Prompt engineering guidelines for LLMs in Requirements Engineering

Bachelor of Science Thesis in Software Engineering and Management

Simon Arvidsson
Johan Axell

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2023

Prompt engineering guidelines for Requirements Engineering
This paper explores Prompt engineering guidelines for generative AI models as well as their advantages and limitations for Requirement Engineering.

Supervisor: Krishna Ronanki
Examiner: Richard Berntsson Svensson

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering University of Gothenburg
Chalmers University of Technology
Gothenburg, Sweden 2023

Prompt engineering guidelines for LLMs in Requirements Engineering

1st Simon Arvidsson
dept. of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gussimoar@student.gu.se

2nd Johan Axell
dept. of Computer Science and Engineering
University of Gothenburg
Gothenburg, Sweden
gusaxelljo@student.gu.se

***Abstract*—The rapid emergence of large generative AI models has demonstrated their utility across a multitude of tasks. Ensuring the quality and accuracy of the models' output is done in different ways. In this study, we focused on prompt engineering. Prompt engineering guidelines for how to utilize large generative AI models in the field of requirements engineering are limited in the literature. The objective of this study was to explore the potential advantages and limitations of the possible application of existing prompt engineering guidelines from the literature in requirements engineering. To achieve this goal, we conducted a systematic literature review on prompt engineering guidelines to gather guidelines which could be applicable to various tasks. Subsequently, we considered different requirements engineering activities and their characteristics before proposing a mapping of our gathered guidelines to requirements engineering activities. Furthermore, we conducted interviews with three requirements engineering experts to gain further perspectives on our findings and mapping suggestions. Through thematic analysis, we extracted the advantages and limitations of the mapping. While our review shows how prompt guidelines for domain-specific tasks still are limited in literature, we did identify prompt guidelines in the current literature which show promise when working with an LLM in the practice of requirements specification. Additionally, we draw the conclusion that large generative AI models as we know them might not be fully ready for certain tasks in requirements engineering and suggest future work to explore how guidelines could be adapted to fit other requirements engineering tasks better.**

***Index Terms*—Requirements Engineering, Prompt Engineering, Generative AI, LLM, Prompt Guidelines**

I. Introduction

Requirements engineering (RE) is an important part of software development. It can positively impact other development processes such as testing, product quality, and project planning [1], [2], while poorly executed RE activities can often drive the failure of projects [3]. RE is not a new topic and discussions regarding the criticality of the issues RE aims to address date back to 1987 [4]. With the increase of research in topic areas of artificial intelligence (AI) and the development of the artificial intelligence for requirements engineering (AI4RE) field, the utilization of AI-powered solutions helps to facilitate the management of different RE activities by reducing time consumption, complexity, and human effort [5], [6]. It can also help with error identification related to RE. These errors are crucial to address while still in the RE phase in order to help mitigate unnecessary expensive corrections of the errors at later phases [7], [8]. Often, the errors arise due to different interpretations or terminology among stakeholders and the written requirements being ambiguous [9]. These requirements are written using natural language (NL), and one sub-fields of AI4RE addresses AI for NL and include natural language processing (NLP). NLP incorporates different computational techniques in order to enable interaction between AI and humans through the use of natural language.

This described subfield of AI4RE, known as NLP4RE, revolves around the application of NLP techniques in activities related to RE like requirement elicitation and classification. The 2018 NLP4RE workshop highlighted the growing significance of NLP as a fundamental component in domains including RE [10].

In the context of NLP4RE, it has been claimed that various NLP techniques provide support for different aspects of requirement elicitation such as in the context of text generation, question-asking- and answering, as well as text-to-speech [11]. Other work in the NLP4RE field, has also presented how NLP can be of assistance to detect ambiguity or other issues [9], [12], [13]. One of the common techniques for accomplishing such NLP tasks is by using language models. As one of the essential aspects of RE is the use of natural language [14], language models could be utilized in order to understand the content of requirements as well as its context [15], [16].

Language models aim to recognize and understand the intent and the context of language in environments such as speech or text. They exist in different forms, such as probabilistic language models or deep neural network-based models. Language models that are trained on a larger amount of data with a significantly larger amount of parameters in comparison to regular language models are referred to as Large language models (LLM). Given the larger training set, LLMs can provide a wider understanding of linguistics which gives direct advantages that enable LLMs to be used in systems for high-quality language generation or learning a wider range of complex domain-specific knowledge through fine-tuning [17], [18]. Within software engineering, some LLM-based tools are assisting in programming activities by predicting the intended functionality of code to improve productivity among developers [19], [20].

The quality of the output generated by a language model is largely dependent on the prompt it received. Prompts in this context can for example be a sentence, a question, or an instruction given to the model in natural language [21]. Prompt Engineering (PE) is a process focusing on creating, optimizing, and refining prompts to ensure the relevance and quality of the output, which encompasses the generation of information that aligns with the intended purpose and is linguistically sound. The usage of PE can strengthen the quality of what a language model interprets as well as any generated output [22]. For LLMs, PE is essential as LLMs' deep understanding of linguistics poses a challenge in formulating high-quality prompts, which in turn affects the quality of any generated output.

The use of PE enables LLMs capabilities of solving NLP tasks at a more advanced level. As NLP has been, and still is used for RE tasks while natural language remains a central component of RE, the idea of using PE for LLMs in RE emerged. However, for LLMs to be of use within RE, it must obtain knowledge of RE activities and learn the applied context in order to create an understanding.

Model training or knowledge transfer can be performed using various techniques. It is often categorized in the stages of pre-training, fine-tuning, and prompt learning [21]–[23]. This work will however focus on prompts in natural language and aims to review the existing PE guidelines for LLMs, which can be used for designing effective natural language prompts to assist in RE. The existing guidelines will provide a starting point for proposed guidelines that can ensure the effective use of LLMs in RE activities.

The emergence of LLMs has introduced various opportunities in a multitude of fields within Software engineering, one of them being RE [24]–[26]. Because of this, there is a need to identify and gather existing literature on PE guidelines for natural language and their application to large generative models, including LLMs. This study focuses on how these guidelines in turn can help in designing prompts for LLMs' usage in RE activities and aim to achieve this through the mapping of said guidelines and the similarities between the guidelines and RE activities. Furthermore, the study contributes by addressing the gap in the current literature in order to provide guidance and a framework for practitioners and researchers alike within the field. Our findings could potentially lead to improved effectiveness in certain RE tasks, help in forming new tailored guidelines for RE specifically, and reduce errors and time consumption. Because of this, the significance of this research lies primarily in the possible utilization of LLMs in RE and secondarily in providing an overview of existing gaps in the literature for opportunities in future work.

II. Related Work

In the literature, RE can be described as the process of eliciting, analyzing, documenting, validating, and managing software requirements in order to ensure that the software being developed meets the needs and expectations set by the respective stakeholders [27]. The following definitions are the ones that will be used for this paper. It is however worth noting that the definition of the following activities within RE can vary between practitioners.

Elicitation revolves around deepening the comprehension of the project along with its limitations, as well as what is expected by the stakeholders. In certain cases, elicitation can also incorporate activities that serve to look into different options for how these expectations could be met [27].

The aspect of requirement analysis emphasizes the understanding of how the implementation of requirements will be conducted [28]. Apart from this, requirement analysis also serves to identify the level of incompleteness and ambiguity within requirements, as well as if they could be conflicting [27].

After analyzing, the process of documenting requirements takes place. Doing so incorporates making sure that they are clear and consistent as well as traceable [27]. Documenting requirements is often defined as part of the specification phase. In the majority of cases, the documentation of requirements is done in natural language. One of the major issues with natural language is how imprecise it is, which results in requirements turning out ambiguous, incomplete, or even inaccurate [29]. According to [8], 28% of all bugs encountered in a software project are due to incomplete or ambiguous requirements.

Requirement validation can be described as the process of making sure the established requirements are complete and correspond to the expectations of the stakeholders [30]. This process serves to make sure that the correct system is being built.

Requirements management can be described as a cycle that proceeds throughout the entirety of the lifespan of a project [31]. It relates to many sub-activities within the field of RE, some of the more prominent ones being documentation and analysis, as well as tracing and prioritizing requirements. Requirement tracing focuses on being able to follow the life cycle of a requirement, not only backward but forward as well [32].

These mentioned papers and books related to RE have been instrumental in shaping our understanding of the different RE activities and how they ought to be implemented. They also play a vital role in the foundation for our ability to achieve some of our research objectives.

When looking at the context of RE and the various activities which RE includes, it has been shown how AI technologies can be of great use [33], especially for analysis and elicitation [5], [6], [9].

The utilization of AI for RE as a field has made a lot of progress in the past decades, especially since the introduction of NLP with the use of machine learning and deep learning which in NLP4RE'18 was mentioned to facilitate utilization of NLP tools and techniques [10].

These papers further enhance the evidence that the application of AI technologies holds significant potential when combined with RE, something which played an integral part in shaping the topic and objectives of this paper.

When looking into the recent literature and the field of PE, a great extent of the research focus has been put into generative models, large generative models in particular. Previously the overall focus in the literature has largely been on just text-to-image generation tasks but recently shifted more toward text-to-text generation since the quick emergence of LLMs. Several studies have suggested PE guidelines for such large generative models [34]–[36], and commercial companies provide these guidelines for prompt design in their product documentation [37]. Other studies have provided guidelines for different prompting strategies and patterns, such as [18], [22], [38]. The literature on PE guidelines and approaches tied specifically to RE remains remarkably limited at the time of this study, with only a few papers partially covering guidelines for tasks and prompt learning within RE activities [24], [25].

These sources were also crucial for shaping and enhancing our understanding of generative AI and its uses. They provide insight to different prompting techniques and an introduction to what guidelines within PE could entail, this was useful as it assisted in our work to identify new guidelines down the line. The identified gaps within the field of PE and their applications within RE, acted as one of the main motivators for conducting this study.

