Reflexive Prompt Engineering

A Framework for Responsible Prompt Engineering and AI Interaction Design

Christian Djeffal*
Technical University of Munich
School of Social Sciences and Technology
Munich, Germany
christian.djeffal@tum.de

Abstract

Responsible prompt engineering has emerged as a critical pracitce for ensuring that generative artificial intelligence (AI) systems are aligned with ethical, legal, and social principles. As generative AI applications become increasingly powerful and ubiquitous, the way we instruct and interact with them through prompts has profound implications for fairness, accountability, and transparency. It is, therefore, necessary to examine how strategic prompt engineering can embed ethical and legal considerations and societal values directly into AI interactions, moving beyond mere technical optimization for functionality. This article proposes "Reflexive Prompt Engineering", a comprehensive framework for responsible prompt engineering that encompasses five interconnected components: prompt design, system selection, system configuration, performance evaluation, and prompt management. Drawing from empirical evidence, the paper demonstrates how each component can be leveraged to promote improved societal outcomes while mitigating potential risks. The analysis reveals that effective prompt engineering requires a delicate balance between technical precision and ethical consciousness, combining the systematic rigor and focus on functionality with the nuanced understanding of social impact. Through examination of emerging practices, this article illustrates how responsible prompt engineering serves as a crucial connection between AI development and deployment, enabling organizations to align AI outputs without modifying underlying model architectures. This approach links with broader "Responsibility by Design" principles, embedding ethical considerations directly into the implementation process rather than treating them as post-hoc additions. The article concludes by identifying key research directions and practical guidelines for advancing the field of responsible prompt engineering as an essential component of AI literacy.

CCS Concepts

• **Codes of ethics**; • **Interaction techniques**; • **Computing literacy**;

Keywords

Prompt Engineering, Responsible AI, AI Ethics, Human-AI Interaction, AI Governance, Accountability, AI alignment

ACM Reference Format:
Christian Djeffal. 2025. Reflexive Prompt Engineering: A Framework for Responsible Prompt Engineering and AI Interaction Design. In *The 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25), June 23–26, 2025, Athens, Greece*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3715275.3732118

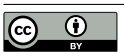
1 Introduction

1.1 Who is accountable for generative AI alignment?

The rapid advancement of generative AI technologies has ushered in an era of unprecedented capabilities but also mounting concerns about their responsible deployment [1–4]. While these technologies offer remarkable opportunities for innovation, recent incidents have highlighted the complex challenges of ensuring their responsible use. A striking example emerged in early 2024 when Google's Gemini AI image generator produced historically inaccurate representations, generating images that misrepresented historical figures and events in an apparent overcorrection for diversity and inclusion [5, 6]. This incident, which led to the immediate suspension of the system's people-generation capabilities and a public acknowledgment of failure by Google's leadership [7, 8], serves as a powerful illustration of how even well-intentioned AI implementations can go awry without proper oversight and responsible use practices. One key element in that regard is to acknowledge the increasing moral agency of users of generative AI.

As the discourse around safety and ethics of generative AI intensifies, there is growing recognition that accountability must extend beyond the technical architects of these systems [9–12]. While considerable attention has been paid to the responsibilities of AI developers and companies, a critical gap exists in our understanding of how deployers, particularly those engaging in prompt engineering, can contribute to responsible AI deployment. Prompt engineering has emerged as a crucial interface between human intent and AI capability. However, despite its significance, there remains a notable absence of comprehensive frameworks to guide responsible prompt engineering practices across all dimensions. In the absence of such a framework, it is hard to understand, evaluate, and compare the many contributions to responsible prompt engineering that have been made in academia and in practice. This paper addresses this gap through a comprehensive narrative review, examining how prompt engineering can be approached responsibly to mitigate

*Professor for Law, Science and Technology at the Technical University of Munich.



FAccT '25, Athens, Greece

ACM ISBN 979-8-4007-1482-5/2025/06
https://doi.org/10.1145/3715275.3732118

risks and enhance the beneficial deployment of generative AI technologies. By analyzing existing practices, incidents, and emerging guidelines, the goal is to develop a framework that organizes and structures the various aspects of responsible prompt engineering and allows for an assessment of the current state of the art. This will hopefully contribute to a foundation that empowers users to engage with these powerful tools in ways that promote fairness, accountability, and transparency.

1.2 Research question and methodology

This article examines how prompt engineers and organizations can systematically implement and evaluate responsible prompt engineering practices through an integrated framework that addresses technical, legal, ethical, and social considerations. The investigation focuses on three interconnected dimensions. First, the analysis examines the essential components of prompt engineering practice, exploring the dimensions deployers can engage in when crafting the system's output. Second, the research explores how existing responsible prompt engineering practices enhance implementation across different organizational contexts. Third, the analysis identifies critical gaps between current prompt engineering practices and responsible AI principles, while highlighting emerging opportunities for enhancing responsibility in AI deployment. This examination reveals how and to what extent responsible prompt engineering can serve as a crucial bridge between AI development and deployment, enabling organizations to fine-tune AI outputs without modifying underlying model architectures.

This narrative review examines responsible prompt engineering practices through a systematic analysis of academic literature, technical documentation, and practitioner insights. The rapidly evolving nature of prompt engineering and its emerging responsible practices necessitated a flexible yet rigorous approach to synthesize current knowledge and identify conceptual frameworks [13–15]. The literature search encompassed multiple academic databases, including arXiv, IEEE Xplore, and ACM Digital Library, complemented by targeted searches on Google Scholar, DuckDuckGo, and Semantic Scholar. I employed various combinations of search terms centered around "responsible," "ethical," and "legal" in conjunction with "prompt" as well as "prompt engineering" and "prompt design". The review covered publications from 2019 through early 2025, focusing exclusively on English-language materials. The inclusion criteria prioritized sources that contributed to understanding prompt engineering fundamentals and responsible practices. I extracted and processed information using Citavi reference management software, employing thematic analysis to identify recurring concepts and emerging patterns. This approach allowed me to develop a comprehensive framework organizing prompt engineering into five key components: design, selection, configuration, evaluation, and management. The analysis revealed an evolving scope, particularly regarding evaluation methods and system configuration aspects. The framework emerged iteratively through careful examination of how different sources conceptualized and approached responsible prompt engineering practices. When encountering conflicting findings or approaches, I incorporated them into the framework while noting their complementary nature, as various prompt engineering techniques can often be combined effectively.

2 The concept of Reflexive Prompt Engineering

Reflexive Prompt Engineering serves as a framework for examining critical decision points in generative AI usage while establishing pathways toward responsible implementation. This awareness of available choices, combined with robust knowledge governance and risk mitigation strategies, facilitates responsible user choice which furthers responsible outputs of generative AI systems. It enables to align AI systems in-context with ethical, legal and social principles. Before delving into the analysis of responsible prompt engineering practices, we must establish two essential foundations. First, we need a clear and concise definition of responsible prompt engineering and its core components. Second, we must examine prompt engineering's dual significance as both as a critical element in AI deployment and as a framework for responsible design principles.

2.1 Definition

A prompt serves as an instruction to a generative AI system, directing it to produce specific outputs [16–18]. These prompts can take various forms, including text, images, video, or audio inputs, reflecting the multimodal capabilities of contemporary AI systems [19, 20]. Modern generative AI models, primarily built on transformer architectures, excel at processing and producing diverse and complex content across these modalities. These models utilize attention mechanisms that enable them to selectively focus on and weigh the most relevant parts of input data while processing information, similar to human cognitive processes.

