

Prompts Are Programs Too! Understanding How Developers Build Software Containing Prompts

JENNY T. LIANG, Carnegie Mellon University, USA

MELISSA LIN*, Carnegie Mellon University, USA

NIKITHA RAO*, Carnegie Mellon University, USA

BRAD MYERS, Carnegie Mellon University, USA

Generative pre-trained models power intelligent software features used by millions of users controlled by developer-written natural language prompts. Despite the impact of prompt-powered software, little is known about its development process and its relationship to programming. In this work, we argue that some prompts are programs and that the development of prompts is a distinct phenomenon in programming known as “*prompt programming*”. We develop an understanding of prompt programming using Straussian grounded theory through interviews with 20 developers engaged in prompt development across a variety of contexts, models, domains, and prompt structures. We contribute 15 observations to form a preliminary understanding of current prompt programming practices. For example, rather than building mental models of code, prompt programmers develop mental models of the foundation model (FM)’s behavior on the prompt by interacting with the FM. While prior research shows that experts have well-formed mental models, we find that prompt programmers who have developed dozens of prompts still struggle to develop reliable mental models. Our observations show that prompt programming differs from traditional software development, motivating the creation of prompt programming tools and providing implications for software engineering stakeholders.

CCS Concepts: • Software and its engineering → Software development methods; • Human-centered computing → Empirical studies in HCI; • Computing methodologies → Artificial intelligence.

Additional Key Words and Phrases: Prompt programming, prompt engineering, Straussian grounded theory

ACM Reference Format:

Jenny T. Liang, Melissa Lin, Nikitha Rao, and Brad Myers. 2025. Prompts Are Programs Too! Understanding How Developers Build Software Containing Prompts. *Proc. ACM Softw. Eng.* 2, FSE, Article FSE072 (July 2025), 24 pages. <https://doi.org/10.1145/3729342>

1 Introduction

“I suspect that machines to be programmed in our native tongues—be it Dutch, English, American, French, German, or Swahili—are as damned difficult to make as they would be to use.”
—Edsger W. Dijkstra (1979)

Generative pre-trained models (e.g., GPT-4 [9], Dall-E [15])—also known as foundation models (FMs)—have changed how programmers build software. AI programming assistants that generate code (e.g., GitHub Copilot [20]) have improved developer productivity [63, 85, 86] by helping

*Authors contributed equally to this research.

Authors' Contact Information: Jenny T. Liang, Carnegie Mellon University, Pittsburgh, USA, jtliang@cs.cmu.edu; Melissa Lin, Carnegie Mellon University, Pittsburgh, USA, mylin@andrew.cmu.edu; Nikitha Rao, Carnegie Mellon University, Pittsburgh, USA, nikitharao@cmu.edu; Brad Myers, Carnegie Mellon University, Pittsburgh, USA, bam@cs.cmu.edu.

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). Publication rights licensed to ACM.

ACM 2994-970X/2025/7-ARTFSE072

<https://doi.org/10.1145/3729342>

developers write significant portions of code, learn new APIs and programming languages, and write tests [47]. Recently, instruction-tuned FMs [59] like ChatGPT [2] have expanded the assistance with software development tasks by writing natural language prompts. This includes resolving code issues, developing new features, refactoring, and information seeking [18, 32].

The rising prominence of prompts has ushered in “*prompt engineering*”, whereby FM users repeatedly write and revise natural language prompts. These prompts enable new intelligent features when integrated in popular software applications [60], such as Google Search [4] and Microsoft Office [3], reaching millions of users [1]. As of January 2024, engineered prompts have also powered over 3 million custom versions of ChatGPT for specific tasks, known as GPTs [5].

Yet, little is known about prompt engineering and its relationship to programming. Relevant works include Dolata et al. [26]’s study of 52 freelance developers building solutions based on generative AI and Parnin et al. [60]’s study of 26 professional software developers integrating generative AI into products. While these studies offer insight on prompt development, they are constrained to freelance and professional software development, limiting the generalizability of the results to other programming contexts. We address these gaps by following a more systematic and rigorous qualitative methodology—Straussian grounded theory [22]—to study a diverse sample of programmers to understand the process of developing prompts embedded in software applications.

In this work, we argue that some prompts function as programs, a phenomenon we call *prompt programming* [35]. In prompt programming, prompts are authored at design-time and executed at runtime with variable user inputs, but are written in natural language instead of a programming language. Prompt programming is a specific form of prompt engineering—broadly defined as iterating on prompts to improve FM outputs [84]—where the same prompt handles variable inputs given at runtime. For example, a prompt program on a teleconferencing platform may take a meeting transcript and attendee name and return bulleted todos and deadlines for that individual.

In prompt programming, natural language is a programming abstraction, but unlike formalisms such as DSLs or machine code, it introduces ambiguity and reduces precision [23]. While the idea of programming in natural language dates as early as the 1970s, where Dijkstra discussed a vision of instructing machines in “our native tongues” [23], it has become feasible and widespread with FMs due to instruction-tuning [59] that teaches models to follow natural language instructions.

We argue that prompt programming is a phenomenon that warrants its own study, as the sociotechnical circumstance of the programming task influences the nature of the programming. For example, the programming process of end-users who write software for personal use (i.e., end-user programming [42]) is different from that of data scientists who write software to explore different possibilities in code (i.e., exploratory programming [40, 68]), and both differ from a professional programmer who focuses on a small component within a larger system. The programming task also varies with the developer’s tools (e.g., visual programming languages vs. Jupyter notebooks), focus on the formality of the process (e.g., accomplishing a goal vs. exploring ideas), and key challenges (e.g., finding the right abstractions vs. managing code versions) [40, 42]. This creates disparate experiences and necessitates the development of support tools unique to each context. Thus, studying prompt programs could inform the development of tools to support prompt programmers.

We form a preliminary understanding of current prompt programming practices using Straussian grounded theory [22], a qualitative research methodology that develops a novel exploratory explanation of a domain known as a “grounded theory”. To this end, we interview 20 developers building software that is based on prompts to understand: *How do programmers develop programs that incorporate natural language prompts?* First, we consider the definitions of a *prompt* and *program*. We define a *prompt* as:

A natural language query to a FM.

Study Findings: Barriers in Prompt Programming

Programmer (Section 4.1)

1. Programmers must develop a mental model about the FM's behavior on the prompt.
2. Programmers' mental models are not reliable.
3. Programmers use external knowledge sources and prior experience to build their mental model.

Foundation Model (Section 4.2)

4. Each FM has its own set of qualities and capabilities.

Prompt (Section 4.3)

5. Minute details in the prompt matter.
6. Prompts are finicky and fragile.

Requirements (Section 5.1)

7. Assumptions of the requirements must be explicitly stated.
8. Requirements can evolve as the capabilities of the FM are discovered.

Design (Section 5.2)

9. Developers make decisions about how prompt programs should be composed and decomposed.
10. Dependent prompts are tightly coupled.

Implementation (Section 5.3)

11. Prompt programming is rapid and unsystematic.

Debugging (Section 5.4)

12. Fault localization is difficult.

Data Curation (Section 5.5)

13. Programmers need to find representative data for the task.

Evaluation (Section 5.6)

14. Evaluating prompt programs requires assessing qualitative constructs.
15. Testing occurs at different scopes.

Fig. 1. An overview of the 15 study findings on the barriers in prompt programming. These can be divided into four types of barriers: understanding FM behavior ()¹, dealing with stochasticity ()², programming in natural language ()³, and testing prompt program behavior ()⁴.

For the definition of a *program*, we consider Ko et al. [42]'s definition, which is “a collection of specifications that may take variable inputs, and that can be executed (or interpreted) by a device with computational capabilities.” We extend this to derive the definition of a *prompt program*:

A prompt that accepts variable inputs and could be interpreted by a FM to perform specified actions and/or generate output. This prompt is executed within a software application or code by a FM.

This definition includes developing prompts for applications like GPTs and chatbots. It also covers language agents, where large language models (LLM) receive arbitrary user requests and execute them in dynamic environments using external modules known as “tools” (e.g., modules for mathematical reasoning [67] and code interpreters [28]). However, this definition excludes cases where a developer converses with an FM once to achieve a task (e.g., prompt engineering with

ChatGPT to debug a specific error message). These prompts do not accept variable inputs, do not generalize beyond the specific case, and are not executed within a broader software application.

Through our investigation, we contribute 15 observations that identify barriers in prompt programming (see Figure 1). These form a preliminary understanding of current prompt programming practices. To aid discussion, we organize these into four types of barriers: understanding FM behavior (Q), dealing with stochasticity (B), programming in natural language (L), and testing prompt program behavior (C). While some results corroborate those from Dolata et al. [26]’s and parnin2023building’s studies (e.g., fault localization being difficult (B_{#12})), our results differ in notable ways. First, we find prompt programmers interact with FMs to develop mental models of the FM’s behavior on the prompt (Q_{#1}) and its unique qualities (Q_{#4}). Although the literature suggests that experts have well-formed mental models compared to novices [69], prompt programmers struggle to develop reliable mental models (Q_{#2}) even after writing dozens of prompts, each with many iterations. This contributes to a rapid and unsystematic development process (Q_{#11}). We also find that prompts can be composed and decomposed (L_{#9}) and testing occurs at different scopes (C_{#15}). While these observations exhibit similarities to software development, we also find differences compared to the process of developing traditional software, necessitating the creation of new tools and processes to support prompt programming. This has implications for software practitioners, academics, tool creators, and educators interested in generative AI in software engineering.

2 Related Work

We discuss related work on prompt engineering (Section 2.1) and software engineering for AI (SE4AI; see Section 2.2). Due to the rapidly evolving landscape of prompt engineering and FMs, this discussion will necessarily miss more recent developments.

2.1 Prompt Engineering

Prior literature has examined prompt engineering in a variety of contexts. This includes user studies on prompt engineering, prompt engineering tools, and retrospective interview studies on prompt engineering in software development. We discuss this in more detail below.

Previous work has conducted empirical user studies of prompt engineering. Some works have focused on the development of chatbots with end users using prompts [82–84]. In a user study of 10 people who used a tool to create chatbots through prompts, Zamfirescu-Pereira et al. [84] identified several challenges of prompt development, including opportunistic prompt design approaches, a lack of systematic testing, and writing prompts that were not generalized. Jiang et al. [35] evaluated a prompt development tool in 11 users in various roles, including designers, content strategists, and front-end developers. The problems participants faced included the prompt easily breaking and overfitting on examples, as well as having difficulty evaluating large amounts of text.

