rather than a purely technical optimisation practice.

The empirical findings show that prompting techniques significantly influence the perceived usefulness of AI-generated outputs, particularly in tasks requiring analysis and synthesis. More structured prompting strategies are generally associated with higher perceived usefulness, supporting the relevance of prompt design for effective human-AI collaboration in business settings. However, the analysis does not provide evidence of a statistically significant moderating effect of generative AI usage frequency. This suggests that, within the scope of this study, the perceived usefulness of prompting techniques does not differ substantially between users with varying levels of AI usage experience.

This research advances the understanding of prompt engineering in business contexts by integrating technical insights from the prompt engineering literature with a user-centred evaluation perspective. It highlights the importance of assessing generative AI systems through human perceptions of usefulness and underscores the role of prompt enhancement techniques as a practical lever for improving AI-supported business tasks. While the study is subject to limitations related to sample composition and task selection, it provides a solid foundation for future research to further investigate how user characteristics, organisational contexts, and task complexity may shape the effectiveness and evaluation of prompting strategies.

Bibliography

Abeysinghe, B., & Circi, R. (2024a). *The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches*. http://arxiv.org/abs/2406.03339

Abeysinghe, B., & Circi, R. (2024b). *The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches*. http://arxiv.org/abs/2406.03339

Abou, B., Karam, E., Fissaa, T., & Marghoubi, R. (2025). AI-Powered Assessment of Resistance to Change in the Context of Digital Transformation. In *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 16, Number 6). www.ijacsa.thesai.org

Agbareia, R., Omar, M., Zloto, O., Glicksberg, B. S., Nadkarni, G. N., & Klang, E. (2025). Multimodal LLMs for retinal disease diagnosis via OCT: few-shot versus single-shot learning. *Therapeutic Advances in Ophthalmology*, *17*. https://doi.org/10.1177/25158414251340569

Anam, R. K. (2025). *Prompt Engineering and the Effectiveness of Large Language Models in Enhancing Human Productivity*. http://arxiv.org/abs/2507.18638

Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, *33*(1). https://doi.org/10.1007/s12525-023-00680-1

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting Linear Mixed-Effects Models using lme4*. http://arxiv.org/abs/1406.5823

Bharti, P. K., Panchal, M., Dalal, V., Agarwal, M., & Ekbal, A. (2026). Not all peer reviews are significant: a dataset of exhaustive vs. trivial scientific peer reviews leveraging chain-of-thought reasoning. *Scientometrics*. https://doi.org/10.1007/s11192-025-05435-7

Bozkurt, A. (2024). Tell Me Your Prompts and I Will Make Them True: The Alchemy of Prompt Engineering and Generative AI. In *Open Praxis* (Vol. 16, Number 2, pp. 111–118). International Council for Open and Distance Education. https://doi.org/10.55982/openpraxis.16.2.661

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920960351

Bukhary, N., Ahmad, M., Rashad, K., Rai, S., Shapsough, S., Kaddoura, Y., Dghaym, D., & Zualkernan, I. (2025). Few-Shot Evaluation of Vision Language Models for Detecting Visual Defects in Autonomous Vehicle Software Requirement Specifications. *IEEE Access*, *13*, 117914–117942. https://doi.org/10.1109/ACCESS.2025.3586554

Byra, M., Rachmadi, M. F., & Skibbe, H. (2025). Few-shot medical image classification with simple shape and texture text descriptors using vision-language models. *Bulletin of the Polish Academy of Sciences: Technical Sciences*, *73*(3). https://doi.org/10.24425/bpasts.2025.153838

Carlson, N. A., & Burbano, V. (2025). The use of LLMs to annotate data in management research: Foundational guidelines and warnings. *Strategic Management Journal*. https://doi.org/10.1002/smj.70023

Castelvecchi D. (2016). Can we open the black box of AI? *Nature*, *538(7623)*(Nature), 20–23. https://doi.org/https://doi.org/10.1038/538020a

Chen, P., & Li, Z. (2025). Length Instruction Fine-Tuning with Chain-of-Thought (LIFT-COT): Enhancing Length Control and Reasoning in Edge-Deployed Large Language Models. *Electronics (Switzerland)*, *14*(8). https://doi.org/10.3390/electronics14081662

Chen, X., Chen, Z., & Cheng, S. (2025). CoTHSSum: Structured long-document summarization via chain-of-thought reasoning and hierarchical segmentation. *Journal of King Saud University - Computer and Information Sciences*, *37*(4). https://doi.org/10.1007/s44443-025-00041-2

Chen, Y., Huang, Y., Chen, X., Shen, P., & Yun, L. (2026). GPTVD: vulnerability detection and analysis method based on LLM's chain of thoughts. *Automated Software Engineering*, *33*(1). https://doi.org/10.1007/s10515-025-00550-4

Darwiyanto, E., Gusnaen, R. A., & Nurtantyana, R. (2025). A comparative study of large language models with chain-of-thought prompting for automated program repair. *IAES International Journal of Artificial Intelligence*, *14*(6), 4579–4589. https://doi.org/10.11591/ijai.v14.i6.pp4579-4589

Elnashar, A., White, J., & Schmidt, D. C. (2025). Enhancing structured data generation with GPT-4o evaluating prompt efficiency across prompt styles. *Frontiers in Artificial Intelligence*, *8*. https://doi.org/10.3389/frai.2025.1558938

Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). Qualitative Content Analysis: A Focus on Trustworthiness. *SAGE Open*, *4*(1). https://doi.org/10.1177/2158244014522633

Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *Business and Information Systems Engineering*, *66*(1), 111–126. https://doi.org/10.1007/s12599-023-00834-7

Filienko, D., Wang, Y., Jazmi, C. El, Xie, S., Cohen, T., De Cock, M., & Yuwen, W. (2024). *Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy*. http://arxiv.org/abs/2409.00112

Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. In *Journal of Information Technology Case and Application Research* (Vol. 25, Number 3, pp. 277–304). Routledge. https://doi.org/10.1080/15228053.2023.2233814

Geroimenko, V. (n.d.). *SpringerBriefs in Computer Science The Essential Guide to Prompt Engineering*.