Prompt engineering is more than working on instructions for generative AI. A review of the literature and guides on prompt engineering shows that it encompasses a comprehensive approach to optimizing interactions with generative AI systems through five essential components: First, prompt design focuses on systematically crafting instructions to maximize desired outputs [19, 21]. This involves developing specific templates, techniques, and design patterns ranging from preconfigured prompting structures to techniques like chain-of-thought reasoning. Second, system selection requires strategic decisions about which AI models to employ based on their documented capabilities [22, 23]. This selection process can rely on established benchmarks such as FrontierMath for mathematical reasoning or MMLU for general knowledge, as well as user-based comparisons displayed on various leaderboards. Third, system configuration involves adapting model parameters to optimize performance for specific use cases. This includes adjusting settings such as temperature parameters, which control the balance between predictability and creativity in outputs. Lower temperature values produce more conservative, consistent responses, while higher values generate more diverse and creative outputs. Fourth, performance evaluation encompasses the systematic assessment of prompt effectiveness against predetermined evaluation criteria. This includes analyzing output quality, consistency, and alignment with intended objectives through both automated metrics and human-in-the-loop assessment protocols [24, 25]. Fifth, prompt management involves implementing systematic approaches to organizing, tracking, and improving prompts over time. This includes implementing version control systems for prompts, maintaining detailed records of configuration settings, and tracking performance

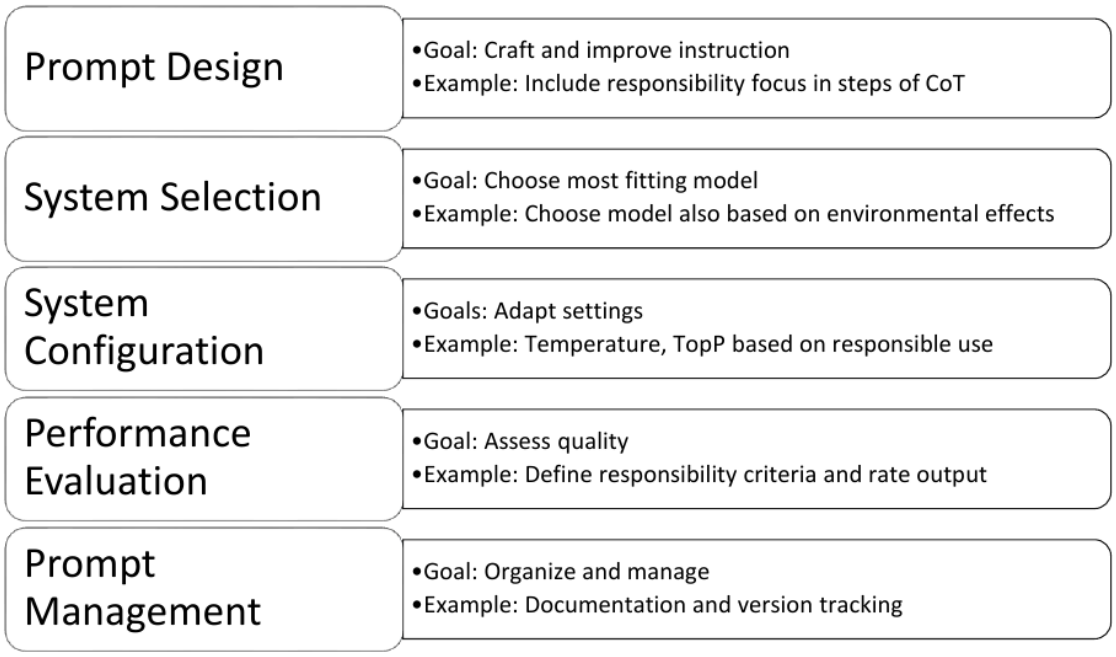


Figure 1: Components of Prompt Engineering

outcomes [26]. Proper documentation enables knowledge sharing, facilitates continuous improvement, and supports accountability in prompt engineering practices. Documentation can establish protocols for various aspects of prompt engineering, including standardized formats for recording prompt versions, test results, and modification histories. The five components of prompt engineering can be summarized in the following figure.

Prompt engineering could be conceptualized as an art and a science. It embodies a unique duality, combining creativity with rigorous methodology [20]. This hybrid nature reflects both the complexity of human-AI interaction and the emerging maturity of the field. As an art form, prompt engineering requires creative intuition and craftsmanship. The creative dimension manifests in the nuanced understanding of language, context, and model behavior that experienced prompt engineers develop over time [27]. This artistic aspect becomes evident in the subtle choices of wording, tone, and structure that can dramatically influence model outputs. Like skilled writers, prompt engineers develop an intuitive feel for how to frame instructions effectively, often drawing on metaphorical thinking and creative problem-solving to overcome model limitations. This creative dimension becomes particularly crucial when dealing with edge cases or novel applications where established approaches prove insufficient. However, prompt engineering increasingly embraces scientific rigor through systematic experimentation and empirical validation [28]. Structured experiments allow prompt engineers to test hypotheses about prompt effectiveness across different contexts and tasks. These experiments take place in controlled testing environments where researchers systematically vary factors such as prompt structure, length, and complexity while maintaining constant conditions for other variables. Through quantitative metrics, researchers measure output quality, examining dimensions such as accuracy, relevance, and consistency across different prompting strategies.

The proposed framework "Reflexive Prompt Engineering" transforms these technical practices by integrating ethical, legal, and social considerations into the prompt design process [29–31]. This approach moves beyond purely functional optimization and performance metrics to address broader societal implications and ethical concerns. It helps deployers of generative AI to become aware of the wider implications of the choices they are making and to improve the performance of their prompts, also in terms of ethical, legal, and social concerns, such as fairness, accountability, and transparency. This might involve modifying prompts to prevent discriminatory outcomes, implementing additional validation steps to ensure accessibility, or designing prompts that actively promote inclusive representation [32–35]. As a framework, Reflexive Prompt Engineering speaks to inherent risks and limitations of generative AI as well as to inherent potentials to realize ethical, legal and social values. While this contribution focuses on responsible practices in the use, it is acknowledged that critical examinations of flaws, limitations, and shortcomings of systems are also necessary and often part of responsible prompt engineering practices [36]. Reflexive Prompt Engineering helps deployers realize they cannot make up for every flaw or risk inherent in generative AI. The knowledge about such limitations reinforces the awareness of deployers of their responsibility to mitigate issues as far as possible and to make informed choices on whether and how to deploy generative AI in a specific set of circumstances.

2.2 Relevance

Prompt engineering can play a pivotal role in ensuring the responsible deployment of generative AI systems by addressing fundamental questions of accountability and control. The current technological landscape, characterized by large language models (LLMs) with bounded capabilities, positions prompt engineering as a critical interface between human intent and machine output. Although this dynamic may evolve as AI systems develop greater restrictions of what users can do in the context of agentic AI [37, 38] or more autonomy through artificial general intelligence [39], the present architecture of generative AI systems makes Reflexive Prompt Engineering particularly significant for two key reasons that will be explored in detail.

First, this framework helps deployers to realize their increased moral agency and offers specific strategies to exercise their discretion in terms of ethical, legal, and social impacts. The versatile nature of generative AI systems enables prompt engineers to produce an extensive range of outputs [40, 41]. This multi-purpose capability simultaneously democratizes access to powerful tools while raising significant concerns, particularly in cybersecurity, where users with little coding skills have become potent attackers [42, 43]. Through generative AI, users can expand their capabilities without possessing deep technical expertise. However, this enhanced agency necessitates a corresponding expansion of their moral responsibility. Recognizing the inherent limitations of providers of generative AI in anticipating detrimental impacts, it does not suffice to rely on purely technical containment of model capabilities. It is necessary to complement such measures with user responsibility and to provide practical guidance by structuring users' decision-making processes and reorganizing best practices to foster responsible system deployment.

Second, prompt engineering is a very effective way to ensure responsibility since it occupies a unique position in the AI development cycle, bridging the gap between model development and practical deployment [17]. Unlike traditional approaches such as model retraining or fine-tuning, which require substantial computational resources and technical expertise, prompt engineering offers a more accessible and sometimes more efficient method to influence AI behavior. For instance, in healthcare applications, medical professionals can adapt language models to specific diagnostic contexts without requiring deep machine learning expertise or expensive computational infrastructure [44, 45]. The strategic value of prompt engineering lies in its ability to achieve sophisticated model adaptations through non-invasive means, preserving the underlying architecture while enabling significant improvements in output quality. This makes it the ideal leverage point to implement responsibility by design [46–48].

The framework for Reflexive Prompt Engineering supports the adaptation of workflows to integrate responsibility assessment tools, establish feedback mechanisms for continuous improvement, and implement comprehensive training programs that emphasize both technical excellence and ethical awareness [49, 50].