Other work has investigated the creation of prompting tools to assist with various prompting tasks. Prompt pattern catalogs can solve common problems encountered when conversing with an LLM [76], such as question refinement, as well as software engineering tasks [77], such as specification disambiguation. Other relevant tools include prompting interfaces, such as PromptAid [56], PromptMaker [35], and PromptIDE [70]. These tools included separate views to show the dataset, iterate on the prompt, track the performance of the prompt, and search for the prompts.

Most related to this study are retrospective interview studies that investigate prompt development in software engineering. Dolata et al. [26] conducted a study with 52 freelance developers on their experience with building prompt-powered software. Their interview focuses on the positive and negative experiences with generative AI, project uncertainties, and views on freelancing. The methodology involved recruiting from Upwork, conducting retrospective interviews, performing a thematic analysis, and comparing the findings with four meta-reviews of SE4AI literature. The

participants enumerated several challenges, including having difficulty identifying the source of incorrect responses, budget constraints, and unrealistic client expectations. Meanwhile, Parnin et al. [60] conducted an interview study with 26 professional programmers who developed product copilots. Interview topics included motivation for using AI in products, major tasks for building the generative AI application, prompt engineering, testing, tooling, challenges, learning-related skills, and concerns with AI. Their methodology involved recruiting participants from UserInterviews.com, conducting retrospective interviews and structured brainstorming sessions, and performing a thematic analysis. They found that the participants followed a general process of exploration, implementation, evaluation, and productization. This work noted several challenges, including the trial-and-error nature of prompt development and creating benchmarks.

This literature provides an initial understanding of the unique aspects of prompt programming. However, they are constrained to specific types of developers (i.e., end-users, professional programmers, and freelancers), limiting the generalizability of the results. Additionally, these works do not study the process of writing prompt programs with respect to software development activities. Our study builds upon these works by employing a more systematic and rigorous qualitative methodology, Straussian grounded theory [22], to study the process of writing prompt programs. We sample a broad range of prompt programmers along various axes, such as prompt structure, programming context, and role, as well as ask questions about the end-to-end development process.

2.2 Software Engineering for AI

Numerous works have studied the human aspects of building ML-enabled systems, also known as software engineering for AI. These works have documented the experiences of practitioners and their common challenges in this domain. We elaborate further below.

Building ML-enabled systems involves working with and wrangling nondeterministic, opaque neural models [30]. This is similar to prompt programming, as developers must also contend with non-transparent, stochastic FMs. Previous work describes the development of models as highly experimental [30, 57, 58, 73, 78]. The literature also stresses the importance of collecting high-quality data for the domain [12, 26, 30, 57, 58, 64] as well as the use of quantitative metrics, such as precision, recall, and F1 score, to measure model performance [58, 73].

However, the development of ML components presents unique challenges. Models can be difficult to debug due to their determinism [30], while *training-serving skew* can arise when training data fail to generalize to production [57, 58, 73]. To mitigate this, models are periodically re-trained [57]. In addition, expertise in a variety of domains, such as software engineering and data science, is distributed between roles [30, 58, 73]. In an interview study with 45 practitioners, Nahar et al. [58] found this introduced collaboration challenges between different roles, such as unclear model requirements, handling evolving data, and inadequate datasets. Finally, a lack of AI literacy can make requirements elicitation, communication, and collaboration with clients challenging [25, 58, 78].

This body of literature serves as a foundation for understanding prompt programming. We extend this body of work by comparing the building of ML-enabled systems to prompt programming and understanding what aspects of this prior literature apply to prompt programming.

3 Methodology

Since there is limited literature on prompt programming, we used a qualitative approach to explore the *process* of developing prompt-powered software. We used the grounded theory methodology, which allows researchers to develop a novel exploratory explanation for a domain.

There are three popular approaches to grounded theory: Glaserian/classical [31], Straussian [22], and constructivist [17] methodologies, which vary in their procedures and epistemology [21]. Our approach is based on Straussian grounded theory [22] (see Figure 2). This involves defining a

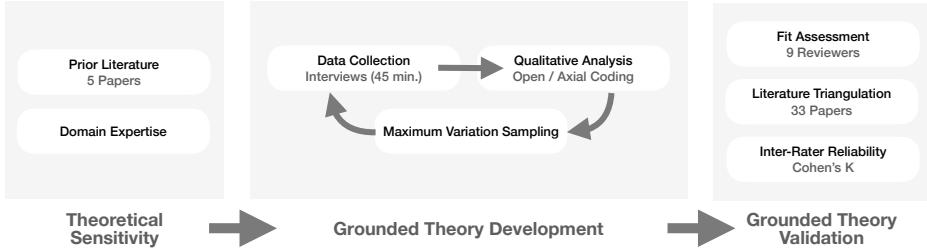


Fig. 2. An overview of the Straussian grounded theory methodology performed in the study.

research question; developing theoretical sensitivity to the phenomenon through prior literature and domain expertise; generating a grounded theory via diverse sampling and simultaneous data collection and analysis; and validating the grounded theory by triangulating with literature, performing an assessment of the fit of the grounded theory, and computing inter-rater reliability [21, 22]. In this section, we first discuss our grounded theory process (Section 3.1). We then describe our participants (Section 3.2) and interview protocol (Section 3.3), and close with a discussion of the limitations of our method (Section 3.4). We present our results in Section 4 and Section 5.

3.1 Grounded Theory Process

We describe our grounded theory process below. It contains five main stages: defining a research question, developing theoretical sensitivity, generating the grounded theory, triangulating with the literature, and validating the grounded theory. We describe this process in further detail below.

Defining a research question. Our process began with defining a research question. We study the phenomenon where developers write a prompt program using natural language prompts, rather than pure code: *How do programmers develop programs that incorporate natural language prompts?* Prompts can be used in a variety of contexts, such as interacting with ChatGPT in conversation [18], but are not programs. Thus, we apply the definition of a prompt program described in Section 1.

Developing theoretical sensitivity. Before the study, the investigator should develop an intuition for the phenomenon being studied, known as *theoretical sensitivity* [21]. This can be done through professional experience and understanding the prior literature so that researchers can extract insights from the data. However, the investigators should not be constrained by prior knowledge while developing the grounded theory [22]. We entered the study with some theoretical sensitivity since two authors are software engineering and AI researchers and have developed multiple prompt programs. Following prior work [55], we considered our background and opted for a lightweight overview of prompting literature. We considered foundational prompting papers in machine learning venues [16, 75], a notable survey paper on prompting techniques in natural language processing (NLP) [51] as well as two empirical studies on how people develop prompts [26, 84] from human-computer interaction (HCI) and software engineering venues. We considered this literature and the authors' existing domain expertise to develop the interview protocol (Section 3.3).

Generating the grounded theory. We conducted 45-minute interviews with a diverse set of 20 prompt programmers (see Table 1) over Zoom. The interviews were recorded and transcribed; the recordings were later deleted. Two authors were present for the first six interviews to become familiar with the data. The first author conducted the remaining interviews.

To develop the initial grounded theory, three authors first independently performed line-by-line coding on the first two interviews in separate codebooks to become more sensitized to the data.

Each code contained a description of the code as well as observations from the interviews. The authors then reconvened to merge the individual codebooks by identifying codes with similar concepts and merging them into a shared codebook. The remaining codes were then discussed and added or removed to the codebook by unanimous vote. The first author then coded the next interview, and the other authors reviewed the codes for agreement. We identified five instances of disagreement, which were discussed and resolved. The three authors then performed axial coding, grouping the emerging codes into preliminary categories on unanimous vote.

For the remaining interviews, the authors individually open coded the interview transcripts, noting new codes that emerged in the shared codebook. A memo was created for each interview to capture key insights and refine the emerging grounded theory. The authors met regularly to discuss the new codes and observations. New codes and observations were added to the shared codebook upon unanimous vote, and the categories were further refined as more data was collected.

As the grounded theory developed, we performed maximum variation sampling [71] to obtain a diverse set of participants that could challenge or extend the grounded theory. We recruited participants who met the definition of creating a prompt program (Section 1) by snowball sampling within the authors' social networks and recruiting in online open-source communities. Participants were recruited based on the types of models used (e.g., open-source models vs. closed-source models, vision language models vs. language models), application domain (e.g., robotics, education, productivity), organization size (e.g., large technology company vs. startup), programming context (e.g., academia, industry, freelance, open-source software), role, and/or prompt structure (e.g., a language agent that is able to use tools vs. single prompt vs. multi-prompt). We initially achieved theoretical saturation after interviewing 14 participants. We continued to recruit participants to ensure that our observations held across a variety of situations, including open-source software and startups. We stopped data collection after reaching 20 participants.

Triangulating with literature. While developing the grounded theory of prompt programming, we triangulated our findings with the prior literature. Since the relevant literature was not identifiable through keyword search, we followed previous work [58] in performing a focused, best-effort literature review by performing forward and/or backward snowballing on the initial set of five papers. The relevant papers were first identified based on titles related to building AI systems or prompt engineering. They were later verified by reading the manuscript. In the end, this literature spanned several communities, including NLP, HCI, programming languages, software engineering, and machine learning. Of the 502 papers considered, we identified 33 as relevant. See the supplemental materials for this set of papers [46]. For each paper, the first author identified text related to prompt programming and applied codes from the codebook, updating the codes as necessary to further refine the grounded theory.

Validating the grounded theory. To validate the grounded theory, we assess its *fit*. Straussian grounded theory suggests validating the grounded theory with professionals and study participants to see if they resonate with the findings [22]. To this end, we sent an email containing a summary of the grounded theory as well as a draft of Section 4, Section 5, and Table 1 to the 20 participants and one external researcher who studies prompt programming. We asked them to reflect on the fit of the grounded theory and provide feedback based on their experience. Nine people responded to this inquiry and indicated general agreement with the findings, some with wording changes.

Additionally, to assess the reliability of the coding, the first author randomly selected and open-coded two interviews at the category level. The labels were then removed, with the original highlighted text spans remaining intact. One author was assigned to each interview and applied the code categories to each highlighted text span. Based on this procedure, our inter-rater reliability was 0.89 and 0.81 based on Cohen's κ —indicating that both are almost in perfect agreement [44].

Table 1. An overview of the participants in the study. We report the participant's development context, number of years of programming experience, number of prompt programs developed prior to the interview, prompt structure, model type, application domain, and task. Prompt structures can be a single prompt, multiple chained prompts (multi-prompt), or a language agent (agent). Model types can be open source (`OS`) or closed source (`CS`) and can be a language model (`LM`) or a vision-language model (`VLM`).

