



What Should We Engineer in Prompts? Training Humans in Requirement-Driven LLM Use

QIANOU MA, Carnegie Mellon University, USA

WEIRUI PENG, Columbia University, USA

CHENYANG YANG, Carnegie Mellon University, USA

HUA SHEN, University of Washington, USA

KENNETH KOEDINGER, Carnegie Mellon University, USA

TONGSHUANG WU, Carnegie Mellon University, USA

Prompting LLMs for complex tasks (e.g., building a trip advisor chatbot) needs humans to clearly articulate customized requirements (e.g., “start the response with a tl;dr”). However, existing prompt engineering instructions often lack focused training on requirement articulation and instead tend to emphasize increasingly automatable strategies (e.g., tricks like adding role-plays and “think step-by-step”). To address the gap, we introduce Requirement-Oriented Prompt Engineering (ROPE), a paradigm that focuses human attention on generating clear, complete requirements during prompting. We implement ROPE through an assessment and training suite that provides deliberate practice with LLM-generated feedback. In a randomized controlled experiment with 30 novices, ROPE significantly outperforms conventional prompt engineering training (20% vs. 1% gains), a gap that automatic prompt optimization cannot close. Furthermore, we demonstrate a direct correlation between the quality of input requirements and LLM outputs. Our work paves the way to empower more end-users to build complex LLM applications.

CCS Concepts: • **Human-centered computing** → **Interactive systems and tools**; • **Applied computing** → **Computer-assisted instruction**; • **Computing methodologies** → Natural language generation.

Additional Key Words and Phrases: LLM, Human-AI Interaction, Prompt Engineering, Requirement Engineering, End-User Programming

1 INTRODUCTION

General-purpose AIs like large language models (LLMs) have evolved from simple next-word predictors [6] to powerful assistants capable of fulfilling complex user needs [50, 55]. This improvement in LLM’s ability to follow instructions has encouraged users to delegate increasingly intricate tasks to these models. In the early days of prompt engineering, users primarily focused on refining the wording of simple instructions to improve LLM output quality [24, 64]. Today, prompts resemble detailed “essays” that define LLM behaviors. Rather than one-off small requests, these prompts start to power the generation of complex applications, and *everyday users* can write complex prompts to either generate code for softwares [72, 76] or tailor general-purpose LLMs into special

Authors’ Contact Information: Qianou Ma, qianouma@cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Weirui Peng, wp2297@columbia.edu, Columbia University, New York, NY, USA; Chenyang Yang, cyang3@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Hua Shen, huashen@uw.edu, University of Washington, Seattle, WA, USA; Kenneth Koedinger, koedinger@cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA; Tongshuang Wu, sherryw@cs.cmu.edu, Carnegie Mellon University, Pittsburgh, PA, USA.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2025 Copyright held by the owner/author(s).

ACM 1557-7325/2025/4-ART

<https://doi.org/10.1145/3731756>

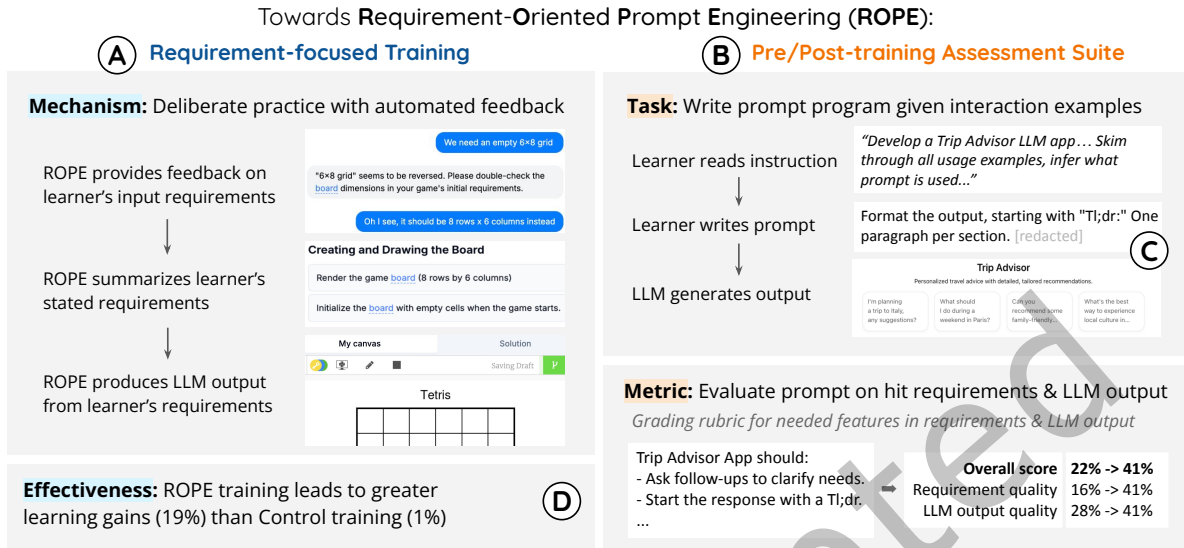


Fig. 1. We propose Requirement-Oriented Prompt Engineering (ROPE) and contribute a training and assessment suite for ROPE. (A) ROPE training helps learners write effective prompt programs by providing automated feedback on adding and clarifying requirements. (B) In our pre-post assessments, we ask learners to write prompts to develop customized applications (e.g., Trip Advisor via a prompt in C) and assess learners' prompt quality on both requirement quality and LLM output quality. (D) In a pre-post randomized experiment, ROPE training significantly improves learners' prompt quality, compared to conventional prompt engineering training.

purpose LLM applications [82]. For instance, an LLM application (or a GPTs) like Trip Advisor¹ can be built solely using prompts similar to Figure 1C. These LLM applications function similarly to traditional software and are accessible through LLM App Stores like GPT Store² and FlowGPT³, where they can be reused with a single click. In just a few months, over 30,000 "LLM app developers" have created millions of GPT apps, with the most popular ones being used more than 5 million times [82]. This trend signals a future of **end-user programming through natural language**, where anyone can create *reusable programs* just by writing prompts [20].

The future of writing prompt programs looks promising, but *status quo* prompt engineering remains a challenge. Various studies show that optimizing LLM outputs requires well-rounded prompts, involving fluent writing, clear instructions, personas, formats, and effective adaption of prompting techniques like Chain-of-Thought [7, 41, 43, 60]. These complexities often confuse novices, causing them to make *ad hoc* revisions to prompts without a clear understanding of what needs improvement [80]. Recognizing the challenge, the NLP community is developing *prompt optimizers* to automate some prompt refinement steps, e.g., incorporating role-plays, applying effective techniques, structuring prompts, and enhancing text fluency (e.g., Figure 2, discussions in Section 2.2).

The rise of optimizers makes us think: What aspects of prompt engineering should humans still master? We argue that the remaining part of manual prompt creation is akin to the core step before engineering in a design process: **requirement specification**. A *requirement* is a description of user goals; it defines and documents how a system should behave, including expected inputs and outputs [32] (e.g., ``Start the response with

¹<https://chatgpt.com/g/g-K38J9feSL-trip-advisor>

²<https://openai.com/index/introducing-the-gpt-store/>

³<https://flowgpt.com/>

a `tl;dr`' for Trip Advisor in Figure 1B). Requirements are the backbone of program functionalities, and user-specified requirements in prompts (e.g., Figure 2) are the driving forces behind complex LLM adaptations. **Customized requirements are hard to predict, and thus difficult to fully automate and must be specified by humans.** We argue that in a future where prompt programs are common, individuals should develop proficiency in requirement articulation. In other words, **we need to shift our focus towards Requirement-Oriented Prompt Engineering, an emerging paradigm that we call “ROPE”.**

Articulating clear and complete requirements is a known challenge in fields such as product design and software engineering [34, 47, 67], and poor requirements often cause inefficiency, frustrations, and software failures [44, 80]. This difficulty is amplified by the unpredictable nature of LLMs, which complicates controlling outputs [3, 12]. Even experts need multiple iterations to refine requirements [62], and novices frequently make requirement mistakes [4]. Recognizing both the importance and difficulty of this task, we pose the research question: **Can we help end-users better instruct LLMs to achieve their goals through ROPE training?**

To answer this question, we present the ROPE paradigm as a future human-LLM partnering strategy (Section 3), and we develop requirement-focused training and assessments to (1) help users practice writing complete and correct requirements, and (2) deepen our understanding of how requirements affect the quality of LLM outputs. Specifically, we formalize *tasks* (Section 4.1) with complex requirements (e.g., game development), and design corresponding *assessments* (Section 4.2) that can measure the requirement quality, the LLM output quality, and their correlations. We also design and implement an interactive training mechanism (Section 4.3) that *disentangles requirement articulation from other adjacent prompting aspects*. Our system supports deliberate practice on requirement refinement by generating immediate feedback grounded on reference requirements.

We conduct a randomized controlled user study (Section 5) with 30 prompting novices, comparing the prompts users write before and after training, and against a control group that received standard prompt engineering training. We confirm that *our requirement-focused training is effective* (Section 6): novices receiving this training significantly improved their requirement performances by two-fold in the pre-to-post test, achieved significantly more gains than conventional prompt engineering training, and learned to iterate on prompts in a more structured way towards requirements. Moreover, we confirm that *requirements are indeed important for the LLM generation*: There is a strong positive correlation between the requirement qualities and the LLM output qualities. Importantly, *the training gains are complementary to potential improvements from optimizers and models*: Using a popular prompt optimizer, we observe that optimizers can indeed improve requirement quality, but they cannot close the gap between the baseline and the training groups; Moreover, when switching from GPT-4o to the more advanced reasoning model o3-mini, the model's improved instruction-following capabilities made its outputs more reflective of users' explicit requirements, rather than rendering requirements obsolete. We conclude with an in-depth discussion

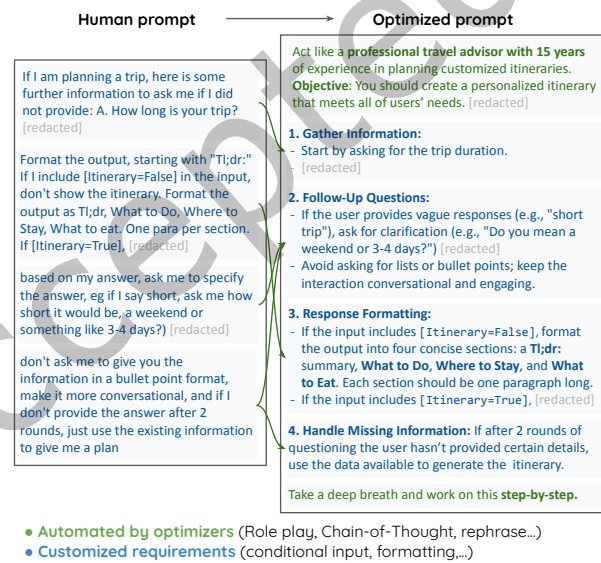


Fig. 2. Automatically refine a user's prompt for Trip Advisor using the optimizer Prompt Maker. Customized requirements still need to come from the human prompt, while the optimizer improves on non-requirement prompting aspects such as fluency, role plays, structures, etc.

(Section 7) on future directions for human training and optimizer design to enhance human-LLM complementarity in ROPE.

In summary, we make the following contributions:

- We identify the significance of requirement-based prompting, and propose the ROPE paradigm as a foundation for future human-LLM task delegation in prompt engineering.
- To the best of our knowledge, we are the first to investigate training specifically for requirement-based prompting. We developed tasks, assessments, and training systems that effectively improve requirement generation skills.
- We provide both quantitative and qualitative insights for the future of prompting, highlight the relationship between the quality of requirements and LLM performance across model advancements, and discuss current optimizers' limitations and potentials.

As natural language prompt programs become more widespread, everyone needs the skill to write good requirements. We believe that with our ROPE paradigm, LLMs can empower anyone to program reusable AI assistants. We open-source our training system, supplemental study materials, and user prompts at <https://github.com/mqo00/rope/>.