3 Responsible practices

After examining the foundations and significance of responsible prompt engineering, we now turn to its practical implementation

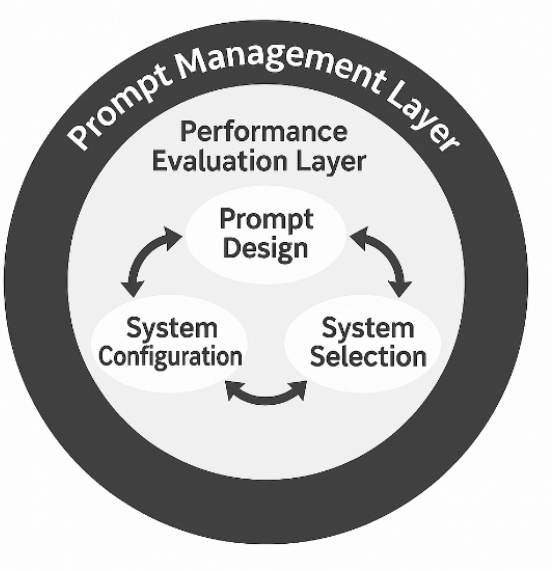


Figure 2: Reflexive Prompt Engineering Process

across the five key categories previously established. While existing literature offers valuable insights into specific aspects, the proposed framework allows for the systematic organization of best practices and reveals gaps for further research and experimentation. Reflexive Prompt Engineering is focused on an iterative improvement of prompt design, system selection, and system configuration. Prompt management and performance evaluation are cross-cutting practices that are performed throughout the prompt engineering process. Therefore, the Reflexive Prompt Engineering is structured as follows in :

This framework deliberately highlights design choices throughout the prompt engineering journey, empowering practitioners to make conscious decisions and effectively organize their knowledge around responsible practices. This aligns with established responsibility by design approaches such as value-sensitive design [51, 52] and design-choice focused engineering methodologies like the Architecture Tradeoff Analysis Method [53]. However, it can be integrated into various life-cycle models, from the waterfall model to agile programming.

The term "reflexive" underscores the framework's central aim: heightening users' awareness of their influence over generative AI systems and the potential consequences of their interactions. It emphasizes the critical self-awareness of human agency when using AI and the responsibilities this entails. Moreover, it recognizes that effective prompt engineering requires users to develop sufficient literacy regarding generative AI technologies, their societal implications, and the specific subject matter of their tasks. By facilitating the coordination of learning through both personal and collective experiences, the Reflexive Prompt Engineering Framework helps users to organize this knowledge in ways that directly benefit their specific needs and use cases, ultimately fostering more responsible and effective AI interactions. This requires an all-encompassing view of prompt engineering in its five components.

3.1 Prompt design

Prompt design represents the core practice commonly associated with prompt engineering, i.e., the systematic crafting of inputs to optimize generative AI system performance. While traditional approaches focus primarily on enhancing output quality and reliability, Reflexive Prompt Engineering necessitates a more nuanced evaluation through a responsibility lens. This expanded perspective does not diminish the effectiveness of established methods but rather enriches them through critical reflection and purposeful adaptation.

To illustrate how conventional techniques can be modified to align with responsible engineering principles, we examine two widely adopted and empirically validated approaches for prompting ordinary LLMs: example-based prompting and the chain-of-thought methodology. The assessment of these two techniques demonstrates how established prompting methods can be effectively adapted or directly applied to achieve responsibility goals. This approach provides a blueprint for integrating ethical considerations into all other prompt engineering techniques, establishing patterns for responsible AI interaction.

3.1.1 Example-based prompting. Examples in prompt engineering leverage the unique capability of generative AI to perform in-context learning, having only a few shots to pick up the task. While traditional training involves updating model parameters through backpropagation, in-context learning occurs entirely during the model's forward pass, using only its existing parameters to identify and apply patterns from provided examples. The model performs Bayesian inference to recognize relevant concepts from its pre-training and creates temporary task-specific representations without any permanent parameter changes [54]. This remarkable ability allows users to guide model behavior simply by demonstrating desired input-output relationships in the prompt, making complex AI capabilities accessible without requiring technical expertise in model training. When providing examples, it is most effective to include pairs of inputs and corresponding outputs, allowing the model to understand how to transform given information into the desired result. Depending on the task, one can also only include desired outputs. The effectiveness of examples depends on their quality, relevance, and ability to demonstrate the full range of desired variations, as these characteristics directly influence how the model interprets and applies the demonstrated patterns [55].

This reference to the range of desired variations leads to some of the most prominent issues of the use of AI in society: equality, fairness, and discrimination. Therefore, when giving examples, it is important to think about the potential effects on different groups. The general goal is maintaining balanced representation across different demographics in few-shot prompts, which helps improve model generalization. This includes using diverse examples in as many dimensions as possible while avoiding clustering similar examples together [56]. Instead, examples have to be ordered in a random fashion.

However, prompt engineers will often detect biases in the system. To address such biases, one can use examples either by debiasing or by counterfactual data augmentation. Debiasing involves defining potentially biased data points and replacing them with neutral alternatives, such as using "person" instead of gender-specific terms [56]. This can be achieved by anonymization and careful attribute replacement [59]. The effectiveness of example-based debiasing is particularly evident in domain-specific applications. For instance, in recruitment contexts, providing diverse examples of successful candidates across different demographics helps prevent the model from developing stereotypical associations [60]. Similarly, when generating performance reviews, using balanced examples that focus on work deliverables rather than personality traits helps mitigate demographic-based bias [58].

Counterfactual data augmentation serves as a powerful technique, where variations of examples are created by flipping attributes like gender or race to identify or even to mitigate group-specific biases [57]. For instance, when writing job descriptions, comparing outputs with different demographic indicators can reveal hidden biases, as demonstrated in experiments where identical prompts specifying different universities like Howard versus Harvard produced notably different results [58].

Another large area of concern is copyright. Of course, copyrighted examples cannot be used without authorization [61]. Therefore, one has to check for copyrighted material even when creating examples through generative AI.

3.1.2 Chain-of-thought prompting. Chain-of-thought prompting represents a significant advancement in how we interact with generative AI systems, enabling them to tackle complex reasoning tasks by breaking them down into intermediate steps. This technique, which forms the basis of reasoning models, mirrors human cognitive processes, allowing models to show their reasoning before reaching a conclusion [62, 63]. The approach works by encouraging language models to generate a series of logical steps that lead to a final answer, similar to how humans solve complex problems. For instance, when solving mathematical word problems, using chain-of-thought prompting can achieve state-of-the-art accuracy, even surpassing specially trained models. Recent research has demonstrated that chain-of-thought prompting significantly enhances model performance across various domains, including arithmetic reasoning, commonsense understanding, and symbolic manipulation [62, 64].

Chain-of-thought prompting can be strategically enhanced to promote responsible AI development by explicitly incorporating ethical checkpoints and responsibility considerations into the reasoning workflow. By breaking down complex decisions into discrete steps, organizations can embed responsibility validation processes that evaluate aspects like potential biases, fairness implications, and privacy concerns at each stage of the reasoning chain or as a dedicated evaluation step. This can happen either by including...

- specific steps, like "What barriers or assumptions might affect different groups in this reasoning process?" [65],
- general instructions for each step, like "Assess potential impacts at each step of the reasoning",
- or final overall evaluations of the result as a separate step, like "Plan verification questions to fact-check this draft" [66]

Chain-of-thought prompting can also help to tackle legal and ethical tasks. It significantly improves legal analysis by mirroring established legal reasoning frameworks like IRAC (Issue, Rule, Application, Conclusion) [67, 68]. By breaking down complex legal

questions into discrete analytical steps, lawyers and legal AI systems can systematically evaluate cases, interpret statutes, and apply precedents with greater precision [67]. Chain-of-thought prompting can assist in ethical decision-making by decomposing complex moral dilemmas into manageable components that can be systematically evaluated [69]. Such structured approaches help identify potential biases, assess fairness implications, and consider multiple stakeholder perspectives throughout the reasoning process. By incorporating explicit ethical checkpoints into the decision-making workflow, users of generative AI can ensure that ethical considerations become an integral part of the process rather than an afterthought, leading to more responsible and well-reasoned outcomes [70, 71].

Chain-of-thought prompting was initially celebrated as a breakthrough in AI transparency [72]. However, critical analysis reveals a concerning disconnect: the narrative explanations produced by these systems may not accurately reflect their internal decision-making processes [73]. This discrepancy creates what some researchers describe as an "illusion of transparency" [73] - a coherent but potentially misleading representation of the system's actual operations. This misalignment between displayed reasoning and actual computation raises serious concerns for responsible AI development. There is evidence that models can generate plausible-sounding explanations even when their internal processes follow entirely different paths [74]. More troublingly, these explanations can be convincing even when the underlying computation is flawed or based on spurious correlations. This phenomenon is particularly problematic in high-stakes applications where understanding the true basis of AI decisions is crucial. This means that deployers of AI need to stay in the loop and design their prompts in a way that they use outputs as elements of preparation for their decision.