III. Research Methodology

The framework presented in [39] was used as the foundation for this systematic literature review. Alongside the review to answer RQ1, research synthesis was conducted to answer RQ2 and interviews were conducted to answer RQ3. The review consisted of three stages – planning, conducting, and reporting. A flowchart illustrating the process used for conducting the review is presented in Fig. 2. Furthermore, research synthesis was performed to gather and synthesize data on the most commonly performed activities within RE. In order to gain additional perspectives on the findings as well as highlight the advantages and limitations of our results, interviews with experts within the RE field were conducted.

A. Research questions and objective

The usage of large generative AI models for NLP tasks has increased dramatically during the last decade. This increase has also resulted in a broader field of use. However, the research regarding how to best utilize these models for domain-specific tasks is limited at best, especially when looking into PE guidelines for downstream tasks in domain areas such as RE. We have looked further into how these models can be utilized through the lens of PE, as well as what guidelines currently exist in the same domain. Furthermore, we explored the possible usage of these guidelines within the domain of RE, as well as what advantages and limitations they may introduce. In order for us to realize these objectives, the following research questions were formed:

RQ1: What does the existing literature regarding PE guidelines for large generative AI models say?

RQ2: What are the relevant guidelines found in **RQ1** that can be used in RE activities?

RQ3: What are the advantages and the limitations the identified guidelines provide for the usage of LLMs in RE?

B. Planning the review

Prior to conducting the review itself, a review protocol was established and evaluated. The establishment of this protocol was done iteratively to make sure new information was always considered and acted on, one example of this was information that indicates the need for revision or certain criteria. Subsequent sections will further describe the different segments included in this protocol.

1) Study selection: Through prior research within related domains as well as snowballing based on this research, we identified keywords and in combination with our objective and research questions, we created a search string used for the selection of papers relevant to this study. During the exploration of studies including these keywords, it was evident that there was a tangible discrepancy regarding the used terminology which was why the resulting string had to be extended and resulted in ("Prompt engineering" OR "Prompt Patterns" OR "Prompt Design" OR "Prompt Catalog" OR "Prompt Guidelines" AND "Large Language Models" OR "Generative AI"). We adapted the search string to fit the interface characteristics of the search engines for each database respectively. This was done without any changes to the included keywords.

We chose to limit our search to only include papers that were published from the year 2018 onward. 2018 was chosen due to transformer architecture being presented along with OpenAI's generative pre-trained transformer (GPT) model, and Google's Bidirectional Encoder Representations from Transformers (BERT) model launching this year, which can be seen as the start of the emergence of large generative models as we know them today [40]–[43]. This limitation also helps ensure relevancy and state-of-the-art within the field of PE to match the rapid development of large generative models.

The 5 databases used for this review were ACM, Scopus, IEEE Explore, Science Direct, and arXiv. The basis for choosing the first 4 was mainly the comprehensive coverage they provide. They are also well-established and respected publishers within fields such as computer science and software engineering, as well as within the academic community. At a later stage in the review, arXiv was added in order to capture newer papers which proved to be a vital part of the review due to how fast the field of PE has been growing recently. Adding arXiv does add another threat to validity which is further discussed in the limitations section.

2) Inclusion and Exclusion criteria: In order to find the primary studies and ensure their relevancy for this SLR, inclusion and exclusion criteria were applied to the search. Table I shows the criteria used for filtering the papers.

The evaluation of each study was based on the title, abstract, introduction, conclusion, and keywords. Whenever further ambiguity was present after this stage, the rest of the study was read and evaluated.

TABLE I
INCLUSION AND EXCLUSION CRITERIA FOR THE REVIEW.

Inclusion criteria	Exclusion criteria
Written in English language	Papers with sections or content in languages other than English
Date of publication from 2018	Published prior to 2018
Emphasizes generative AI models	Unrelated to generative AI models
Focus on Natural language prompts	Emphasis on model tuning
Relevant to RQ1	Does not contain PE guidelines

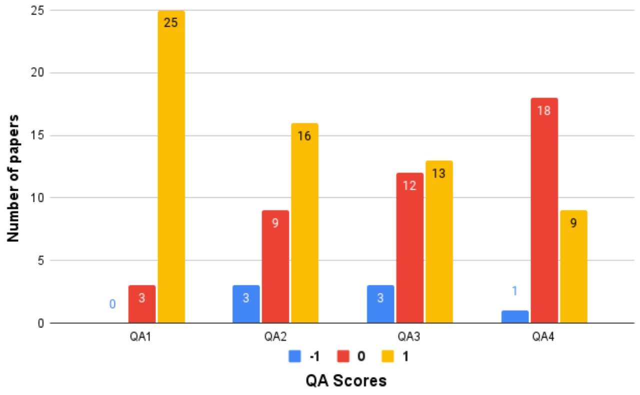


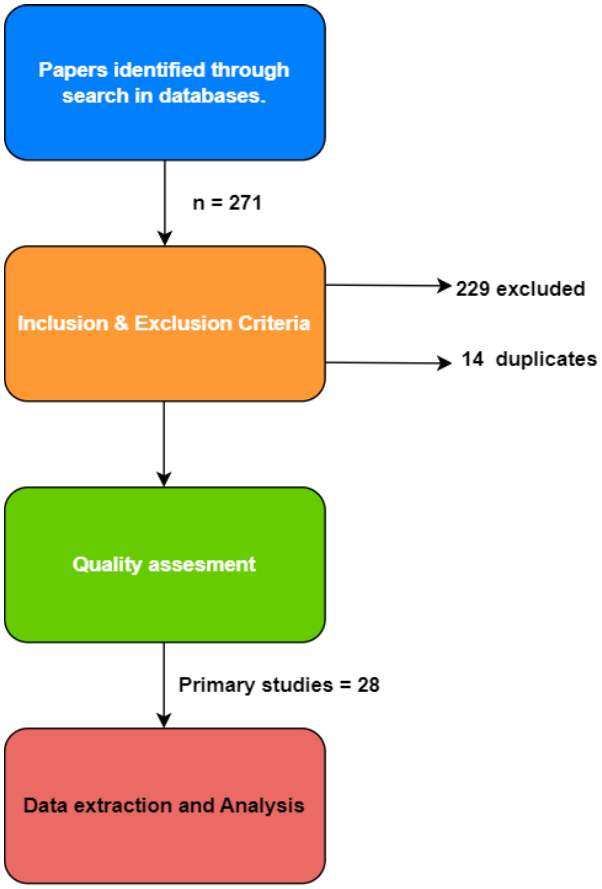
Fig. 1. An overview of the scores per quality assessment (QA) criteria and the number of papers that received each score in the quality assessment. It serves to provide transparency regarding the quality of the primary studies and relevance to the review. -1 indicating a disagreement with the QA criteria, 0 implying an unsure state, and 1 indicating an agreement.

C. Conducting the review

1) Study selection result: After applying the search string to the chosen databases, 271 studies were included for the initial screening. Post the evaluation using inclusion and exclusion criteria (Table I) and removing duplicates, 28 (10,3%) were recognized as primary studies. The result can be seen in Table III.

2) Criteria for quality assessment: When the primary studies had been identified, further evaluation through quality assessment was conducted. This evaluation served solely to provide transparency regarding the quality of the primary studies used in the data extraction, as well as their relevance to our SLR. This was done by reading the papers in their entirety while applying the pre-established quality assessment criteria and scoring each paper -1, 0, or 1. -1 indicating a disagreement with the criteria, 1 implying an agreement, and 0 representing an unsure state in between 1 and -1. These criteria can be seen in Table II, and the result of the quality assessment is displayed in Fig. 1.

3) Data collection and synthesis: In order to conduct extraction of relevant data, the form depicted in Table IV was used. This form was established in order to ensure an organized as well as standardized data collection process. It also facilitated the process of distinguishing relevant data for our research question.



Papers identified through search in databases.
n = 271
Inclusion & Exclusion Criteria
229 excluded
14 duplicates
Quality assesment
Primary studies = 28
Data extraction and Analysis

Fig. 2. A flowchart of the stages in the review process and how the filtering of studies was conducted.

TABLE II
QUALITY ASSESSMENT CRITERIA

Item	Assessment Criteria	Score Scale
QA1	To what extent does the study demonstrate that it has achieved its stated objective(s) in a concrete and detailed way?	1
		0
		-1
QA2	Are the limitations of the study clearly described and discussed?	1
		0
		-1
QA3	Does the study provide contribution to the field of prompt engineering?	1
		0
		-1
QA4	Does the study provide insight to better understand how to prompt large generative AI models?	1
		0
		-1

D. Mapping through thematic synthesis

After extracting the data from the literature review in order to identify and categorize the PE guidelines, thematic synthesis was conducted with the goal of suggesting a mapping between the guidelines and activities within RE. Given the discrepancy in terminology, the definition of these categories varied. The definitions used for this paper are described in the related work section.

Research synthesis can be described as a collection of

TABLE III
SEARCH RESULTS

Database	Count of studies	Count after applying inclusion-exclusion criteria
ACM	129	6
Scopus	17	1
IEEE Xplore	7	1
Science Direct	6	0
arXiv	112	20
Total	271	28

TABLE IV
DATA EXTRACTION

Data	Description	Relevant RQ
DOI	The unique document identifier	General
Year		General
Model type	Text-to-image, text-to-text, GPT-3, BLOOM, Codex, etc.	RQ1
Prompt method	What PE techniques were studied?	RQ1
Guidelines	What guidelines are presented?	RQ1
Findings	Strengths or limitations of the guidelines presented	RQ1

methods used for summarizing, integrating, combining, and comparing findings collected on a specific topic. The methods themselves serve to create something new, such as a theory, conclusion, or framework out of parts gathered from research. Inspiration for the steps used in our synthesis was collected from [44].

The process was initiated by reading and extracting data from various papers and books which emphasized an overview of the activities often conducted within RE. Apart from these sources, we also included research focused on aspects that have proven to be important in order to conduct these activities successfully. Google Scholar was used as the main search engine for locating these papers.