2 RELATED WORKS

We review three key areas relevant to our work: prompt engineering, prompt optimization in natural language processing, and requirements engineering in end-user software engineering. We highlight a significant gap in the literature: the lack of requirement-oriented prompt engineering training for end users, which we propose to address in this paper.

2.1 Current Prompt Engineering Practices and Challenges

Prompt engineering (PE) has been essential for crafting effective input instructions to guide LLMs toward generating desired outputs [5, 36, 80]. However, the non-deterministic nature of LLMs makes it challenging for humans to predict LLMs' behavior across prompts [15, 51, 65]. This leads to wasted time on unproductive strategies, such as trivial wording changes [15, 19, 80]. While some challenges can be mitigated through automation (see Section 2.2), e.g., automating word choices, human requirements (some also refer to as constitutions [52] or principles [38] in literature) remain crucial for customized or specialized tasks [36]. We define these tasks as *LLM-hard*, as a simple prompt cannot produce a satisfactory response. Developing reusable prompts for customized chatbots or GPTs [5, 80] is a common LLM-hard task. While a *prompt* is an input to LLMs, a *reusable prompt* is a set of instructions for recurring goals, essentially functioning as a *prompt program*. As more users create reusable prompts, the ability to write good prompt programs with clear requirements becomes increasingly important.

However, prompt creation can be complex and inaccessible to non-experts who lack technical knowledge of LLMs, highlighting the need for more end-user prompt engineering (EUPE) scaffolding and training [80]. Various strategies have been developed to support PE, such as prompt technique catalog [7, 37, 69], chat-based tools like GPT-builder,⁴ and toolkit for orchestrating complex tasks [9, 25]. Each tool has trade-offs, balancing flexibility, precision, and technicality. For example, chat-based tools may ask narrow clarifications without fully understanding the context, leaving users to self-refine requirements when LLM responses are open-ended [30, 54].

Existing PE training emphasizes the importance of explicitly stating requirements within prompts [43, 59, 60], yet there is still a lack of focused training on requirement articulation skills for everyday users. Without training, non-experts often make mistakes such as missing requirements and writing conflicting requirements in prompts [19, 40, 47, 53]. Even experts need many iterations to improve their requirements for complicated LLM tasks [62]. However, training novices on PE is difficult; for instance, novice programmers may not improve at prompting

⁴<https://chatgpt.com/create>

within a 75 minute study when provided with test cases and generated code as feedback [49]. We aim to address this gap by proposing a ROPE training for end-users to improve their requirements engineering skills in the context of prompt creation.

2.2 Instruction Following for Foundation Models

Optimizing LLM performance on customized tasks typically follows two approaches: improving the models directly, or refining prompts with models fixed. The former has enabled the versatility of models like GPT-4,⁵ Gemini,⁶ and LLaMA 3,⁷ which are fine-tuned through instruction tuning [50] and further aligned with human preferences using Reinforcement Learning from Human Feedback (RLHF) [13]. For the latter, the NLP community has been exploring LLMs' capabilities to follow requirements (or "constraints") with various techniques such as decomposition (least-to-most prompting [84] and AI chains [73]) for complex tasks. Various datasets have been proposed to assess whether LLMs can fulfill constraints, such as IFEval [85], INFOBench [55], and FollowBench [23]. These datasets focus on evaluating models' abilities to handle compositional constraints, and have demonstrated promising advancements.

Our work hypothesizes that requirements can be central to human-LLM interactions, as we have observed the LLM proficiency in following instructions. However, existing requirement datasets are often synthetic, containing prompts that are intentionally tricky or unrealistic, with requirements generated automatically via templates and limited to certain categories (e.g., sentence or paragraph length constraints, format constraints like output in JSON). While useful for identifying *LLM-hard* problems for requirement-focused training, these datasets do not fully capture how LLMs respond to real-world user prompts. Human prompts may have more diverse requirements across various tasks and less standardized language — a gap we seek to fill in this paper.

Building on this foundation of instruction following, *prompt optimizers* have been proposed to improve the performance of frozen LLMs on user-defined downstream tasks by automatically refining user-written prompts. Automation ranges from simple adjustments like adding role-playing or chain-of-thought prompting [70] (e.g., Prompt Maker in Figure 2) to more advanced methods like TextGrad [79], which searches for optimal wording, PromptCharm [69], which enables users to adjust diffusion models' attention to keywords in prompts, and DSPy [25], which extends the search space to few-shot prompts and various prompting techniques. While these sophisticated methods make more tailored changes, they are harder to apply and often require access to labeled datasets. Despite being in early stages, these optimizers aim to relieve humans from exhausting all possibilities in the natural language prompting space, as automated experiments with semantically-preserving edits can be more effective. We share their vision and foresee a future where optimizers play a crucial role in human-LLM interaction.

2.3 Requirements in Software Engineering and Design

In software engineering and end-user software engineering, *requirement* describes what a human wants to achieve and how a program should behave in the world [27]. In product and engineering design, *requirement* is a description of the desired solution to a problem [42, 67]. Requirements encompass all the necessary conditions, constraints, and desired outcomes to ensure the output aligns with the human's needs and expectations, and specifying requirements is often a crucial step in both the design and engineering process [16]. *Requirements engineering* stems as a field that focuses on generating and documenting requirements for software [66]. Good quality requirements need to be *accurate* and *complete*, without *commission* (inclusion of irrelevant or incorrect details) and *omission* (exclusion of necessary details) defects [1]. Previous studies have identified requirements

⁵<https://openai.com/index/gpt-4/>

⁶<https://gemini.google.com/>

⁷<https://www.llama.com/>

engineering as a challenging skill to master. While training mechanisms have been developed to help students avoid commission and omission errors [48], a significant skill gap remains between graduates and professional engineers [18, 56].

Parallels can be drawn between prompt engineering (PE) and requirements engineering. For example, adapting LLMs to diverse scenarios demands well-specified requirements, and prompt authors often need to iterate on prompts that are too ambiguous for the LLM to interpret [52]. While PE training stresses the importance of clear requirements [43, 59, 60], this is often presented just as one of many prompting principles (e.g., one of 26 [7]), limiting its impact on prompt authors. In fact, little attention has been given to explicitly training end-users on requirements engineering in the context of prompt programs, including both natural-language-to-code and natural-language driven GPT applications. While several tools like EvalLM [26] and EvalGen [63] have explored extracting user criteria to support prompt *evaluation*, the concept of requirements has not yet been fully emphasized in the context of prompt *construction*. We aim to address this gap by introducing the ROPE paradigm and developing a requirement-focused training mechanism to help end-users more effectively instruct LLMs to achieve their desired goals.

3 THE ROPE PARADIGM

We offer our definition of Requirement-Oriented Prompt Engineering (ROPE) and describe why, when, and who need it, and why we propose a training toward ROPE in this work. ROPE is a human-LLM partnering strategy where humans maintain agency and control of the goals via specifying requirements for LLM prompts. ROPE represents an **emerging paradigm in how we interact with LLMs, focusing on the importance of crafting accurate and complete requirements to achieve better results, especially for complex, customized tasks.**

The definition of requirements. We define a *requirement* as a skeleton instruction that communicates an essential condition or constraint on desired LLMs output (e.g., ``Response is less than 100 words``). Requirements instruct LLMs to perform tasks that may deviate from their default behavior, guiding LLMs to align with user’s goals.

In our work, we focus on evaluating requirements quality rather than quantity, and we operationalize requirement quality in our work (defined in Section 4.2) by adopting a taxonomy from the requirements engineering literature (Section 2.3). Ensuring requirements are accurate and complete is essential for effective LLM prompting, as poor-quality requirements can lead to harmful outcomes – e.g., missing requirement ``anonymize the data`` may make LLMs keep identifiable information in data, and vague requirement ``delete harmful content`` without a clear definition of “harmful” may cause LLMs to misinterpret sensitive topics like mental health or race as harmful.

Here we also focus on *natural language requirements* due to its universal understanding and tight connection to LLM prompting. While multi-modal requirements can be compiled into prompts for different generative AI models, we leave this exploration for future work.

The relation between requirements and prompts. We view a prompt as a *super-set* of requirements. It contains not only users’ customized requirements, but also other (orthogonal) factors like text fluency and standard prompting tricks. Among these factors, requirements are more *user-centered* and *LLM-agnostic* – users’ goals generally remain consistent across models (e.g., GPT-4, Gemini). Fully articulating these requirements is essential, as omitted customized details can be difficult to recover automatically (explored further below).

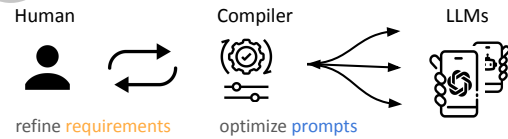


Fig. 3. Our envisioned ROPE paradigm.

In contrast, other factors tend to be more *LLM-centered*: different models may prefer different writing styles, example types, or requirement order. While optimizing these factors can enhance LLM performance, they are better suited for automation rather than humans due to their (1) formulaic nature (e.g., wrapping prompts in a template consistently improves Llama-3-8b’s performance on any task [17]) and (2) idiosyncratic behavior (e.g., GPT-3 can be sensitive to semantically invariant edits that humans do not expect to make drastic differences [80]). This division is crucial given the rapid evolution of LLMs, as developing a mental model for a specific LLM can be a waste of effort in the long run [37].

With this distinction, we view ROPE as a task-delegation strategy. In crafting executable prompt programs, users should focus on *iterating and refining requirement-related components*, leaving other aspects automated. We find recent advancements in prompt optimizers (Section 2.2) particularly promising. With a consistent set of user requirements, we hope future optimizers could function like program synthesizers, searching through possible “implementations” – in terms of phrasing, examples, and structure – to generate prompts that best align with each LLM’s preferences.

The applicability of ROPE: LLM-hard tasks and prompt programmers. In theory, any task can have countless requirements; for instance, implicit needs like “reply in the same language” are almost always assumed in LLM applications. However, not all requirements need to be explicitly stated. For standard tasks well-represented in LLM training data, e.g., “correct the grammar in this email”, LLMs can meet expectations relying on their parametric knowledge without explicit clarification (e.g., on what grammatical rules exist).

We argue that explicitly articulating requirements becomes more crucial for what we refer to as **LLM-hard** tasks. These tasks require significant customization where users must intentionally *deviate LLMs from their default behaviors* [23, 55, 62], making autofulfillment impossible (e.g., how many rounds of questions should be asked in Figure 2). At the time of writing, tasks that involve multi-step processes, decision-making, or context knowledge are likely more *LLM-hard* and thus benefit more from explicit requirements. However, this is a dynamic concept that evolves as LLMs improve.

Nowadays, LLMs are widely used by *end user to create reusable prompt programs*, and these prompt programs are often LLM-hard, as users frequently seek customization with various requirements, anticipating diverse inputs and expected outputs. From content creators crafting prompts for creative graphics, educators tailoring LLMs as tutors, to office workers automating workflows with GPT, people across various fields are all building natural language prompt programs.⁸ We argue that *all these potential prompt programmers would benefit from articulating better requirements.*⁹

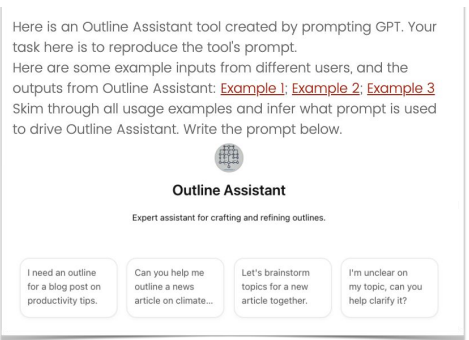
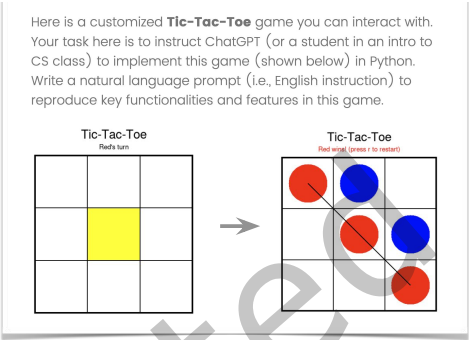
The challenge of ROPE and the call for requirement-focused training. Ideally, we hope LLM users would naturally refine requirements in prompts based on unsatisfactory model outputs. In reality, this is challenging. Novices struggle to *understand what constitutes a requirement* in prompt engineering [80], while even experienced users struggle at *extracting and expressing requirements appropriately*. For example, users find it difficult to translate raw observations on model outputs (“I don’t like how the chatbot didn’t introduce itself”) to clear requirements (“Introduce yourself at the start of the conversation, and state what you can help with”) [35, 52]; Users also cannot decide on the right level of specificity and abstraction [34], e.g., use overly general keywords where more granular, domain-specific requirements are needed for LLM-hard tasks [47, 49].