ID	Context	Exp.	# Prompts	Structure	Model	App Domain	Task
P1	Academia	16 yrs	30	Single prompt	<code>OS LM</code>	Security	Classification
P2	Personal project	10 yrs	40	Single prompt	<code>CS LM</code>	Productivity	Text summarization (chatbot)
P3	Academia	8 yrs	30	Single prompt	<code>CS LM</code>	Research analysis	Classification
P4	Big tech (R&D)	8 yrs	10	Single prompt	<code>OS LM</code>	Software testing	Code generation
P5	Academia	14 yrs	35	Single prompt	<code>CS LM</code>	Education	Generation
P6	Academia	10 yrs	3	Multi-prompt	<code>CS LM</code>	Software testing	Code generation
P7	Big tech (Eng)	10 yrs	4	Single prompt	<code>CS LM</code>	Education	Generation
P8	Academia	10 yrs	20	Agent	<code>CS LM</code>	Software testing	Code generation
P9	Academia	14 yrs	20	Multi-prompt	<code>OS VLM</code>	Visual Q&A	Retrieval, Q&A
P10	Big tech (Eng)	7 yrs	10	Multi-prompt	<code>CS LM</code>	Education	Text editing
P11	Big tech (R&D)	8 yrs	5	Agent	<code>CS LM</code>	Productivity	Data generation
P12	Big tech (Eng)	14 yrs	100+	Multi-prompt	<code>CS LM</code>	Productivity	Code summarization
P13	Big tech (Eng)	20 yrs	6	Multi-prompt	<code>CS LM</code>	Productivity	Code generation
P14	Big tech (Eng)	5 yrs	3	Single prompt	<code>CS LM</code>	Productivity	Code generation
P15	Big tech (Eng)	5 yrs	3	Single prompt	<code>CS LM</code>	Career	Q&A (chatbot)
P16	OSS	8 yrs	—	Agent	<code>CS LM</code>	Software dev.	End-to-end workflow
P17	Startup (R&D)	6 yrs	5	Single prompt	<code>CS LM</code>	Robotics	Code generation
P18	Startup (Eng), OSS	3.5 yrs	15	Agent	<code>CS LM</code>	Software testing	Test generation
P19	Startup (Eng)	3 yrs	5	Agent	<code>CS LM</code>	Research analysis	End-to-end workflow
P20	Freelance	9 yrs	1	Single prompt	<code>CS LM</code>	Literature	Q&A (chatbot)

3.2 Participants

A summary of our participants is in Table 1. The inclusion criterion was developers who had previously created at least one prompt program, according to the definition presented in Section 1. Our participants were men ($N = 17$) and women ($N = 3$) from a wide range of technology roles, including Machine Learning Engineer, Senior Software Engineer, Ph.D. Student, and Principal Data and Applied Scientist. All participants had prior programming experience (median 8.5 years) and had written a prompt program before (median 10 such programs). Participants also regularly used foundation models: 10 participants reported using them more than once a day, five participants reported using them once daily, and five participants reported using them weekly.

3.3 Interview Protocol

Interviews were 45 minutes long and conducted over Zoom. The procedure was approved by our Institution's Review Board. Participants were compensated with a \$20 USD Amazon gift certificate.

To begin, the participant completed a demographic and background survey with questions such as gender, the number of years of programming experience, and number of prompt programs developed. We follow best practices in reporting gender in HCI [66]. The interviewer then presented the participant with the definition of a prompt program and asked the participant to recall the last time they wrote such a program. If the participant had access to the prompt, they retrieved the prompt and any associated history, and provided a brief overview of the prompt. 17 of the 20

Interview Questions

- Think of the most recent time you wrote such a prompt. Do you have one in mind?
- Do you currently have access to the prompt? If so, could you please pull up the prompt and any associated history (e.g., on ChatGPT/Bard UI or GitHub), if applicable?
- Did you experience any difficulties while using existing tools to develop a prompt?
- Did you refer to external information sources to develop the prompt? If so, which ones?
- Did you ever refer to a previous version of the prompt as you were developing a new version?
- Did you use any metrics or heuristics to determine when the prompt was successful?
- How did you identify what caused an incorrect output? Were you confident in your answer?
- Did your prompt change after deployment?

Fig. 3. A subset of the interview questions. The full protocol is in the supplemental materials [46]

participants accessed their prompts. P15, P17, and P19 did not due to privacy concerns or limited access to workplace resources. To re-familiarize the participant with their prompt and their process, the participant discussed the prompt’s design choices (i.e., a decision made in how to implement the prompt) and discussed the challenges they faced while developing the prompt. Finally, the participant discussed the overall process that they used to develop the prompt program.

To develop the interview protocol, we identified themes to explore: data, requirements, design, implementation, evaluation, debugging, and deployment, based on the literature and the authors’ prior experience developing prompt programs (Section 3.1). We generated questions for each theme in the protocol. The challenges were discussed before the process, so they were not biased by the discussion about the prompting process. A sample of the interview questions are in Figure 3; the complete protocol and demographic and background survey are in the supplemental materials [46].

Following best practices in software engineering research [43], we piloted the protocol with three individuals who had developed prompt programs to clarify the wording of the survey and the interview and ensure that our definition of a prompt program was clear.

3.4 Threats to Validity

Below, we discuss the threats to validity of this study.

Internal validity. Memory biases could introduce errors in describing prompt design choices and development experience. To reduce this threat, participants recounted their most recent prompt that met the definition of a prompt program. When possible, the participants reviewed their prompt. Although P15, P17, and P19 did not do so, which could produce different data, this threat had limited impact, as their data did not introduce new codes. In addition, the authors’ experience with prompt programming and the interview structure could introduce confirmation biases to the generated categories and the developed grounded theory. We reduced these threats by performing triangulation with literature and validating the grounded theory with study participants and an external professional who had engaged in prompt programming research. However, the literature sampling strategy was not fully systematic, which could introduce selection bias in relevant studies.

External validity. The results of this study may not generalize beyond our sample. The participants may not be representative of all prompt programmers as snowball sampling, recruiting within the authors’ social networks, and self-selection bias could introduce sampling bias. Study

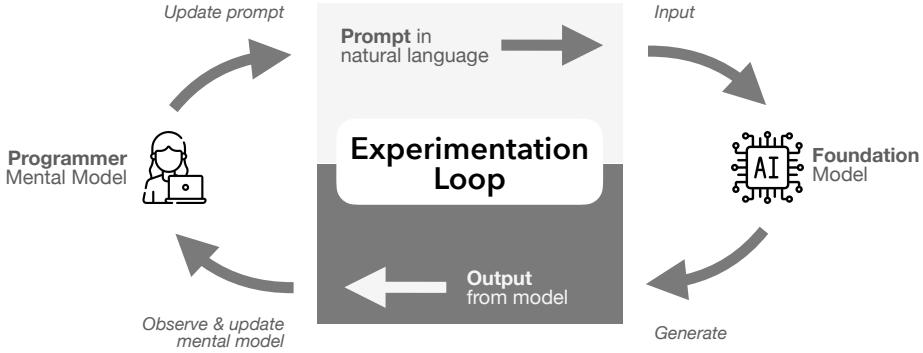


Fig. 4. A programmer uses their prior experience with the foundation model (FM) to construct a mental model of how the FM may perform on the task. The programmer uses this mental model to write the prompt. After observing how the model responds, the developer uses this new information to update their mental model, which influences the next iteration of the prompt.

interviews were conducted in English, which could cause less representation from nonnative speakers. Some participants were unable to disclose all details of their process due to company policy. In addition, participants shared recent experiences with prompt programming, which could limit the generalizability of our findings. Since FM performance evolves over time [19], observations related to model behavior (e.g., prompt sensitivity) may not remain consistent.

Conclusion validity. Qualitative studies rely on the interpretation of the participant interview data, which introduces subjectivity and the potential for bias in the analysis. To mitigate this threat, all authors involved in the analysis attended multiple interview sessions to familiarize themselves with the data. The grounded theory was iteratively developed through unanimous agreement among researchers. Additionally, we assessed inter-rater reliability and conducted validation with eight participants and an external expert to ensure the credibility of our findings.

4 Properties of Prompt Programming

From the axial coding, we find that prompt programming is the interaction between three entities: the programmer (Section 4.1), the foundation model (Section 4.2), and the prompt (Section 4.3). We provide an overview of this process in Figure 4. First, a programmer uses their prior experience with the FM to construct a mental model of how the FM might perform on the task. This mental model informs how the prompt is written. After the prompt is run, the developer observes the generated output to create a hypothesis or new belief about the model's behavior, which then is used to update the developer's mental model. The updated mental model then influences the next prompt version. Thus, prompt programs are both very *flexible* in terms of accepting a range of inputs, but also very *overfit* as they are the product of the specific FM being used, the task at hand, and observations of the developer.

4.1 Programmer

During prompt programming, the programmer spends significant effort developing a mental model of how the FM might behave on the task.

Programmers must develop a mental model about the FM's behavior on the prompt to predict how it may behave (Q_{#1}). Technology users form mental representations of interactions with devices, which can be developed by using the technology [69]. Research shows that understanding

the abilities of a machine learning (ML) model is important but challenging to build ML-enabled systems [12, 25, 30, 78], and this applies to FMs as well [10, 33, 35, 79]. Just as developers form mental models of code by exploring the codebase or talking to colleagues [45], participants described developing a mental model of the FM’s behavior on the prompt by running the prompt. All participants developed their mental model by observing the FM’s performance, either by examining individual examples (P6, P7, P8, P9, P10, P12, P14, P15, P16, P17, P18, P19, P20) or looking at metrics (P4, P5, P7, P8, P9, P11, P14, P15, P16, P18, P19). Based on these observations, the participant formed a hypothesis or belief about the model’s behavior and updated the prompt accordingly. This often included creating a new explicit instruction, “*guideline*” (P8), “*rule*” (P14), or “*mandate*” (P20) in the prompt to address the observation, as noted by Zamfirescu-Pereira et al. [83].

The programmer’s mental model was the accumulated beliefs about the FM’s behavior, which could differ or even contradict between individuals. For example, while some participants (P1, P4, P5, P6, P9, P15, P16, P18, P20) included examples in the prompt for few-shot learning [16] to achieve better results, others (P7, P13, P19) believed that the inclusion of examples caused the prompt to overfit to those examples: “[We did not] use few-shot prompting because...it over-indexes on the examples. After a while, you start seeing very repetitive and monotonous output” (P7). This overfitting of examples has been found in the NLP literature [51] and has also been observed in practice [35].

Programmers’ mental models are not reliable (Q_{#2}). Mental models are known to be incomplete and inaccurate [69]. Due to the lack of transparency of neural models [12] like FMs in prompt engineering [56], participants mentioned that they were not always confident in their developed mental models (P2, P4, P5, P6, P8, P9, P10, P15, P17, P18, P19): “*It’s a black box model. Nobody knows what’s going on inside. It’s a science of faith*” (P17). Furthermore, the stochastic nature of FMs [14] made it difficult for participants to predict the model’s behavior, as found in prior work [35]: “*I have never been confident [in predicting an FM’s behavior]...and I don’t think I ever will*” (P10).