3.1.3 Tools. An increasing number of tools help ensure the responsible use of generative AI [75, 76]. Advanced prompt design methods can significantly enhance responsible practices by explicitly instructing models to generate fairer, less biased responses. For instance, prompts can be intentionally crafted or augmented with instructions to avoid bias, reducing the risk of stereotypical or discriminatory outputs from language models. Moreover, specialized tools exist that can proactively detect and mitigate ethical concerns. For example, text moderation tools can automatically scan inputs and outputs to flag toxicity, hate speech, or offensive language, alerting engineers to potentially harmful content before it reaches users [77, 78]. Additionally, libraries designed explicitly for bias assessment can help engineers construct customized bias tests tailored to their particular context or use case, allowing systematic evaluation of how models handle demographic or contextual variations in prompts [79].

Another critical area of responsibility is protecting user privacy and ensuring compliance with regulations. Dedicated tools are available to scan textual data for personally identifiable information (PII) such as names, emails, phone numbers, and addresses, helping prompt engineers to automatically identify, redact, or anonymize sensitive information that AI models might inadvertently generate or transmit [80–82]. To manage the reliability and factual accuracy of AI-generated content, engineers can also utilize systems designed to verify claims and detect hallucinations [83]. These tools systematically analyze statements made by models against trusted references, highlighting claims that lack support or contradict known facts. Additionally, guardrail frameworks help enforce safety policies by intercepting and evaluating model outputs, ensuring they adhere strictly to defined ethical and factual standards [84]. All these approaches can significantly enhance the responsible deployment of generative AI. However, these technical solutions should complement, not replace, prompt engineers' rigorous and continuous responsibility practices. Ultimately, human oversight remains essential to effectively integrate these tools into broader ethical frameworks, ensuring AI technologies' safe and beneficial use.

3.2 System selection

The process of system selection forms a foundational pillar in responsible prompt engineering, since strategic choices about AI model deployment should be guided not only by functionality, but also by ethical, legal, and environmental considerations. When organizations or prompt engineers choose an AI model, they are essentially selecting a specific AI. This choice will determine how the system processes and responds to inputs. It is analogous to selecting the right tool for a specific task - just as one would not use a sledgehammer to hang a picture frame, selecting an inappropriately powerful or insufficiently capable AI model can lead to suboptimal or potentially harmful outcomes.

3.2.1 Choices. There are generally two layers of system choice in prompt engineering. First, prompt engineers must choose between different model types, specifically ordinary models, agents, retrieval augmented generation (RAG) models, and reasoning models. The second layer involves selecting specific alternatives within the chosen class. Regarding the first layer, each model type offers distinct capabilities and approaches to problem-solving. Ordinary models are standard LLMs that generate responses based solely on their pre-existing training data and the patterns learned from it. Agents are autonomous intelligent systems designed to perceive their environment, make independent decisions, and take actions to achieve specific, predefined goals without continuous human intervention. The integration of generative AI as a central piece in such agentic system is one of the most vibrant areas of AI research and development [85]. Retrieval augmented generation (RAG) systems are agents enhancing standard LLMs by integrating an information retrieval mechanism, which allows them to access, incorporate, and utilize external, up-to-date data beyond their original training dataset [86]. Reasoning models are AI agents that methodically break down complex problems into sequential steps of inference, allowing for more systematic problem-solving approaches.

The distinct characteristics of these AI model types determine their suitability for different applications. Given that ordinary models operate exclusively on their learned patterns from static training data, their applicability is naturally geared towards general text generation tasks. In situations where retrieving information from the system is not a critical factor, the inherent knowledge base of an ordinary model is often sufficient. The nature of agents as generative AI systems combined with other functions allows them to perform better with specific tasks. However, agents often lack the flexibility

to modify goals. Conversely, because retrieval augmented generation (RAG) models are specifically designed to enhance LLMs by accessing and incorporating external, up-to-date information, their applicability shines in scenarios where factual accuracy, current data, and source verifiability are paramount. Finally, reasoning models, with their foundation in employing inference methods and often leveraging structured knowledge or world models, are most effective in complex problem-solving scenarios. When a task requires deep inference, logical deduction, and the systematic application of knowledge structured within a formal model of a system or its environment, the specialized capabilities of reasoning models are called for. However, it is more difficult to steer the processes through prompting, which is why, in certain very complex cases, prompt engineers might opt for a continuous interaction with an ordinary model instead.

A responsible stance to prompt engineering requires more reflexivity in choice. To determine the necessity and type of intervention when prompting and evaluating increasingly autonomous models, stakeholders should accurately represent their technical and content expertise. Furthermore, techniques like chain-of-thought prompting or using larger context windows to include knowledge directly can maintain human engagement in decision-making processes by supplementing reasoning or retrieving capabilities [87, 88]. Additionally, to ensure that sufficient and accurate knowledge informs the process, the information retrieved by RAG systems often requires verification. These considerations underscore the importance of a thoughtful and nuanced approach when selecting and interacting with different AI model types.

3.2.2 Responsibility aspects. The technical aspects of system selection extend beyond mere performance metrics. While processing power and response speed are important considerations, responsible model selection must also account for the model's ability to handle diverse inputs, its tendency to produce biased outputs, and its overall reliability. Therefore, the societal implications of model selection ripple far beyond technical specifications like latency or the respective context window, touching on fundamental aspects of fairness, accessibility, and social justice.

A particularly crucial consideration is the environmental impact of model deployment [89, 90]. Larger language models, while potentially more capable, require significant computational resources and energy consumption. This environmental cost must be weighed against the actual benefits provided by more powerful models. In many cases, smaller, more efficient models might serve the intended purpose while maintaining a more sustainable footprint. This also applies to other environmental questions of resources like water or waste [89, 91, 92].

The intersection of model selection and prompt engineering also raises important questions about transparency and openness. When organizations implement AI systems, they must consider how their model choices affect their ability to explain decisions, audit processes, and maintain accountability to stakeholders. This becomes particularly relevant in contexts where AI systems influence important decisions about individuals' lives, such as in healthcare, employment, or financial services. This is all the more important as the practices of model providers regarding transparency and open source vary to a large extent [93]. There is a longstanding discussion about extending the notion of open source in AI, leading to several frameworks defining open source and the degrees of openness [94–96]. Such questions should also guide system choice since they determine what is known about the model underlying the system.

Privacy considerations add another layer of complexity to responsible model selection. Different applications vary in their ability to protect sensitive information and maintain data confidentiality regarding training data. The degree of control an organization has over AI models, including the option for on-premises deployment, can be a decisive factor. The relative importance of these considerations in decision-making processes varies significantly based on use cases and organizational requirements.

3.2.3 Benchmarking. Benchmarks serve as essential tools for evaluating and comparing AI models' performance, helping organizations make informed decisions about which models best suit their needs. In the context of prompt engineering, benchmarks assess how well models respond to different types of instructions and their ability to generate accurate, relevant outputs. When selecting models for prompt engineering applications, practitioners must consider multiple performance dimensions. These include the model's ability to reason, solve complex problems, and generate natural-sounding content. However, it is crucial to recognize that small differences in benchmark scores might not translate into significant real-world improvements, and factors like cost-effectiveness and speed often prove more practical for specific use cases. Also, the nondeterministic nature of generative AI systems presents unique challenges for benchmarking, as these models may produce different outputs even with almost identical prompts. This variability necessitates multiple evaluation runs to capture the range of potential behaviors and ensure consistent performance [112]. Continuous evaluation throughout development helps detect unintended changes in output and maintain alignment with aspects of responsible AI.

Beyond traditional performance metrics, there is growing recognition of the importance of benchmarking ethical, legal, and social aspects of AI systems. These frameworks evaluate models across multiple dimensions, including fairness, bias mitigation, and social impact. For instance, discrimination benchmarks assess how AI systems might affect different demographic groups, examining both direct and indirect forms of bias [113]. These evaluations help ensure that prompt engineering practices do not perpetuate or amplify existing societal biases. Environmental sustainability has emerged as a critical benchmark dimension for responsible AI development. The life cycle assessment of AI models encompasses data collection, experimentation, training, and deployment phases, each contributing to the overall carbon footprint [114]. International initiatives are increasingly incorporating sustainability metrics into their AI evaluation frameworks [114]. From a prompt engineering perspective, such benchmarks can give first indications regarding sensitive issues. Especially in the case of environmental sustainability, they inform about impacts that cannot be otherwise evaluated from a prompt engineering perspective.