When deciding on what papers to include, we looked primarily at the number of citations as well as how recently the paper or book was published. The data extraction process followed a similar pattern to the one used for the SLR. In this case, we extracted title, year of publication, amount of citations, RE activities described as well as any aspects pointing to what would enhance the success rate of these activities.

When the data was extracted, the focus shifted to identifying the most commonly used activities within RE. This was done by looking at the number of mentions of each activity as well as their respective description. Apart from this, we also noted the characteristics and aspects which were described as the most crucial ones for exercising these activities as successfully as possible. These in turn were categorized as themes.

These activities and identified themes, combined with the PE guideline themes provided the foundation for our proposed mapping. The mapping was conducted by looking at these themes and then finding connections by identifying common denominators to the themes created for the PE guidelines. Lastly, we assessed the trustworthiness and plausibility of the proposed mapping by discussing looking into whether or not we have considered any misinterpretations or assumptions, as well as if the suggested mapping can be considered a good fit compared to what the identified evidence shows.

E. Interviews and thematic analysis

In order to further explore the plausibility, as well as limitations and advantages of the suggested mapping, we reached out to 3 experts at different academic institutions, specializing in different areas within RE.

The primary expertise sought for this evaluation was in the field of Requirement Engineering, however, knowledge within AI4RE and related areas were considered an added advantage. The experts were selected based on this expertise in combination with recommendations from our academic supervisor. Consideration was also given to the experts' experience and publication history. While the emphasis of our research was on requirements engineering, which inherently has a narrower scope, the backgrounds of the experts do provide a variety of perspectives. These backgrounds include different European academic institutions and include but are not limited to the fields of Requirements Engineering for Machine Learning, Requirement Engineering for AI, and NLP4RE.

The participants were contacted through email with a short introduction of the thesis subject itself along with details regarding how the interview would be conducted. During this exchange, it was agreed that the interviews would be conducted online over Zoom. The length of the interviews varied slightly but on average lasted around 20 minutes. Sound and video were recorded and consent for this was collected from each participant before beginning. As our main goal was to identify and extract advantages and limitations, the questions asked were closed ones with minimal room for follow-up questions or elaborations outside of what was asked for. Due to the nature of the subject as well as the anonymity provided when participating, no particular considerations regarding ethics had to be taken.

In order to analyze the interviews which were used to gain perspectives on the suggested mapping, and find answers to RQ3, a thematic analysis was conducted. We chose this method due to it being flexible and adaptable to capture nuanced data while providing a suitable way of working collaboratively[45].

In this analysis, we opted initially for a deductive as well as exploratory approach. It is deductive in the sense that we knew beforehand how wanted to capture advantages and limitations. However, it also contains inductive elements in the sense that we discovered two new themes during the analysis. The analysis can be categorized as exploratory as we are not attempting to prove or disprove a hypothesis, but rather explore the different perspectives of the experts on a specific topic[46].

The overall process used for this analysis was derived from [45].

The process was initiated by re-reading the transcript collaboratively but also re-listening to the audio recordings to enhance our understanding of the data and what was mentioned during interviews. To highlight the content of the importance of the interviews, the second step of the analysis was coding the transcripts. By doing this step collaboratively, the codes were developed and agreed upon with both our interpretations. An example of a generated code is “knowledge of domain context needed” from the transcript data “if you elicitate requirements, you first also need to elicitate the context in which these requirements are valid.”.

Another example from the analysis is the code “PE can help find contradicting requirements” that was generated from the following part of the transcript

“it can help you find inconsistencies among the requirements, because contradictions can be found by large language models, so they have somehow understood also some logical aspects”.

As we sought two main points based on RQ3, advantages, and limitations, they also served as two distinctive themes in the analysis. This was possible due to the nature of the interview questions, which asked about the advantages and limitations of the guidelines. The data in the themes of advantages and limitations serve as the basis of our answer to RQ3. We present the results of advantages and limitations in detail in Table VIII and IX, and in sub-section C of the results.

However, we additionally included two more themes. Firstly, “Suggestions”, as many ideas for further application of guidelines were present in the data. This theme included data that suggested further ideas beyond the asked questions’ advantages and limitations. Besides the results, we included and considered this theme also when mentioning possible further research in section VI.

Secondly, the theme “Uncertainty”. The theme encapsulated codes that capture perspectives which can be perceived as incertitude or doubt.

Based on these themes, we found some interesting commonalities, and these can be found in sub-section D of the results section.

F. Limitations

A study of this character has different limitations which affect the possible conclusions of the study.

After carefully assessing and deciding on a time frame for our review, the time frame of the review that ranged 2018-present still may introduce a time frame bias with a risk that studies that could have been valuable for the review have been disregarded. Although the limitation may be a lesser risk as most LLMs are based on the transformer architecture which was first presented at the end of 2017 [41]. And as our results show after this study was conducted, we did not find PE guidelines for natural language prompts in literature form between 2018-2020, and instead it appeared first in 2021 while gaining greater traction during 2022. This may mean a lower risk for a time frame bias existing in our review.

However, another aspect to consider in regard to the limitations of this study is the interviewed experts. They may have individual beliefs they favor about certain aspects of the questions which could mean a risk of introducing confirmation bias in our interviews.

1) Construct validity: Construct validity describes the correctness of the study’s assessment in comparison to what is sought to be assessed. This study’s exclusion and inclusion criteria, as well as the quality assessment criteria, may have affected the construct validity of this study as described by Zhou *et al.* in [47]. In the interviews, the experts may all naturally have a better understanding of specific areas within RE, favoring or disfavoring certain mappings based on a better or worse understanding of specific activities but also guidelines. This threat could have been mitigated by developing systematic selection criteria for the selection of experts to interview.

2) Internal validity threats: Causal relationships being credible and not affected by other factors within the study is referred to as internal validity. As described in [47], the internal validity may be affected by several factors. One of the described factors of internal validity threats mentioned in [47] that may affect this study is the limited number of samples currently available in the literature on the topic, and the selection of studies to review could then pose a selection bias.

Further, the area of research and the literature is developing rapidly which may pose a threat to the internal validity as inconsistencies of terminology in studies exist. Considering this issue, it is important to acknowledge that there is a risk of creating misinterpretations based on personal perspectives. By utilizing a continuous and open dialog discussing cases of uncertainty around some studies’ content and, at times, specific terminology used, we aimed to mitigate these risks.

Additionally, the selection of RE activities and the suggested mapping to prompt guidelines could also be viewed as an internal validity threat as only the RE activities with the most similar characteristics as the guidelines were considered at the mapping stage.

3) External validity threats: The generalizability of the study to other contexts is referred to as external validity.

In a review, a reviewed study may suffer from limited generalizability. The limited generalizability may be caused by the study’s research focus or research method being specific to a certain context. Another threat to reviews, identified by Zhou *et al.* in [47], is the risk of reviewing primary studies in which research information is insufficient for the review as a whole. This may be a threat to external validity as the state of the literature on the reviewed topic is limited.

Additionally, the limited size of our study’s time frame and rapid development within this area of research both pose threats to this study’s external validity. Because of the rapid development in the area, our review includes unpublished and non-peer-reviewed studies from arXiv which introduce an additional risk of bias and threat to external validity. However, by including this source of studies we allow the review to capture the new unpublished but peer-reviewed studies of high quality with state-of-the-art prompt guidelines.

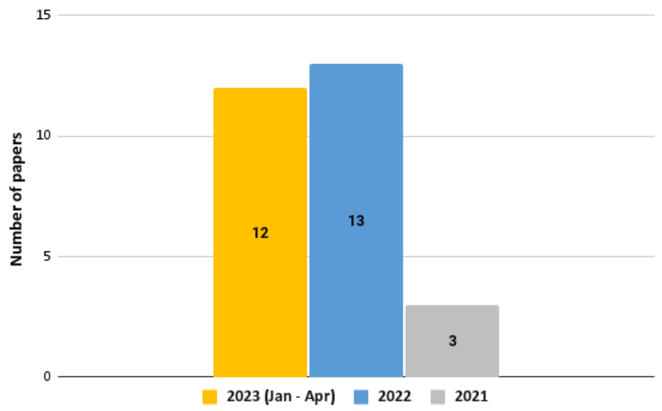


Fig. 3. The distribution of the paper publications years among the studies from which the PE guidelines were extracted.

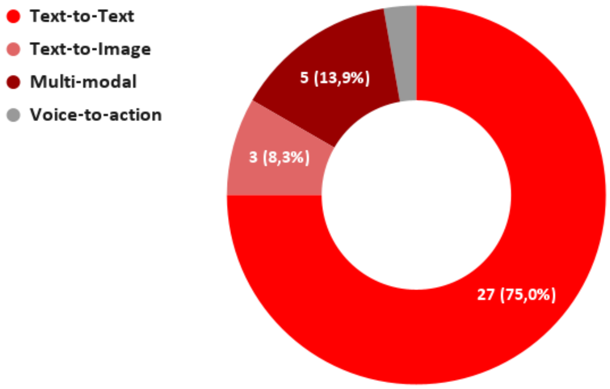


Fig. 4. The number of guidelines per generative model type found in the review.

IV. Results

The results of this study are divided into three subsections. The first subsection, subsection A, presents our findings, its' details, and answers for **RQ1** with a non-exhaustive list of PE guidelines, categorized into the 10 most occurring themes among the identified guidelines.

In the second subsection, subsection B, we present our findings and answer **RQ2** by suggesting a mapping of guidelines and themes from RQ1 to various components within RE activities of similar nature.

The third subsection, subsection C, presents possible advantages and limitations when applying the PE guidelines in RE. By conducting interviews with 3 RE experts, the advantages and limitations of mapped guidelines and RE activities from RQ2 were discussed and established as the answer to **RQ3** as well as additional perspectives on **RQ2**.