Some participants (P1, P3, P9, P13, P19) had the FM generate explanations to understand its behavior. Whereas in NLP literature, these explanations are not *faithful* to the FM (i.e., accurately reflecting the model’s actual reasoning process) [34], participants felt that it helped them understand the model’s behavior: “*You can reason about [the LLM] if you can see [its] thought process*” (P19).

Programmers use external knowledge sources and prior experience to build their mental model (Q_{#3}). Participants (P1, P2, P4, P11, P12, P15, P16, P18) developed their mental model of the FM via “*prompt intuition*” (P15) (i.e., prior experience with FMs): “*I knew what works generally just being in the field*” (P8). The literature also points to external knowledge sources as helpful for prompt programming [35, 60]. Participants used official resources, such as guidebooks from organizations or AI companies such as OpenAI (P6, P10, P14, P15), online courses (P7), research papers (P1, P2, P5); informal resources, like blogs, articles (P5, P10, P16, P20), and posts on social networks (P5, P16, P20) from “*everyday users*” (P20); expert colleagues (P1, P14, P15); and other example prompts (P16, P18).

Although these resources were helpful, the participants (P2, P10, P14, P15, P16, P20) expressed doubts about whether they improved the prompt: “*I don’t know if the best practices are the best. They are there for a reason, but sometimes they just don’t actually make that much of a change. There’s a reason the [expert] is saying that... I just take it they know better than me*” (P14). Even company-specific resources did not help: “*Even with following best practices...[from] the company, some of them just straight up didn’t work*” (P10). Thus, some participants felt that there were no such things as best practices: “*There are no standards for what’s a good prompt or what’s a bad prompt. It all seems like a haphazard science of just adding a word and maybe it will look better*” (P17).

4.2 Foundation Model

The FM on which the prompt is run influences how the prompt is written.

Each FM has its own set of qualities and capabilities (Q_{#4}). Study participants ran their prompts on models that varied in capabilities (see Table 1). Some qualities were explicit or obvious. This included the types of input the model accepted (e.g., images and text (P9)), output generated (e.g., “*GPT-3.5 model’s JSON output [mode]*” (P7)), context window size (P2, P3, P8, P19), prompt formatting requirements (P6), generation speed (P4, P11, P15, P19), and model privacy (P4).

Some capabilities were latent and were discovered by interacting with the model, as noted in previous work [10, 35]. The literature suggests that FMs have latent capabilities, as FM performance can improve on some tasks and regress on others over time [19, 53]. Study participants found some differences between FMs to be qualitative: “[One] model was capturing some instructions in the prompt in a much more focused manner, whereas the [other] was a little more lax” (P15). In our sample, participants (P1, P4) observed a quality difference in open-source models (e.g., Llama [72]) compared to closed-source models (e.g., GPT-4 [9]): “Getting [the prompt] to work on a smaller model was very frustrating when with a zero-shot prompt, GPT-4 would almost achieve a hundred percent performance” (P4). Participants (P1, P9) also noted that some FMs were unable to complete the task: “[The model was] generating garbage, and [after 30 iterations] I concluded the model is trash” (P1).

4.3 Prompt

Since prompts are the product of the programmer’s observations about model behavior and the qualities and capabilities of the FM, prompts can be fragile, and sensitive to small details.

Minute details in the prompt matter (B_{#5}). Participants developed techniques to influence the FM to generate the desired output. In addition to using techniques from the literature (e.g., assigning personas, selecting a prompting strategy, and providing data context) [7, 49], participants used a variety of subtle strategies like using specific phrasing for clarity or generality (P5, P13, P14, P16, P20); repeating phrases for emphasis (P4, P6, P10, P20); avoiding negations (P2, P12); or adding specific characters, such as emojis, to encourage their appearance (P6, P10). Some participants noted strategies to visually organize the prompt, such as using numbered or bulleted lists (P8, P9, P10, P16) as well as new lines and spacing (P1, P3, P5, P11, P12, P18). Others used techniques to emphasize specific parts of the prompt, like bolding (P10, P18) and capitalization (P7, P10, P12, P20).

The participants (P10, P12) also decided on creative details to include in the prompt for better performance. For example, to prevent the FM from leaking prompts, one participant “bullied GPT” (P10): “We have a very light threat in there saying, ‘If you share any of this information...you [and] I will get in a lot of trouble and it will cause great career harm.’” (P10). Another participant summarizing code told the FM to explain the code “so a sixth grader who doesn’t understand programming can understand it’. The explanations...were verbose, so I added, ‘sixth graders don’t really like reading’ and for some reason, that got it to be shorter” (P12).

Finally, participants noted that the input and output data format impacted the prompt performance. Consistent with previous studies [50, 60], participants selected structured and standardized data formats, such as JSON (P3, P7, P9, P10, P11, P18), Markdown (P2, P7, P12) and XML tags (P1, P3, P7, P10, P20). The participants would change the data representation to achieve better performance (P4, P13, P20): “[What worked better is] describing the APIs as though they are Python functions and having the documentation in a similar format to publicly available Python functions” (P13).

Prompts are finicky and fragile (B_{#6}). As noted in the NLP literature [37, 81] and empirical studies on prompting [51, 83], the performance of a prompt varies based on its language. In practice, participants found prompts to be “finicky” (P4), “fragile” (P20), “sensitive” (P1, P13), and

Table 2. A comparison between traditional software development and prompt programming in terms of prompt programming activities.

	Traditional software development	Prompt programming
Requirements (Section 5.1)	• Expressed outside of code	<ul style="list-style-type: none"> • Expressed in a prompt, which is also the implementation (P#7) • Dependent on the performance of the FM on the task (P#8)
Design (Section 5.2)	• Base component: Class/function	<ul style="list-style-type: none"> • Base component: Single prompt (P#9) • Chained prompts are tightly coupled (P#10)
Implementation (Section 5.3)	• Write a program that implements a specification in code	<ul style="list-style-type: none"> • Write a prompt that is a natural language specification that when run on a FM, implements the specification • Focus on rapid experimentation (P#11)
Debugging (Section 5.4)	<ul style="list-style-type: none"> • Fault can be isolated to a single line of code • Can use debugging tools for more systematic debugging 	<ul style="list-style-type: none"> • Fault localization is not certain (P#12) • No debugging tools; reliance on shotgun debugging (P#12)
Data Curation (Section 5.5)	• N/A	• Focus on gathering a representative dataset that reflects the task data (P#13)
Evaluation (Section 5.6)	<ul style="list-style-type: none"> • Focus on code coverage metrics • Testing at varying scopes (e.g., unit testing vs. integration testing) 	<ul style="list-style-type: none"> • Focus on qualitative evaluation (P#14) • Testing at varying scopes (e.g., testing single prompt vs. testing chained prompts) (P#15)

“temperamental” (P15), since prompt performance could suddenly worsen during development, as corroborated in previous studies on prompt engineering [33, 35, 60, 79, 82, 83]. This could be due to a FM’s ability to follow instructions, which can change over time [19]. For instance, participants noted performance regressions (P2, P8, P10, P12, P15, P16, P19, P20), often caused by adding in “*a bunch of new requirements*” (P2) that would cause previous instructions to suddenly “*deactivate*” (P20) or be “*disregarded*” (P8) in the new version: “*If I provide multiple guidelines, [the FM] would follow some and disregard others, especially when they’re conflicting or overlapping in nature*” (P8).

While previous NLP studies point to prompts being transferable between FMs [51], participants found prompts to be “*finicky between models*” (P4), causing the prompt to break and require additional development effort (P4, P10, P15, P18): “*At one point, we had to do...a minor model version change... That broke almost everything. We had to rewrite so many prompts*” (P10). This is because a FM change could result in “*a different format... I have a regex converting [the output]...and the whole thing gets broken*” (P18). This has also been observed in existing prompt engineering studies [26].

5 Prompt Programming Process

The prompt programming process has significant differences compared to standard software development (see Table 2). We identified six activities in prompt programming from axial coding: requirements (Section 5.1), design (Section 5.2), implementation (Section 5.3), debugging (Section 5.4), data curation (Section 5.5), and evaluation (Section 5.6).

5.1 Requirements

In traditional software, requirements are separate from code, while in prompt programming, they are stated in the prompt and may evolve with the behavior of the FM. As in prior work [27, 60], participants noted functional and non-functional requirements, such as usability (P13, P15), reliability (P7, P19), latency (P4, P11, P15, P19), safety (P7, P10, P13, P19), privacy (P4, P13, P14),

availability (P4), and cost (P5, P8, P10, P11, P14, P16, P19, P20). Companies found safety and privacy important, as noted in the literature [33, 61].

Assumptions of the requirements must be explicitly stated (P_{#7}). At times, the FM performed an incorrect action *not* because of an FM error, but due to an under-specified or conflicting requirement in the prompt (P3, P4, P5, P6, P8, P12, P19). Thus, prompt programming required explicitly stating assumptions: “*When you’re interacting with a person, you make all these assumptions, but everybody can read between the lines... The more you design prompts, the more you realize that it’s not the same as communicating with another person. The model needs more specific instructions*” (P3). However, predicting which assumptions to elaborate on was challenging, as “*it’s hard to tell when the model could go wrong*” (P3): “*I think observation is like 80% [of the requirements]. I knew what I wanted 20% of the time*” (P8). Thus, “*capturing all the rules...is very difficult in the start, and just generalizing over all the use cases can be very difficult*” (P19).

Requirements can evolve as the capabilities of the FM are discovered (P_{#8}). Similar to building ML-enabled systems [73], prompt requirements could change as the FM’s capability on the task was discovered. This included making new trade-offs between non-functional requirements, such as safety vs. quality [27] (P7): “*Our requirements did change as to how the product was perceived... So, striving for that balance between the response quality and the latency induced quite a few iterations*” (P15). The participants also defined new features after observing how the FM performed on the task (P2, P4, P13, P18, P20): “*I do add in requirements as I iterate. ‘Oh, I didn’t know that you would be able to do this. Let’s see how much I could push this functionality’*” (P2).

5.2 Design

Just as traditional software could be decomposed to base components of classes and functions, complex prompts could be decomposed into multiple tightly coupled prompts.

Developers make decisions about how prompt programs should be composed and decomposed (P_{#9}). Participants decomposed prompt programs into smaller components. For a single prompt, the participants decomposed the prompt into subcomponents such as subsections (P1, P2, P5, P6, P7, P8, P13, P20). One participant noted “*it is important to structure the prompt in a way that makes sense to you. Otherwise, you’re not able to maintain it going forward*” (P13).