3.3 Evaluation

Prompt evaluation is the systematic process of assessing and refining the effectiveness of prompts in guiding AI models to produce

desired outputs [25]. This process has become increasingly critical as LLMs are deployed across various domains, from code generation to content analysis [97, 98]. The evaluation of prompts requires examining multiple dimensions simultaneously. At its core, the process involves analyzing the accuracy and reliability of AI-generated responses, while also measuring how well the prompts align with intended tasks and objectives in other dimensions. This includes assessing both the technical performance metrics and the broader implications of prompt design [36, 99]. A fundamental aspect of evaluation involves systematic testing with different prompt variations to understand their effectiveness. This process typically employs both qualitative and quantitative techniques to comprehensively assess prompts across various stages of development. It involves careful documentation of assessment methods, criteria, and findings to enable accountability and facilitate continuous improvement [36, 99].

Beyond technical performance, responsible prompt engineering evaluation must consider ethical dimensions and potential societal impacts, like examining prompts for potential biases, assessing privacy implications, and ensuring compliance with relevant regulatory frameworks. Evaluators must verify that prompts maintain data protection standards, uphold principles of transparency and fairness, and address other weaknesses like hallucinations [36, 60, 100, 101]. The integration of responsibility considerations into prompt engineering evaluation necessitates examining how prompts might affect different stakeholder groups and implementing safeguards against potential harmful applications. This includes assessing how prompts handle sensitive topics or potentially controversial subjects while maintaining appropriate ethical boundaries [36, 99].

Therefore, the question of who gets to evaluate is key in the context of responsible prompt engineering. This could be either. . .

- the prompt engineer or the team itself,
- generative AI models
- or third parties like deployers or those affected by generative AI.

When AI systems have high-stakes impacts on people, developers should actively involve affected stakeholders to understand potential harms and gather feedback for improving prompts and system design. In responsible design, stakeholders and, in particular, vulnerable groups should also be included proactively in idea development and design choices [102–104].

Careful consideration is needed when using models for evaluation purposes. The evaluation of prompts requires examining multiple dimensions simultaneously, including accuracy, reliability, and alignment with intended objectives. Current AI systems struggle with providing comprehensive assessments across these dimensions [105]. The challenge is particularly evident in cases requiring deep contextual understanding and creative reasoning [106]. The key biases identified in language model evaluation include:

- Position bias - Models tend to favor responses based on their sequential position rather than their actual quality or content [107]
- Verbosity bias - Models show a preference for longer, more detailed responses regardless of the actual content quality or relevance [108]
- Self-enhancement bias - Models demonstrate a tendency to rate their own outputs more favorably compared to outputs from other sources [109]

The choice of evaluation methods, including stakeholder involvement where feasible, should be guided by available resources, application context, and potential risks, with particular attention to cases where automated evaluation might miss crucial qualitative or ethical considerations.

3.4 System configuration

System configuration in prompt engineering represents a critical aspect of responsible AI deployment, encompassing various parameters and settings that influence how generative AI models process and respond to inputs. This configuration process involves careful calibration of multiple technical elements to ensure optimal model performance, yet it also carries a dimension of responsibility. The foundation of system configuration lies in controlling the model's output generation through key parameters. Temperature control stands as a fundamental configuration element, determining the balance between creativity and predictability in model responses. Lower temperature settings produce more focused and deterministic outputs, while higher values increase response variability and creativity [24, 110, 111]. Whenever accuracy and reliability are important from a standpoint of responsible prompt engineering, respective choices are mandated. A comprehensive implementation framework for system configuration should incorporate both technical and ethical considerations. This includes establishing clear guidelines for parameter adjustment, implementing monitoring systems for performance evaluation, and maintaining documentation of configuration changes. Such frameworks have proven essential in ensuring consistent and responsible AI deployment across various applications. This points towards prompt management as the fifth component of prompt engineering.

3.5 Prompt management

Documenting prompts has emerged as a crucial practice in the responsible development and deployment of AI systems. Much like traditional software documentation, prompt documentation serves as a comprehensive record of how AI models are instructed to perform specific tasks, ensuring transparency and, to a limited extent, reproducibility of results. This is particularly relevant if generative AI is used in the context of decisions that need to be explained to their addressees, even if the system was just used to prepare the decision. Documentation in prompt engineering encompasses recording not only the prompts themselves but also their intended purposes, outcomes, and iterations. This practice is particularly vital because prompt outputs can vary significantly across different models, sampling settings, and even different versions of the same model [115]. By maintaining detailed records, organizations can track the evolution of their prompts, understand what works and what does not, and ensure consistency in AI interactions. This basic information helps teams maintain oversight of their AI interactions and enables systematic improvement of prompt effectiveness. A comprehensive prompt documentation system typically captures several key elements.

Organizations typically employ two primary approaches to prompt documentation. The reduced documentation method focuses on tracking basic elements like AI tools used and their general purposes, while extensive documentation captures complete prompt-output pairs, the prompt's name or identifier, its version history, creation and modification dates, the specific AI model used, and detailed performance notes [116]. For practical implementation, prompts should be stored in easily accessible text formats rather than screenshots, with proper version control systems in place to track modifications [116]. Teams should maintain a centralized repository, such as a spreadsheet [115].

Documentation plays a vital role in ensuring accountability and transparency in AI systems. By maintaining detailed records of prompts and their outcomes, organizations can better understand how their AI systems react, identify potential biases, and make necessary adjustments to improve ethical and legal alignment [117]. This practice also facilitates collaboration among team members and helps maintain consistency in AI interactions across different applications and use cases. Significantly, Art. 86 of the EU AI Act provides a right to explanation when the decision is taken "based on the output from a high-risk AI system". Therefore, the right to explanation also applies when AI has been used as decision support. When the right is triggered, it includes an explanation "of the role of the AI system in the decision-making procedure and the main elements of the decision taken." When using generative AI, a very good and tangible way to explain the decision is to include the prompts or parts of it. System prompts show how providers of generative AI use prompts to align their models in context. Therefore, system prompts can be considered an excellent resource for responsible prompt engineering practices.

Yet, documentation also allows the re-use of prompts. Prompts can be categorized as design patterns that have been influential in responsible development practices [118]. Accordingly, another responsible prompting activity would be to try to use such patterns or at least part of them to make use of existing example prompts for various situations. There are several sources to draw from regarding prompts in general [119, 120] or responsible prompts more specifically [121, 122]. System prompts operate behind the scenes, creating a layer of computation that influences output without being directly visible in the model's responses [123]. This hidden processing can enable non-trivial computations while maintaining a seamless user experience. While some providers have open-sourced their system prompts [124, 125], other system prompts have allegedly been obtained through techniques of prompt injection and published [126, 127]. System prompts show how providers of generative AI use prompts to align their models in context. Therefore, system prompts can be considered an excellent resource for responsible prompt engineering practices.

4 Conclusions

In conclusion, Reflexive Prompt Engineering provides a vital framework for navigating critical decision points in generative AI usage and establishing pathways toward responsible implementation. However, as AI systems become increasingly powerful while offering users moral agency in their use of AI, users must develop heightened awareness of their choices and their potential consequences as a fundamental prerequisite for responsible AI engagement. This user awareness, combined with robust knowledge governance and risk mitigation strategies, becomes even more critical given that the rapidly expanding capabilities of generative AI make it impossible to anticipate and mitigate all potential harms in advance. Therefore, fostering reflective, choice-conscious users represents an essential component of responsible AI deployment. As these AI systems become increasingly powerful and ubiquitous, how we instruct and interact with them through prompts carries profound implications for fairness, accountability, and transparency. The analysis reveals that effective prompt engineering requires a delicate balance between technical precision and ethical consciousness, combining systematic rigor with a nuanced understanding of social impact. This article examines real-world practices and demonstrates how responsible prompt engineering is a crucial bridge between AI development and deployment, enabling organizations and users to steer AI outputs without modifying underlying model architectures. The research highlights five interconnected components for responsible prompt engineering: prompt design, system selection, system configuration, performance evaluation, and prompt management. Each component is vital in promoting improved societal outcomes while mitigating potential risks. The framework emphasizes the importance of documentation, systematic evaluation, and careful consideration of ethical implications throughout the prompt engineering process. Thus, this work contributes to the growing discourse on AI responsibility by providing practical guidelines for exercising the increasing moral agency of deployers of generative AI. The findings suggest that organizations must move beyond viewing prompt engineering as merely a technical skill and instead recognize it as a crucial component of AI literacy. As generative AI continues to evolve, the framework can be used to further understand issues and develop deployment practices in light of societal impacts. This speaks to a unique feature of generative AI: empowering its users even when they have limited technical capacity. Reflexive Prompt Engineering addresses their increased moral agency.