A. SLR and guideline categorization(RQ1)

The review was conducted as documented in section III and Table I, II, III, and IV. The flow chart in Fig. 2 displays the workflow of the filtering process of primary studies in the review, and also the final number of studies used for data extraction.

The studies included in the data extraction phase (n=28) are all published within the range of the past three years of this study as displayed in Fig. 3. These show a somewhat varying but centered model type variation for the extracted PE guidelines. These application areas include four different model types. The more significant majority of the studies present and apply the PE guidelines to Text-to-text models while a minority apply them to Text-to-image, Multimodal, and Voice-to-action models as shown in Fig. 4.

From these studies, we identified and extracted a total of 36 PE guidelines for natural language prompts. We categorized these 36 guidelines into 10 different themes based on the characteristics of the extracted guidelines. Each of the themes that were used to categorize the guidelines was created with the following definitions in order to outline the underlying concepts and group similar guidelines for a clearer overview and understanding of the guidelines' overall intent.

1) Context: The "Context" theme revolves around guidelines that relate to contextual information in any manner for prompts.

2) Persona: The "Persona" theme is a high-level abstraction of guidelines that revolves around strategies for LLMs to take on specific or different perspectives on specified tasks by using prompts. Persona in this instance can be compared to perspectives or points of view.

3) Templates: The "Templates" theme encapsulates guidelines that only provide an explicit structure of prompts, known as a template.

4) Disambiguation: The "Disambiguation" theme refers to guidelines that aim to address ambiguity, clarification, or understanding of intent.

5) Reasoning: The "Reasoning" theme captures guidelines that aim to affect reasoning capabilities or the ability to think through complex problems or tasks in a generated output.

6) Analysis: The theme "Analysis" revolve around guidelines examining, evaluating, or analyzing information or tasks.

7) Keywords: The theme "Keywords" represent guidelines that involve any use of single-word modifiers to prompts.

8) Wording: The theme "Wording" refers to guidelines that relate to choices of words, text formatting, writing styles, or inclusion and exclusion of text.

9) Shorten: The theme "Shorten" captures guidelines that highlight summarizing, paraphrasing, or text shortening.

10) Few-shot Prompts: The theme "Few-shot Prompts" categorizes guidelines that are intended for any form of few-shot prompting.

Each guideline was categorized into one of the themes and additionally given a number as an identifier and listed in Table VI. The theme categorization gives insight into what types of guidelines are common in literature. As can be seen in Fig. 5, the themes "Context", "Wording", and "Few-shot prompts" guidelines are among the more common types. Further, when referring to specific guidelines in this paper, the initial letter of the theme in combination with the identifier of the guideline

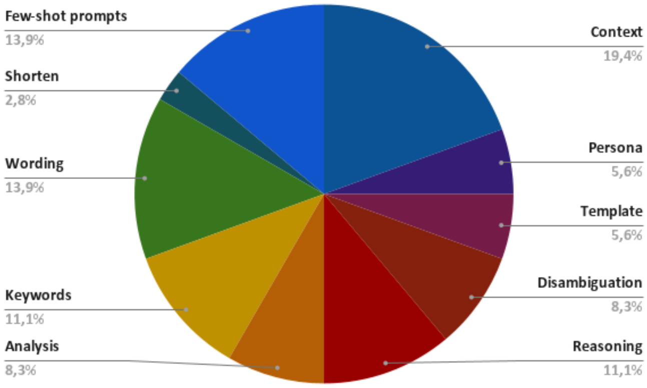


Fig. 5. Theme distribution of the extracted guidelines from the studies in the review.

TABLE V
MODEL TYPES AND MODELS FROM THE REVIEW MAPPED TO RESPECTIVE GUIDELINES

Model Type	Models	Guidelines
Text-to-text	GPT-2, GPT-3, GPT-3.5, T0, BLOOM, OPT, InstructGPT, EleutherAI, GPT-J, Galactica, BioBERT Comet, Codex, GitHubCopilot, PaLM-540B	C1, C2, C3, C4, C5, P1, P2, T1 D1, D2, D3, R1, R2, R3, A1, A2, A3, K2 W1, W3, W4, W5 F1, F2, F3, F4, F5
Text-to-image	DALL-E, Midjourney, Imagen, (VQGAN)	T2, K1, K3
Multimodal	CLIP, GPT-4	C6, C7, T1, K4, W2
Voice-to-Action	Undisclosed	S1

will be used. For example, C1 refers to the first listed guideline within the "Context" theme.

Table V displays the mapping of the themed guidelines to the respective model type to give a summary of what guidelines originate from which model type. In the same table, we additionally include a mapping of the large generative models that were used in the studies. As Table V shows, there is a wide variety of models used in the reviewed studies. Encoder-only, Decoder-only, and Encoder-Decoder models are all present in the review with guidelines for Encoder-only models such as BioBERT (E.g. R1) [48], Decoder-only models such as GPT-3 (E.g. P1) [49], and Encoder-Decoder models such as the model T0 (E.g. C4) [50].

The presented findings in the reviewed studies pose a wide agreement that the context and structure of prompts have a significant impact on generated output. At the same time, among the studies in the review, there are a multitude of guidelines of approaches to how to achieve an output of good quality. Some studies report results with the presented guidelines in few-shot prompting showing improved output compared to their studied alternatives ([52], [62], [70]–[72]), at the same time other studies report the opposite showing results of cases where other various guidelines in few-shot prompting fall short [60], [63].

Moreover, some studies in the review show results of the specific models GPT-3, GPT-3.5, and ChatGPT being incapable of performing well at tasks involving emotions or mathematical reasoning regardless of the usage of PE [18], [64]. Studies also show how the same models can confidently perform tasks providing incorrect facts and inaccurate data, often referred to as hallucinations, and suggests approaches as attempts of mitigating such fallacies [18], [52], [57], [73], [74].

Some of the more recently published, among our reviewed studies, apply approaches and guidelines from other reviewed studies to various new domains. Art, software engineering, and healthcare are common domains in our review and often provide guidelines and ideas for the respective domains [22], [24], [52], [58], [60], [62]–[64], [68], [70].

B. RE activities and applicable guidelines (RQ2)

What follows is the proposed mapping between each RE activity and the identified guideline themes, along with the justifications for it. A table of this mapping can be found in Table V.

1) Requirement Elicitation: We theorize that the guidelines found under the theme "context" could provide value to this activity. If elicitation is conducted with generative AI as an assistive tool, incorporating context into the prompts could prove to be a vital parameter considering the importance of extensive knowledge about the system as well as the stakeholders' needs when eliciting requirements.

2) Requirement Analysis: When analyzing requirements, we propose that guidelines found within the themes of Reasoning, Templates, and Analysis could be suitable. Reasoning would help in order to provide a clearer and more distinct picture of the model's thought process, this in turn could be valuable when identifying areas of uncertain or ambiguous nature within the requirements themselves. Using the guidelines found in the analysis theme could prove useful in order to break down requirements that are more complex into components that are more manageable in order to spot conflicts or inconsistencies. Applying templates is another technique that is established to be a viable strategy when prompting. In the case of requirement analysis, they provide a structure that is predefined which could reduce the risk of leaving out vital information, they are also reusable which could save time and effort when analyzing a larger number of requirements.

3) Requirement Specification: Considering the importance of reducing or removing ambiguity, incompleteness, and inaccuracy when documenting requirements, we theorize that the guidelines contained in the theme of "Disambiguation" would be a good fit. These guidelines aim to identify weaknesses such as these previously mentioned which is the main reason for this mapping.

4) Requirement Validation: When conducting requirement validation, using guidelines included in the theme "Persona" could potentially provide value. One example of this would be if the large language model was given the requirements, and then the ones validating would be asked to interact with the

TABLE VI
CLASSIFICATION OF PE GUIDELINE THEMES

Theme	Description
Context (C)	1. Adding context to examples in prompts produce more efficient and informative output. [51] 2. Provide context to all prompts to avoid output hallucinations. [52] 3. Provide context of the prompt to ensure a closely related output. [53] 4. Use open-ended prompts to generate context before providing the intended question(s). [38] 5. Provide context to the topic of the prompt before describing a task. [54] 6. Adding context tokens to enhance the prompt, improves the related output. [55] 7. The more context tokens pre-appended to prompts, the more fine-grained output.[55]
Persona (P)	1. Improves the generation quality by conditioning the prompt with an identity, such as "Python programmer" or "Math tutor" [56] 2. To explore the requirements of a software-reliant system, include: - "I want you to act as the system", - "Use the requirements to guide your behavior", - "I will ask you to do X, and you will tell me if X is possible given the requirements.", - "If X is possible, explain why using the requirements.", - "If I can't do X based on the requirements, write the missing requirements needed in the format Y." [24]
Templates (T)	1. To improve reasoning and common sense in output, follow a template such as: - "Reason step-by-step for the following problem. [Original prompt inserted here]" [57] 2. The following prompt template has shown an impressive quality of AI art: - "[Medium] [Subject] [Artist(s)] [Details] [Image repository support]" [58]
Disambiguation (D)	1. Ensure any areas of potential miscommunication or ambiguity are caught, by providing a detailed scope: - "Within this scope", - "Consider these requirements or specifications" [24] 2. To find points of weakness in a requirements specification, consider including: - "Point out any areas of ambiguity or potentially unintended outcomes" [24] 3. The persona prompt method can be used to consider potential ambiguities from different perspectives. [22]
Reasoning (R)	1. Prepending "Let's think step by step" improves zero-shot performance. [59] 2. Extending the previously known "Let's think step by step", with "to reach the right conclusion," to highlight decision-making in the prompt. [60] 3. Factual inconsistency evaluation can be significantly boosted using chain-of-thought prompting. [61] 4. Chain Of Thought (CoT) prompting improves LLM performance compared to Zero-shot and without CoT. [62]
Analysis (A)	1. Prepend a prompt in a Zero-shot setting: "Please analyze if the hypothesis is true or false" and use the following template for an analytical output: prompt + approach + premise + hypothesis + "True or False?" [63] 2. ChatGPT models are not "mature enough" for emotional evaluations. [64] 3. Emotion-enhanced CoT prompting is an effective method to leverage emotional cues to enhance the ability of ChatGPT on mental health analysis. [64]
Keywords (K)	1. When picking the prompt, focus on the subject and style keywords instead of connecting words. [34] 2. Pre-appending keywords to prompts are shown to greatly improve performance by providing the language model with appropriate context. [65] 3. Modifiers/Keywords can be added to the details or image repository sections of a template such as: - "[Medium] [Subject] [Artist(s)] [Details] [Image repository support]" [58] 4. The inclusion of multiple descriptive keywords tends to align results closer to expectations. [35]
Wording (W)	1. In translation tasks, adding a newline before the phrase in a new language increases the odds that the output sentence is still English. [51] 2. A complete sentence definition with stop words performs better as a prompt than a set of core terms that were extracted from the complete sentence definition after removing the stop words. [66] 3. Words such as "well-known" and "often used to explain" are successful for analogy generation. [67] 4. Modifying prompts to resemble pseudocode tend to be the most successful in coding tasks. [24], [68] 5. Prompts to contain explicit algorithmic hints in engineering tasks perform better. [68]
Shorten (S)	1. For summarization or text-shortening tasks, the prompt should be written results- and information-oriented, leaving out unnecessary elements. [69]
Few-shot Prompts (F)	1. Inclusion of "Question:" and "Answer:" improves the response, but rarely gives a binary answer. [70] 2. For easier understanding, number examples in few-shot prompting. [71] 3. The format of [INPUT] and [OUTPUT] should linguistically imply the relationship between them. [71] 4. Specifications can be added to each [INPUT] and [OUTPUT] pair to give extra insight into complicated problems. [71] 5. In Few-shot prompting include a rationale in each shot (Input-rationale-output). [72]