Ghimire, P., Kim, K., Stentz, T., & Roy, T. (2026). Modular Chain-of-Thought (CoT) for LLM-Based Conceptual Construction Cost Estimation. *Buildings*, *16*(2), 396. https://doi.org/10.3390/buildings16020396

Giannilias, T., Papadakis, A., Nikolaou, N., & Zahariadis, T. (2025). Classification of Hacker's Posts Based on Zero-Shot, Few-Shot, and Fine-Tuned LLMs in Environments with Constrained Resources. *Future Internet*, *17*(5). https://doi.org/10.3390/fi17050207

Gibreel, O., & Arpaci, I. (2025). Development and validation of the prompt engineering competence scale (PECS). *Information Development*. https://doi.org/10.1177/02666669251336455

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., Wirrell, E. C., Knupp, K. G., Patel, M., Loeb, J. A., Sullivan, J. E., Harte-Hargrove, L., Fureman, B. E., Buchhalter, J., & Sahoo, S. S. (2025). Ontology accelerates few-shot learning capability of large language model: A study in extraction of drug efficacy in a rare pediatric epilepsy. *International Journal of Medical Informatics*, *201*. https://doi.org/10.1016/j.ijmedinf.2025.105942

Gozzi, M., & Di Maio, F. (2024). Comparative Analysis of Prompt Strategies for Large Language Models: Single-Task vs. Multitask Prompts. *Electronics (Switzerland)*, *13*(23). https://doi.org/10.3390/electronics13234712

Gu, X., Chen, X., Lu, P., Li, Z., Du, Y., & Li, X. (2024). AGCVT-prompt for sentiment classification: Automatically generating chain of thought and verbalizer in prompt learning. *Engineering Applications of Artificial Intelligence*, *132*. https://doi.org/10.1016/j.engappai.2024.107907

Guney, G., Yigin, B. O., Guven, N., Alici, Y. H., Colak, B., Erzin, G., & Saygili, G. (2021). An overview of deep learning algorithms and their applications in neuropsychiatry. In *Clinical Psychopharmacology and Neuroscience* (Vol. 19, Number 2, pp. 206–219). Korean College of Neuropsychopharmacology. https://doi.org/10.9758/cpn.2021.19.2.206

Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687

Huang, Y., Ma, T., Yang, K., & Zhang, Z. (2026). FinSent-DistillQ: A distilled large language model with chain-of-thought fine-tuning for financial sentiment analysis. *Journal of Intelligent Information Systems*. https://doi.org/10.1007/s10844-025-01020-9

Indira Kumar, A. K., Sthanusubramoniani, G., Gupta, D., Nair, A. R., Alotaibi, Y. A., & Zakariah, M. (2025). Multi-task detection of harmful content in code-mixed meme captions using large language models with zero-shot, few-shot, and fine-tuning approaches. *Egyptian Informatics Journal*, *30*. https://doi.org/10.1016/j.eij.2025.100683

Jager, J., Putnick, D. L., & Bornstein, M. H. (2017). II. MORE THAN JUST CONVENIENT: THE SCIENTIFIC MERITS OF HOMOGENEOUS CONVENIENCE SAMPLES. *Monographs of the Society for Research in Child Development*, *82*(2), 13–30. https://doi.org/10.1111/mono.12296

Janiesch, C., Zschech, P., & Heinrich, K. (n.d.). *Machine learning and deep learning*. https://doi.org/10.1007/s12525-021-00475-2/Published

Jeon, S., & Kim, H. G. (2025). A comparative evaluation of chain-of-thought-based prompt engineering techniques for medical question answering. *Computers in Biology and Medicine*, *196*. https://doi.org/10.1016/j.compbiomed.2025.110614

Köksal, A., & Alatan, A. A. (2026). SAMChat: Introducing Chain-of-Thought Reasoning and GRPO to a Multimodal Small Language Model for Small-Scale Remote Sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *19*, 795–804. https://doi.org/10.1109/JSTARS.2025.3637115

Lee, A. V. Y., Teo, C. L., & Tan, S. C. (2024). Prompt Engineering for Knowledge Creation: Using Chain-of-Thought to Support Students' Improvable Ideas. *AI (Switzerland)*, *5*(3), 1446–1461. https://doi.org/10.3390/ai5030069

Mao, W., Wu, J., Chen, W., Gao, C., Wang, X., & He, X. (2025). Reinforced Prompt Personalization for Recommendation with Large Language Models. *ACM Transactions on Information Systems*, *43*(3). https://doi.org/10.1145/3716320

Mayring, P. (n.d.). *Qualitative Content Analysis Theoretical Foundation, Basic Procedures and Software Solution*. Retrieved www.beltz.de

Meshkin, H., Zirkle, J., Arabidarrehdor, G., Chaturbedi, A., Chakravartula, S., Mann, J., Thrasher, B., & Li, Z. (2024). Harnessing large language models' zero-shot and few-shot learning capabilities for regulatory research. *Briefings in Bioinformatics*, *25*(5). https://doi.org/10.1093/bib/bbae354

Meteyard, L., & Davies, R. A. I. (2020). Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, *112*. https://doi.org/10.1016/j.jml.2020.104092

Miao, J., Thongprayoon, C., Suppadungsuk, S., Krisanapan, P., Radhakrishnan, Y., & Cheungpasitporn, W. (2024). Chain of Thought Utilization in Large Language Models and Application in Nephrology. In *Medicina (Lithuania)* (Vol. 60, Number 1). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/medicina60010148

Mohanty, A., Parthasarathy, V. B., & Shahid, A. (2025). *The Future of MLLM Prompting is Adaptive: A Comprehensive Experimental Evaluation of Prompt Engineering Methods for Robust Multimodal Performance*. http://arxiv.org/abs/2504.10179