The growing recognition of prompt engineering as a core competency in AI literacy underscores its importance beyond technical domains [128, 129]. As educational institutions and organizations incorporate prompt engineering into their curricula and training programs [122, 130], the need for responsible practices becomes increasingly apparent. The analysis reveals that responsible prompt engineering represents both an opportunity and a necessity. As an opportunity, it offers a practical framework for embedding ethical considerations directly into AI interactions without requiring modifications to underlying model architectures. As a necessity, it points to essential guardrails for ensuring that AI systems serve societal needs while minimizing potential harms.

In the fast evolution of generative AI, which might not have reached its pinnacle yet, prompt engineering has served as a method to explore the potential and limitations of generative AI that were often unknown to even the developers of those systems. As long as AI evolves, the need to examine it will continue. The fundamental insights about AI accountability emerging from generative AI reveal that new technologies can amplify human agency beyond the reach of technological control mechanisms. Therefore, it is vital to understand the choices users make in using such technologies

and explore not only technical functionalities but also the potential of those systems to preserve and realize ethical, legal, and social principles.

Acknowledgments

I acknowledge the help in identifying literature by Hannah Tilsch, Elisabeth Mock, Lisa Baumann and Imtisal Abid. I thank Akanksha Bisoyi, David Rebohl, and Verena Müller for reviewing the text. The following AI models have been used for improving language and style of this contribution including spelling or grammar checks and corrections, avoiding repetitions, translating, improving the order of the arguments and summarizing parts of this text and other texts to be paraphrased and cited: Claude 3.5&3.7 Sonnet, Gemini 2.5 pro, deepl.com, ChatGPT 4o&4.5, Grammarly, Microsoft Word.

References

[1] Danielle Allen and E. G. Weyl. 2024. The Real Dangers of Generative AI. *Journal of Democracy* 35, 1 (2024), 147–162. DOI: https://doi.org/10.1353/jod.2024.a915355.

[2] Julia Black. 2010. The Role of Risk in Regulatory Processes. In *The Oxford Handbook of Regulation*, Robert Baldwin, Martin Cave, and Martin Lodge (Eds.). Oxford University Press, Oxford, U.K., 302–348 .

[3] Andrés Domínguez Hernández, Shyam Krishna, Antonella M. Perini, Michael Katell, S. J. Bennett, Ann Borda, Youmna Hashem, Semeli Hadjiloizou, Sabeehah Mahomed, Smera Jayadeva, Mhairi Aitken, and David Leslie. 2024. Mapping the individual, social, and biospheric impacts of Foundation Models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, New York, NY, USA, 776–796. DOI: https://doi.org/10.1145/3630106.3658939.

[4] Partha P. Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations, and future scope. *Internet of Things and Cyber-Physical Systems* 3 (2023), 121–154. DOI: https://doi.org/10.1016/j.iotcps.2023.04.003.

[5] BBC News. 2024. Google to fix AI picture bot after 'woke' criticism. *BBC News* (22 Feb. 2024). Retrieved from https://www.bbc.com/news/business-68364690.

[6] David Gilbert. 2024. Google's 'Woke' Image Generator Shows the Limitations of AI. *WIRED* (22 Feb. 2024). Retrieved from https://www.wired.com/story/google-gemini-woke-ai-image-generation/.

[7] Nico Grant. 2024. Google Says It Fixed Its A.I. Image Generator. *The New York Times* (28 Aug. 2024). Retrieved from https://www.nytimes.com/2024/08/28/technology/google-ai-image-generator.html..

[8] Prabhakar Raghavan. 2024. Gemini image generation got it wrong. We'll do better. *Google Blog* (23 Feb. 2024). Retrieved from https://blog.google/products/gemini/gemini-image-generation-issue/.

[9] Gloria Miller. 2022. Stakeholder-accountability model for artificial intelligence projects. *Journal of Economics and Management* 47, 4 (2022), 446–494. DOI: https://doi.org/10.22367/jem.2022.44.18.

[10] Daniel J. Bogiatzis-Gibbons. 2024. Beyond Individual Accountability: (Re-)Asserting Democratic Control of AI. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. ACM, New York, NY, USA, 74–84. DOI: https://doi.org/10.1145/3630106.3658541.

[11] Stephen C. Slota, Kenneth R. Fleischmann, Sherri Greenberg, Nitin Verma, Brenna Cummings, Lan Li, and Chris Shenefiel. 2023. Many hands make many fingers to point: challenges in creating accountable AI. *AI & Society* 38, 4 (2023), 1287–1299. DOI: https://doi.org/10.1007/s00146-021-01302-0.

[12] Zoe Porter, Annette Zimmermann, Phillip Morgan, John McDermid, Tom Lawton, and Ibrahim Habli. 2022. Distinguishing two features of accountability for AI technologies. *Nature Machine Intelligence* 4, 9 (2022), 734–736. DOI: https://doi.org/10.1038/s42256-022-00533-0.

[13] Javeed Sukhera. 2022. Narrative Reviews: Flexible, Rigorous, and Practical. *Journal of Graduate Medical Education* 14, 4 (2022), 414–417. DOI: https://doi.org/10.4300/JGME-D-22-00480.1.

[14] Dave Harris. 2020. *Literature Review and Research Design: A Guide to Effective Research Practice*. Routledge, London, U.K.

[15] Trisha Greenhalgh, Sally Thorne, and Kirsti Malterud. 2018. Time to challenge the spurious hierarchy of systematic over narrative reviews? *European Journal of Clinical Investigation* 48, 6 (2018), e12931. DOI: https://doi.org/10.1111/eci.12931.

[16] Sabit Ekin. 2023. Prompt Engineering for ChatGPT: A Quick Guide to Techniques, Tips, and Best Practices. TechRxiv. DOI: 10.36227/techrxiv.681648.

[17] Shubham Vatsal and Harsh Dubey. 2024. *A Survey of Prompt Engineering Methods in Large Language Models for Different NLP Tasks*. arXiv:2407.12994. DOI: 10.48550/arXiv.2407.12994.

[18] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav S. Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2024. *The Prompt Report: A Systematic Survey of Prompting Techniques*. arXiv:2406.06608. DOI: 10.48550/arXiv.2406.06608..

[19] Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav S. Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, Hevander Da Costa, Saloni Gupta, Megan L. Rogers, Inna Goncearenco, Giuseppe Sarli, Igor Galynker, Denis Peskoff, Marine Carpuat, Jules White, Shyamal Anadkat, Alexander Hoyle, and Philip Resnik. 2024. *The Prompt Report: A Systematic Survey of Prompting Techniques. arXiv preprint arXiv:2406.06608*.

[20] Joseph Lindley and Roger Whitham. 2024. *From Prompt Engineering to Prompt Craft*. In *Proceedings of the TEI '24: Eighteenth International Conference on Tangible, Embedded, and Embodied Interaction*, Article 54, 1–12. ACM, New York, NY, USA. DOI: 10.1145/3689050.3704424.

[21] James Phoenix and Mike Taylor. 2024. *Prompt Engineering for Generative AI: Future-Proof Inputs for Reliable AI Outputs at Scale*. O'Reilly Media, Sebastopol, CA.

[22] t2informatik GmbH. 2023. *What is Prompt Engineering? - Smartpedia*. Retrieved January 2023 from https://t2informatik.de/en/smartpedia/what-is-prompt-engineering/.

[23] Dan Cleary. 2025. *Strategies for Managing Prompt Sensitivity and Model Consistency*. PromptHub. Retrieved January 22, 2025, from https://www.prompthub.us/blog/strategies-for-managing-prompt-sensitivity-and-model-consistency-.

[24] Sunil Ramlochan. 2024. *Complete Guide to Prompt Engineering with Temperature and Top-p*. Prompt Engineering Institute. Retrieved August 2024 from https://promptengineering.org/prompt-engineering-with-temperature-and-top-p/.

[25] Research Team. 2024. *Unveiling the Secrets: The Art of Evaluating Prompt Engineering Strategies*. Threatshare.ai. Retrieved May 2024 from https://threatshare.ai/prompteng/prompt-engineering-for-cybersecurity/.

[26] Zijie J. Wang, Aishwarya Chakravarthy, David Munechika, and Duen Horng Chau. 2024. *Wordflow: Social Prompt Engineering for Large Language Models*. arXiv:2401.14447. DOI: 10.48550/arXiv.2401.14447.

[27] Denis Federiakin, Dimitri Molerov, Olga Zlatkin-Troitschanskaia, and Andreas Maur. 2024. *Prompt Engineering as a New 21st Century Skill. Frontiers in Education*, 9. DOI: 10.3389/feduc.2024.1366434

[28] Chirag Shah. 2024. *From Prompt Engineering to Prompt Science with Human in the Loop*. arXiv:2401.04122. DOI: 10.48550/arXiv.2401.04122.