model as if they were end-users. By having the language model act as a persona or a system, the evaluators can assess whether the model's responses align with the desired user experience and meet the specified requirements.

5) Requirement Management: As mentioned in the related work section, this phase of the requirements life cycle inhabits several activities. The activity used in this mapping is requirement tracing. The guideline theme of keywords could prove useful when utilizing keywords for searches in design documents and code in order to find relevant artifacts. Using guidelines within this theme may also help in establishing traceability links between requirements and other artifacts.

C. Interviews (RQ3)

Interviews were conducted with three different RE experts, enquiring about their views on the advantages and the limitations of guidelines from our review and their possible usage for LLMs in the mapped RE activities (Table VII).

The results from the interviews are presented in three parts. Two tables include quotes from the interviewed experts. Table

TABLE VII
MAPPING OF GUIDELINES AND RE

RE Activity	Guidelines
Elicitation	C1-C7
Validation	P1-P2
Analysis	R1-R4, T1-T2, A1-A3
Specification	D1-D3
Management	K1-K4

VIII presents advantages mentioned in the interviews and Table IX presents limitations brought up in the interviews.

Following is an overview of the experts' views on each of the mappings (Table VII) and further context to some quotes from Table VIII and Table IX.

1) Elicitation and Context Guidelines: Each of the interviewed experts mentioned that context is useful for prompts in practice, but also for elicitation.

For advantages, Expert 3 mentioned that the theme could be useful for brainstorming when eliciting requirements, and could be useful for stakeholders. Expert 2 mentioned similar thoughts after discussing how to utilize these for LLMs when eliciting requirements, stating "It may be creative requirements, like to say, hey, have you thought of this? And then a person says, oh, that's a good idea, or no, that's a bad idea." as an advantage.

The three experts all stated limitations regarding the term "context" included in the guidelines, arguing how it could be ambiguous as it is a quite general word. Expert 3 stated "I would say that context may need to be decomposed in somehow because otherwise your guideline may be too generic.", and Expert 1 mentioned that the context that the guidelines refer to, applied to elicitation purposes, would first need to be elicited from somewhere too, likely a stakeholder. They further mentioned that context may vary depending on different made assumptions.

Expert 2 brought up another limitation, on how the context may not fully cover what the stakeholders want, stating "So you can use it to elicit requirements, but they're not necessarily requirements that anyone wants to implement.".

2) Validation and Persona guidelines: The experts presented more limitations than advantages regarding the persona guidelines. Expert 2 presented their view on the guidelines and usage of LLMs for requirements validation stating "So I'll just say that I think it's a bad idea to use the prompt idea for validation." and explained that the LLM could not know if a system is correct or not and wouldn't be appropriate to use for requirements validation.

Experts 1 and 3 however provided similar possible advantages, besides additional limitations. Expert 3 mentioned that the idea of LLMs in validation is possibly a relevant area for further research and stated "So like a user that you want, for example, to validate the requirements against the need of a certain user. So not just 'I want to act as the system' but 'I want to act as a certain type of user' could be surely helpful." while they many times brought up the issue of context as mentioned about elicitation, and the need of providing extensive descriptions about the system. Expert 1 also claims that a possible advantage is being able to carefully explore the validity of certain specific goals or targets from various points of view.

Expert 1 further brought up limitations about the context among personas. In this case, the signification of context is what the personas know about the system, and the mentioned limitation is defining what context the personas should know. Expert 1 further stated "Well, if you limit the LLM only to look from a certain dimension or certain perspective on the problem, then you scope down your validation a lot to be only valid for that certain perspective that you define beforehand." as an additional limitation. Expert 3 also explained that validation is the last step in the requirements process, and express their uncertainty about the LLMs of today being sufficiently accurate in handling such a substantial amount of technical information as systems typically have at that stage of development.

3) Requirements Analysis and Analysis guidelines: The experts presented primarily limitations regarding the guidelines mapped to requirements analysis. Expert 1 explained that one major limitation of the analysis guidelines is the lack of confidence in feedback and uncertainty around it, and Expert 3 stated "I'm not sure that this type of prompts really capture what is needed for requirements analysis. So I'm not sure this really fits for requirements analysis context." while they also suggested that the theme would fit better for requirements elicitation. Expert 2 expressed uncertainty about whether language models are capable of performing requirements analysis and stated "it's a language model, it's not a formal method." but at the same time brought up a possible advantage "if you're asking is this maintainable, is this verifiable, is this unambiguous? Then you could come back and say yes or no and that might be useful." and followed up with once again stating their doubt.

Expert 1 pointed out the possibility of using the guidelines for analysis saying "An advantage maybe is that it's rather easy to do requirement analysis this way, but then it might be too easy" but further mentioned limitations and explained how the uncertainty around output puts it in a state where the output would not useful and the lack of confidence in the feedback from the LLM.

4) Requirements Analysis and Templates guidelines: For the template guidelines, the experts presented various advantages and limitations with certain further ideas to them as well. Expert 3 started by explaining that templates could be useful in any type of prompt activities and any RE task that profit from LLMs, and mentioned requirements elicitation as one of the RE activities. Suggesting an advantage of LLMs being able to implement common requirements templates such as ROPS or EARS templates. However, they additionally stated a limitation saying "This specific case is not so convincing of prompts that can be applied to requirements cases and in particular to requirements analysis." which also Expert 2 did, they stated "It's really hard to imagine it part of requirements analysis at all" while also mentioning that templates could be

TABLE VIII
ADVANTAGES MENTIONED IN THE INTERVIEWS

Mapping	Expert 1	Expert 2	Expert 3
Req. Elicitation Context guidelines	“adding all these assumptions when elicitating requirements makes it much easier to get the output in one.”	“obviously the more you give, the more information you put in a prompt, the better output you’re going to get.” ”It may be creative requirements, like to say, hey, have you thought of this? And then a person says, oh, that’s a good idea, or no, that’s a bad idea.”	“adding context is beneficial for sure.” “large language model is something that is in principle reasonable because it can help you brainstorming. For example, you can do a sort of focus group with two subjects. One is the subject yourself and the second subject is the large language model.”
Req. Validation Persona guidelines	“An advantage, of course, is that you can explore the validity of requirements very targeted to different targets or goals based on the perspective that you want”	–	“So like a user that you want, for example, to validate the requirements against the need of a certain user. So not just ’I want to act as the system’ but ’I want to act as a certain type of user’ could be surely helpful.”
Req. Analysis Analysis guidelines	“An advantage maybe is that it’s rather easy to do requirement analysis this way, but then it might be too easy”	“if you’re asking is this maintainable, is this verifiable, is this unambiguous? Then you could come back and say yes or no and that might be useful.”	“it can help you find inconsistencies among the requirements, because contradictions can be found by large language models, so they have somehow understood also some logical aspects”
Req. Analysis Template guidelines	“templates are always nice because you can validate that templates actually work nicely beforehand and then people can just use it without having to analyze if the template is good.”	“I guess completeness you can analyze because the form is incomplete”	“I believe Chat GPT could surely transform a requirement in these desired templates. So templates may be better connected to this, this towards specification in general”
Req. Analysis Reasoning guidelines	“you get a much better traceability or track why a certain decision has been made by the LLM or a certain output came from the LLM.”	–	“it could be useful for the analysis of requirements towards the generation of system architecture for example, this could be reasonable.” “’let’s think step by step’ for sure it’s good for making the reasoning linear, this is also embedded like this is, from a practical standpoint, in Auto GPT.”
Req. Specification Disambiguation guidelines	“you can explore maybe even iteratively any unclear aspects of your requirements and as I say, any ambiguities where there could be potential misunderstanding.” “if you combine it with the previous theme of persona, that you have different perspective from which you look at a requirement and identify if from any of these perspectives that are relevant, there are no ambiguities left.”	“it could help to point out weaknesses in a requirement specification” “prompt engineering could help point out ambiguity”	“This surely presents a sure advantage in requirement specification and also in requirements review. Requirements review is performed in, typically in safety-critical context, when you want to be sure that no ambiguities exist in the requirements.” “So this is a useful prompt for sure, it’s going to change the way we make requirements reviews.”
Req. Management Keyword guidelines	“by adding certain keywords you kind of define the context and maybe even some assumptions about the context. And this of course limits the scope basically out of which the output can come and this makes it more precise”	“You sort your requirements in categories. So if you can get prompts to help you with sorting it, but I guess that’s part probably more of requirement specification. But I would be really surprised if that actually works.”	“keywords when they represent classes can actually be helpful for the requirements classification that helps the requirement management.” “It could be also helpful, but I’m a bit less for requirements management related to similarity analysis or tracing.”

a way of analyzing completeness.