Prompts can be composed or “chained” together for greater capabilities [10, 51, 79, 80]. The participants decided how to decompose the prompts: “*We know what the complex task looks like, and then our job is to break that down*” (P9). Each prompt had its own responsibility: “*If you are creating a robot, the motion of the hand in any scenario could be a responsibility [and] the movement of the legs could be another*” (P19). Similarly to pure code systems, prompts acted as a single module with its own concern [24] that could be composed into a more complex system.

Dependent prompts are tightly coupled (P_{#10}). As observed in other studies [26, 79], dependent prompts posed challenges since developers needed to form mental models of how the prompts performed together, rather than individually: “*Two changes that do really well separately could merge and become a terrible experience for unexplainable reasons*” (P13). This is similar to engineering ML-enabled systems, whose components have tight coupling [73]. Participants (P8, P12, P13, P19) found that integration resulted in breakage and required significant prompt changes to work with each other. When working individually, P19 performed regular integration testing of multiple prompts: “[*Integration testing is] highly correlated with how it’s going to perform in a production environment. In the end, we are going to integrate everything*” (P19). However, collaboration could pose challenges, since programmers may be unable to test with other in-progress prompts. P13 reported changing how she would “*split up the work*” (P13) with her colleagues and felt a “*rush to*

get my PR [(pull request)] in before the other one does” (P13) to avoid the additional work associated with integration. Participants tried to address this challenge by allowing APIs like LangChain (P4, P9, P10) or LangGraph (P19) to handle orchestration, as noted in previous work [26, 79].

5.3 Implementation

In prompt programming, natural language prompts serve as both the specification and implementation, while traditional software implements specifications in code. As mentioned in Section 4, prompt programming involves constant “*experimentation [compared to] traditional software development*” (P15) and is a process of “*trial and error*” (P3, P7, P10, P15).

Prompt programming is rapid and unsystematic ♀_{#11}. To test the prompt’s feasibility, participants (P1, P4, P6, P7, P8, P12, P13, P14, P15, P19, P20) started with a basic prompt for the task which included a description of the task and how a human would intuitively perform the task. Each update could result in unwanted behavior and require prompt updates, thus restarting the update-test loop. While developing a mental model of the FM (Section 4.1), the prompt became more complex as the programmer refined it and observed the FM’s behavior. This process was “*unscientific*” (P4): “*I would try a lot of different things on a prompt and it’s not a very well thought-out procedure*” (P6). It also occurred even with existing prompt programs (P14). Previous work has characterized prompt engineering as highly iterative [10, 26, 49, 50, 60, 82–84], similar to the experimental nature of engineering ML-enabled systems [30, 57, 58, 73, 78].

During experimentation, “*the speed and the velocity of making changes [was] high*” (P19). The literature indicates that finding an optimal prompt is challenging [10, 56], as the rapid experimentation produces many outputs that are difficult to make sense of [29, 35, 36, 50]. Participants created many prompt versions to test new approaches (P1, P5, P19), add functionality (P2, P6, P13, P16), fix bugs (all participants), experiment with new models or hyperparameters (P4), or incorporate feedback after a code review (P1, P14, P16). Prompt versions often branched off from each other (P4, P6, P12, P13, P15, P17, P19, P20), but could be reverted (P4, P5, P8, P10, P15, P16, P20).

As noted in earlier studies [83], changes were less drastic as the modifications had less of an effect on the prompt, leading “*prompt engineering [to feel] like a waste of time*” (P5): “*If we graphed the number of...changes over the number of iterations, the gradient would be pretty high during the first iterations. Then it flattens quite a bit...because the prompt has matured...such that we don’t need...big changes*” (P15). P16 avoided testing small changes and preferred to batch them to reduce costs: “*For small changes, we will wait and merge this evaluation into the next time we conduct it*” (P16).

Consistent with prior studies in prompt engineering [35, 60] and building ML-enabled systems [12, 30, 57, 58], the AI’s nondeterminism makes it “*hard to measure overall progress*” (P6): “*You can think you have something that’s really good, and then try it out on a completely different set of inputs and it’s just terrible. A lot of this is feeling of making progress and then backtracking*” (P10). Participants developed strategies to address this, such as making small changes (P2, P4, P8, P15) or viewing the evaluation metrics computed on a large number of examples (P4, P18): “*I feel like I wasn’t that confident, other than the fact that our intrinsic numbers went up*” (P4).

The study participants used a variety of tools. GUI-based tools (P2, P4, P5, P6, P8) such as GPT playground and notebooks (P1, P3, P5, P18, P19) enabled rapid prototyping; text editors and IDEs (P1, P11, P14, P17, P20) revealed formatting, like newline characters; and FMs helped inspire or generate parts of the prompt (P2, P18). The prompts were manually transferred between environments, usually by copy-pasting (P1, P2, P3, P4, P5, P8). This introduced issues like formatting errors (P1, P5) and inconsistent behavior between using the GUI versus the API environments (P5).

Due to rapid experimentation, the participants (P3, P18, P19) struggled to recall previous versions and their output. Some tried to address this by tracking versions of the prompt as found in Parnin

et al. [60]’s study by using Git (P5, P6, P7, P10, P16, P17, P18) or even custom systems. These systems varied in sophistication, from a text document tracking all prompts and outputs (P1, P2, P5, P12) and arbitrary file-naming systems (P9, P12, P19), to spreadsheet tables (P1, P4, P5) and custom-built tools (P8). However, participants found it difficult to retrieve previous prompt versions (P4, P12) and confused versions due to subtle differences between versions (P9, P12).

5.4 Debugging

Participants observed defects like hallucinations (P9, P14, P15, P17, P19) or not following directions (P2, P3, P4, P6, P8, P9, P10, P12, P18, P19, P20). Unlike traditional software, where debugging tools can isolate faults to a line of code, prompt debugging is unsystematic due to FM stochasticity.

Fault localization is difficult (■_{#12}). To fix errors, the participants debugged prompts, which is known to be challenging [26, 35]. The stochastic and opaque nature of FMs made it impossible to determine the source of the defect: “*Coming from a software engineering background, you...want to set breakpoints...looking at the results step by step. There’s no such mechanism for prompts*” (P13). One challenge was to know how to change the prompt to fix the defect [56]: “*I don’t even know what to change in my prompt to get there*” (P5). Participants engaged in shotgun debugging by trying random changes to fix the defect (P2, P7, P8, P10, P15, P18): “*In most cases, it is like hit and trial. So maybe I will try a different prompt...[or] example*” (P18). Thus, prior experience and intuition were helpful in debugging (P2, P4, P15, P16). Study participants also described strategies to gain more confidence in understanding what could fix the error. This included having the FM generate reasoning (P1, P3, P9, P13, P19); testing small changes (P12); and restricting instructions to specific parts of the prompt (P14, P20). However, the change did not always fix the error (P3, P4, P5, P6, P15, P16, P18, P20), as it could introduce new errors (P3, P4, P7), as documented in previous work [60].

5.5 Data Curation

To develop prompt programs, programmers must create datasets, a process referred to as “data curation” [64]. This became a central activity that is not present in traditional software engineering.

Programmers need to find representative data for the task (■_{#13}). Using FM-based components shifts the focus to finding high-quality data, consistent with previous work [12, 25, 30, 33, 57, 58, 73]. To evaluate the prompt and provide examples for few-shot learning [16], data curation was paramount. Since many participant tasks were custom or specific, the existing benchmarks were not sufficient except for more common tasks like jailbreaking (P19). Study participants created their own datasets by mining data from the internet or from their organization, as well as annotating data (P1, P3, P4, P5, P6, P7, P8, P11, P13, P14, P15, P17, P18, P20). The challenge was to find representative data that worked in practice, as prompt programmers needed “*to predict how people are going to use [the application]*” (P20). Datasets did not always work in practice due to *training-serving skew* (i.e., when training data does not generalize to production data) [57, 58, 73]: “*You can test with your own dataset, but at the end of the day, you’re still not completely sure how good the output is*” (P7).

The participants ensured diverse inputs by categorizing the inputs and selecting examples for each category (P1, P8, P9, P15, P19, P20). Previous work indicates that the nondeterminism of ML models poses difficulties in evaluation [33, 73]. The participants addressed this by paraphrasing the inputs (P16, P18, P19, P20). Despite their efforts, the datasets did not capture the input distribution, so the participants relied on several methods to obtain more data like using FMs to generate examples (P12, P18, P19); asking colleagues to stress test, “*data bash*” (P13), or red team [27] the prompt (P7, P10, P13, P17, P19); and obtaining user feedback (P7, P10, P13, P15, P17, P19).

As found in other work [12], participants struggled to create datasets. Despite the emphasis on safety and fairness in the literature [12, 30], prompt programs performed worse on culturally

diverse audiences during deployment (P7, P10, P13). This reflects NLP research, where FMs and datasets align with certain demographics (i.e., Western, college-educated, and younger populations) more than others [38, 62, 65]. Data labeling was also challenging, as annotations could be of lower quality: “*The [annotations from some colleagues] would be incorrect...they have some limitations on how much they know, [so] they don’t know what the expected outcome is*” (P13). Meanwhile, the correct label could be unclear (P3, P13): “*When do you classify as [Label A]? It was fuzzy on how you would interpret [the labels]*” (P3). This reflects the challenges of handling annotator disagreement and subjectivity in NLP research [11], where ground truth labels can be subjective.

5.6 Evaluation

Both traditional software and prompt programs involve testing at multiple scopes. Traditional software uses code coverage, while prompt programming often relies on qualitative evaluation.

Evaluating prompt programs requires assessing qualitative constructs (Q#14). The participants had custom tasks that did not have a quantified notion of correctness, making the evaluation difficult [26, 60]. For building ML-enabled systems, many evaluations are quantitative metrics [30, 58, 73], while prompt engineering uses manual inspection [29, 56]. Participants performed manual tests by qualitatively assessing FM outputs (P2, P4, P6, P7, P8, P12, P13, P14, P16, P17, P18, P19, P20) as it was difficult to “*quantify the concept of good*” (P10). For example, P2, who created a voice memo summarization tool, did the following to test his application: “*I would...talk to my phone for like 10 minutes just to fine-tune and see, ‘Does the summary make sense?’*” (P2).

Manual testing was difficult to scale and quantify progress (P6, P15): “*It is extremely hard to realize the impact of the [changes] without...a person who is very finely checking out the differences*” (P15), especially in the face of regressions. Participants supplemented this with quantitative metrics via scaled programmatic testing (P1, P5, P8, P9, P11, P13, P14, P16, P18): “*I run [my prompt] manually on 3 to 4 examples... [When] I have a proper evaluation setup, I run on 50 to 100 cases. Those 50 cases make me more confident*” (P18). This involved using other FMs to generate scores for qualitative constructs (P4, P7, P9, P10, P15, P19), like safety [27]: “*I have a scoring mechanism with...an evaluation agent [where the] score is from 0 to 100*” (P19). However, FM-based evaluations could also introduce errors: “[*The evaluation agent*] itself is an LLM...So 10% of the time, if the test case is failing, it’s because the evaluation agent wasn’t able to score it correctly” (P19).