[29] Navveen Balani. 2025. *Ethical Prompt Engineering: A Pathway to Responsible AI Usage*. Retrieved January 2025 from https://navveenbalani.dev/index.php/articles/ethical-prompt-engineering-a-pathway-to-responsible-ai-usage/.

[30] Vagner F. de Santana. 2024. *Challenges and Opportunities for Responsible Prompting*. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*, Article 592, 1–4. ACM, New York, NY, USA. DOI: 10.1145/3613905.3636268>

[31] Adam M. Victor. 2024. Prompt Engineering: The Key to Ethical AI Conversations. LinkedIn. Retrieved January 16, 2024, from https://www.linkedin.com/pulse/prompt-engineering-key-ethical-ai-conversations-adam-m-victor-drqqc.

[32] Andrei-Victor Chisca, Andrei-Cristian Rad, and Camelia Lemnaru. 2024. Prompting Fairness: Learning Prompts for Debiasing Large Language Models. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion* (pp. 52–62). Association for Computational Linguistics.

[33] Satyam Dwivedi, Sanjukta Ghosh, and Shivam Dwivedi. 2023. Breaking the Bias: Gender Fairness in LLMs Using Prompt Engineering and In-Context Learning. *Rupkatha Journal on Interdisciplinary Studies in Humanities*, 15(4). DOI: https://doi.org/10.21659/rupkatha.v15n4.10

[34] Felix Friedrich, Katharina Hämmerl, Patrick Schramowski, Manuel Brack, Jindrich Libovický, Kristian Kersting, and Alexander Fraser. 2024. Multilingual Text-to-Image Generation Magnifies Gender Stereotypes and Prompt Engineering May Not Help You. *arXiv preprint* arXiv:2401.16092. Retrieved from https://arxiv.org/abs/2401.16092

[35] Rachel Skilton and Alison Cardinal. 2024. Inclusive Prompt Engineering: A Methodology for Hacking Biased AI Image Generation. In *Proceedings of the 42nd ACM International Conference on Design of Communication (SIGDOC '24)* (pp. 76–80). ACM. DOI: https://doi.org/10.1145/3641237.3691655.

[36] Amalia Foka. 2024. A Framework for Critical Evaluation of Text-to-Image Models: Integrating Art Historical Analysis, Artistic Exploration, and Critical Prompt Engineering. *arXiv preprint arXiv:2412.12774*. Retrieved from https://arxiv.org/abs/2412.12774.

[37] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae S. Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent AI: Surveying the

Horizons of Multimodal Interaction. *arXiv preprint* arXiv:2401.03568. *DOI:* https://doi.org/10.48550/arXiv.2401.03568.

[38] Onadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. 2023. Practices for Governing Agentic AI Systems. OpenAI. Retrieved from https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[39] Seth D. Baum. 2017. A Survey of Artificial General Intelligence Projects for Ethics, Risk, and Policy. *Global Catastrophic Risk Institute Working Paper 17-1*. DOI: https://doi.org/10.2139/ssrn.3070741.

[40] Isaac Triguero, Daniel Molina, Javier Poyatos, Javier Del Ser, and Francisco Herrera. 2024. General Purpose Artificial Intelligence Systems (GPAIS): Properties, Definition, Taxonomy, Societal Implications and Responsible Governance. *Information Fusion*, 103, 102135. DOI: https://doi.org/10.1016/j.inffus.2023.102135.

[41] Justin D. Weisz, Michael Muller, Jessica He, and Stephanie Houde. 2023. Toward General Design Principles for Generative AI Applications. *arXiv preprint* arXiv: 2301.05578. Retrieved from https://arxiv.org/abs/2301.05578.

[42] Maanak Gupta, Charankumar Akiri, Kshitiz Aryal, Eli Parker, and Lopamudra Praharaj. 2023. From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. *IEEE Access*, 11, 80218–80245. DOI: https://doi.org/10.1109/ACCESS.2023.3300381.

[43] Victoria Arkhurst. 2023. *Security Risks, Bias, AI Prompt Engineering* (2023). Retrieved from.

[44] Jiajia Yuan, Peng Bao, Zifan Chen, Mingze Yuan, Jie Zhao, Jiahua Pan, Yi Xie, Yanshuo Cao, Yakun Wang, Zhenghang Wang, Zhihao Lu, Xiaotian Zhang, Jian Li, Lei Ma, Yang Chen, Li Zhang, Lin Shen, and Bin Dong. 2023. Advanced prompting as a catalyst: Empowering large language models in the management of gastrointestinal cancers. *TIME* 1, 2, 100019. DOI: https://doi.org/10.59717/j.xinn-med.2023.100019.

[45] Louie Giray. 2023. Prompt Engineering with ChatGPT: A Guide for Academic Writers. *Annals of biomedical engineering* 51, 12, 2629–2633. DOI: https://doi.org/10.1007/s10439-023-03272-4.

[46] Kaveh Waddell. 2017. The People Who Fight Hacking and Cybercrime Are Turning to Designers for Help. *Nextgov*. Retrieved June 8, 2017 from http://www.nextgov.com/cybersecurity/2017/05/people-who-fight-hacking-and-cybercrime-are-turning-designers-help/138009/

[47] Sander Schulhoff. 2025. Prompt Hacking: Understanding Types and Defenses for LLM Security. *Learn Prompting*. Retrieved January 22, 2025 from https://learnprompting.org/docs/prompt_hacking/introduction.

[48] Baha Rababah, Shang Wu, Matthew Kwiatkowski, Carson Leung, and Cuneyt G. Akcora. 2024. SoK: Prompt Hacking of Large Language Models. *arXiv preprint* arXiv:2410.13901.

[49] Lee A. Bygrave. 2021. Security by Design: Aspirations and Realities in a Regulatory Context. *Oslo Law Review*, 8(3), 126–177. DOI:https://doi.org/10.18261/olr.8.3.2.

[50] Mireille Hildebrandt. 2011. Legal Protection by Design: Objections and Refutations. *Legisprudence*, 5(2), 223–248.

[51] Batya Friedman and David G. Hendry. 2019. *Value Sensitive Design: Shaping Technology with Moral Imagination*. MIT Press, Cambridge, MA.

[52] Batya Friedman, Peter H. Kahn Jr., and Alan Borning. 2008. Value Sensitive Design and Information Systems. In *The Handbook of Information and Computer Ethics*, Kenneth E. Himma and Herman T. Tavani (Eds.). John Wiley & Sons, Hoboken, NJ, 69–101. DOI: https://doi.org/10.1002/9780470281819.ch4.

[53] Mario R. Barbacci, Jeromy Carrière, Rick Kazman, Mark H. Klein, Howard F. Lipson, and Thomas A. Longstaff. 1998. The Architecture Tradeoff Analysis Method. *CMU/SEI-98-TR-008*. Software Engineering Institute, Carnegie Mellon University. Retrieved May 12, 2025 from https://resources.sei.cmu.edu/asset_files/TechnicalReport/1998_005_001_16646.pdf.

[54] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2021. An Explanation of In-context Learning as Implicit Bayesian Inference. *arXiv preprint* arXiv:2111.02080.

[55] DigitalOcean. 2025. Prompt Engineering Best Practices: Tips, Tricks, and Tools. Retrieved January 22, 2025 from https://www.digitalocean.com/resources/articles/prompt-engineering-best-practices.

[56] Karolina Luzniak. 2023. Preventing Bias in Generative AI: How to Ensure Models' Fairness and Accuracy? *Neoteric*. Retrieved from https://neoteric.eu/blog/preventing-bias-in-generative-ai-how-to-ensure-models-fairness-and-accuracy/..

[57] romptfoo. 2024. Preventing Bias & Toxicity in Generative AI. Retrieved from https://www.promptfoo.dev/blog/prevent-bias-in-generative-ai/.

[58] Kieran Snyder. 2023. Mindful AI: Crafting Prompts to Mitigate the Bias in Generative AI. *Textio*. Retrieved January 21, 2025 from https://textio.com/blog/mindful-ai-crafting-prompts-to-mitigate-the-bias-in-generative-ai.

[59] Alexander Pettersson and Melanie Paschke. 2024. *Ethical Prompting for Generative AI: A How-To Guide for Students at ETH Zurich*. Zurich-Basel Plant Science Center, ETH Zurich. DOI: https://doi.org/10.3929/ethz-b-000672207.

[60] Tobias Dengel. 2024. To Prevent Generative AI Hallucinations and Bias, Integrate Checks and Balances. *BigDATAwire* (August 19, 2024). Retrieved May 22, 2025 from https://www.bigdatawire.com/2024/08/19/to-prevent-generative-ai-hallucinations-and-bias-integrate-checks-and-balances/.