Expert 1 explained the inconsistencies of LLMs as well as the fact that output is not always the same even when the exact same prompt is used. Then explained their view on templates in requirements analysis and stated “So in LLMs, I would be careful to claim too much into using templates.”. Further, expert 1 stated the advantage of the usage of templates saying “templates are always nice because you can validate that templates actually work nicely beforehand, and then people can just use it without having to analyze if the template is good.”, and then described the risk of inconsistencies produced by LLMs even when using templates again.

5) Requirements Analysis and Reasoning guidelines: The answers regarding this mapping stated by the experts consisted primarily of limitations. Expert 1 expressed the need of pointing out the fact of not knowing what a step in reasoning mean for an LLM. Expert 2 expressed a similar limitation by stating “It’s only capturing one very narrow type of reasoning if you’re trying to get it to reason in a process view, like step one, step two, step three, I’m not even sure that’s reasoning.” and pointing out that it’s unsure if PE can help with the type of reasoning used in requirements engineering at all. Expert

TABLE IX
LIMITATIONS MENTIONED IN THE INTERVIEWS

Mapping	Expert 1	Expert 2	Expert 3
Req. Elicitation Context guidelines	"If you elicitate requirements, you first also need to elicitate the context in which these requirements are valid." "And from a LLM is the output that comes out of it that then rather fits to the assumptions that you had in your mind when writing the prompt."	"I'm not even sure you can use it to elicit requirements because requirements should come from stakeholders." "you can use it to elicit requirements, but they're not necessarily requirements that anyone wants to implement." "I'm not convinced that PE can be used to elicit or should be used to elicit requirements."	"adding context is beneficial for sure." "large language model is something that is in principle reasonable because it can help you brainstorming. For example, you can do a sort of focus group with two subjects. One is the subject yourself and the second subject is the large language model."
Req. Validation Persona guidelines	"if you limit the LLM only to look from a certain dimension or certain perspective on the problem, then youscope down your validation a lot to be only valid for that certain perspective that you define beforehand."	"So I'll just say that I think it's a bad idea to use the prompt idea for validation." "It's never going to pay you to make a system, it is not your stakeholder. it is not appropriate to validate whether a requirement is needed or not, or good or bad."	"I would say that I'm not sure that large language models as we know them now are sufficiently accurate to handle this type of technical information." "I would say that it could work. I don't know if really I don't know how feasible in practice it is but probably, it's a thing to further research for sure." "In principle you can say that it could be useful, but you need to specify the entire system very carefully."
Req. Analysis Analysis guidelines	"A huge limitation that I see from the beginning is that you do not get any confidence of feedback"	"So I have no idea if GPT or other prompt analysis can do that type of analysis because it's a language model, it's not a formal method."	"I'm not sure that this type of prompts really capture what is needed for requirements analysis. So I'm not sure this really fits for requirements analysis context."
Req. Analysis Template guidelines	"It's not necessarily reproducible, but the use of templates kind of assumes that something is reproducible. So in LLMs, I would be careful to claim too much into using templates."	"Using templates in requirements engineering is good, but that is not for requirements analysis, that is requirements elicitation or specification." "It's really hard to imagine it part of requirements analysis at all"	"templates is something that is useful for any type of prompt activities, prompt days activities, and so for any requirements engineering task that can profit from LLM usage. This specific case is not so convincing of prompts that can be applied to requirements cases and in particular to requirements analysis."
Req. Analysis Reasoning guidelines	"There's nothing that really tells the LLM what kind of details you are expecting for each reasoning step."	"I'm not sure that prompt engineering can do the type of reasoning that we've done in the past." "It's only capturing one very narrow type of reasoning if you're trying to get it to reason in a process view, like step one, step two, step three, I'm not even sure that's reasoning."	"I wouldn't say that a guideline, simple as that, is sufficient." "The requirements analysis may be system specific, domain specific, and you need to refine the prompt that you're presenting here. Of course, these are prompt templates, but maybe even the template needs to be better refined."
Req. Specification Disambiguation guidelines	"I think what would be important is that you also include in the prompt that the LLM should give you a reason why this requirement is ambiguous or potentially not targeting what you really want to target."	"Either the AI gives you a wrong answer or it points out ambiguity or other requirements formatting or other requirements quality problems that don't actually matter in practice"	–
Req. Management Keyword guidelines	"Limits the scope basically out of which the output can come"	"I'm skeptical that these are useful guidelines and I don't think it's part of requirements management"	"I'm not sure that keywords alone are helpful for this very complex task that requires domain knowledge and project knowledge. This is probably quite limited to the case in which you want to do some requirement specification." "Overall the feeling is that the guidelines need to be refined for the specific cases of requirements engineering"

3 described how the guidelines in this format would not be sufficient for requirements analysis and that guidelines would need refining to a specific domain or a system.

Expert 3 also mentioned possible advantages of the guidelines too, stating "it could be useful for the analysis of requirements towards the generation of system architecture for example, this could be reasonable" in addition to their explanation of how these guidelines could be used while embedded in the LLMs to solve sub-tasks aiding analysis, referring to the implementation details of AutoGPT. Expert 1 explains that an advantage would be to have a chain of thought to follow which would be an advantage in requirements analysis and stated "you get a much better traceability or track why a certain decision has been made by the LLM or a certain output came from the LLM.".

6) Requirements Specification and Disambiguation guidelines: This mapping of guidelines was the only mapping where all experts agreed on it being useful and their answers to a larger extent revolved around advantages rather than limitations.

Expert 1 stated "you can explore maybe even iteratively any unclear aspects of your requirements and as I say, any ambiguities where there could be potential misunderstanding." and specified further that this mapping could be expanded by including the previously mentioned persona guidelines to explore ambiguity from different perspectives. Expert 2 stated the advantages that "it could help to point out weaknesses in a requirement specification" and "prompt engineering could help point out ambiguity" but also pointed out the aspect of ambiguity not mattering referring to previous studies made on smaller teams.

Expert 3 only brought up advantages. They explained that not only will these guidelines have advantages in requirement specification, but in requirements review as well, and put guideline D2 in an example context stating "So this is a useful prompt for sure, it's going to change the way we make requirements reviews." and elaborated on requirements reviews being an important aspect of safety-critical contexts.

Experts 1 and 2 presented one limitation each. Both are related to the confidence of the response. Expert 1 explained a need for justifications or reasons as to why some requirement is deemed ambiguous or not and why it has a weakness. And Expert 2 stated "Either the AI gives you a wrong answer or it points out ambiguity or other requirements formatting or other requirements quality problems that don't actually matter in practice" after they pointed out that ambiguous requirements might not become an issue in certain settings.

7) Requirements Management and Keywords guidelines: Experts 2 and 3 both brought up the advantages of keywords guidelines and their potential to be helpful in sorting and classifying requirements into categories. However, Expert 2 also stated "But I would be really surprised if that actually works." leaving them in an unsure state regarding the mapping. Expert 3 also mentioned the possibility of these guidelines being helpful for two other requirements management tasks, similarity analysis, and tracing. Expert 1 focused mostly on context and explained the keywords could help with providing assumptions and make the output more precise. The output being more precise was not only a benefit, they also explained its limitation that it "Limits the scope basically out of which the output can come" and risks leaving important aspects out of the picture. Expert 2 stated that the mapped guidelines would not be useful and was unsure about whether it could be a part of requirements management specifically.

Expert 3 expressed their thoughts on two limitations. They were unsure if the use of keywords alone could be helpful as the management tasks require in-depth project and domain knowledge. Suggesting these guidelines could be a better fit for requirements elicitation. Lastly, they stated "Overall the feeling is that the guidelines need to be refined for the specific cases of requirements engineering" as a remark regarding all the mapped guidelines to their respective activity.

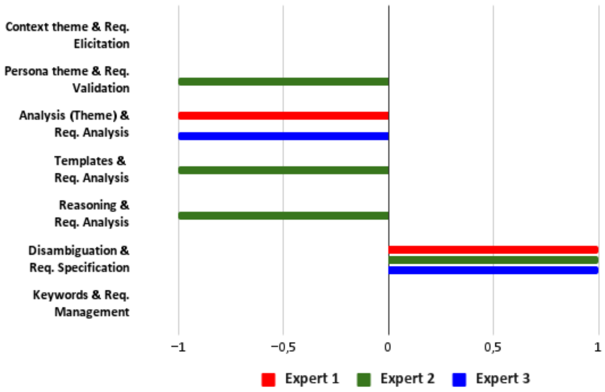


Fig. 6. A visual overview of advantages versus limitations based on conducted interviews. The questions in the interviews are based on the mapping between guideline themes and RE activities. A scale of –1, 0, or 1 is used for responses from the interviewed experts. –1 = disagree with the mapping, 0 = unsure about mapping, 1 = agree with the mapping. (Req. short for Requirements)

D. Interview Analysis

After doing a comparison of the data gathered from the different interviews, a few common denominators were identified. Overall, the viewpoint regarding the mappings was quite unanimous, which can be seen in Fig. 6. Looking at the guideline themes proposed to be applied when conducting Requirement analysis, the experts expressed uncertainty regarding how applicable these guidelines would be. In general, the opinion was that the task of doing requirement analysis was most likely too complex of a task in regard to how developed current generative AIs are.