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A Comprehensive Overview of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *16*(5). https://doi.org/10.1145/3744746

of Melbourne, U., & International, K. (2025). *Trust, attitudes and use of artificial intelligence A global study 2025*. https://doi.org/10.26188/28822919

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., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, *134*, 178–189. https://doi.org/10.1016/j.jclinepi.2021.03.001

Panneer Selvam Viswanathan. (2025). Prompt Engineering for Conversational AI Systems: A Systematic Review of Techniques and Applications. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *11*(1), 733–741. https://doi.org/10.32628/cseit25111276

Pollock, A., & Berge, E. (2018). How to do a systematic review. In *International Journal of Stroke* (Vol. 13, Number 2, pp. 138–156). SAGE Publications Inc. https://doi.org/10.1177/1747493017743796

Polo, F. M., Xu, R., Weber, L., Silva, M., Bhardwaj, O., Choshen, L., de Oliveira, A. F. M., Sun, Y., & Yurochkin, M. (2024). *Efficient multi-prompt evaluation of LLMs*. http://arxiv.org/abs/2405.17202

Qi, X., Yang, B., Wang, S., Zhang, Z., Zhang, Y., & Du, K. (2026). Few-shot and chain-of-thought prompting for equipment maintenance knowledge graph construction via large language models. *Knowledge-Based Systems*, *335*. https://doi.org/10.1016/j.knosys.2026.115266

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. (2024). A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. *IEEE Access*, *12*, 26839–26874. https://doi.org/10.1109/ACCESS.2024.3365742

Razumovskaia, E., Vulić, I., & Korhonen, A. (2025). Analyzing and Adapting Large Language Models for Few-Shot Multilingual NLU: Are We There Yet? *Transactions of the Association for Computational Linguistics*, *13*, 1096–1120. https://doi.org/10.1162/TACL.a.33

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S., & Chadha, A. (2025). *A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications*. http://arxiv.org/abs/2402.07927

Santana, E. G., Benjamin, G., Araujo, M., Santos, H., Freitas, D., Almeida, E., Neto, P. A. da M. S., Li, J., Chun, J., & Ahmed, I. (2025). *Which Prompting Technique Should I Use? An Empirical Investigation of Prompting Techniques for Software Engineering Tasks*. http://arxiv.org/abs/2506.05614

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., Réale, D., Dochtermann, N. A., Garamszegi, L. Z., & Araya-Ajoy, Y. G. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods in Ecology and Evolution*, *11*(9), 1141–1152. https://doi.org/10.1111/2041-210X.13434

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., … Resnik, P. (2025). *The Prompt Report: A Systematic Survey of Prompt Engineering Techniques*. http://arxiv.org/abs/2406.06608

Setyo Nugroho, H., & Shaferi, I. (2025). *ANALYSIS OF PROMPT ENGINEERING EFFECTIVENESS IN STOCK RECOMMENDATION BY CHATGPT: AN EXPERIMENTAL STUDY IN THE INDONESIAN MARKET* (Vol. 01, Number 01).

Shah, K., Xu, A. Y., Sharma, Y., Daher, M., McDonald, C., Diebo, B. G., & Daniels, A. H. (2024). Large Language Model Prompting Techniques for Advancement in Clinical Medicine. In *Journal of Clinical Medicine* (Vol. 13, Number 17). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/jcm13175101

Sikha, V. K. (2023). Mastering Prompt Engineering: Optimizing Interaction with Generative AI Agents. *Journal of Engineering and Applied Sciences Technology*, 1–8. https://doi.org/10.47363/JEAST/2023(5)E117

Stratton, S. J. (2021). Population Research: Convenience Sampling Strategies. In *Prehospital and Disaster Medicine* (Vol. 36, Number 4, pp. 373–374). Cambridge University Press. https://doi.org/10.1017/S1049023X21000649

Šuster, S., Baldwin, T., & Verspoor, K. (2024). Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Research Synthesis Methods*, *15*(6), 988–1000. https://doi.org/10.1002/jrsm.1749

Teng, S., Liu, J., Jain, R. K., Chai, S., Hou, R., Tateyama, T., Lin, L., & Chen, Y. (2025). *Enhancing Depression Detection with Chain-of-Thought Prompting: From Emotion to Reasoning Using Large Language Models*. http://arxiv.org/abs/2502.05879

Thanasi-Boçe, M., & Hoxha, J. (2024). From ideas to ventures: building entrepreneurship knowledge with LLM, prompt engineering, and conversational agents. *Education and Information Technologies*, *29*(18), 24309–24365. https://doi.org/10.1007/s10639-024-12775-z

Tony, C., Díaz Ferreyra, N. E., Mutas, M., Dhif, S., & Scandariato, R. (2025). Prompting Techniques for Secure Code Generation: A Systematic Investigation. *ACM Transactions on Software Engineering and Methodology*, *34*(8). https://doi.org/10.1145/3722108

Viswanathan, V., Gashteovski, K., Lawrence, C., Wu, T., & Neubig, G. (2023). *Large Language Models Enable Few-Shot Clustering*. http://arxiv.org/abs/2307.00524

Wan, T., & Chen, Z. (2024). Exploring generative AI assisted feedback writing for students' written responses to a physics conceptual question with prompt engineering and few-shot learning. *Physical Review Physics Education Research*, *20*(1). https://doi.org/10.1103/PhysRevPhysEducRes.20.010152

Wang, J. (2024). Improving ChatGPT's Competency in Generating Effective Business Communication Messages: Integrating Rhetorical Genre Analysis into Prompting Techniques. *Journal of Technical Writing and Communication*, *54*(4), 369–395. https://doi.org/10.1177/00472816241260033

Wang, X. (2024). Use of proper sampling techniques to research studies. *Applied and Computational Engineering*, *57*(1), 141–145. https://doi.org/10.54254/2755-2721/57/20241324

Yang, G., Zhou, Y., Chen, X., Zhang, X., Zhuo, T. Y., & Chen, T. (2024). *Chain-of-Thought in Neural Code Generation: From and For Lightweight Language Models*. http://arxiv.org/abs/2312.05562