We hypothesize that *users need explicit guidance to prioritize requirements during prompt creation*. Effective training should help users recognize the importance of clear, complete requirements and develop skills that connect abstract ideas with concrete model behaviors – goals we address in our training design.

⁸While a prompt program can output code like traditional natural language programs [22], it can also produce other outputs such as videos or GPTs applications. Note that when we refer to natural language program later in the paper, we are generally describing prompt programs.

⁹In contrast, more ad-hoc interactions with LLMs, such as casual conversations with the model, may require a different set of considerations, which we consider beyond the scope of this discussion.

Table 1. An overview and comparison on the task types in our training.

	GPTs Task	Game Task
Task Type	Customized LLM powered directly by natural language	Natural Language to code (from CS assignment)
Example & Requirements (Rubrics)	<p>Here is an Outline Assistant tool created by prompting GPT. Your task here is to reproduce the tool's prompt. Here are some example inputs from different users, and the outputs from Outline Assistant: Example 1; Example 2; Example 3. Skim through all usage examples and infer what prompt is used to drive Outline Assistant. Write the prompt below.</p>  <p>1. Use follow-up questions to let the users clarify and specify their needs, e.g., target audience, length, focus, tone, and style. 2. Offer advice for potential modification direction. (6 in total)</p>	<p>Here is a customized Tic-Tac-Toe game you can interact with. Your task here is to instruct ChatGPT (or a student in an intro to CS class) to implement this game (shown below) in Python. Write a natural language prompt (i.e., English instruction) to reproduce key functionalities and features in this game.</p>  <p>1. Render the title Tic-Tac-Toe on top of the board. 2. Display the corresponding message (e.g. Red's turn) 3. Keypress 'r' to restart the game & reinitialize the board. (9 in total)</p>
Specificity	Higher-level requirements covering diverse categories, but can have subjective interpretations	Lower level requirements akin to implementation specification but more verifiable
Modality	Convey requirements through textual examples – users are given 2-3 chat histories to control for task time and interaction exposure. It is representative of <i>status-quo</i> prompt iteration.	Convey requirements visually – users are given an interactive game. The interaction reduces ambiguity that is common with few examples and requires less English reading skills.

4 TRAINING AND EVALUATION DESIGN FOR ROPE

We develop a **training and assessment suite** to help users improve their ability to write *accurate and complete* requirements for instructing LLMs. We adopt a backward design method [71], a well-established instructional design approach that starts by identifying the desired learning outcomes – in our case, writing effective requirements for LLMs – then works backward to develop assessments and training aligned with goals. Aligned assessments and training are critical to ensure that what is taught directly prepares participants for the skills they are expected to demonstrate.

Through multiple pilot studies with novices and experts ($n = 10$), we refine three key components of design: (1) realistic tasks that mirror real-world prompting challenges (Section 4.1); (2) assessments that connect requirement quality to LLM outcomes (Section 4.2); and (3) deliberate practices with feedback in a system (Section 4.3). Beyond training (Section 6), our ROPE assessment and materials also enable us to build more nuanced understandings on how requirements affect LLM outputs.

4.1 Task Design: LLM-Hard Prompt Programs for Replication

We aim for novices to eventually develop the ability to articulate requirements for their own prompt programs. However, to begin their training, we need to provide a set of concrete sample tasks for them to practice and for us to evaluate their progress. Below, we outline the six carefully designed tasks.

Task setup. To ensure the training and evaluation process is both realistic and representative, our primary objective is to have users *write natural language prompts to instruct LLMs in generating applications*. Instead of open-ended tasks — which are difficult to assess and provide consistent feedback on — we aim to *provide clear ground truths for evaluation and training*. To this end, users are asked to write prompts to **develop an LLM app given examples**.

Each task consists of two key components: (1) A set of *interaction examples* from which users can deduce requirements; and (2) A set of *gold requirements* that precisely reflect the app’s behavior. Ideally, an expert skilled in requirement-driven prompt engineering should be able to express all the gold requirements within their prompts, guiding the LLM to generate an app that produces all the interaction examples.

This setup requires users to derive requirements from provided examples, which represents a realistic form of communication from a client to a developer. It captures actual challenges of requirement articulation, especially for LLM-hard tasks in the real world (as described in Section 3): Users as LLM app developers often have access to some text inputs or visual examples like a demo or a Figma mockup, often need to refine their requirements by analyzing available outputs, but do not always know how to connect model outputs with abstract requirements. We discuss potential future work of open-ended tasks in Section 7.1.

Task selection and customization. To cover broad types of realistic prompt programs users may write, we incorporate both GPTs applications directly powered by natural languages, as well as natural-language-to-code applications. We prioritize tasks most suitable for targeted and engaging training, through two dimensions: (1) we prioritize *generic tasks* over specialized ones (e.g., data science tasks), to ensure that the challenge lies in specifying requirements rather than applying domain-specific expertise, preventing construct irrelevance; and (2) we opt for *tasks with visual interfaces*, enabling users to interact with and visualize outputs in a manner that is accessible to novices.

With these criteria in mind, we select three GPT-driven tasks —Outline Assistant, Trip Advisor, and Email Proofreader — along with three coding tasks presented as visual game interfaces: Connect4, Tic-Tac-Toe, and Tetris (see examples in Table 1). The GPTs tasks provide interaction examples in terms of textual chat histories, and the Game tasks provide interaction examples via an interactive visual interface. All these tasks are inspired by real-world applications and have reference requirements, though we make necessary adaptations to ensure they are *LLM-hard* problems. The Game tasks are sourced from the Introduction to CS course at our institution, each with reference requirements provided by the instructor. They require minimal customization, as instructors have already incorporated LLM-hard requirements to prevent students from cheating using large language models. For the GPTs tasks, we adapt real-world GPTs prompts,¹⁰ and make them more LLM-hard by using frameworks like IFEval to introduce customized requirements such as format constraints [85]. We then derived the requirements by parsing the underlying prompts through expert annotation and pilot testing.¹¹

While these tasks are aligned for the training and assessment design [71], they span a variety of application domains with unique requirements, preventing novices from relying on memorization and promoting meaningful learning. As shown in Table 1, the GPTs and Game tasks also offer complementary coverage on requirement specificity and modality.

Task validation. We distribute the six tasks in assessments and the training session. We use Tetris and Email Proofreader in training, and the rest for pre- and post- assessments before and after training in a counterbalanced design. In pilot studies, we confirmed that (1) we could distinguish prompting novices and experts using the tasks, with the average novice performance being 28% and expert achieving full scores, (2) all tasks could be

¹⁰Writing Assistant prompt: link, Trip Advisor prompt: link, Email Proofreader prompt: link

¹¹Detailed instructions of the task, as well as the requirement breakdown, are provided in supplemental materials.

implemented by LLMs, with the average highest score being 91%, and (3) the assessment tasks were all comparably challenging for users.

4.2 Assessment Design: Requirement-Focused Intrinsic and Extrinsic Evaluation

To evaluate users' ability to instruct LLMs effectively, we assess both the quality of the user prompts (requirements within the prompt) and the prompt's impact on the quality of the LLM's output. Thus, for each user task completion, we calculate its *Overall Score* as the average of two components, *Requirement Quality* and *LLM Output Quality*:

- *Requirement Quality*: This intrinsic metric assesses, "Does the user's prompt accurately and comprehensively cover all requirements?" It measures the correctness and completeness of the requirements by comparing those extracted from the user's prompt to expert-defined reference requirements for each task (described in 4.1). We operationalize requirement quality as the percentage of reference requirements that are correctly described in the user's prompt. Drawing from the requirements defect taxonomy [1, 45], we mark both *commission errors* (incorrect, inconsistent, or ambiguous requirements) and *omission errors* (missing reference requirements) in users' prompt, so a requirement is only correctly described if it is free from commission and omission defects.
- *LLM Output Quality*: This extrinsic metric evaluates, "Can the user's prompt successfully guide the LLM to achieve the intended goals?" It measures the proportion of desired features implemented in the LLM-generated output that align with the reference (Section 4.1). We pass users' prompts to GPT-4o and compare the generated output to the reference program.¹² Since many features are difficult to evaluate automatically (e.g., GPTs behaviors cannot always be checked via rules), we rely on manual evaluation. For games, we interact with the generated program to verify whether the requirement features are correctly implemented. For GPT interactions, we grade whether the GPTs' replies display expected behaviors (e.g., asking a follow-up question when the user's input is ambiguous). We grade 10 complete conversations per GPTs.

Assessment validation. Three of the authors discuss and iterate the grading rubrics on the tasks during the pilot study. We confirm that an expert can measure requirement quality using the percentage of correct requirement clauses in novice prompts, and that this correlates positively with the expert's judgment of overall requirement quality (Spearman's $\rho = 0.66$). For grading the *Overall Score* in the user study (Section 5), two authors (one of whom did not participate in the development of the rubrics during the pilot) independently grade 10% of the randomly chosen pre-post test responses. We check the inter-rater reliability between the authors' *Overall Scores* by calculating the Intraclass Correlation Coefficient (ICC) [29], finding strong reliability ($ICC = 0.9$, 95% Confidence Interval = [0.7, 0.98]). Any discrepancies in scoring are resolved through discussion, after which the authors individually grade half of the participants' responses. Detailed rubrics are provided in supplemental materials.

4.3 Interactive Training Mechanism: Dedicated Practice and Feedback on Requirement Defects

We design a training mechanism to support deliberate practice on requirement elicitation and refinement by *disentangling requirement articulation from peripheral tasks like persona crafting or paraphrasing in existing prompt engineering instructions*. We apply key learning principles like scaffolding and worked example [28], and we implement the training into an interactive interface as shown in Figure 4.

In this training, a user (a prompt novice) is asked to develop one Game (Tetris) and one GPTs (Email Proofreader) given the interaction examples, *solely by describing the requirements of the app as accurate and complete as possible*. For example, to develop the customized Tetris game in Figure 4, the user will first start by outlining the main milestones (e.g., creating the game board, handling piece placement, etc.), and then provide

¹²The output is generated using a prompt similar to ``Follow this prompt: {user_prompt}``. Further details are in the supplemental materials.