In regard to Requirement elicitation, our mapping proposed the application of context guidelines. Even if all participants agreed on the importance of context itself along with proposed guidelines, they were hesitant regarding in what context they would be applied, mainly considering the importance of actual stakeholders when eliciting requirements. However, 2 out of the 3 experts suggested using the guidelines when brainstorming using LLMs in order to possibly come up with new or missing requirements which might not have been considered previously.

For requirements validation in conjunction with the "Persona"-theme, the participants expressed uncertainty towards the idea, primarily because of the model's need for a strong understanding of the system and its context.

How suitable the mapping between the themes "Keywords" and Requirement Management was, was harder to determine due to the varying definitions of what the term management entails. However, all participants viewed the mapping and guidelines as ambiguous and in need of further elaboration and refinement.

The most positive response was towards the disambiguation theme and its application in combination with requirement specification. All participants provided positive feedback and various advantages of this mapping and 2 of the 3 experts even suggested additional possible activity tasks where these guidelines could be useful.

V. Discussion

One main objective of this study has been to explore PE guidelines in natural language and their application within the realm of large language models. PE is a field that is expanding rapidly. This becomes very apparent when looking into the search results just a few months after this study was first initiated. As of May 2023, the amount of papers written on the subject is close to equal to the amount of all papers within the same field written during 2022. Apart from these papers, various communities and websites such as [75], are gaining traction. The guidelines and their respective themes we present in this study originate from a large variety of models. The most common ones are text-to-text models, such as GPT-3.5 and BERT. Another substantial amount of the guidelines and themes were extracted from papers looking at text-to-image and multi-modal models, such as DALL-E and the combination of CLIP and VQGAN. It is apparent that even though these guidelines do share some overlap, based on our findings there still are noteworthy differences which may indicate that the importance of PE will only grow from here. A majority of the guidelines found in our systematic literature review were only mentioned once across all studies which further shows how broad the field is as well as how many gaps still exist, something that clearly points to the need for further research. The set of guidelines that were mentioned in the literature the most, were those regarding the importance of context. This was also pointed out by the interviewed experts which makes it apparent that generative AI models in most cases perform better in a setting where an increased amount of context is provided through prompts or pre-training and fine-tuning.

The other objective of our study was to look into the suitability of these guidelines in various RE activities. We did so by mapping our established guideline themes to said activities, based on the steps that are usually performed along with the characteristics that are of importance in each activity. Considering that the literature review looked into a wide range of guidelines, some, such as those targeting only the few-shot prompting approach, was not as applicable, which is why they were excluded from the mapping. Clear indications of multiple alternatives to these mappings were presented in the results of the interviews. However, the expressed opinions towards the guideline themes themselves were relatively unanimous. The results of the interview suggest that the least impact of using LLMs would be in the requirements analysis phase. The reason for this could lie within the capabilities needed for generative AI in order to perform analysis. The interviewees pointed out that analysis itself is too broad of a term and would have to be broken down further in order to better elaborate on the possibilities for the guidelines within the same theme. The participants expressed the most optimism towards the guideline theme of Disambiguation and its applicability within requirement specification. This could also indicate that guideline themes used with RE activities that are of a less complex nature are the ones that are most probable to see use as of now.

VI. Future Work

In order to further explore PE guidelines and establish their applicability within RE activities and other domains, we suggest alternative studies and fields of research. As part of this study, our secondary goal was to present possible future work and after conducting the study we have found a multitude of possible topics for further research.

We found that in the current state of the literature, overall the PE guidelines are still limited and that further research in domain-specific tasks and related prompting guidelines is needed in this popular field.

Additionally, from the extracted prompt guidelines in our review, guidelines intended for domain-specific tasks and general guidelines can be seen. A distinction between the guidelines that are in nature more general and domain-specific guidelines can be made. But going one step further and taking advantage of guidelines that are more general in nature, such as C1-C7, P1, or R1 when developing guidelines or approaches for a specific task domain, the guidelines being developed may suffer less from fallacies as the foundation already has been laid in the initial prompt guidelines. Can general, non-domain-specific, guidelines from the literature somehow be integrated with new prompt guidelines for LLM tasks in specific domains such as RE? Would it require further model fine-tuning? These are questions that can be explored in future work.

As expressed by one of the experts in the interviews, there is uncertainty about what technical information LLMs can accurately handle and to what extent. Would an LLM be capable of handling a substantial amount of technical documentation and requirements in large complex systems at the stage of requirements validation? That is another question that is left for future work to answer.

For an alternative way of conducting this study, we suggest administering a survey to a broad range of experts and practitioners within RE to look into the perspective on identified guidelines and their application. Questions included in the survey could emphasize the participants' different experiences and challenges as well as their preferences regarding the guidelines. A rigorous quantitative and qualitative analysis could then be conducted to identify patterns and themes commonly found in the collected data.

VII. Conclusion

This study has provided an overview of the current state of the literature regarding PE guidelines for the field of RE. We presented the advantages of found guidelines in RE, as well as the limitations. We also identified gaps in the literature and where further research efforts could be beneficial.

The topic explored through our systematic literature review showed to be rapidly growing in the literature, including almost the same number of papers from the first quarter of 2023 as from the whole year of 2022. Through our review of the literature, we presented PE guidelines serving as our answer to RQ1. Another observation was that guidelines for domain-specific tasks, including RE, are still limited.

Based on our review and related work about RE activities, we explored a way of mapping these guidelines to RE activities and proposed the mapping of what guidelines could be useful in RE, answering our RQ2. These mappings were used in interviews with three RE experts to answer RQ3 regarding the advantages and limitations of the guidelines when applied to LLMs in RE activities. The results showed that the experts saw benefits with guidelines for disambiguation in conjunction with requirements specification.

However, we also found indications pointing to activities where LLMs may not be capable enough as of now, in particular requirement analysis. This is another aspect emphasizing a need for further research efforts in the field.

2 experts emphasized LLMs' ability to perform certain RE tasks and the multitude of opportunities which can be explored within the area of PE for LLMs in RE. Based on one of the interviewed expert's concerns and suggestions, we further suggested future work to explore the capabilities of LLMs to accurately handle substantial amounts of technical documentation present in large complex systems, in the context of requirements validation.

We believe that by reaching a point where AI models, such as LLMs, are properly equipped and are sufficiently accurate in performing certain RE activities, the number of RE-related errors could be lowered, and the errors could be identified at an earlier stage. Sequentially, this would help avoid expensive corrections of errors at later stages and lead fewer projects to failure.

References

[1] D. Damian and J. Chisan, "An empirical study of the complex relationships between requirements engineering processes and other processes that lead to payoffs in productivity, quality, and risk management," *IEEE Transactions on Software Engineering*, vol. 32, no. 7, pp. 433–453, 2006. DOI: 10.1109/TSE.2006.61.

[2] J. Drew Procaccino, J. M. Verner, S. P. Overmyer, and M. E. Darter, "Case study: Factors for early prediction of software development success," *Information and Software Technology*, vol. 44, no. 1, pp. 53–62, 2002, ISSN: 0950-5849. DOI: https://doi.org/10.1016/S0950-5849(01)00217-8. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0950584901002178.

[3] J. M. Verner, S. P. Overmyer, and K. W. McCain, "In the 25 years since the mythical man-month what have we learned about project management?" *Information and Software Technology*, vol. 41, no. 14, pp. 1021–1026, 1999.

[4] F. P. Brooks and N. S. Bullet, "Essence and accidents of software engineering," *IEEE computer*, vol. 20, no. 4, pp. 10–19, 1987.

[5] K. Kaur, P. Singh, and P. Kaur, "A review of artificial intelligence techniques for requirement engineering," *Computational Methods and Data Engineering: Proceedings of ICMDE 2020, Volume 2*, pp. 259–278, 2020.

[6] J. Winkler and A. Vogelsang, "Automatic classification of requirements based on convolutional neural networks," in *2016 IEEE 24th International Requirements Engineering Conference Workshops (REW)*, IEEE, 2016, pp. 39–45.

[7] B. W. Boehm, "Software engineering economics," *IEEE transactions on Software Engineering*, no. 1, pp. 4–21, 1984.

[8] G. Mogyorodi, "Requirements-based testing: An overview," Feb. 2001, pp. 286–295, ISBN: 0-7695-1251-8. DOI: 10.1109/TOOLS.2001.941681.

[9] A. Ferrari, A. Esuli, and S. Gnesi, "Identification of cross-domain ambiguity with language models," in *2018 5th International Workshop on Artificial Intelligence for Requirements Engineering (AIRE)*, IEEE, 2018, pp. 31–38.

[10] F. Dalpiaz, A. Ferrari, X. Franch, and C. Palomares, "Natural language processing for requirements engineering: The best is yet to come," *IEEE software*, vol. 35, no. 5, pp. 115–119, 2018.

[11] C. Cheligeer, J. Huang, G. Wu, N. Bhuiyan, Y. Xu, and Y. Zeng, "Machine learning in requirements elicitation: A literature review," *AI EDAM*, vol. 36, e32, 2022.

[12] F. Dalpiaz and N. Niu, "Requirements engineering in the days of artificial intelligence," *IEEE software*, vol. 37, no. 4, pp. 7–10, 2020.

[13] A. Ferrari and A. Esuli, "An nlp approach for cross-domain ambiguity detection in requirements engineering," *Automated Software Engineering*, vol. 26, no. 3, pp. 559–598, 2019.

[14] C. Rolland and C. Proix, "A natural language approach for requirements engineering," in *Advanced Information Systems Engineering: 4th International Conference CAiSE'92 Manchester, UK, May 12–15, 1992 Proceedings 4*, Springer, 1992, pp. 257–277.

[15] T. Hey, J. Keim, A. Koziolek, and W. F. Tichy, "Norbert: Transfer learning for requirements classification," in *2020 IEEE 28th International Requirements Engineering Conference (RE)*, IEEE, 2020, pp. 169–179.