Zhang, W., Wu, M., Zhou, L., Shao, M., Wang, C., & Wang, Y. (2025). A sepsis diagnosis method based on Chain-of-Thought reasoning using Large Language Models. *Biocybernetics and Biomedical Engineering*, *45*(2), 269–277. https://doi.org/10.1016/j.bbe.2025.04.002

Zhen, H., Shi, Y., Huang, Y., Yang, J. J., & Liu, N. (2024). Leveraging Large Language Models with Chain-of-Thought and Prompt Engineering for Traffic Crash Severity Analysis and Inference. *Computers*, *13*(9). https://doi.org/10.3390/computers13090232

Appendix A.1 Baseline Prompting Techniques

Authors	Technique	Task / Domain	Evidence type	Reported Effects
Ghimire, P.; Kim, K.; Stentz, T.; Roy, T.	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 realiability and indicating necessity of structured prompting strategies
Bharti, P.K.; Panchal, M.; Dalal, V.; Agarwal, M.; Ekbal, A.	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.	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. (2025)	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. (2025)	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. (2025)	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. (2025)	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. (2025)	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 (2026)	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.

Shafikuzzaman et al. (2025)	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. (2025)	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. (2024)	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.; Teo, C.L.; Tan, S.C. (2024)	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. (2024)	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. (2024)	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. (2024)	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 >20k sentences.
Meshkin et al. (2024)	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 >700k FDA label sentences without fine-tuning.
Ono, Dickson & Koga (2024)	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. (2023)	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.

Lee et al. (2024)	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. (2024)	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.; Praas, R.; Kiesewetter, L. P.; Samwald, M. (2024)	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.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. (2024)	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.; Wang, Y.; El Jazmi, C.; Xie, S.; Cohen, T.; De Cock, M.; Yuwen, W. (2024)	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; Hoxha, J	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	Zero-shot	Business process modeling	Empirical (experimental / applied)	Prompt without examples used to generate BPM elements from textual input
Carlson, NA; Burbano, V	Zero-shot	Data annotation / text classification	Empirical (mixed-methods)	Direct task instruction without examples
Abou El Karam, B; Fissaa, T; Marghoubi, R	Zero-shot	Change management / employee attitude assessment	Empirical (experimental)	Classification of open-ended employee responses without examples
Wang, J.	Zero-shot	Business communication writing (negative message)	Empirical (experimental)	Direct prompt without additional guidance or interaction

Jovic, M; Papakonstantinidis, S; Kirkpatrick, R	Zero-shot	Written feedback generation	Empirical (experimental)	Minimal instruction without contextual or iterative guidance
Jovic, M; Papakonstantinidis, S; Kirkpatrick, R	Zero-shot	Written feedback generation	Empirical (experimental)	Baseline AI feedback without structured context
Tocev, T.; Atanasovski, A.	Zero-shot	IFRS advisory / financial reporting	Empirical (experimental)	Single-pass query without examples or stepwise decomposition
Tocev, T.; Atanasovski, A.	Zero-shot	IFRS advisory / financial reporting	Empirical (experimental)	Direct prompt describing the case once
Anam, RK	Zero-shot	Knowledge work / general productivity	Empirical (survey-based)	Generic prompts without examples or structure
Ghimire, P.; Kim, K.; Stentz, T.; Roy, T.	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 realiability and indicating necessity of structured prompting strategies
Bharti, P.K.; Panchal, M.; Dalal, V.; Agarwal, M.; Ekbal, A.	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.	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. (2025)	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. (2025)	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. (2025)	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. (2025)	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. (2025)	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 (2026)	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.
Shafikuzzaman et al. (2025)	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. (2025)	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. (2024)	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.; Teo, C.L.; Tan, S.C. (2024)	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. (2024)	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. (2024)	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. (2024)	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 >20k sentences.

Meshkin et al. (2024)	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 >700k FDA label sentences without fine-tuning.
Ono, Dickson & Koga (2024)	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. (2023)	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.
Lee et al. (2024)	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. (2024)	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.; Praas, R.; Kiesewetter, L. P.; Samwald, M. (2024)	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.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. (2024)	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.; Wang, Y.; El Jazmi, C.; Xie, S.; Cohen, T.; De Cock, M.; Yuwen, W. (2024)	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; Hoxha, J	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	Zero-shot	Business process modeling	Empirical (experimental / applied)	Prompt without examples used to generate BPM elements from textual input
Carlson, NA; Burbano, V	Zero-shot	Data annotation / text classification	Empirical (mixed-methods)	Direct task instruction without examples
Abou El Karam, B; Fissaa, T; Marghoubi, R	Zero-shot	Change management / employee attitude assessment	Empirical (experimental)	Classification of open-ended employee responses without examples
Wang, J.	Zero-shot	Business communication writing (negative message)	Empirical (experimental)	Direct prompt without additional guidance or interaction
Jovic, M; Papakonstantinidis, S; Kirkpatrick, R	Zero-shot	Written feedback generation	Empirical (experimental)	Minimal instruction without contextual or iterative guidance
Jovic, M; Papakonstantinidis, S; Kirkpatrick, R	Zero-shot	Written feedback generation	Empirical (experimental)	Baseline AI feedback without structured context
Tocev, T.; Atanasovski, A.	Zero-shot	IFRS advisory / financial reporting	Empirical (experimental)	Single-pass query without examples or stepwise decomposition
Tocev, T.; Atanasovski, A.	Zero-shot	IFRS advisory / financial reporting	Empirical (experimental)	Direct prompt describing the case once
Anam, RK	Zero-shot	Knowledge work / general productivity	Empirical (survey-based)	Generic prompts without examples or structure

Appendix A.2 Task Alignment Prompting Techniques

Authors	Technique	Task / Domain	Evidence type	Reported Effects
Golnari et al. (2025)	One-shot	Medical		One-shot offers an improvement in respect to zero-shot
Giannilias et al. (2025)	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. (2025)	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. (2024)	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. (2025)	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.
Razumovskaia et al. (2025)	Few-shot	Multilingual NLU	F1, Accuracy	Using demonstrations in the target language yields inconsistent improvements and does not reliably enhance multilingual generalization.
Byra et al. (2025)	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. (2023)	Few-shot	Text clustering	Macro F1, Micro F1, Pairwise F1, Accuracy, NMI	Few-shot prompting for keyphrase expansion is the most effective LLM integration strategy, outperforming classical and neural baselines on most datasets.
Lee et al. (2024)	Few-shot	Automatic scoring (science education)	Accuracy, Precision, Recall, F1, QWK	Few-shot prompting improves scoring accuracy by constraining output structure and aligning model behavior with human scoring patterns.