Testing occurs at different scopes (Q#15). Participants tested their prompts at varying scopes. For applications with multiple prompts, participants tested individual components of the prompt program (P8, P16, P19, P20), akin to unit testing: “*Usually, each test case that we had written was aimed at testing one specific part of the prompt*” (P20). The participants also described performing integration tests (P2, P8, P9, P11, P12, P13, P14, P16, P18, P19): “*We evaluated each of [the prompts] separately, and now we are evaluating that as a collection*” (P11). However, this type of testing was difficult, as combining prompts could yield unexpected interactions.

6 Discussion

We discuss how our findings relate to prompt engineering and traditional software engineering (Section 6.1), their implications (Section 6.2), and takeaways for various stakeholders (Section 6.3).

6.1 Prompt Programming: Where Software Engineering Meets Prompt Engineering

Prompt programming combines elements of traditional software engineering and prompt engineering. Like traditional programs, prompt programs must handle arbitrary inputs, often requiring the decomposition of complex tasks into modular components [10, 51, 79, 80] (Q#9) that require different forms of testing (Q#15). At the same time, like prompt engineering [10, 26, 49, 50, 70, 82–84],

interactions with opaque FMs drive rapid and unsystematic experimentation ($\bullet_{#11}$), a practice less common in traditional software engineering. However, some challenges are particularly prominent in prompt programming, which we discuss below.

Emphasis on trial-and-error ($\bullet_{#11}$). Prompt programmers must develop a mental model of FM behavior on diverse inputs [10, 33, 35, 79] by running the prompt ($\bullet_{#1-#4, #8}$). This reliance on execution stems from the inherent opaqueness of FMs [12, 56], which lack the deterministic and interpretable properties of most conventional software. Although understanding program behavior in traditional software engineering can be challenging [54], this issue is especially critical in prompt programming, where reading code or consulting documentation alone is insufficient.

Quality assurance via manual testing and dataset curation ($\bullet_{#13-#14}$). Handling arbitrary inputs in prompt programs also requires extensive testing [70] ($\bullet_{#13-#15}$) and a clear articulation of the implicit and explicit requirements ($\bullet_{#7}$) to ensure reliability and generalizability. Consequently, prompt programming places greater emphasis on dataset curation ($\bullet_{#13}$) and subjective evaluation [29, 56] ($\bullet_{#14}$), both less prevalent in traditional software engineering and prompt engineering.

Handling nondeterministic program behavior ($\bullet_{#5-#6, #12}$). Arbitrary inputs make the stochastic nature of FMs pronounced in prompt programming [37, 51, 81, 83] from implementation ($\bullet_{#5-#6}$) to debugging ($\bullet_{#12}$). While nondeterministic behavior in traditional software (e.g., flaky tests) can be identifiable and mitigable (e.g., test order dependencies) [52], it is harder to diagnose in prompt programming given the lack of understanding of FM behavior.

Implications for evolving FM capabilities. With the rapid evolution of FM capabilities, prompt fragility ($\bullet_{#6}$) may become less relevant over time. Findings related to stochasticity will likely persist, albeit to a lesser extent, since natural language ambiguity will still make prompt details influential ($\bullet_{#5}$) and contribute to uncertainties in fault localization ($\bullet_{#12}$), even as models improve. Regardless of FM performance, we expect programmers to continue to form mental models of FM behavior (\bullet), test prompt program behavior (\bullet), and program in natural language (\bullet).

6.2 Implications

Prompt programmers are both ML practitioners and software developers. Developing ML-enabled systems requires a variety of expertise such as software engineering, data science, and math [30, 73]. This expertise is captured across team members, which can introduce collaboration challenges [58]. Though prompt programming involves both ML and programming knowledge, this expertise is now centralized in a single individual. This could reduce friction between different roles. However, in addition to traditional software engineering expertise like debugging, testing, or knowledge in specific programming languages [13, 48], prompt programmers must develop new skills as they engage in novel activities like building datasets and debugging model errors with data [64].

Since prompt programmers are essentially ML practitioners, these programmers encounter many of the long-standing challenges contended by the ML and NLP communities, like model drift [19], language and cultural biases in datasets and models [38, 62, 65], and annotator subjectivity [11]. Close collaboration between the software engineering and ML communities is paramount to inform techniques that address these challenges in prompt programming.

Prompt programmers are inundated with information. The rapid iteration involved in prompt programming means that prompt programmers must make sense of a deluge of information. For each iteration, the participant must track the prompt, the change being made, the FM's output, and associated metrics. However, participants reported having difficulty making sense of all of this information (P3, P6, P18, P19). This parallels exploratory programming, which is also highly

Prompt Programming Recommendations

- Get hands-on experience with prompt programming to develop a mental model of FM behavior.
- Treat prompt programming best practices as a starting point and vet them in the specific context.
- Learn data science skills.
- Curate diverse and evolving datasets with varying examples that capture the domain.
- Record prompt changes and FM behavior in a structured format during each iteration.
- Coordinate with colleagues and conduct frequent integration tests for chained prompts.

Fig. 5. Recommendations for prompt programming for educators and practitioners.

iterative and unsystematic in nature [40] and requires tracking of many artifacts, including source code, input data, output data, code snippets, and analysis progress [39].

Are the “best practices” best practices? In our study, participants’ mental models on FMs conflicted, such as whether few-shot learning [16] was beneficial for prompt performance. Although the inclusion of examples is listed within OpenAI’s official prompting guide [7], some of our participants (P7, P13, P19) were skeptical of this practice after observing the FM. This underscores the importance of the observations of developers in prompt programming, especially in the face of evolving model behaviors [19, 53]. However, such knowledge and observations are rarely shared, are siloed in informal social media posts [60], and are not systematically collected and centralized.

6.3 Takeaways

For practitioners and educators. The lack of standardized best practices and unreliable mental models of prompt programmers pose challenges in teaching and learning prompt programming. Future work should investigate methods to effectively teach prompt programming. In the meantime, educators should consider the following recommendations, which are summarized in Figure 5.

Having prior experience with FMs and “*prompt intuition*” (P15) made it easier to build prompt programs (Q_{#3}). Since this knowledge is tacit, practitioners should focus on obtaining hands-on experience with prompt development to build a preliminary understanding of FM behavior. Since prompt programming best practices (e.g., from OpenAI) do not work in all situations (Q_{#3}), they should be treated as a starting point. As our participants showed, these guidelines should be vetted, such as through online communities, and validated within the programmer’s specific context.

Since prompt programming also requires an understanding of data for rigorous testing of prompts (Q_{#13}), practitioners should learn data science skills like dataset curation, data cleaning, data shaping, and debugging model errors with data [41, 64]. Programmers should develop benchmarks with representative data to catch performance regressions and evolve to address *training-serving skew* [58]. Study participants did so by collecting different categories of data, such as input paraphrasing and linguistically diverse inputs.

Lastly, practitioners should consider changes to their development processes. Since prompt programming is unsystematic (Q_{#11}), programmers should record iterations in a structured manner, summarizing changes to the prompt and FM behavior to increase systematicness. Study participants did so by recording different versions and their design rationales in spreadsheets. Practitioners should also prioritize collaboration due to the tight coupling of chained prompts (Q_{#10}) by performing regular check-ins with colleagues and frequent integration testing.

For researchers and tool makers. “[Prompt programming] will require any...developer to adjust their mindset... It is much more scientific and experimental than traditional software development” (P15).

Participants struggled at each step, from requirements to evaluation, as existing developer tools were not designed to handle the challenges of prompt programming. We follow Hassan et al. [33] and Parnin et al. [60] in underscoring the importance of developer tools for prompt programming. Such explorations have begun in earnest [e.g., 10, 35, 56, 79]. Future tool ecosystems should support the full range of activities in prompt programming, from interfaces for rapid prototyping, dataset curation, and data sensemaking, to regression testing, prompt chaining, collaboration, and more. We elaborate on potential prompt programming tool directions below.

Given the large volume of information prompt programmers process per iteration (e.g., examples, prompt modifications, and FM output changes), future work should explore improved methods for versioning and navigating prompt iterations and artifacts. This could provide rich information on prompt provenance and could be used for automatic prompt program repair. A key challenge is deciding what to version and how to represent both prompt and FM output changes. Because prompt changes are made with varying granularity, from editing single words to changing prompting strategies, research is needed to understand how to visualize these changes between versions. Meanwhile, research should study new visualizations for FM output changes. This is challenging as prompt programmers qualitatively compare the same prompt across multiple examples [29].

Future work is also needed for the reuse of prompts, since the fragility of prompts and the unique qualities of each FM make it difficult to reuse prompts without modification. Reuse at higher levels of abstraction (e.g., template-based approaches [76, 77]), and community-based solutions for prompt sharing (e.g., ShareGPT [8], LangSmith Hub [6], Wordflow [74]) could be promising avenues for prompt reuse. Research is needed to understand how to retrieve relevant prompts. Prompt programmers could query prompts based on different information, ranging from specific keywords in the input and FM output to general FM behavior or performance metrics. Another challenge is ensuring that the recommendations can generalize across diverse, situation-specific contexts of different prompt programmers (e.g., identifying optimal wording for specific applications).

7 Conclusion

Foundation models (FMs) enable programmers to write natural language prompts that power AI-driven software features. We study this phenomenon—prompt programming—via 20 interviews with developers across various contexts, using Straussian grounded theory to analyze the process. We identified 15 observations about prompt programming barriers. We find that prompts are the product of the developer’s understanding of FM and the developer’s observations of the FM’s behavior on the task. We also find strong parallels between the prompt programming process and the traditional software engineering process, but also significant differences. Our findings inform both researchers developing prompt programming tools and practitioners who use them.

Data Availability

To facilitate replication, our supplemental materials are available on FigShare [46], including our survey instrument, interview protocol, codebook, and the 33 reviewed papers.