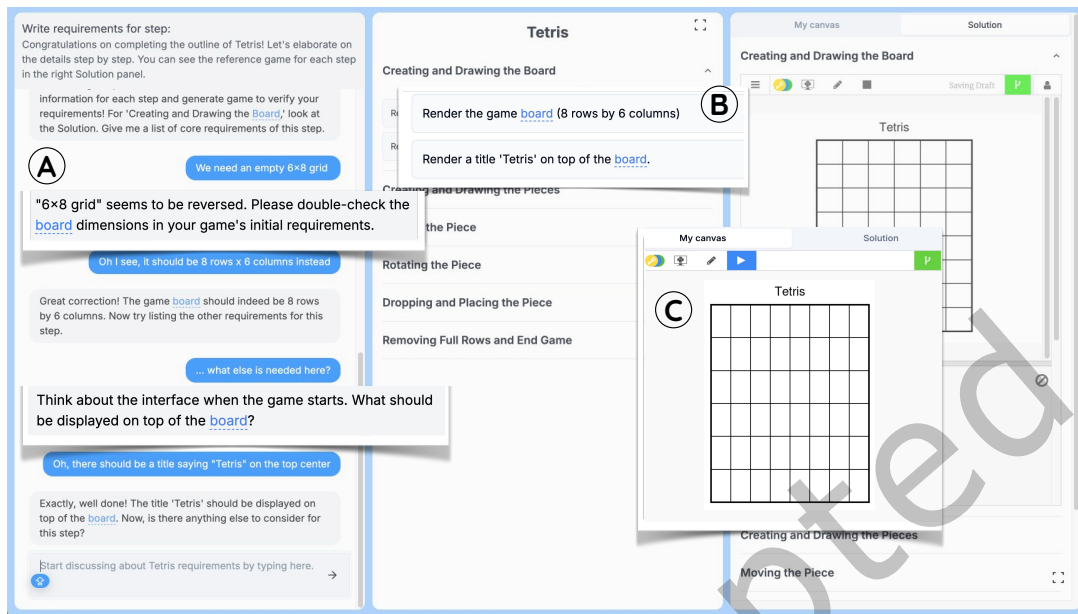


Fig. 4. Our ROPE training interface, with three types of feedback on requirement defects: (A) *Chat-based hints* on incomplete or inaccurate requirements, (B) *Reference requirement examples* to reinforce appropriately identified and expressed requirements, and (C) *LLM output counterfactual* to support user’s reflection on incorrect or ambiguous requirements.

more detailed specification per step (e.g., define the size of the game board). Novices are not expected to achieve full success on their first try; we use three types of **requirement-focused feedback** to guide users in continuously refining their requirements:

- *Conversational hint and clarification*, via chatbot (Figure 4A): We create a tutor chatbot to provide feedback on users’ requirement mistakes. For instance, the chatbot may ask “What’s on top of the board?”, when a user *misses* the requirement for “Render a title ‘Tetris’ on top of the board” (*omission error*). As shown in Figure 4A, the chatbot may suggest the user to “double-check the board dimensions” if the user *incorrectly* reverses the number of rows and columns for the requirement “Render the game board (8 rows by 6 columns)” (*commission error*). The textual feedback encourages critical thinking on *missing* or *incorrect* requirements and offers a natural, conversational experience to discuss and reflect on requirements.
- *Reference requirement example*, via requirement working document (Figure 4B): We progressively reveal reference requirements as feedback when users correctly identify them during interaction with the chatbot. The expert-written reference examples are provided to *reinforce correct requirements*, helping users understand how to formalize and organize requirements.
- *LLM output counterfactual*, via generated visualization (Figure 4C): If novices produce incorrect requirements that can be visually demonstrated (e.g., wrong board dimensions), we provide visual feedback by generating flawed programs that implement the *incorrect* requirements, or maliciously misinterpret *ambiguous* requirements. For example, a vague requirement like “Use keys to move pieces” might result in a flawed Tetris game where pieces can move upward, exposing the unspecified allowed movement in requirement. The visual counterfactual is currently limited to the controllable game code generation; as GPTs outputs tend to be less predictable, we display static chat histories as illustrative examples for GPTs.

Note that our assessment tasks require users to write complete prompts similar to those in Figure 2, we deliberately avoided having users write prompts during training. Instead, users engage in conversational interactions designed to continuously encourage them to think about requirements. Additionally, users’ written responses do not directly trigger LLM output generations. We use reference requirements to guide the LLM generation, ensuring that all feedback reflects the quality of the requirements alone. This approach ensures that novices remain focused on improving their requirement articulation skills, without distractions from other factors.

Interaction and feedback implementation. We use OpenAI GPT-4o to power the interface,¹³ leveraging its interactive nature to adaptively provide feedback on the diverse set of possible user inputs. However, feedback generation does not rely on the model’s inherent reasoning capabilities but is anchored in our predefined reference requirements. Specifically, we always have GPT-4o to compare the current user requirements to the reference, and select the most critical defect to provide feedback on (roughly, mismatched requirements are considered most critical, then missing requirements, then ambiguous ones). Then, we have the LLM respond targetedly to the defect. For example, visual counterfactuals on game tasks are generated by having the LLM minimally edit reference Python code based on the identified incorrect requirement. To enhance interaction, we also implement features like highlighting special variables (e.g., board, keystrokes) in conversations and as hyperlinks for easy cross-referencing across the interface.

Validation on feedback generation. We analyze the interface log data to evaluate the feedback quality for all types of feedback (Figure 4A, B, C). From our user study (Section 5), we collected a total of 635 interaction turns from Tetris (ranging from 11 to 73 turns per user) and 180 turns from Email Proofreader (ranging from 4 to 20 turns per user). From this, we randomly selected 10 different users and annotated conversations for 5 users per task, resulting in 151 Tetris turns, 66 Email Proofreader turns, and a total of 113 annotations per feedback type. For each LLM turn, we assess: (1) Is the feedback needed? (2) Is the feedback provided? (3) Is the provided feedback correct? Given the answers to the three questions, each feedback was categorized as correct, incorrect, irrelevant, missing, or not provided (Table 2). For example, *irrelevant feedback* are those that are (1) not needed but (2) provided. Note that the right feedback includes both *correct feedback* that was (1) needed, (2) provided, and (3) correct, and *not provided feedback* that was (1) not needed and (2) not provided.

Table 2. Proportions of feedback types across different feedback sources and IRR (Krippendorff’s α) of human annotations

Feedback Source	Right Feedback			Wrong Feedback				IRR (α)
	Correct	Not Provided	Total	Incorrect	Irrelevant	Missing	Total	
Chat (e.g., Fig. 4A)	69.03%	20.35%	89.38%	10.62%	0.00%	0.00%	10.62%	0.86
Ref Example (Fig. 4B)	48.67%	38.05%	86.73%	1.77%	5.31%	6.19%	13.27%	0.88
LLM Output (Fig. 4C)	3.75%	86.25%	90.00%	0.00%	5.00%	5.00%	10.00%	0.68
All Feedback	44.44%	44.12%	88.56%	4.58%	3.27%	3.59%	11.44%	0.87

Two authors independently annotated one participant’s chat log for each task, and we measured Inter-Rater Reliability (IRR) by calculating Krippendorff’s α [8], achieving a high agreement rate ($\alpha = 0.87$) across all feedback types (see Table 2 for the detailed breakdown). Any discrepancies were discussed and resolved, after which one author completed the remaining annotations. The final analysis revealed a correct feedback rate of 88.6%.

¹³The state-of-the-art LLM at the time of writing: <https://openai.com/index/hello-gpt-4o/>. We used a temperature of 0.3 for code generation and 0.7 for other tasks. Refer to supplemental materials for our prompts.

5 USER STUDY DESIGN

We conducted a user study to understand whether our requirement-focused training is effective in improving novices on writing requirements in prompts, and whether requirement-focused training is *more* effective than *status-quo* prompt engineering training. We also examined how automatic prompt optimization affects performance.

Study procedure. To capture the requirement-focused training gains, we utilized a **pre-test and post-test approach**, a standard instruction evaluation method [57], comparing participants’ performance on isomorphic tasks before and after ROPE training (within-subject). To compare requirement-focused training with standard training, we utilized a **randomized-control design** (between-subject). We randomly assigned participants to either the experimental condition (ROPE group), where they receive requirement-focused training in our interface, or the *control* condition (PE group), where they receive a standard prompt engineering tutorial and then self-practice using ChatGPT.

Our user study – for either the ROPE or the PE group – is 1.5 hours long. In the study, participants first completed a two-minute *pre-survey*, providing demographic information and rating their prompting experiences (e.g., confidence and self-perceived productivity using ChatGPT) with 7-level Likert Scale questions. Then, they started a *pre-test* (20 minutes) to write prompts to replicate key features on existing programs (Table 1) – one Game (Tic-Tac-Toe or Connect4) and one GPTs (Outline Assistant or Trip Advisor). They proceeded to the 40-minute *training session* (different interfaces depending on experiment group) on two tasks (Tetris and Email Proofreader), followed by a *post-test* that contained different Game and GPTs tasks (20 minutes).¹⁴ To further capture participants’ iterative prompting behaviors, for the post-test GPTs task, participants were also asked to feed their prompt into ChatGPT and iterate their original prompt by observing the ChatGPT outputs. They completed the study with another two-minute *post-survey*, providing feedback with open-ended or Likert Scale questions on their experience and perceptions during the training. Participants were compensated with a \$20 Amazon Gift Card. Please refer to supplemental materials for our study materials.

Control group (PE) design. We designed the PE group to replicate the “business-as-usual” scenario, where novices learn prompt engineering through publicly available resources and self-practice. Specifically, PE group participants first watched a 20-minute YouTube tutorial,¹⁵ covering best prompting practices including clear instructions, adopting personas, specifying format, limiting the scope, and few-shot prompting. The tutorial included various prompting examples such as poem writing, code-based data processing, essay summarizing, etc. Afterward, participants self-practiced writing prompts for the two training tasks, iterating with ChatGPT directly. For the game task, we offered participants an online code compiler so participants could see games rendered in real-time, similar to the visual feedback in our interface (Figure 4C). Similar to the ROPE group setup, our PE group offers practices on coding and GPTs tasks to establish task familiarity. However, unlike our requirement-focused training where users practice writing only requirements without actually writing prompts, the conventional practice for PE group directly train participants on writing prompts, which is closer to what users actually write in pre-post tests, and the direct interaction with ChatGPT might help users build a better mental model on LLM behaviors.

Data collection. To understand participants’ perceived experiences, prompting behaviors, and learning outcomes, we collected participants’ prompt responses, survey answers, time on task, as well as interaction log data during the training phase. Afterward, we collected the LLM outputs by feeding participant prompts

¹⁴The pre- and post- assessments each contained one Game task and one GPTs task (Section 4.1), counterbalanced to mitigate problem sequencing bias or different task difficulties. We found that the two versions of the tests have comparable difficulties from pre-test results (an average overall score of 21.0% and 25.6% for each version).

¹⁵Relevant sections from Prompt Engineering Tutorial – Master ChatGPT and LLM Responses

into OpenAI GPT-4o.¹⁶ Using the evaluation method and rubrics we developed for each of the assessment tasks Section 4.2, we graded the pre- and post- tests using users' *original prompts* for all tasks from the test, on a scale of 0% to 100%. We calculated the *Overall Scores* as the average of the *Requirement Quality* and *LLM Output Quality* scores. To further understand the effect of model advancement, we collected LLM outputs with the newer, more powerful reasoning model OpenAI o3-mini¹⁷ and conducted the same analysis. The results and comparisons against GPT-4o are presented in Section 6.2.

To understand the effect of an optimizer on a prompt (as discussed in Section 2.2), we further use the Prompt Maker,¹⁸ a popular optimizer with 50k usages in GPT Store, to refine participants' *original prompts* into *optimized prompts*. This optimizer is easily applicable to any kind of prompts (Figure 2 shows an example of its output), though it is a rather preliminary version, as it rephrases user prompts in a rule-based manner without criticizing the LLM outputs of the original prompt. We evaluated the optimized prompts in the same way as the human prompts (Section 4.2), including generating and grading corresponding LLM outputs.

Participants. We recruited users with limited or no NLP background from different institutions in the United States. We conducted studies with 32 participants randomly assigned to the ROPE or PE group, and we removed two participants who disengaged during the study from the analysis. In total, we have 30 participants (S1-30, $n = 15$ for each condition) — 19 female, 10 male, 1 agender, and 18 non-native English speakers (9 in each condition), with an average age of 26. Our participants have a diverse range of backgrounds including HCI, education technology, communication, graphic design, mechanical engineering, information systems, literature, psychology, and economics. We asked the participants to rate their familiarity with LLMs and complex prompts in the pre-survey, and the average rating is 3 out of 7.