Jiang, Y.; De Raedt, M.; Deleu, J.; Demeester, T.; Develder, C.	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.; Yang, B.; Wang, S.; Zhang, Z.; Zhang, Y.; Du, K.	Few-shot	Construction of Knowledge Grapghs in the domain of equiment 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.; Panchal, M.; Dalal, V.; Agarwal, M.; Ekbal, A.	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.	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. (2025)	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. (2025)	Few-shot	power equipment defect grading	accuracy, ecs	Few-shot examples improve performance over zero-shot but show limited gains in domain-specific reasoning.
Indira Kumar et al. (2025)	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. (2025)	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. (2025)	Few-shot	Software requirements classification (FR/NFR, NFR subclasses, security)	Precision, Recall, F1 (macro)	Few-shot prompting reliably improves classification performance across all tasks.
Agbareia et al. (2025)	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. (2024)	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.; Teo, C.L.; Tan, S.C. (2024)	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. (2024)	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. (2024)	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. (2024)	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 (2024)	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. (2023)	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.; Chen, X.; Lu, P.; Li, Z.; Du, Y.; Li, X. (2024)	Few-shot	Sentiment classification	Accuracy, F1	Few-shot AGCVT prompting achieves the highest accuracy while preserving interpretability and low data requirements.

Choi, J. (2024)	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. (2024)	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.
Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. (2024)	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.; Wang, Y.; El Jazmi, C.; Xie, S.; Cohen, T.; De Cock, M.; Yuwen, W. (2024)	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	Few-shot	Business process modeling	Empirical (experimental / applied)	Prompt includes examples to guide BPM model generation
Carlson, NA; Burbano, V	Few-shot	Data annotation / text classification	Empirical (mixed-methods)	Including examples in prompt
Wang, J.	Few-shot	Business communication writing	Empirical (experimental)	Providing genre-based outlines or examples
Tocev, T.; Atanasovski, A.	Few-shot	IFRS advisory / financial reporting	Empirical (experimental)	Prompt includes worked IFRS examples before the target case
Tocev, T.; Atanasovski, A.	Few-shot	IFRS advisory / financial reporting	Empirical (experimental)	Example-driven prompt formulation
Anam, RK	Few-shot	Pattern-based tasks (writing, coding)	Empirical (survey-based)	Use of examples to illustrate desired output

Carlson, NA; Burbano, V	Role	Data annotation / text classification	Empirical (mixed-methods)	Assigning an identity to focus responses on relevant domain knowledge
Abou El Karam, B; Fissaa, T; Marghoubi, R	Role	Change management / decision support	Empirical (experimental)	Implicit expert framing via allies strategy categories
Wang, J.	Role	Business communication writing	Empirical (experimental)	Assigning ChatGPT an expert professional role
Tocev, T.; Atanasovski, A.	Role	IFRS advisory / professional decision support	Empirical (experimental)	Assigning GPT-4 the role of an IFRS expert
Anam, RK	Role	Business and professional communication	Empirical (survey-based)	Assigning professional or expert roles to the AI
Mao, W.; Wu, J.; Chen, W.; Gao, C.; Wang, X.; He, X.	Role	Recommendation systems	Empirical (experimental)	Assigning expert or recommender roles to the LLM
Cheung, K.S.	Role	Professional decision support	Conceptual (viewpoint)	Framing the LLM as a professional valuer