[16] A. Sainani, P. R. Anish, V. Joshi, and S. Ghaisas, "Extracting and classifying requirements from software engineering contracts," in *2020 IEEE 28th international requirements engineering conference (RE)*, IEEE, 2020, pp. 147–157.

[17] J. Wei, Y. Tay, R. Bommasani, *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[18] Y. Bang, S. Cahyawijaya, N. Lee, *et al.*, "A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity," *arXiv preprint arXiv:2302.04023*, 2023.

[19] M. Jaworski and D. Piotrkowski, "Study of software developers' experience using the github copilot tool in the software development process," *arXiv preprint arXiv:2301.04991*, 2023.

[20] S. Peng, E. Kalliamvakou, P. Cihon, and M. Demirer, "The impact of ai on developer productivity: Evidence from github copilot," *arXiv preprint arXiv:2302.06590*, 2023.

[21] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[22] J. White, Q. Fu, S. Hays, *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[23] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[24] J. White, S. Hays, Q. Fu, J. Spencer-Smith, and D. C. Schmidt, "Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design," *arXiv preprint arXiv:2303.07839*, 2023.

[25] X. Luo, Y. Xue, Z. Xing, and J. Sun, "Prcbert: Prompt learning for requirement classification using bert-based pretrained language models," in *37th IEEE/ACM International Conference on Automated Software Engineering*, 2022, pp. 1–13.

[26] W. Alhoshan, A. Ferrari, and L. Zhao, "Zero-shot learning for requirements classification: An exploratory study," *Information and Software Technology*, vol. 159, p. 107 202, 2023.

[27] S. Sharma and S. Pandey, "Revisiting requirements elicitation techniques," *International Journal of Computer Applications*, vol. 75, no. 12, 2013.

[28] S. A. Fricker, R. Grau, and A. Zwingli, "Requirements engineering: Best practice," in *Requirements Engineering for Digital Health*, S. A. Fricker, C. Thümmler, and A. Gavras, Eds. Cham: Springer International Publishing, 2015, pp. 25–46, ISBN: 978-3-319-09798-5. DOI: 10.1007/978-3-319-09798-5_2. [Online]. Available: https://doi.org/10.1007/978-3-319-09798-5_2.

[29] C. Denger, D. Berry, and E. Kamsties, "Higher quality requirements specifications through natural language patterns," in *Proceedings 2003 Symposium on Security and Privacy*, 2003, pp. 80–90. DOI: 10.1109/SWSTE.2003.1245428.

[30] H. A. Bilal, M. Ilyas, Q. Tariq, and M. Hummayun, "Requirements validation techniques: An empirical study," *International Journal of Computer Applications*, vol. 148, no. 14, 2016.

[31] M. N. A. Khan, M. Khalid, and S. ul Haq, "Review of requirements management issues in software development," *International Journal of Modern Education and Computer Science*, vol. 5, no. 1, p. 21, 2013.

[32] M. Jarke, "Requirements tracing," *Communications of the ACM*, vol. 41, no. 12, pp. 32–36, 1998.

[33] F. Dalpiaz and N. Niu, "Requirements engineering in the days of artificial intelligence," *IEEE Software*, vol. 37, pp. 7–10, Jul. 2020. DOI: 10.1109/MS.2020.2986047.

[34] V. Liu and L. B. Chilton, "Design guidelines for prompt engineering text-to-image generative models," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–23.

[35] Y. Hao, Z. Chi, L. Dong, and F. Wei, "Optimizing prompts for text-to-image generation," *arXiv preprint arXiv:2212.09611*, 2022.

[36] P. Maddigan and T. Susnjak, "Chat2vis: Generating data visualisations via natural language using chatgpt, codex and gpt-3 large language models," *arXiv preprint arXiv:2302.02094*, 2023.

[37] OpenAI, *Openai prompt design guidelines*, https://platform.openai.com/docs/guides/completion/prompt-design, Accessed: 2023-03-10.

[38] S. Arora, A. Narayan, M. F. Chen, *et al.*, "Ask me anything: A simple strategy for prompting language models," in *The Eleventh International Conference on Learning Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=bhUPJnS2g0X.

[39] B. Kitchenham and S. Charters, "Guidelines for performing systematic literature reviews in software engineering," vol. 2, Jan. 2007.

[40] J. Yang, H. Jin, R. Tang, *et al.*, "Harnessing the power of llms in practice: A survey on chatgpt and beyond," 2023. arXiv: 2304.13712 [cs.CL].

[41] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[42] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[43] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training," 2018.

[44] D. S. Cruzes and T. Dybå, "Research synthesis in software engineering: A tertiary study," *Information and Software Technology*, vol. 53, no. 5, pp. 440–455, 2011, Special Section on Best Papers from XP2010, ISSN: 0950-5849. DOI: https://doi.org/10.1016/j.infsof.2011.01.004. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S095058491100005X.

[45] C. Willig and W. S. Rogers, *The SAGE handbook of qualitative research in psychology*. Sage, 2017.

[46] G. Guest, K. M. MacQueen, and E. E. Namey, *Applied thematic analysis*. sage publications, 2011.

[47] X. Zhou, Y. Jin, H. Zhang, S. Li, and X. Huang, "A map of threats to validity of systematic literature reviews in software engineering," in *2016 23rd Asia-Pacific Software Engineering Conference (APSEC)*, IEEE, 2016, pp. 153–160.

[48] J. Lee, W. Yoon, S. Kim, *et al.*, "Biobert: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[49] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[50] V. Sanh, A. Webson, C. Raffel, *et al.*, "Multitask prompted training enables zero-shot task generalization," *arXiv preprint arXiv:2110.08207*, 2021.

[51] L. Reynolds and K. McDonell, "Prompt programming for large language models: Beyond the few-shot paradigm," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–7.

[52] K. Kumar, "Geotechnical parrot tales (gpt): Harnessing large language models in geotechnical engineering," *arXiv e-prints*, arXiv–2304, 2023.

[53] T. Lubiana, R. Lopes, P. Medeiros, *et al.*, "Ten quick tips for harnessing the power of chatgpt/gpt-4 in computational biology," *arXiv preprint arXiv:2303.16429*, 2023.

[54] T. Sorensen, J. Robinson, C. M. Rytting, *et al.*, "An information-theoretic approach to prompt engineering without ground truth labels," *arXiv preprint arXiv:2203.11364*, 2022.

[55] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[56] J. Wei, S. Kim, H. Jung, and Y.-H. Kim, "Leveraging large language models to power chatbots for collecting user self-reported data," *arXiv preprint arXiv:2301.05843*, 2023.

[57] P. Koralus and V. Wang-Maścianica, "Humans in humans out: On gpt converging toward common sense in both success and failure," *arXiv preprint arXiv:2303.17276*, 2023.

[58] J. Oppenlaender, R. Linder, and J. Silvennoinen, "Prompting ai art: An investigation into the creative skill of prompt engineering," *arXiv preprint arXiv:2303.13534*, 2023.

[59] Y. Zhou, A. I. Muresanu, Z. Han, *et al.*, "Large language models are human-level prompt engineers," *arXiv preprint arXiv:2211.01910*, 2022.

[60] B. Clavié, A. Ciceu, F. Naylor, G. Soulié, and T. Brightwell, "Large language models in the workplace: A case study on prompt engineering for job type classification," *arXiv preprint arXiv:2303.07142*, 2023.

[61] Z. Luo, Q. Xie, and S. Ananiadou, "Chatgpt as a factual inconsistency evaluator for abstractive text summarization," *arXiv preprint arXiv:2303.15621*, 2023.

[62] S. Chen, Y. Li, S. Lu, *et al.*, "Evaluation of chatgpt family of models for biomedical reasoning and classification," *arXiv preprint arXiv:2304.02496*, 2023.

[63] F. Yu, L. Quartey, and F. Schilder, "Legal prompting: Teaching a language model to think like a lawyer," *arXiv preprint arXiv:2212.01326*, 2022.

[64] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," *arXiv preprint arXiv:2304.03347*, 2023.

[65] K. I. Gero, V. Liu, and L. Chilton, "Sparks: Inspiration for science writing using language models," in *Designing Interactive Systems Conference*, 2022, pp. 1002–1019.

[66] G. Yong, K. Jeon, D. Gil, and G. Lee, "Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model," *Computer-Aided Civil and Infrastructure Engineering*, 2022.

[67] B. Bhavya, J. Xiong, and C. Zhai, "Analogy generation by prompting large language models: A case study of instructgpt," *arXiv preprint arXiv:2210.04186*, 2022.

[68] P. Denny, V. Kumar, and N. Giacaman, "Conversing with copilot: Exploring prompt engineering for solving cs1 problems using natural language," in *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*, 2023, pp. 1136–1142.

[69] A.-M. Meck and L. Precht, "How to design the perfect prompt: A linguistic approach to prompt design in automotive voice assistants–an exploratory study," in *13th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 2021, pp. 237–246.

[70] D. Trautmann, A. Petrova, and F. Schilder, "Legal prompt engineering for multilingual legal judgement prediction," *arXiv preprint arXiv:2212.02199*, 2022.

[71] P. West, C. Bhagavatula, J. Hessel, *et al.*, "Symbolic knowledge distillation: From general language models to commonsense models," *arXiv preprint arXiv:2110.07178*, 2021.

[72] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, and D. Zhou, "Rationale-augmented ensembles in language models," *arXiv preprint arXiv:2207.00747*, 2022.

[73] A. Liew and K. Mueller, "Using large language models to generate engaging captions for data visualizations," *arXiv preprint arXiv:2212.14047*, 2022.

[74] E. Perez, S. Huang, F. Song, *et al.*, "Red teaming language models with language models," *arXiv preprint arXiv:2202.03286*, 2022.

[75] *Prompt engineering guide*, https : / / www . promptingguide.ai, Accessed: 2023-05-15.