6 USER STUDY RESULTS

6.1 Learning Gain: Requirement-focused Training Helps Students Instruct LLMs

We start by answering our central research questions: *Can we help end-users better instruct LLMs through requirement-focused training?* We quantitatively measured learning gains by capturing participant performances from the pre-test to post-test within each experiment group, and evaluated the effectiveness of the two training approaches by comparing the ROPE and PE group. We further unpacked the quantitative results through analyses of participants' Likert Scale ratings,¹⁹ as well as the transcribed comments during their training sessions and in post surveys.

Requirement-focused training is effective. Using a two-tailed paired t-test,²⁰ we found that the overall scores for participants in the ROPE group significantly improved by 19.1%²¹ from pre- to post-test ($p < 0.001$, from $21.9\% \pm 14.6\%$ to $40.9\% \pm 18.4\%$), an almost two-fold increase (as shown in Figure 5a). This demonstrated that *requirement-focused training enabled participants to instruct LLMs more effectively*, and we achieved our desired goal: *to concentrate users' attention on only requirement iterations during the training, but make sure users can*

¹⁶We used a temperature of 0 for GPT-4o to generate deterministic outputs, please refer to our supplemental materials for more details.

¹⁷<https://openai.com/index/openai-o3-mini/> with the default settings.

¹⁸<https://chatgpt.com/g/g-hhh4w3eov-prompt-maker>: from a simple prompt to an optimized prompt.

¹⁹We investigated whether the learning gains are associated with any demographic factors like age and did not find any significant indicator. There is a weak negative correlation between self-perceived familiarity with LLM and learning gains, which means that the lower LLM familiarity the participants rated, the more they learn from the training ($\rho = -0.2$).

²⁰We did a linear mixed effects regression to account for the correlation across participants and task (both as random effects) of the repeated measures for the pre- and post-test times and for the requirement and LLM output scores. We found the same pattern of significant results as revealed by the t-tests.

²¹*Requirement Quality* score improved by 25.4% and *LLM Output Quality* score improved by 12.7%, both significant ($p < 0.05$). Pre-test performances were not correlated with learning gains.

still apply their learning in overall prompt writing. Analyzing requirement defects in ROPE participants' prompts, we observed a noticeable decrease in omission errors (from 5.6 to 3.2 per participant), but a slight increase in commission errors (from 0.5 to 0.7 per participant). This indicated that participants *became more aware of requirements, but need additional training for requirements clarity and accuracy.*

ROPE participants' self-reflection highlighted that they understand the value of *emphasizing and iterating* on clear and complete requirements, which may contribute to their improvement. For example, 7 out of 15 PE group and 13 out of 15 ROPE group participants mentioned "be specific" in their answers to prompting strategies in post-survey. When asked about what they learned during the instruction, four PE participants noted to "give exact instruction," as described by S23, "we should give as much information as we can to ChatGPT". Meanwhile, some ROPE participants developed a more in-depth understanding of articulating requirements in prompts. For example, six ROPE participants explicitly noted the connection between requirements and prompts. S30 described requirements as a way "to be more clear and not confuse the system by giving too much unnecessary information", and S8 noted that a requirement-focused approach helped them *organize their thoughts*: "I learned to organize my requirements logically so we can easily revise and improve them". S10 neatly described their shift towards requirement-focused prompting strategy: "sometimes when I write prompts, [I find] steps are hard to be clearly divided, or I didn't consider to divide them that detailed. However, it's important to do so, as it appears to give LLM more direct and clear instruction. When the steps are divided, it's easier to see the missing details in my original prompts too".

Requirement-focused training is more effective than standard prompt engineering practice. From the post-test prompt data, we observed that requirement-focused training and standard practice had distinct effects on how participants wrote prompts. ROPE group spent more time (19 minutes vs. 14 minutes) and wrote longer prompts (796 vs. 458 characters) than the PE group. In 100 randomly selected pairwise comparisons, ROPE participants produced more structured prompts 87% of the time, consistent with their self reflections reported above.

These prompts also revealed learning gain differences. A two-sample t-test showed that the ROPE achieved significantly higher learning gains compared to the PE group ($p < 0.001$), as shown in Figure 5b. In fact, while ROPE significantly improved (19.1% as mentioned above), we did not observe much gain in PE group's pre-to-post score (0.7%). The contrast suggested that *novices indeed could not acquire requirement articulation skills through standard prompting training alone.*

Participant feedback supported our hypothesis that *requirement articulation skills do not naturally emerge.* While we hoped interacting with LLMs would help participants build mental models of LLM behaviors, PE participants found the LLM response too unpredictable for effective self-practice. Similar to prior work reporting the challenges of prompting nondeterministic LLM [49, 80], many got frustrated during their unfruitful self-practice sessions: "I was confident with my communication ability at first but later during the tasks I was frustrated with my communication skill with AI" (S11). The unpredictability distracted participants from applying the best practices taught in the tutorial, as S19 noted: "I lost some of the concepts at the end that looked very minor in determining the effectiveness of the prompt".

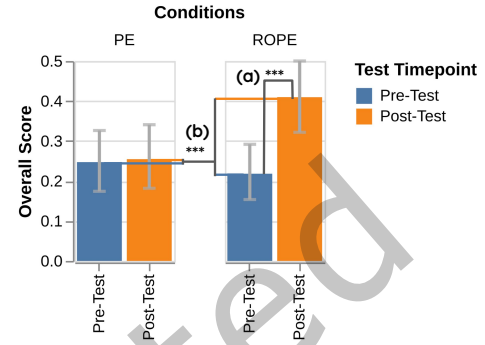


Fig. 5. The pre-test and post-test overall scores between PE and ROPE conditions. (a) Overall scores for ROPE significantly improved from pre- to post-test; (b) ROPE achieved significantly higher learning gains (post – pre) than PE (***) denotes $p < 0.001$).

Requirement-focused training encouraged more targeted and iterative prompting. The aforementioned pre- and post-tests analyses revealed participants’ learning gains in *writing initial prompts for natural language programs*. Beyond initial prompts, we analyzed participants’ prompt iteration behaviors on the GPTs tasks at the end of post-tests. We observed that *requirement-focused training led users’ to make more requirement-related changes during iterations*.

In the PE group, most participants either chose not to make any iterations (5 out of 15) or only made superficial edits (e.g., altering wording or grammar, 8 out of 15), with only 2 adjusting prompt requirements. In contrast, after ROPE training, participants *were more likely to engage in substantive requirement engineering*, updating or adding specific requirements instead of making random edits typical of end-user prompt engineering. Eight of 15 ROPE participants made meaningful requirement adjustments (5 were fully successful, 2 made partial progress, and 1 introduced an incorrect requirement). Participants were also aware of their requirement-centric iterations: “[My strategies were to] break down my requirement into several key points, use examples, iterate, self-check if more details are needed, or if more steps should be elaborated, ..., see the test result using examples” (S10). Overall, ROPE training effectively equipped users with a more structured and systematic approach to prompt generation.

Key to success: deterministic feedback as a jump start. Looking at both learning gains and behavioral changes, we speculate that the success of requirement-focused training stemmed from our use of *requirement-focused feedback*.

Requirement articulation was challenging for all participants, as shown by the low pre-test scores (21.9% for the ROPE and 24.8% for the PE). This poor starting point likely hindered the PE’s ability to receive useful feedback during self-practice, as their requirements were too weak for the LLM to generate “good enough” model output with clear indicators on what to improve. Consequently, PE participants were trapped and frustrated by LLM unpredictability — S6 noted that “ChatGPT is a very malleable tool and can change responses pretty drastically depending on the prompt”, and S7 commented that “if ChatGPT doesn’t get it, I might not be patient”.

In contrast, our feedback loop for ROPE (Section 4.3) provided tightly controlled *feedback and program output* based on the requirements users specified, regardless of grammatical issues or natural language variances. S8 highlighted that: “It’s easy to understand (very logical) and can see how the changes and revises influence the system immediately”. This deterministic feedback reinforced ROPE participants’ belief in the importance of clear requirements, leading to them writing better initial prompts with more requirements in the post-test. We hypothesize that these better initial prompts might also become *more suitable for further iteration*. Although still imperfect, ROPE participants’ prompts might allow the LLM to generate partially correct outputs that contain enough signal towards future improvements.

We suspect that effective prompt engineering training does require deterministic scaffolding. Novices may need reinforcement outside of LLM feedback until they reach a level where interpreting LLM responses becomes valuable.

6.2 In-depth Analysis: The Validity of ROPE Paradigm

Going beyond learning gains, we dive deeper into the connections between requirements in prompts and LLM output quality, the role of optimizers, as well as the differences brought by model advancement.

Requirement quality vs. LLM output quality: promising correlation with nuances. To examine whether more correct and complete requirements led to better LLM outputs, we calculated the correlation between the two components of our *Overall Scores: Requirement Quality* and *LLM Output Quality* scores using two LLMs generating the outputs — GPT-4o as in our main experiment, and o3-mini as explained below. We found a strong positive correlation, with a Spearman’s correlation coefficient of $\rho = 0.71$ [61]. However, task-specific analysis revealed nuances. As shown in Table 3, while most tasks achieved $\rho \geq 0.7$, Connect4 showed no correlation when LLM

output is generated using GPT-4o. Upon further inspection, we found that interestingly, GPT-4o tends to produce two implementation types for Connect4: command line and pygame. Prompts that include interactive or visual requirements such as ``Highlight the hovered column`` often led to pygame implementations. If we split the Connect4 correlations by the LLM output implementation type, we get reasonable numbers for both cases: $\rho = 0.89$ (command line) and $\rho = 0.58$ (pygame), suggesting that implementations may respond to requirements differently. For example, the requirement ``Display player's turn message`` was always implemented in command line Connect4, even if omitted in prompts. However, only 1 pygame Connect4 implemented the same requirement, out of 6 participants who correctly described it. This suggests that requirements can be interdependent, with *LLM-hardness* varying by implementations.

Table 3. Spearman correlations (ρ) between LLM Output Quality and Requirement Quality scores by tasks and models.

Task		ρ (GPT-4o)	ρ (o3-mini)
GPTs	TripAdvisor	0.79	0.75
	OutlineAssistant	0.75	0.83
Game	TicTacToe	0.90	0.90
	Connect4	0.10	0.67
Overall		0.71	0.80

research [23, 85], which tend to rely on synthetic prompts with perfectly written requirements of a fixed set of common constraint types (Section 2.2). Our results highlight the need for further investigation into how imperfect user requirements interact with inherent LLM biases.

Optimizer improves prompts but introduces biases too. To assess how the optimizer affected prompt quality, we experimented on a single optimizer to explore the feasibility, without generalizing to all optimizers (Section 7.1).

We first noticed that the Prompt Maker optimizer effectively enhanced user prompts by adding elements such as role-playing and chain-of-thought, making prompts more structured and fluent, as shown in Figure 2. Besides formatting, the optimizer also made direct modifications to requirements; on average, it added 0.5 missing requirements and corrected 0.3 inaccurate requirements, reducing omission errors from 4.4 to 3.9 and commission errors from 0.6 to 0.3 per participant across all prompts. These changes significantly improved overall scores, as shown by a paired t-test for all users' prompts before and after optimization ($p < 0.001$).

It seemed like there existed some requirements that were *LLM-hard* but not *optimizer-hard*. For example, GPT-4o would not implement ``Display messages for a tie game`` unless explicitly prompted, but the optimizer would automatically add this requirement to user prompts. However, highly customized requirements, such as ``Cross out the winning cells``, were never automatically added (0 out of 51) without explicit user input.

Importantly, *the optimizer could not close the performance gap between the ROPE and PE group*. We computed the gains of PE participants after using the optimizer (7.3% = optimized prompts' post-test scores – original prompts' pre-test scores). Using a two-sample t-test, the ROPE's 19.1% pre- to post-test gains remained significantly higher

We further analyzed how different types of requirement defects impact LLM output quality. We found that the number of omission errors (incomplete requirements) had a stronger negative impact on *LLM Output Quality Score* ($\rho = -0.49$) compared to commission errors (inaccurate requirements) ($\rho = 0.05$). This suggests that while LLMs can correct inaccuracies in requirements, they struggle to fill in missing information – further highlighting *the importance of training humans to express all their necessary requirements*.