Appendix A.3 Output Transparency Prompting Techniques

Authors	Technique	Task / Domain	Evidence type	Reported Effects
Chen, Y.; Huang, Y.; Chen, X.; Shen, P.; Yun, L.	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, X.; Yang, B.; Wang, S.; Zhang, Z.; Zhang, Y.; Du, K.	Chain-of-Thought	Construction of Knowledge Grapghs in the domain of equiment 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, P.; Kim, K.; Stentz, T.; Roy, T.	Chain-of-Thought	Cost estimation in construction	human evaluation	2.52 (84%) av confidence score marks 20% improvement from ZS. his 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, P.K.; Panchal, M.; Dalal, V.; Agarwal, M.; Ekbal, A.	Chain-of-Thought	Peer reviews assessment	human evaluation, recall, precision, f1	Most detailed and interpreatable outputs and provided well structured, evidence based justifications
Köksal, A.; Alatan, A.A.	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, E.; Gusnaen, R.A.; Nurtantyana, R.	Chain-of-Thought	automated program repair	plausible patches	The designed CoT prompting structure has been generally shown to improve the ability of LLM models to generate solutions for APR tasks.
Qiao, J.; Li, S.; Liu, J.; Yu, H.; Xiao, Y.; Yu, H.; Zheng, Y.	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.	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 (2025)	Chain-of-Thought	Medical	Cohen's d effect sizes calculated against this Control condition (baseline prompt)	Most conistent 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)
Du et al. (2025)	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. (2025)	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. (2025)	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 (2025)	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. (2025)	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. (2025)	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 (2026)	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. (2025)	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. (2025)	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.
Bukhary et al. (2025)	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. (2025)	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. (2024)	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. (2024)	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. (2024)	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, X.; Chen, X.; Lu, P.; Li, Z.; Du, Y.; Li, X. (2024)	Chain-of-Thought	Sentiment classification	Accuracy, F1	Chain-of-thought improves reasoning transparency but manual construction is costly and non-scalable.
Bentham, O.; Stringham, N.; Marasović, A. (2024)	Chain-of-Thought	Multiple-choice QA; arithmetic addition	Accuracy; answer-change rate	CoT prompting often improves accuracy, but unchanged answers cannot be interpreted as evidence of post-hoc or unfaithful reasoning.
Yang, G.; Zhou, Y.; Chen, X.; Zhang, X.; Zhuo, T.Y.; Chen, T.	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, K.; Praas, R.; Kiesewetter, L. P.; Samwald, M. (2024)	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.
Hebenstreit, K.; Praas, R.; Kiesewetter, L. P.; Samwald, M. (2024)	Chain-of-Thought	Multi-domain multiple-choice question answering	Krippendorff's alpha; accuracy	The CoT prompt discovered through automated prompt engineering by Zhou et al. yields the strongest and most stable performance gains, especially for larger models.
Miao, J.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. (2024)	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, J.; Thongprayoon, C.; Suppadungsuk, S.; Krisanapan, P.; Radhakrishnan, Y.; Cheungpasitporn, W. (2024)	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, D.; Wang, Y.; El Jazmi, C.; Xie, S.; Cohen, T.; De Cock, M.; Yuwen, W. (2024)	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	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	Chain-of-Thought	Business process modeling	Empirical (experimental / applied)	Prompt requests step-by-step reasoning before generating BPM models
Ayad, S; Alsayoud, F	Chain-of-Thought	Business process modeling	Empirical (experimental / applied)	Prompt requests step-by-step reasoning before generating BPM models
Carlson, NA; Burbano, V	Chain-of-Thought	Data annotation / text classification	Empirical (mixed-methods)	Explicit step-by-step reasoning process
Abou El Karam, B; Fissaa, T; Marghoubi, R	Chain-of-Thought	Organizational analysis	Empirical (experimental)	Stepwise reasoning via explicit class descriptions
Tocev, T.; Atanasovski, A.	Chain-of-Thought	IFRS advisory / financial reporting	Empirical (experimental)	Sequential decomposition into subquestions with stepwise reasoning
Tocev, T.; Atanasovski, A.	Chain-of-Thought	IFRS advisory / financial reporting	Empirical (experimental)	Multi-step reasoning-oriented prompting
Anam, RK	Chain-of-Thought	Analytical and problem-solving tasks	Empirical (survey-based)	Explicit step-by-step reasoning requests
Mao, W.; Wu, J.; Chen, W.; Gao, C.; Wang, X.; He, X.	Chain-of-Thought	Personalized recommendation	Empirical (experimental)	Step-by-step reasoning to infer user preferences
Cheung, K.S.	Chain-of-Thought	Property valuation reporting	Conceptual (viewpoint)	Structured step-by-step prompt aligned with RICS "Red Book" valuation standards
Jeon at al (2025)	Self Conistency	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	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	Self Consistency	Data annotation / text classification	Empirical (mixed-methods)	Multiple independent attempts at same task

Appendix B.1 H1 Testing

- Cronbach's Alpha

	Scale	Cronbach_alpha
1	Output usefulness (QID112)	0.947

- LmerTest, Linear mixed-effects model

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: usefulness ~ technique + task_type + (1 | ResponseId)
   Data: usefulness_long

REML criterion at convergence: 466.2

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.8528 -0.3315  0.0644  0.4373  3.5531

Random effects:
 Groups     Name        Variance Std.Dev.
 ResponseId (Intercept) 0.03316  0.1821
 Residual               0.14280  0.3779
Number of obs: 420, groups:  ResponseId, 105

Fixed effects:
                      Estimate Std. Error        df t value Pr(>|t|)
(Intercept)            3.28459    0.05345 406.60916  61.452   <2e-16 ***
techniquecot           0.84251    0.06658 378.12497  12.653   <2e-16 ***
techniquefew_shot      0.82317    0.06712 379.27601  12.265   <2e-16 ***
techniquefew_shot_cot  1.52224    0.06694 379.99802  22.742   <2e-16 ***
techniqueone_shot      0.74002    0.06708 369.70219  11.033   <2e-16 ***
techniqueone_shot_cot  1.55228    0.06785 383.16249  22.878   <2e-16 ***
task_typesynthesis    -0.03317    0.03691 310.05832  -0.899    0.369
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) tchnqc tchnqf_ tchnqf__ tchnqn_ tchnqn__
techniquect -0.629
tchnqfw_sht -0.623  0.493
tchnqfw_sh_ -0.634  0.506  0.502
technqn_sht -0.608  0.487  0.480   0.495
tchnqn_sht_ -0.617  0.493  0.485   0.502    0.491
tsk_typsynt -0.351  0.013  0.014   0.017   -0.014  -0.005
```

- Lmer, Interaction model

```
> summary(model_int)
Linear mixed-effects model fit by REML
  Data: usefulness_long
       AIC      BIC    logLik
  501.7998 557.9576 -236.8999

Random effects:
 Formula: ~1 | ResponseId
        (Intercept)  Residual
StdDev:   0.1805242 0.3783835