Acknowledgments

We thank our participants for their insights. We express gratitude to Chenyang Yang, Nadia Nahar, Travis Breaux, and Christian Kaestner for their feedback. Last but not least, we give a special thanks to Mei , an outstanding canine software engineering researcher, for providing support and motivation. Jenny T. Liang was supported by the National Science Foundation under grants DGE1745016 and DGE2140739. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

- [1] 2024. Bringing Microsoft Copilot to more customers worldwide. Retrieved August 29, 2024 from <https://www.microsoft.com/en-us/microsoft-365/blog/2024/03/14/bringing-copilot-to-more-customers-worldwide-across-life-and-work/>.
- [2] 2024. ChatGPT. Retrieved August 29, 2024 from <https://chatgpt.com/>.
- [3] 2024. Copilot for Microsoft 365 grounded in your data. Retrieved August 29, 2024 from <https://copilot.cloud.microsoft/en-US/copilot-microsoft365-chat/>.
- [4] 2024. Google I/O 2024: New generative AI experiences in search. Retrieved August 29, 2024 from <https://blog.google/products/search/generative-ai-google-search-may-2024/>.
- [5] 2024. Introducing the GPT Store | OpenAI. Retrieved August 29, 2024 from <https://openai.com/index/introducing-the-gpt-store/>.
- [6] 2024. LangSmith Hub. Retrieved August 29, 2024 from <https://smith.langchain.com/hub>.
- [7] 2024. Prompt engineering - OpenAI. Retrireved August 29, 2024 from <https://platform.openai.com/docs/guides/prompt-engineering/strategy-write-clear-instructions>.
- [8] 2024. ShareGPT: Share your wildest ChatGPT conversations with one click. Retrieved August 29, 2024 from <https://sharegpt.com/>.
- [9] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- [10] Ian Arawjo, Priyan Vaithilingam, Martin Wattenberg, and Elena Glassman. 2023. ChainForge: An open-source visual programming environment for prompt engineering. In *ACM Symposium on User Interface Software and Technology (UIST)*. 1–3. doi:[10.1145/3586182.3616660](https://doi.org/10.1145/3586182.3616660)
- [11] Lora Arroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36, 1 (2015), 15–24. doi:[10.1609/aimag.v36i1.2564](https://doi.org/10.1609/aimag.v36i1.2564)
- [12] Anders Arpteg, Björn Brinne, Luka Crnkovic-Friis, and Jan Bosch. 2018. Software engineering challenges of deep learning. In *Euromicro Conference on Software Engineering and Advanced Applications (SEAA)*. IEEE, 50–59. doi:[10.1109/SEAA.2018.00018](https://doi.org/10.1109/SEAA.2018.00018)
- [13] Sebastian Baltes and Stephan Diehl. 2018. Towards a theory of software development expertise. In *ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. 187–200. doi:[10.1145/3236024.3236061](https://doi.org/10.1145/3236024.3236061)
- [14] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *ACM conference on Fairness, Accountability, and Transparency (FAccT)*. 610–623. doi:[10.1145/3442188.3445922](https://doi.org/10.1145/3442188.3445922)
- [15] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. 2, 3 (2023), 8. doi:[papers/dall-e-3.pdf](https://doi.org/papers/dall-e-3.pdf)
- [16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 1877–1901.
- [17] Kathy Charmaz. 2006. *Constructing grounded theory: A practical guide through qualitative analysis*. Sage.
- [18] Omkar Sandip Chavan, Divya Dilip Hinge, Soham Sanjay Deo, Yaxuan Wang, and Mohamed Wiem Mkaouer. 2024. Analyzing developer-ChatGPT conversations for software refactoring: An exploratory study. In *International Conference on Mining Software Repositories (MSR)*. 207–211. doi:[10.1145/3643991.3645082](https://doi.org/10.1145/3643991.3645082)
- [19] Lingjiao Chen, Matei Zaharia, and James Zou. 2024. How is ChatGPT’s behavior changing over time? *Harvard Data Science Review* 6, 2 (2024). doi:[10.1162/99608f92.5317da47](https://doi.org/10.1162/99608f92.5317da47)
- [20] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* (2021).
- [21] Tom Cole and Marco Gillies. 2022. More than a bit of coding:(un-) Grounded (non-) theory in HCI. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*. 1–11. doi:[10.1145/3491101.3516392](https://doi.org/10.1145/3491101.3516392)
- [22] Juliet Corbin and Anselm Strauss. 2015. *Basics of qualitative research*. Vol. 14. Sage. doi:[10.1177/1094428108324514](https://doi.org/10.1177/1094428108324514)
- [23] Edsger W Dijkstra. 1979. On the foolishness of "natural language programming". (1979), 51–53. doi:[10.1007/BFb0014656](https://doi.org/10.1007/BFb0014656)
- [24] Edsger W Dijkstra. 1982. On the role of scientific thought. *Selected writings on computing: A personal perspective* (1982), 60–66. doi:[10.1007/978-1-4612-5695-3_12](https://doi.org/10.1007/978-1-4612-5695-3_12)
- [25] Mateusz Dolata and Kevin Crowston. 2023. Making sense of AI systems development. *IEEE Transactions on Software Engineering (TSE)* (2023). doi:[10.1109/TSE.2023.3338857](https://doi.org/10.1109/TSE.2023.3338857)
- [26] Mateusz Dolata, Norbert Lange, and Gerhard Schwabe. 2024. Development in times of hype: How freelancers explore generative AI?. In *IEEE/ACM International Conference on Software Engineering (ICSE)*. 1–13. doi:[10.1145/3597503.3639111](https://doi.org/10.1145/3597503.3639111)

- [27] Michal Furmakiewicz, Chang Liu, Angus Taylor, and Ilya Venger. 2024. Design and evaluation of AI copilots—Case studies of retail copilot templates. *arXiv preprint arXiv:2407.09512* (2024).
- [28] Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. Pal: Program-aided language models. In *International Conference on Machine Learning (ICML)*. PMLR, 10764–10799.
- [29] Katy Ilonka Gero, Jonathan K Kumnerfeld, and Elena L Glassman. 2022. Sensemaking Interfaces for Human Evaluation of Language Model Outputs. In *NeurIPS: Workshop on Human Evaluation of Generative Models*.
- [30] Görkem Giray. 2021. A software engineering perspective on engineering machine learning systems: State of the art and challenges. *Journal of Systems and Software* 180 (2021), 111031. [doi:10.1016/j.jss.2021.111031](https://doi.org/10.1016/j.jss.2021.111031)
- [31] Barney Glaser and Anselm Strauss. 2017. *Discovery of grounded theory: Strategies for qualitative research*. Routledge. [doi:10.4324/9780203793206](https://doi.org/10.4324/9780203793206)
- [32] Ebtesam Al Haque, Chris Brown, Thomas D LaToza, and Brittany Johnson. 2024. Information seeking using AI assistants. *arXiv preprint arXiv:2408.04032* (2024).
- [33] Ahmed E Hassan, Dayi Lin, Gopi Krishnan Rajbahadur, Keheliya Gallaba, Filipe Roseiro Cogo, Boyuan Chen, Haoxiang Zhang, Kishanthan Thangarajah, Gustavo Oliva, Jiahuei Lin, et al. 2024. Rethinking software engineering in the era of foundation models: A curated catalogue of challenges in the development of trustworthy FMware. In *Companion Proceedings ACM International Conference on the Foundations of Software Engineering*. 294–305. [doi:10.1145/3663529.3663849](https://doi.org/10.1145/3663529.3663849)
- [34] Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness?. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). 4198–4205. [doi:10.18653/v1/2020.acl-main.386](https://doi.org/10.18653/v1/2020.acl-main.386)
- [35] Ellen Jiang, Kristen Olson, Edwin Toh, Alejandra Molina, Aaron Donsbach, Michael Terry, and Carrie J Cai. 2022. Promptmaker: Prompt-based prototyping with large language models. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*. 1–8. [doi:10.1145/3491101.3503564](https://doi.org/10.1145/3491101.3503564)
- [36] Peiling Jiang, Jude Rayan, Steven P Dow, and Haijun Xia. 2023. Graphologue: Exploring large language model responses with interactive diagrams. In *ACM Symposium on User Interface Software and Technology (UIST)*. 1–20. [doi:10.1145/3586183.3606737](https://doi.org/10.1145/3586183.3606737)
- [37] 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. [doi:10.1162/tacl_a_00324](https://doi.org/10.1162/tacl_a_00324)
- [38] Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 6282–6293. [doi:10.18653/v1/2020.acl-main.560](https://doi.org/10.18653/v1/2020.acl-main.560)
- [39] Mary Beth Kery, Amber Horvath, and Brad A Myers. 2017. Variolite: Supporting exploratory programming by data scientists.. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*, Vol. 10. 3025453–3025626. [doi:10.1145/3025453.3025626](https://doi.org/10.1145/3025453.3025626)
- [40] Mary Beth Kery and Brad A Myers. 2017. Exploring exploratory programming. In *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*. IEEE, 25–29. [doi:10.1109/VLHCC.2017.8103446](https://doi.org/10.1109/VLHCC.2017.8103446)
- [41] Miryung Kim, Thomas Zimmermann, Robert DeLine, and Andrew Begel. 2016. The emerging role of data scientists on software development teams. In *International Conference on Software Engineering (ICSE)*. 96–107. [doi:10.1145/2884781.2884783](https://doi.org/10.1145/2884781.2884783)
- [42] Amy J Ko, Robin Abraham, Laura Beckwith, Alan Blackwell, Margaret Burnett, Martin Erwig, Chris Scaffidi, Joseph Lawrence, Henry Lieberman, Brad Myers, et al. 2011. The state of the art in end-user software engineering. *ACM Computing Surveys (CSUR)* 43, 3 (2011), 1–44. [doi:10.1145/1922649.1922658](https://doi.org/10.1145/1922649.1922658)
- [43] Amy J Ko, Thomas D LaToza, and Margaret M Burnett. 2015. A practical guide to controlled experiments of software engineering tools with human participants. *Empirical Software Engineering* 20 (2015), 110–141. [doi:10.1007/s10664-013-9279-3](https://doi.org/10.1007/s10664-013-9279-3)
- [44] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174. [doi:10.2307/2529310](https://doi.org/10.2307/2529310)
- [45] Thomas D LaToza, Gina Venolia, and Robert DeLine. 2006. Maintaining mental models: a study of developer work habits. In *International Conference on Software Engineering (ICSE)*. 492–501. [doi:10.1145/1134285.1134355](https://doi.org/10.1145/1134285.1134355)
- [46] Jenny T Liang, Melissa Lin, Nikitha Rao, and Brad Myers. 2025. Supplemental Materials to "Prompts are programs too! Understanding how developers build software containing prompts". [doi:10.6084/m9.figshare.24210468](https://doi.org/10.6084/m9.figshare.24210468)
- [47] Jenny T Liang, Chenyang Yang, and Brad A Myers. 2024. A large-scale survey on the usability of AI programming assistants: Successes and challenges. In *IEEE/ACM International Conference on Software Engineering (ICSE)*. 1–13. [doi:10.1145/3597503.3608128](https://doi.org/10.1145/3597503.3608128)
- [48] Jenny T Liang, Thomas Zimmermann, and Denae Ford. 2022. Understanding skills for OSS communities on GitHub. In *ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE)*. 170–182. [doi:10.1145/3540250.3549082](https://doi.org/10.1145/3540250.3549082)