Our findings complement existing instruction following re-

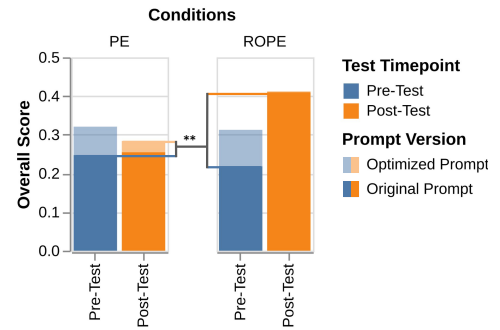


Fig. 6. The pre-test and post-test overall scores between PE and ROPE conditions and the original and optimized prompt versions. ROPE's learning gain remains significantly higher than PE's optimized gains (** denotes $p \leq 0.01$).

($p = 0.01$) than PE group’s optimized gains (7.3%), as shown in Figure 6. This supports the finding that missing, highly customized requirements are unlikely to be added automatically.

While the optimizer generally had a positive impact, *it might misinterpret user inputs or add irrelevant details*. For instance, for four out of five participants who requested an alphabetic list (A, B, C), the optimizer changed it to a numbered list. In some cases, these misinterpretations drastically reduced LLM performance. For example, two participants used the word “interactive” in their prompt, leading the optimizer to incorrectly add the role of “intensive interactive web application developer” and request the use of “Flask and HTML” in the optimized prompts, leading to poorly implemented code since it is much harder to generate a functional web app.

These issues highlight potential *gaps in user-expressed requirements and LLM-interpreted requirements*, suggesting that future training should address the risks of misinterpretation by either LLMs or optimizers (Section 7.2). On a more positive note, such misinterpretations could provide useful feedback, alerting users to potential misalignment between their requirements and the LLM’s understanding.

ROPE’s effectiveness generalizes to more advanced (reasoning) LLMs. Noticing the continuous evolution of LLMs, we naturally ask the question: Would ROPE’s effectiveness generalize beyond the GPT-4o we experimented, to models with better capabilities and alignments to human intents? Recent advancements in LLMs have introduced reasoning models, which have improved instruction-following capabilities and exhibit better performance in complex reasoning tasks [14]. We conducted an additional experiment using OpenAI’s o3-mini, a representative model of the emerging reasoning LLMs that has been shown to surpass GPT-4o [11]. We had o3-mini generate outputs using the same prompts (both original and optimized ones as above), and repeated all the grading and analyses as done on GPT-4o.

The results showed that *ROPE-trained prompts not only remained effective, but in many cases performed even better with o3-mini*. Specifically, when using the overall scores calculated on o3-mini (which impacted the *LLM Output Quality* score), we saw a larger learning gains compared to GPT-4o (23% > 19% for the ROPE group, 2% > 1% for PE group). This learning gap again could not be closed by the optimizer. Digging deeper, we found that o3-mini improved the *LLM Output Quality* score by 9.7% compared to GPT-4o (paired t-test, $p < 0.001$), which means that o3-mini better implemented the prompts. This suggests that improved model capabilities do not render explicit human requirements obsolete, but rather amplify their benefits, making learning on requirements more reflective on the model output quality.

Furthermore, we also found a stronger alignment between user-specified requirements and generated outputs for o3-mini, indicated by the stronger correlation between *Requirement Quality* and *LLM Output Quality* scores when using o3-mini compared to GPT-4o ($\rho = 0.80 > 0.71$, as shown in Table 3). The largest improvement in correlation came from the Connect4 task, where o3-mini showed better abilities to implement customized requirements like “Highlight the hovered column.” These findings underscore the continued relevance of human-authored requirements even as LLMs evolve (more discussions in Section 7.2).

7 DISCUSSION

In this work, we introduce Requirement-Oriented Prompt Engineering (ROPE), a human-AI collaborative prompting paradigm where humans focus on effective requirements specification. Our evaluation shows that requirement-focused training significantly enhances novices’ ability to extract and articulate requirements in prompts, leading to more goal-aligned LLM outputs; we also notice a key communication barrier between humans’ under-specified requirements and LLMs’ misinterpretations (Section 6.2).

While our work made an important step toward ROPE, several limitations remain, and there are many interesting questions to explore. Here, we reflect on possible immediate extensions on requirement-focused training, as well as the longer-term evolution of the ROPE paradigm.

7.1 Limitations and Next Steps on Requirement-Focused Training

Our training program significantly improved participants' ability to extract and articulate requirements, though the skill remains challenging. Post-training assessments showed an average score of 41% (max=63%), and the top ROPE learner gained 43% in scores from pre- to post-test. To further improve the effectiveness of requirement-focused training, we propose several next steps.

One easy extension is to make the training session longer. We observed that the 40-minute training time was difficult for non-native English speakers, and participants had limited opportunities for iteration due to time constraints. Future work should explore longer training sessions and adaptations for users with varying language proficiencies.

In addition to extending length, our training can also be broadened for better generalizability. Our study demonstrated that the requirement specification skills users acquired through training generalized to new GPTs and Game development tasks. However, an open question remains: how broadly do these requirement skills transfer to other natural language programming tasks (e.g., data analysis or creative content generation)? Does training on easier tasks generalize to harder tasks? In preparation for generalizability, our ROPE interface is adaptable to different tasks, as the training and feedback loop can be easily generated based on provided reference requirements and LLM outputs. We open-source ROPE system and encourage its use for customized tasks across domains. Moreover, beyond human-AI interactions, ROPE skills — such as requirement extraction and iterative refinement — can be broadly applicable to fields like software engineering and design. Future research should explore further transfer tasks in assessments, and also how ROPE skills may enhance communication, problem decomposition, and computational thinking across various fields.

Another dimension of generalizability is towards users' different prompting behaviors. Our study focused on extracting fixed requirements, which allowed for controlled feedback and simulated the scenarios when interaction examples are available. However, this setup did not fully capture all real-world prompting scenarios: users often engage in *one-off* interactions (e.g., direct question-answering) or *iterative and open-ended* prompting (e.g., chatting with LLMs on a rough idea) [26, 62]. Future studies should investigate more self-directed tasks, compare different types of prompting, and tailor training or support accordingly. While providing feedback without clear requirements may be challenging, pre-generated materials on common novice approaches could offer useful scaffolding.

One primary bottleneck for broadening the scope is our lack of automatic assessment methods for the LLM-hard tasks. Our time-consuming manual grading effort limited our ability to conduct more experiments by e.g., grading multiple LLM outputs using different temperatures per user prompt, experimenting with multiple optimizers, etc. Future work should explore automatic assessments and their trade-offs. For example, LLM-as-a judge paired with few-shot prompting might provide reasonable estimations [83] and might be suitable for tasks that contain substantial numbers of requirements where wrong grading on a single requirement does not drastically affect the assessment results. Future work can also explore alternative instructional designs such as using human experts or incorporate additional learning principles [28] to maximize the efficacy of ROPE training.

7.2 Reflection on the ROPE Paradigm

Along with prior work on requirement-oriented evaluation [26], we believe that ROPE moves us towards a future where **requirements serve as the central interface between AI and humans** especially as models become more aligned with humans — as we have seen in Section 6.2, more advanced models better implement user requirements. Here, we further discuss key skills humans and LLMs need to develop to prepare for a ROPE future. While our paper focused on a ROPE training system, our insights can also be useful for designing better prompting support tools.

What should humans be equipped with for ROPE? End-to-end prompting skills. To ensure the success of ROPE, multiple skills need to be cultivated among humans. Our training successfully helped users develop the ability to *extract requirements from given examples and express them completely and correctly*, but our study also revealed more opportunities in requirement iteration and testing skills of an end-to-end prompting procedure. Specifically, users need to identify examples to compare LLM outputs and adjust the requirements accordingly.

For example, users should *create diverse and representative examples to identify and test requirements*, particularly for ambiguous or open-ended tasks involving edge cases or complex interactions. In our study, many users, especially those unfamiliar with games like Connect4, missed the testing scenarios like a tie game, which led them to omit the requirement ```Display a ``Tie Game!`` message if no player wins``` in their prompt. This is consistent to findings in debugging training [39] and other end-user prompt engineering work [80], as novices often create too few test examples. Future studies should support users to generate more nuanced testing examples, supporting comprehensive hypothesis testing in prompt engineering [2].

Additionally, users should *iterate requirement granularity based on task “LLM-hardness.”* This involves observing LLM failures and refining requirements, essentially developing a theory of mind for LLMs [68]. Some of our participants were already thinking about “what is obvious enough for ChatGPT to already understand/perform well on” (S28) and adjust their prompt accordingly (although not necessarily correctly). In our training, we introduced the concept of requirement specificity in Tetris, where users generate requirements in two levels of granularity (i.e., main steps and details, Section 4.3). However, some tasks required even greater specificity, such as generating non-standard Connect4 gameplay elements ```Cross out winning cells```. This ability to adjust specificity in requirements can be influenced by domain expertise, for example, expert developers can iterate prompts with more accurate details to successfully generate code while students often struggle [47]. Future studies should offer adaptive support for adjusting requirement granularity based on task complexity and user expertise, such as via interactive dialog and automated diagnosis.

We believe that *improving humans’ requirement specification skills will remain important* despite rapid model advancements, since it is grounded in a fundamental challenge in human-LLM interaction: LLMs need access to user intent, but specific intent may not be available until users articulate it. As mentioned before, requirements might be ambiguous or entirely missing. The former might be mitigated by more capable models (or prompting support tools) that are better at e.g., making personalized guesses on user needs, asking clarification questions, or providing requirement-specific feedback [46, 81]. However, these can only be complementary to humans’ own requirement elicitation — a model that always asks for clarification on vague instructions can quickly become tedious and overwhelming to interact with. More importantly, missing requirements will remain a challenge, especially in *LLM-hard tasks*. When humans actually *need to deviate the model’s behavior but forget to articulate it*, models may already have a strong default behavior that prevents them from noticing anything missing. As a result, missing requirements risk becoming *unknown unknowns* [31, 58] to LLMs, and models may default to “common sense” completions that conflict with users’ actual needs (e.g., defaulting to circles instead of X and O’s in customized Connect4 games). Given decades of research in design and software engineering show that requirement elicitation is non-trivial for humans (Section 2.3), we anticipate ROPE training will remain essential, even when LLMs and optimizers become more powerful.

What should optimizers and LLMs be capable of for ROPE? Support complementary task delegation. For a successful ROPE future, optimizers and LLMs must develop complementary capabilities to support human skills.

First, as humans improve on articulating good requirements, optimizers should *test different “implementations” of requirements*. This includes varied wordings, alternative expressions (e.g., zero-shot descriptions or few-shot examples), and the order, hierarchy, or strictness of requirement compositions. Optimizers should also dynamically

group the same requirements presented in multiple formats to reveal which (combinations of) requirements are missing or ambiguous.

Second, optimizers can also *assist test data synthesis*. More advanced optimizers need to iterate prompts on validation datasets and automatic objective functions, similar to machine learning. Insights from software engineering test automation or NLP approaches to extract or synthesize test cases from natural language requirements could be helpful [21, 74, 75, 77].

Finally, LLMs or optimizers should *explicitly communicate their interpretations and reveal biases in their implementation*. As we observed in our study, optimizers can misinterpret requirements or add unnecessary details (e.g., “interactive game” can lead to unwanted features like Flask integration). LLMs should explicitly communicate their understandings of user intent, such as translating ambiguous natural language requirements into more formal specifications [33, 34], asking clarifying questions [10, 30, 54], or extracting and refining generated criteria [26, 69, 78]. This is not a perfect solution, as models’ translations may still reflect their biases and models may miss to ask about *LLM-hard* requirements. Nonetheless, to correctly implement requirements in LLM outputs when what is *LLM-hard* keeps evolving, humans need to learn when and how to specify requirements with more granularity, and models need to keep communicating what requirements might be *LLM-hard* or *LLM-conflict*.