Fixed effects:  usefulness ~ technique * task_type
                                         Value  Std.Error  DF  t-value
(Intercept)                           3.276038 0.06902763 304 47.45981
techniquecot                          0.869655 0.09404503 304  9.24722
techniquefew_shot                     0.875708 0.09434457 304  9.28202
techniquefew_shot_cot                 1.446359 0.09404197 304 15.37993
techniqueone_shot                     0.754815 0.09652355 304  7.82001
techniqueone_shot_cot                 1.589695 0.09709052 304 16.37332
task_typesynthesis                   -0.019584 0.09367808 304 -0.20906
techniquecot:task_typesynthesis      -0.052762 0.13308435 304 -0.39646
techniquefew_shot:task_typesynthesis -0.102974 0.13335861 304 -0.77216
techniquefew_shot_cot:task_typesynthesis  0.164014 0.13399136 304  1.22406
techniqueone_shot:task_typesynthesis -0.024367 0.13457147 304 -0.18107
techniqueone_shot_cot:task_typesynthesis -0.068791 0.13506436 304 -0.50932
                                        p-value
(Intercept)                              0.0000
techniquecot                             0.0000
techniquefew_shot                        0.0000
techniquefew_shot_cot                    0.0000
techniqueone_shot                        0.0000
techniqueone_shot_cot                    0.0000
task_typesynthesis                       0.8345
techniquecot:task_typesynthesis          0.6920
techniquefew_shot:task_typesynthesis     0.4406
techniquefew_shot_cot:task_typesynthesis  0.2219
techniqueone_shot:task_typesynthesis     0.8564
techniqueone_shot_cot:task_typesynthesis  0.6109
 Correlation:
                                        (Intr) tchnqc tchnqf_ tchnqf__
techniquecot                             0.701
```

```
 Correlation:
                                     (Intr) tchnqc tchnqf_ tchnqf__
techniquecot                         -0.701
techniquefew_shot                    -0.693  0.514
techniquefew_shot_cot                -0.701  0.519  0.514
techniqueone_shot                    -0.667  0.490  0.479   0.488
techniqueone_shot_cot                -0.682  0.503  0.499   0.505
task_typesynthesis                   -0.689  0.513  0.510   0.520
techniquecot:task_typesynthesis       0.501 -0.706 -0.371  -0.373
techniquefew_shot:task_typesynthesis  0.492 -0.364 -0.702  -0.369
techniquefew_shot_cot:task_typesynthesis  0.484 -0.353 -0.351  -0.701
techniqueone_shot:task_typesynthesis  0.482 -0.354 -0.348  -0.356
techniqueone_shot_cot:task_typesynthesis  0.487 -0.358 -0.358  -0.363
                                     tchnqn_ tchnqn__ tsk_ty tchn:_
techniquecot
techniquefew_shot
techniquefew_shot_cot
techniqueone_shot
techniqueone_shot_cot                 0.481
task_typesynthesis                    0.489   0.504
techniquecot:task_typesynthesis      -0.350  -0.365   -0.719
techniquefew_shot:task_typesynthesis -0.341  -0.357   -0.712  0.509
techniquefew_shot_cot:task_typesynthesis -0.333  -0.349   -0.712  0.501
techniqueone_shot:task_typesynthesis -0.718  -0.346   -0.701  0.496
techniqueone_shot_cot:task_typesynthesis -0.340  -0.715   -0.708  0.505
                                     tchnqf_:_ tchnqf__:_ tchnqn_:_
techniquecot
techniquefew_shot
techniquefew_shot_cot
techniqueone_shot
techniqueone_shot_cot
task_typesynthesis
techniquecot:task_typesynthesis
techniquefew_shot:task_typesynthesis
techniquefew_shot_cot:task_typesynthesis  0.499
techniqueone_shot:task_typesynthesis  0.490     0.489
techniqueone_shot_cot:task_typesynthesis  0.498     0.497      0.486
```

```
Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-4.66232810 -0.34392835  0.03387248  0.42450298  3.54311148

Number of Observations: 420
Number of Groups: 105
```

```
> anova(model_int)
                   numDF denDF   F-value p-value
(Intercept)            1   304 26804.081  <.0001
technique              5   304   144.640  <.0001
task_type              1   304     0.806  0.3700
technique:task_type    5   304     0.985  0.4271
```

- Posthoc, emmeans

```
> emm_tech
 technique    emmean     SE  df lower.CL upper.CL
 zero_shot      3.27 0.0501 393     3.17     3.37
 cot            4.11 0.0501 389     4.01     4.21
 few_shot       4.09 0.0509 389     3.99     4.19
 few_shot_cot   4.79 0.0500 396     4.69     4.89
 one_shot       4.01 0.0512 396     3.91     4.11
 one_shot_cot   4.82 0.0513 394     4.72     4.92

Results are averaged over the levels of: task_type
Degrees-of-freedom method: kenward-roger
Confidence level used: 0.95
```

```
> posthoc_all
 contrast                    estimate     SE  df t.ratio p.value
 zero_shot - cot              -0.8425 0.0668 378 -12.619 <0.0001
 zero_shot - few_shot         -0.8232 0.0673 379 -12.231 <0.0001
 zero_shot - few_shot_cot     -1.5222 0.0671 380 -22.679 <0.0001
 zero_shot - one_shot         -0.7400 0.0672 369 -11.007 <0.0001
 zero_shot - one_shot_cot     -1.5523 0.0680 383 -22.812 <0.0001
 cot - few_shot                0.0193 0.0675 386   0.286  0.9997
 cot - few_shot_cot           -0.6797 0.0665 373 -10.223 <0.0001
 cot - one_shot                0.1025 0.0679 384   1.510  0.6583
 cot - one_shot_cot           -0.7098 0.0679 384 -10.455 <0.0001
 few_shot - few_shot_cot      -0.6991 0.0671 376 -10.419 <0.0001
 few_shot - one_shot           0.0832 0.0686 388   1.211  0.8312
 few_shot - one_shot_cot      -0.7291 0.0687 389 -10.612 <0.0001
 few_shot_cot - one_shot       0.7822 0.0675 374  11.581 <0.0001
 few_shot_cot - one_shot_cot  -0.0300 0.0675 373  -0.445  0.9978
 one_shot - one_shot_cot      -0.8123 0.0683 375 -11.898 <0.0001

Results are averaged over the levels of: task_type
Degrees-of-freedom method: kenward-roger
P value adjustment: tukey method for comparing a family of 6 estimates
```

Appendix B.2 H2 Testing

- Cronbach's alpha

	Scale	Cronbach_alpha
1	Prompt usefulness	0.994

- LmerTest, Linear mixed-effects model

```
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: prompt_quality ~ technique + frequency_of_use + (1 | ResponseId)
   Data: prompt_long

REML criterion at convergence: 469

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.6762 -0.1429  0.1081  0.3810  3.0007