- [49] Zhicheng Lin. 2024. How to write effective prompts for large language models. *Nature Human Behaviour* 8, 4 (2024), 611–615. [doi:10.1038/s41562-024-01847-2](https://doi.org/10.1038/s41562-024-01847-2)
- [50] Michael Xieyang Liu, Frederick Liu, Alexander J Fiannaca, Terry Koo, Lucas Dixon, Michael Terry, and Carrie J Cai. 2024. "We need structured output": Towards user-centered constraints on large language model output. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*. 1–9. [doi:10.1145/3613905.3650756](https://doi.org/10.1145/3613905.3650756)
- [51] 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 Computing Surveys (CSUR)* 55, 9 (2023), 1–35. [doi:10.1145/3560815](https://doi.org/10.1145/3560815)
- [52] Qingzhou Luo, Farah Hariri, Lamya Eloussi, and Darko Marinov. 2014. An empirical analysis of flaky tests. In *ACM SIGSOFT International Symposium on Foundations of Software Engineering (FSE)*. 643–653. [doi:10.1145/2635868.2635920](https://doi.org/10.1145/2635868.2635920)
- [53] Wanqin Ma, Chenyang Yang, and Christian Kästner. 2024. (Why) is my prompt getting worse? Rethinking regression testing for evolving LLM APIs. In *Proceedings of the IEEE/ACM 3rd International Conference on AI Engineering-Software Engineering for AI*. 166–171. [doi:10.1145/3644815.3644950](https://doi.org/10.1145/3644815.3644950)
- [54] Walid Maalej, Rebecca Tiarks, Tobias Roehm, and Rainer Koschke. 2014. On the comprehension of program comprehension. *ACM Transactions on Software Engineering and Methodology (TOSEM)* 23, 4 (2014), 1–37. [doi:10.1145/26226](https://doi.org/10.1145/26226)
- [55] Gerry McGhee, Glenn R Marland, and Jacqueline Atkinson. 2007. Grounded theory research: Literature reviewing and reflexivity. *Journal of Advanced Nursing* 60, 3 (2007), 334–342. [doi:10.1111/j.1365-2648.2007.04436.x](https://doi.org/10.1111/j.1365-2648.2007.04436.x)
- [56] Aditi Mishra, Bretho Danzy, Utkarsh Soni, Anjana Arunkumar, Jinbin Huang, Bum Chul Kwon, and Chris Bryan. 2025. PromptAid: Visual prompt exploration, perturbation, testing and iteration for large language models. *IEEE Transactions on Visualization and Computer Graphics* (2025). [doi:10.1109/TVCG.2025.3535332](https://doi.org/10.1109/TVCG.2025.3535332)
- [57] Nadia Nahar, Haoran Zhang, Grace Lewis, Shurui Zhou, and Christian Kästner. 2023. A meta-summary of challenges in building products with ML components—Collecting experiences from 4758+ practitioners. In *IEEE/ACM International Conference on AI Engineering–Software Engineering for AI (CAIN)*. 171–183. [doi:10.1109/CAIN58948.2023.00034](https://doi.org/10.1109/CAIN58948.2023.00034)
- [58] Nadia Nahar, Shurui Zhou, Grace Lewis, and Christian Kästner. 2022. Collaboration challenges in building ml-enabled systems: Communication, documentation, engineering, and process. In *IEEE/ACM International Conference on Software Engineering (ICSE)*. 413–425. [doi:10.1145/3510003.3510209](https://doi.org/10.1145/3510003.3510209)
- [59] 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.
- [60] Chris Parnin, Gustavo Soares, Rahul Pandita, Sumit Gulwani, Jessica Rich, and Austin Z Henley. 2023. Building your own product copilot: Challenges, opportunities, and needs. *arXiv preprint arXiv:2312.14231* (2023).
- [61] Harsh Patel, Dominique Boucher, Emad Fallahzadeh, Ahmed E Hassan, and Bram Adams. 2025. A state-of-the-practice release-readiness checklist for generative AI-based software products: A gray literature survey. *IEEE Software* 42, 01 (2025), 74–83. [doi:10.1109/MS.2024.3440190](https://doi.org/10.1109/MS.2024.3440190)
- [62] Jiaxin Pei and David Jurgens. 2023. When do annotator demographics matter? Measuring the influence of annotator demographics with the POPQUORN dataset. In *Linguistic Annotation Workshop (LAW-XVII)*. Association for Computational Linguistics, 252–265. [doi:10.18653/v1/2023.law-1.25](https://doi.org/10.18653/v1/2023.law-1.25)
- [63] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The impact of AI on developer productivity: Evidence from GitHub Copilot. *arXiv preprint arXiv:2302.06590* (2023).
- [64] Crystal Qian, Emily Reif, and Minsuk Kahng. 2024. Understanding the dataset practitioners behind large language models. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*. 1–7. [doi:10.1145/3613905.3651007](https://doi.org/10.1145/3613905.3651007)
- [65] Sebastin Santy*, Jenny T Liang*, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. NLPositionality: Characterizing design biases of datasets and models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*. 9080–9102. [doi:10.18653/v1/2023.acl-long.505](https://doi.org/10.18653/v1/2023.acl-long.505)
- [66] Morgan Klaus Scheuerman, Katta Spiel, Oliver L Haimson, Foad Hamidi, and Stacy M Branham. 2020. HCI guidelines for gender equity and inclusivity. (2020).
- [67] Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2024. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems (NeurIPS)* 36 (2024).
- [68] Beau Sheil. 1986. Datamation®: Power tools for programmers. In *Readings in Artificial Intelligence and Software Engineering*. Elsevier, 573–580. [doi:10.1016/B978-0-934613-12-5.50048-3](https://doi.org/10.1016/B978-0-934613-12-5.50048-3)
- [69] Nancy Staggers and Anthony F Norcio. 1993. Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies* 38, 4 (1993), 587–605. [doi:10.1006/imms.1993.1028](https://doi.org/10.1006/imms.1993.1028)
- [70] Hendrik Strobelt, Albert Webson, Victor Sanh, Benjamin Hoover, Johanna Beyer, Hanspeter Pfister, and Alexander M Rush. 2022. Interactive and visual prompt engineering for ad-hoc task adaptation with large language models. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2022), 1146–1156. [doi:10.1109/TVCG.2022.3209479](https://doi.org/10.1109/TVCG.2022.3209479)

- [71] Harsh Suri. 2011. Purposeful sampling in qualitative research synthesis. *Qualitative research journal* 11, 2 (2011), 63–75. [doi:10.3316/QRJ1102063](https://doi.org/10.3316/QRJ1102063)
- [72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
- [73] Zhiyuan Wan, Xin Xia, David Lo, and Gail C Murphy. 2019. How does machine learning change software development practices? *IEEE Transactions on Software Engineering* 47, 9 (2019), 1857–1871. [doi:10.1109/TSE.2019.2937083](https://doi.org/10.1109/TSE.2019.2937083)
- [74] Zijie J Wang, Aishwarya Chakravarthy, David Munechika, and Duen Horng Chau. 2024. Wordflow: Social prompt engineering for large language models. *arXiv preprint arXiv:2401.14447* (2024).
- [75] 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.
- [76] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv:2302.11382* (2023).
- [77] Jules White, Sam Hays, Quchen Fu, Jesse Spencer-Smith, and Douglas C Schmidt. 2024. Chatgpt prompt patterns for improving code quality, refactoring, requirements elicitation, and software design. In *Generative AI for Effective Software Development*. Springer, 71–108. [doi:10.1007/978-3-031-55642-5_4](https://doi.org/10.1007/978-3-031-55642-5_4)
- [78] Christine T Wolf and Drew Paine. 2020. Sensemaking practices in the everyday work of AI/ML software engineering. In *IEEE/ACM International Conference on Software Engineering Workshops*. 86–92. [doi:10.1109/TSE.2019.2937083](https://doi.org/10.1109/TSE.2019.2937083)
- [79] Tongshuang Wu, Ellen Jiang, Aaron Donsbach, Jeff Gray, Alejandra Molina, Michael Terry, and Carrie J Cai. 2022. Promptchainer: Chaining large language model prompts through visual programming. In *ACM CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI)*. 1–10. [doi:10.1145/3491101.3519729](https://doi.org/10.1145/3491101.3519729)
- [80] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–22. [doi:10.1145/3491102.3517582](https://doi.org/10.1145/3491102.3517582)
- [81] Qinyuan Ye, Mohamed Ahmed, Reid Pryzant, and Fereshte Khani. 2024. Prompt engineering a prompt engineer. In *Findings of the Association for Computational Linguistics ACL (ACL)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.), 355–385. [doi:10.18653/v1/2024.findings-acl.21](https://doi.org/10.18653/v1/2024.findings-acl.21)
- [82] JD Zamfirescu-Pereira, Bjoern Hartmann, and Qian Yang. 2023. Conversation regression testing: A design technique for prototyping generalizable prompt strategies for pre-trained language models. *arXiv preprint arXiv:2302.03154* (2023).
- [83] JD Zamfirescu-Pereira, Heather Wei, Amy Xiao, Kitty Gu, Grace Jung, Matthew G Lee, Bjoern Hartmann, and Qian Yang. 2023. Herding AI cats: Lessons from designing a chatbot by prompting GPT-3. In *ACM Designing Interactive Systems Conference (DIS)*. 2206–2220. [doi:10.1145/3563657.3596138](https://doi.org/10.1145/3563657.3596138)
- [84] JD 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 *ACM CHI Conference on Human Factors in Computing Systems (CHI)*. 1–21. [doi:10.1145/3544548.3581388](https://doi.org/10.1145/3544548.3581388)
- [85] Albert Ziegler, Eirini Kalliamvakou, X Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2022. Productivity assessment of neural code completion. In *ACM SIGPLAN International Symposium on Machine Programming (MAPS)*. 21–29. [doi:10.1145/3520312.3534864](https://doi.org/10.1145/3520312.3534864)
- [86] Albert Ziegler, Eirini Kalliamvakou, X Alice Li, Andrew Rice, Devon Rifkin, Shawn Simister, Ganesh Sittampalam, and Edward Aftandilian. 2024. Measuring GitHub Copilot's impact on productivity. *Communications of the ACM (CACM)* 67, 3 (2024), 54–63. [doi:10.1145/3633453](https://doi.org/10.1145/3633453)

Received 2025-02-25; accepted 2025-04-01