8 CONCLUSION

In this work, we advocate for focusing prompt engineering effort on human-centered tasks, ensuring users include all necessary requirements in their prompt to achieve goals. We introduce the Requirement-Oriented Prompt Engineering (ROPE) paradigm, and design training materials where prompting novices practice requirement articulation on complex prompting tasks. We also develop aligned assessment metrics capturing both the intrinsic quality of user prompts in terms of requirement quality, as well as the extrinsic quality of prompt effectiveness in achieving intended outcomes. By providing targeted feedback on requirements, our ROPE system helps users produce higher-quality requirements and prompts more effectively than traditional prompt engineering training. As we look to the future, it is clear that the demand for LLM-based applications will only grow. As LLMs become more integrated into complex task-solving for more users, the ability to clearly articulate requirements will be key to effectively guiding LLMs. We believe that users should be equipped with the foundational skills to prepare for the ROPE future, and the right training will empower users to harness the full potential of LLMs.

ACKNOWLEDGMENTS

Thanks to all the participants for the pilot, interview, and user study in this work. Thanks to Kelly Rivers, Michael Taylor, Michael Hilton, Michael Xieyang Liu, Xinran Zhao, Lauren Sands, Austin Schick, David Kosbie, and Daniel Anderson for all the insights, advice, and help. Thanks to the CMU HCII faculty and Ken’s lab for feedback. Thanks to the National Science Foundation (award CNS-2213791, 2414915) and gift fund from Google for support of this work. Thanks to the OpenAI research credit program and Amazon AI research gift fund. Thanks to all reviewers of this work.

REFERENCES

- [1] Amira A Alshazly, Ahmed M Elfataty, and Mohamed S Abougabal. 2014. Detecting defects in software requirements specification. *Alex. Eng. J.* 53, 3 (Sept. 2014), 513–527. doi:10.1016/j.aej.2014.06.001
- [2] Ian Arawjo, Chelse Swoopes, Priyan Vaithilingam, Martin Wattenberg, and Elena L Glassman. 2024. ChainForge: A visual toolkit for prompt engineering and LLM hypothesis testing. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–18. doi:10.1145/3613904.3642016
- [3] Anne Arzberger, Maria Luce Lupetti, and Elisa Giaccardi. 2024. Reflexive data curation: Opportunities and challenges for embracing uncertainty in human-AI collaboration. *ACM Trans. Comput. Hum. Interact.* 31, 6 (Aug. 2024), 1–33. doi:10.1145/3689042

- [4] Hannah Babe, Sydney Nguyen, Yangtian Zi, Arjun Guha, Molly Feldman, and Carolyn Anderson. 2024. StudentEval: A Benchmark of Student-Written Prompts for Large Language Models of Code. In *Findings of the Association for Computational Linguistics: ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 8452–8474. doi:10.18653/v1/2024.findings-acl.501
- [5] Stephen Bach, Victor Sanh, Zheng Xin Yong, Albert Webson, Colin Raffel, Nihal V Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Evry, Zaid Alyafei, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-david, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Fries, Maged Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Dragomir Radev, Mike Tian-Jian Jiang, and Alexander Rush. 2022. PromptSource: An integrated development environment and repository for natural language prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Vol. abs/2202.01279. Association for Computational Linguistics, Stroudsburg, PA, USA, 93–104. doi:10.18653/v1/2022.acl-demo.9
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., USA, 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf
- [7] Sondos Mahmoud Bsharat, Aidar Myrzakhan, and Zhiqiang Shen. 2023. Principled instructions are all you need for questioning llama-1/2, gpt-3.5/4. *arXiv preprint arXiv:2312.16171* (2023).
- [8] Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff’s alpha agreement measure. <https://github.com/pln-fing-udelar/fast-krippendorff>.
- [9] Harrison Chase. 2022. *LangChain*. LangChain. <https://github.com/langchain-ai/langchain>
- [10] John Chen, Xi Lu, Yuzhou Du, Michael Rejtig, Ruth Bagley, Mike Horn, and Uri Wilensky. 2024. Learning agent-based modeling with LLM companions: Experiences of novices and experts using ChatGPT & NetLogo chat. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 13. ACM, New York, NY, USA, 1–18. doi:10.1145/3613904.3642377
- [11] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph Gonzalez, and Ion Stoica. 2024. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML ’24)*. JMLR.org, USA, 30. doi:10.48550/arXiv.2403.04132
- [12] Bhavya Chopra, Ananya Singha, Anna Fariha, Sumit Gulwani, Chris Parnin, Ashish Tiwari, and Austin Z Henley. 2023. Conversational Challenges in AI-Powered Data Science: Obstacles, Needs, and Design Opportunities. *arXiv [cs.HC]* (Oct. 2023). arXiv:2310.16164 [cs.HC] <http://arxiv.org/abs/2310.16164>
- [13] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) (NIPS’17). Curran Associates Inc., Red Hook, NY, USA, 4302–4310.
- [14] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, R Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z F Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, A Liu, Bing Xue, Bing-Li Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, C Deng, Chenyu Zhang, C Ruan, Damai Dai, Deli Chen, Dong-Li Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J Cai, J Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, K Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, M Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R J Chen, R L Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S S Li, Shuang Zhou, Shao-Kang Wu, Tao Yun, Tian Pei, T Sun, T Wang, Wangding Zeng, Wanbiao Zhao, Wen Liu, W Liang, W Gao, Wen-Xia Yu, Wentao Zhang, W L Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, X Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X Q Li, Xiangyu Jin, Xi-Cheng Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y K Li, Y Q Wang, Y X Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yi Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yu-Jing Zou, Yujia He, Yunfan Xiong, Yu-Wei Luo, Yu-Mei You, Yuxuan Liu, Yuyang Zhou, Y X Zhu, Yanping Huang, Yao Li, Yi Zheng, Yuchen Zhu, Yunxiang Ma, Ying Tang, Yukun Zha, Yuting Yan, Z Ren, Z Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhen-Guo Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zi-An Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. (Jan. 2025). arXiv:2501.12948 <http://arxiv.org/abs/2501.12948>

- [15] Paul Denny, Juho Leinonen, James Prather, Andrew Luxton-Reilly, Thezyrie Amarouche, Brett A Becker, and Brent N Reeves. 2024. Prompt Problems: A New Programming Exercise for the Generative AI Era. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1 (SIGCSE 2024)*. Association for Computing Machinery, New York, NY, USA, 296–302. doi:10.1145/3626252.3630909
- [16] H Dubberly. 2004. *How do you design*. Dubberly Design Office, USA.
- [17] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [18] Anna Eckerdal, Robert McCartney, Jan Erik Moström, Mark Ratcliffe, and Carol Zander. 2006. Can graduating students design software systems? *SIGCSE Bull.* 38, 1 (March 2006), 403–407. doi:10.1145/1124706.1121468
- [19] Molly Q Feldman and Carolyn Jane Anderson. 2024. Non-expert programmers in the generative AI future. In *Proceedings of the 3rd Annual Meeting of the Symposium on Human-Computer Interaction for Work (CHIWORK '24)*. ACM, New York, NY, USA, 1–19. doi:10.1145/3663384.3663393
- [20] Li Feng, Ryan Yen, Yuzhe You, Mingming Fan, Jian Zhao, and Zhicong Lu. 2024. CoPrompt: Supporting prompt sharing and referring in collaborative natural language programming. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–21. doi:10.1145/3613904.3642212
- [21] Vahid Garousi, Sara Bauer, and Michael Felderer. 2020. NLP-assisted software testing: A systematic mapping of the literature. *Inf. Softw. Technol.* 126, 106321 (Oct. 2020), 106321. doi:10.1016/j.infsof.2020.106321
- [22] Ellen Jiang, Edwin Toh, Alejandra Molina, Kristen Olson, Claire Kayacik, Aaron Donsbach, Carrie J Cai, and Michael Terry. 2022. Discovering the Syntax and Strategies of Natural Language Programming with Generative Language Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 386, 19 pages. doi:10.1145/3491102.3501870
- [23] Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 4667–4688. doi:10.18653/v1/2024.acl-long.257
- [24] Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics* 8 (2020), 423–438.
- [25] Omar Khattab, Arnab Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. *arXiv preprint arXiv:2310.03714* (2023).
- [26] Tae Soo Kim, Yoonjoo Lee, Jamin Shin, Young-Ho Kim, and Juho Kim. 2024. EvalLM: Interactive evaluation of large language model prompts on user-defined criteria. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 35. ACM, New York, NY, USA, 1–21. doi:10.1145/3613904.3642216
- [27] Amy J Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrance, Henry Lieberman, Brad Myers, Mary Beth Rosson, Gregg Rothermel, Mary Shaw, and Susan Wiedenbeck. 2011. The state of the art in end-user software engineering. *ACM Comput. Surv.* 43, 3 (April 2011), 1–44. doi:10.1145/1922649.1922658
- [28] Kenneth R Koedinger, Julie L Booth, and David Klahr. 2013. Instructional Complexity and the Science to Constrain It. *Science* 342, 6161 (2013), 935–937. doi:10.1126/science.1238056 arXiv:https://www.science.org/doi/pdf/10.1126/science.1238056
- [29] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J. Chiropr. Med.* 15, 2 (June 2016), 155–163. doi:10.1016/j.jcm.2016.02.012
- [30] Shuvendu K Lahiri, Aaditya Naik, Georgios Sakkas, Piali Choudhury, Curtis von Veh, Madanlal Musuvathi, Jeevana Priya Inala, Chenglong Wang, and Jianfeng Gao. 2022. Interactive code generation via test-driven user-intent formalization. *ArXiv* (2022). doi:10.48550/ARXIV.2208.05950 arXiv:2208.05950
- [31] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Eric Horvitz. 2017. Identifying unknown unknowns in the open world: representations and policies for guided exploration. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence* (San Francisco, California, USA) (AAAI'17). AAAI Press, 2124–2132.
- [32] A van Lamsweerde. 2009. *Requirements engineering: from system goals to UML models to software specifications*. John Wiley & Sons, Ltd, USA.
- [33] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable generation of scaffolding prompts through human-AI hybrid knowledge graph annotation. In *CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, NY, USA, 1–18. doi:10.1145/3491102.3502087
- [34] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. “What It Wants Me To Say”: Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23, Article 598)*. Association for Computing Machinery, New York, NY, USA, 1–31. doi:10.1145/3544548.3580817