Random effects:
 Groups     Name        Variance Std.Dev.
 ResponseId (Intercept) 0.03822  0.1955
 Residual               0.14079  0.3752
Number of obs: 420, groups:  ResponseId, 105

Fixed effects:
                       Estimate Std. Error        df t value Pr(>|t|)
(Intercept)             3.07781    0.11130 134.58173  27.654   <2e-16 ***
techniquecot            1.05104    0.06641 375.07959  15.825   <2e-16 ***
techniquefew_shot       1.08081    0.06694 376.36500  16.145   <2e-16 ***
techniquefew_shot_cot   1.77718    0.06678 376.78132  26.612   <2e-16 ***
techniqueone_shot       0.94825    0.06688 366.72740  14.179   <2e-16 ***
techniqueone_shot_cot   1.87593    0.06771 380.04232  27.706   <2e-16 ***
frequency_of_use       -0.03670    0.02648 103.53913  -1.386    0.169
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) tchnqc tchnqf_ tchnqf__ tchnqn_ tchnqn__
techniquect -0.283
tchnqfw_sht -0.298  0.492
tchnqfw_sh_ -0.322  0.506  0.502
techniqn_sht -0.268  0.487  0.479   0.494
tchnqn_sht_ -0.282  0.493  0.484   0.502    0.491
freqncy_f_s -0.892 -0.018  0.002   0.024   -0.028  -0.016
```

- LmerTest, Interaction model

```
> summary(model_prompt_mod)
Linear mixed model fit by REML. t-tests use Satterthwaite's method [
lmerModLmerTest]
Formula: prompt_quality ~ technique * frequency_of_use + (1 | ResponseId)
   Data: prompt_long

REML criterion at convergence: 485.3

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7398 -0.1664  0.1217  0.3746  3.0061

Random effects:
 Groups     Name        Variance Std.Dev.
 ResponseId (Intercept) 0.03778  0.1944
 Residual               0.14176  0.3765
Number of obs: 420, groups:  ResponseId, 105

Fixed effects:
                                          Estimate Std. Error         df
(Intercept)                              3.154e+00  1.896e-01  3.798e+02
techniquecot                             1.053e+00  2.625e-01  3.589e+02
techniquefew_shot                        8.516e-01  2.458e-01  3.640e+02
techniquefew_shot_cot                    1.905e+00  2.569e-01  3.805e+02
techniqueone_shot                        8.156e-01  2.591e-01  3.497e+02
techniqueone_shot_cot                    1.667e+00  2.620e-01  3.736e+02
frequency_of_use                        -5.706e-02  4.900e-02  3.799e+02
techniquecot:frequency_of_use            1.649e-05  6.734e-02  3.611e+02
techniquefew_shot:frequency_of_use       6.183e-02  6.357e-02  3.664e+02
techniquefew_shot_cot:frequency_of_use -3.613e-02  6.752e-02  3.809e+02
techniqueone_shot:frequency_of_use       3.507e-02  6.611e-02  3.548e+02
techniqueone_shot_cot:frequency_of_use   5.505e-02  6.703e-02  3.729e+02

                                       t value Pr(>|t|)
(Intercept)                             16.629  < 2e-16 ***
techniquecot                             4.012 7.34e-05 ***
techniquefew_shot                        3.464 0.000595 ***
techniquefew_shot_cot                    7.417 7.88e-13 ***
techniqueone_shot                        3.148 0.001788 **
techniqueone_shot_cot                    6.364 5.76e-10 ***
frequency_of_use                        -1.164 0.245002
techniquecot:frequency_of_use            0.000 0.999805
techniquefew_shot:frequency_of_use       0.973 0.331432
techniquefew_shot_cot:frequency_of_use  -0.535 0.592874
techniqueone_shot:frequency_of_use       0.531 0.596083
techniqueone_shot_cot:frequency_of_use   0.821 0.412001
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) tchnqc tchnqf_ tchnqf__ tchnqn_ tchnqn__ frqn__ tch:__
techniquect -0.618
tchnqfw_sht -0.666  0.476
tchnqfw_sh_ -0.654  0.482  0.510
technqn_sht -0.614  0.436  0.468   0.456
tchnqn_sht_ -0.630  0.440  0.488   0.476    0.455
freqncy_f_s -0.964  0.597  0.642   0.631    0.595   0.607
tchnqct:f__  0.601 -0.967 -0.463  -0.469   -0.427  -0.426   -0.626
tchnqfw_:__  0.641 -0.460 -0.962  -0.490   -0.452  -0.469   -0.666  0.482
tchnqf__:__  0.620 -0.459 -0.483  -0.965   -0.434  -0.451   -0.644  0.480
tchnqn_s:__  0.602 -0.429 -0.458  -0.447   -0.966  -0.446   -0.628  0.453
tchnqn__:__  0.613 -0.428 -0.474  -0.464   -0.445  -0.966   -0.636  0.447
            tchnqf_:__ tchnqf__:__ tchnqn_:__
techniquect
tchnqfw_sht
tchnqfw_sh_
technqn_sht
tchnqn_sht_
freqncy_f_s
tchnqct:f__
tchnqfw_:__
tchnqf__:__  0.498
tchnqn_s:__  0.477      0.459
tchnqn__:__  0.490      0.473      0.470
```

- Likelihood Ratio Test (LRT)

```
> anova(model_prompt_base, model_prompt_mod)
```

```
refitting model(s) with ML (instead of REML)
```

```
Data: prompt_long
Models:
model_prompt_base: prompt_quality ~ technique + frequency_of_use + (1 | ResponseI
d)
model_prompt_mod: prompt_quality ~ technique * frequency_of_use + (1 | ResponseId)
                  npar    AIC    BIC  logLik -2*log(L)  Chisq Df Pr(>Chisq)
model_prompt_base    9 456.64 493.01 -219.32    438.64
model_prompt_mod    14 463.40 519.96 -217.70    435.40 3.2447  5     0.6623
```