- [35] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Benjamin Zorn, Jack Williams, Neil Toronto, and Andrew D. Gordon. 2023. “What It Wants Me To Say”: Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 598, 31 pages. doi:10.1145/3544548.3580817
- [36] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.* 55, 9 (Sept. 2023), 1–35. doi:10.1145/3560815 arXiv:2107.13586 [cs.CL]
- [37] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *CHI Conference on Human Factors in Computing Systems* (CHI '22). ACM, New York, NY, USA, 1–23. doi:10.1145/3491102.3501825
- [38] Ryan Louie, Ananjan Nandi, William Fang, Cheng Chang, Emma Brunskill, and Diyi Yang. 2024. Roleplay-doh: Enabling Domain-Experts to Create LLM-simulated Patients via Eliciting and Adhering to Principles. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 10570–10603. doi:10.18653/v1/2024.emnlp-main.591
- [39] Qianou Ma, Hua Shen, Kenneth Koedinger, and Sherry Tongshuang Wu. 2024. How to teach programming in the AI era? Using LLMs as a teachable agent for debugging. In *25th International Conference on Artificial Intelligence in Education (AIED) (Lecture notes in computer science)*. Springer Nature Switzerland, Cham, 265–279. doi:10.1007/978-3-031-64302-6_19
- [40] Atefeh Mahdavi Goloujeh, Anne Sullivan, and Brian Magerko. 2024. Is it AI or is it me? Understanding users’ prompt journey with text-to-image generative AI tools. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–13. doi:10.1145/3613904.3642861
- [41] Ggaliwango Marvin, Nakayiza Hellen, Daudi Jjingo, and Joyce Nakatumba-Nabende. 2024. Prompt Engineering in Large Language Models. In *Algorithms for Intelligent Systems*. Springer Nature Singapore, Singapore, 387–402. doi:10.1007/978-981-99-7962-2_30
- [42] Janet McDonnell and Peter Lloyd. 2014. Beyond specification: A study of architect and client interaction. *Des. Stud.* 35, 4 (July 2014), 327–352. doi:10.1016/j.destud.2014.01.003
- [43] Bertalan Meskó. 2023. Prompt engineering as an important emerging skill for medical professionals: Tutorial. *J. Med. Internet Res.* 25, 1 (Oct. 2023), e50638. doi:10.2196/50638
- [44] Lynette I Millett, Martyn Thomas, and Daniel Jackson. 2007. *Software for dependable systems: Sufficient evidence?* National Academies Press, Washington, DC, USA.
- [45] Lloyd Montgomery, Davide Fucci, Abir Bouraffa, Lisa Scholz, and Walid Maalej. 2022. Empirical research on requirements quality: a systematic mapping study. *Requir Eng* 27, 2 (Feb. 2022), 183–209. doi:10.1007/s00766-021-00367-z
- [46] Fangwen Mu, Lin Shi, Song Wang, Zhuohao Yu, Binqun Zhang, ChenXue Wang, Shichao Liu, and Qing Wang. 2024. ClarifyGPT: A Framework for Enhancing LLM-Based Code Generation via Requirements Clarification. *Proc. ACM Softw. Eng.* 1, FSE, Article 103 (July 2024), 23 pages. doi:10.1145/3660810
- [47] Daye Nam, Andrew Macvean, Vincent Hellendoorn, Bogdan Vasilescu, and Brad Myers. 2024. Using an LLM to Help With Code Understanding. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering (ICSE '24, Article 97)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3597503.3639187
- [48] Vicente Lustosa Neto, Roberta Coelho, Larissa Leite, Dalton S Guerrero, and Andrea P Mendonca. 2013. POPT: A Problem-Oriented Programming and Testing approach for novice students. In *2013 35th International Conference on Software Engineering (ICSE)*. IEEE, USA, 1099–1108. doi:10.1109/icse.2013.6606660
- [49] Sydney Nguyen, Hannah Mclean Babe, Yangtian Zi, Arjun Guha, Carolyn Jane Anderson, and Molly Q Feldman. 2024. How Beginning Programmers and Code LLMs (Mis)read Each Other. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (CHI '24, Article 651). Association for Computing Machinery, New York, NY, USA, 1–26. doi:10.1145/3613904.3642706
- [50] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems* 35 (2022), 27730–27744.
- [51] Shuyin Ouyang, Jie M. Zhang, Mark Harman, and Meng Wang. 2025. An Empirical Study of the Non-Determinism of ChatGPT in Code Generation. *ACM Trans. Softw. Eng. Methodol.* 34, 2, Article 42 (Jan. 2025), 28 pages. doi:10.1145/3697010
- [52] Savvas Petridis, Benjamin D Wedin, James Wexler, Mahima Pushkarna, Aaron Donsbach, Nitesh Goyal, Carrie J Cai, and Michael Terry. 2024. ConstitutionMaker: Interactively Critiquing Large Language Models by Converting Feedback into Principles. In *Proceedings of the 29th International Conference on Intelligent User Interfaces* (Greenville, SC, USA) (IUI '24). Association for Computing Machinery, New York, NY, USA, 853–868. doi:10.1145/3640543.3645144
- [53] James Prather, Brent N Reeves, Paul Denny, Brett A Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. 2024. “it’s weird that it knows what I want”: Usability and interactions with Copilot for novice programmers. *ACM Trans. Comput. Hum. Interact.* 31, 1 (Feb. 2024), 1–31. doi:10.1145/3617367
- [54] Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Tell me more! Towards implicit user intention understanding of language model driven agents. In *Proceedings of*

- the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 1088–1113. doi:10.18653/v1/2024.acl-long.61
- [55] Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, Vol. Findings of the Association for Computational Linguistics: ACL 2024. Association for Computational Linguistics, Stroudsburg, PA, USA, 13025–13048. doi:10.18653/v1/2024.findings-acl.772
- [56] Alex Radermacher, Gursimran Walia, and Dean Knudson. 2014. Investigating the skill gap between graduating students and industry expectations. In *Companion Proceedings of the 36th International Conference on Software Engineering (ICSE Companion 2014)*. Association for Computing Machinery, New York, NY, USA, 291–300. doi:10.1145/2591062.2591159
- [57] Jekaterina Rogaten, Bart Rienties, Rhona Sharpe, Simon Cross, Denise Whitelock, Simon Lygo-Baker, and Allison Littlejohn. 2019. Reviewing affective, behavioural and cognitive learning gains in higher education. *Assessment & Evaluation in Higher Education* 44, 3 (April 2019), 321–337. doi:10.1080/02602938.2018.1504277
- [58] Donald Rumsfeld. 2011. *Known and unknown: a memoir*. Penguin, USA.
- [59] Vagner Figueredo De Santana. 2024. Challenges and opportunities for responsible prompting. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–4. doi:10.1145/3613905.3636268
- [60] Douglas C Schmidt, Jesse Spencer-Smith, Quichen Fu, and Jules White. 2024. Towards a catalog of prompt patterns to enhance the discipline of prompt engineering. *ACM SIGAda Ada Lett.* 43, 2 (June 2024), 43–51. doi:10.1145/3672359.3672364
- [61] Patrick Schober, Christa Boer, and Lothar A Schwarte. 2018. Correlation coefficients: appropriate use and interpretation. *Anesthesia & analgesia* 126, 5 (2018), 1763–1768.
- [62] Shreya Shankar, Haotian Li, Parth Asawa, Madelon Hulsebos, Yiming Lin, J D Zamfirescu-Pereira, Harrison Chase, Will Fu-Hinthorn, Aditya G Parameswaran, and Eugene Wu. 2024. SPADE: Synthesizing Data Quality Assertions for Large Language Model Pipelines. *Proceedings VLDB Endowment* 17, 12 (Jan. 2024), 4173–4186. doi:10.14778/3685800.3685835 arXiv:2401.03038 [cs.DB]
- [63] Shreya Shankar, J D Zamfirescu-Pereira, Bjoern Hartmann, Aditya Parameswaran, and Ian Arawjo. 2024. Who validates the validators? Aligning LLM-assisted evaluation of LLM outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, Vol. 4. ACM, New York, NY, USA, 1–14. doi:10.1145/3654777.3676450
- [64] Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 4222–4235. doi:10.18653/v1/2020.emnlp-main.346
- [65] Jiao Sun, Q Vera Liao, Michael Muller, Mayank Agarwal, Stephanie Houde, Kartik Talamadupula, and Justin D Weisz. 2022. Investigating explainability of generative AI for code through scenario-based design. In *27th International Conference on Intelligent User Interfaces*. ACM, New York, NY, USA, 212–228. doi:10.1145/3490099.3511119
- [66] Gursimran Singh Walia and Jeffrey C Carver. 2009. A systematic literature review to identify and classify software requirement errors. *Information and Software Technology* 51, 7 (July 2009), 1087–1109. doi:10.1016/j.infsof.2009.01.004
- [67] Min Wang and Yong Zeng. 2009. Asking the right questions to elicit product requirements. *Int. J. Comput. Integr. Manuf.* 22, 4 (April 2009), 283–298. doi:10.1080/09511920802232902
- [68] Qiaosi Wang, Sarah Walsh, Mei Si, Jeffrey Kephart, Justin D. Weisz, and Ashok K. Goel. 2024. Theory of Mind in Human-AI Interaction. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 493, 6 pages. doi:10.1145/3613905.3636308
- [69] Zhijie Wang, Yuheng Huang, Da Song, Lei Ma, and Tianyi Zhang. 2024. PromptCharm: Text-to-image generation through multi-modal prompting and refinement. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Vol. 29. ACM, New York, NY, USA, 1–21. doi:10.1145/3613904.3642803
- [70] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [71] Grant P Wiggins and Jay McTighe. 2005. *Understanding by Design*. ASCD, USA. <https://play.google.com/store/books/details?id=N2EfKlyUN4QC>
- [72] Ben Wodecki. 2024. Devin: AI Software Engineer that Codes Entire Projects from Single Prompt. <https://aibusiness.com/automation/ai-software-engineer-devin-codes-entire-projects-from-single-prompt>. Accessed: 2024-9-2.
- [73] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22, Article 385)*. Association for Computing Machinery, New York, NY, USA, 1–22. doi:10.1145/3491102.3517582
- [74] Chenyang Yang, Yining Hong, Grace Lewis, Tongshuang Wu, and Christian Kästner. 2024. What Is Wrong with My Model? Identifying Systematic Problems with Semantic Data Slicing. In *Proceedings of the 39th IEEE/ACM International Conference on Automated Software Engineering* (Sacramento, CA, USA) (ASE '24). Association for Computing Machinery, New York, NY, USA, 306–318. doi:10.1145/3691620.3695033

- [75] Chenyang Yang, Rishabh Rustogi, Rachel Brower-Sinning, Grace Lewis, Christian Kaestner, and Tongshuang Wu. 2023. Beyond Testers' Biases: Guiding Model Testing with Knowledge Bases using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13504–13519. doi:10.18653/v1/2023.findings-emnlp.901
- [76] John Yang, Carlos Jimenez, Alexander Wettig, Kilian Lieret, Shunyu Yao, Karthik Narasimhan, and Ofir Press. 2024. Swe-agent: Agent-computer interfaces enable automated software engineering. *Advances in Neural Information Processing Systems* 37 (2024), 50528–50652.
- [77] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: a tale of diversity and bias. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (New Orleans, LA, USA) (NIPS '23)*. Curran Associates Inc., Red Hook, NY, USA, Article 2433, 51 pages.
- [78] Weizhe Yuan, Pengfei Liu, and Matthias Gallé. 2024. LLMCrit: Teaching large language models to use criteria. In *Findings of the Association for Computational Linguistics ACL 2024*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Stroudsburg, PA, USA, 7929–7960. doi:10.18653/v1/2024.findings-acl.472
- [79] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. 2024. TextGrad: Automatic “Differentiation” via Text. *arXiv [cs.CL]* (June 2024). arXiv:2406.07496 [cs.CL] <http://arxiv.org/abs/2406.07496>
- [80] J D Zamfirescu-Pereira, Richmond Y Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23, Article 437)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3544548.3581388
- [81] Michael JQ Zhang and Eunsol Choi. 2023. Clarify when necessary: Resolving ambiguity through interaction with lms. *arXiv preprint arXiv:2311.09469* (2023).
- [82] Zejun Zhang, Li Zhang, Xin Yuan, Anlan Zhang, Mengwei Xu, and Feng Qian. 2024. A first look at gpt apps: Landscape and vulnerability. *arXiv preprint arXiv:2402.15105* (2024).
- [83] Liping Zhao, Waad Alhoshan, Alessio Ferrari, Keletso J Letsholo, Muideen A Ajagbe, Erol-Valeriu Chioasca, and Riza T Batista-Navarro. 2022. Natural Language Processing for Requirements Engineering: A systematic mapping study. *ACM Comput. Surv.* 54, 3 (April 2022), 1–41. doi:10.1145/3444689
- [84] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and Ed Chi. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv [cs.AI]* (May 2022). arXiv:2205.10625 [cs.AI] <http://arxiv.org/abs/2205.10625>
- [85] Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for Large Language Models. *arXiv [cs.CL]* (Nov. 2023). arXiv:2311.07911 [cs.CL] <https://github.com/google-research/>

Received 8 December 2024; revised 4 April 2025; accepted 15 April 2025