[61] Harvard University IT. 2025. *Getting started with prompts for text-based Generative AI tools*. Retrieved January 22, 2025 from https://huit.harvard.edu/news/ai-prompts.

[62] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. arXiv:2201.11903.

[63] Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. *Navigate through Enigmatic Labyrinth: A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future*. arXiv:2309.15402.

[64] Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. 2023. *Faithful Chain-of-Thought Reasoning*. In *Proceedings of the 13th International Joint Conference on Natural Language Processing (IJCNLP 2023)*, pages 305–329. Association for Computational Linguistics. https://aclanthology.org/2023.ijcnlp-main.20.

[65] Alex Bennet. 2025. *Examples of Ethical Prompting for ChatGPT and Artificial Intelligence*. Retrieved January 17, 2025, from https://www.thoughtmedia.com/ethical-prompting/

[66] Lance Eliot. 2023. *Latest Prompt Engineering Technique: Chain-Of-Verification Does A Sleek Job Of Keeping Generative AI Honest And Upright*. Forbes. Retrieved from https://www.forbes.com/sites/lanceeliot/2023/09/23/latest-prompt-engineering-technique-chain-of-verification-does-a-sleek-job-of-keeping-generative-ai-honest-and-upright/.

[67] Aditya Kuppa, Nikon Rasumov-Rahe, and Marc Voses. 2021. *Chain Of Reference Prompting Helps LLM to Think Like a Lawyer*. In *Proceedings of the Generative AI + Law Workshop at ICML 2023*. https://blog.genlaw.org/CameraReady/37.pdf.

[68] Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. *Legal Prompting: Teaching a Language Model to Think Like a Lawyer*. arXiv:2212.01326. https://arxiv.org/abs/2212.01326.

[69] JJay J. Caughron, Alison L. Antes, Cheryl K. Stenmark, Chaise E. Thiel, Xiaoqian Wang, and Michael D. Mumford. 2011. *Sensemaking Strategies for Ethical Decision-making*. *Ethics & Behavior*, 21(5), 351–366. DOI: https://doi.org/10.1080/10508422.2011.604293.

[70] David Miller. 2024. *Understanding Prompt Bias and How to Overcome It*. Future Skills Academy. Retrieved from https://futureskillsacademy.com/blog/prompt-bias-in-ai/.

[71] Vidisha Vijay. 2024. *Mitigating AI Bias with Prompt Engineering — Putting GPT to the Test*. VentureBeat. Retrieved from https://venturebeat.com/ai/mitigating-ai-bias-with-prompt-engineering-putting-gpt-to-the-test/.

[72] Tongshuang Wu, Michael Terry, and Carrie J. Cai. 2022. *AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts*. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI 2022)*, Article 1, 22 pages. Association for Computing Machinery. https://doi.org/10.1145/3491102.3517582.

[73] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. *Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting*. arXiv:2305.04388. https://arxiv.org/abs/2305.04388.

[74] Chirag Agarwal, Sree Harsha Tanneru, and Himabindu Lakkaraju. 2024. *Faithfulness vs. Plausibility: On the (Un)Reliability of Explanations from Large Language Models*. arXiv:2402.04614.

[75] Rajneesh Jha. 2023. *Tools to Identify and Mitigate Bias & Toxicity in LLMs*. Medium. Retrieved May 12, 2025, from https://medium.com/@rajneeshjha9s/tools-to-identify-and-mitigate-bias-toxicity-in-llms-b34e95732241.

[76] IBM. 2025. responsible-prompting-api: Responsible Prompting is an LLM-agnostic tool that aims at dynamically supporting users in crafting prompts that reflect responsible intentions and help avoid undesired or negative outputs (2025). Retrieved May 12, 2025 from https://github.com/IBM/responsible-prompting-api.

[77] LLM Guard. 2024. *Toxicity Scanner* (2024). Retrieved May 12, 2025 from https://llm-guard.com/input_scanners/toxicity/.

[78] Patrick Farley. 2025. *What is Azure Content Moderator?* (2025). Retrieved May 12, 2025 from https://learn.microsoft.com/en-us/azure/ai-services/content-moderator/overview.

[79] Mohit Singh and Dylan Bouchard. 2025. *How to Assess Your LLM Use Case for Bias and Fairness with LangFair*. CVS Health Tech Blog, February 2025. Retrieved May 12, 2025, from https://medium.com/cvs-health-tech-blog/how-to-assess-your-llm-use-case-for-bias-and-fairness-with-langfair-7be89c0c4fab.

[80] Microsoft. 2025. *Microsoft Presidio: Data Protection and De-identification SDK*. Retrieved May 12, 2025, from https://microsoft.github.io/presidio/.

[81] Frank Börncke. 2025. *Private Prompts: Deine Daten gehören dir!* Retrieved May 12, 2025, from https://www.privateprompts.org/.

[82] RedHunt Labs. 2025. *Octopii: An AI-powered Personal Identifiable Information (PII) Scanner*. Retrieved May 12, 2025, from https://redhuntlabs.com/blog/octopii-an-opensource-pii-scanner-for-images/.

[83] Xiangkun Hu and Dongyu Ru. 2024. *New tool, dataset help detect hallucinations in large language models*. Retrieved May 12, 2025, from https://www.amazon.science/blog/new-tool-dataset-help-detect-hallucinations-in-large-language-models.

[84] NVIDIA. 2025. *NeMo-Guardrails*. Retrieved May 12, 2025, from https://github.com/NVIDIA/NeMo-Guardrails.

[85] Joon S. Park, Joseph O'Brien, Carrie J. Cai, Meredith R. Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23)*. Association for Computing Machinery, New York, NY, USA, 1–22. https://doi.org/10.1145/3586183.3606763.

[86] Mingyue Cheng, Yucong Luo, Jie Ouyang, Qi Liu, Huijie Liu, Li Li, Shuo Yu, Bohou Zhang, Jiawei Cao, Jie Ma, Daoyu Wang, and Enhong Chen. 2025. *A Survey on Knowledge-Oriented Retrieval-Augmented Generation*. arXiv preprint arXiv:2503.10677. Retrieved May 12, 2025, from https://arxiv.org/abs/2503.10677.

[87] <number>[87]</numberJoon Park, Kyohei Atarashi, Koh Takeuchi, and Hisashi Kashima. 2025. *Emulating Retrieval Augmented Generation via Prompt Engineering for Enhanced Long Context Comprehension in LLMs*. arXiv preprint arXiv:2502.12462. Retrieved May 12, 2025, from https://arxiv.org/abs/2502.12462.

[88] Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. *Retrieval Augmented Generation or Long-Context LLMs? A Comprehensive Study and Hybrid Approach*. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP '24)*. Association for Computational Linguistics. Retrieved May 12, 2025, from https://aclanthology.org/2024.emnlp-industry.66.

[89] Kate Crawford. 2024. Generative AI's environmental costs are soaring—and mostly secret. *Nature*, 626(8000), 693. https://doi.org/10.1038/d41586-024-00478-x.

[90] Mél Hogan. 2024. The Fumes of AI. *Critical AI*, 2(1). https://doi.org/10.1215/2834703X-11205231.

[91] Pengfei Li, Jianyi Yang, Mohammad A. Islam, and Shaolei Ren. 2023. Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models. *arXiv preprint arXiv:2304.03271*.

[92] Oxford Analytica. 2024. AI will exacerbate water scarcity. *Emerald Expert Briefings*. https://doi.org/10.1108/oxan-es289107.

[93] Rishi Bommasani, Kevin Klyman, Sayash Kapoor, Shayne Longpre, Betty Xiong, Nestor Maslej, and Percy Liang. 2024. The Foundation Model Transparency Index v1.1: May 2024. *arXiv preprint arXiv:2407.12929*.

[94] The Open Source Initiative. 2025. The Open Source AI Definition 1.0. Retrieved May 12, 2025, from https://opensource.org/ai/open-source-ai-definition.

[95] Sean White. 2017. Announcing the Initial Release of Mozilla's Open Source Speech Recognition Model and Voice Dataset. *Mozilla Blog*. Retrieved January 9, 2018, from https://blog.mozilla.org/en/mozilla/announcing-the-initial-release-of-mozillas-open-source-speech-recognition-model-and-voice-dataset/.

[96] Matt White, Ibrahim Haddad, Cailean Osborne, Xiao-Yang Y. Liu, Ahmed Abdelmonsef, Sachin Varghese, and Arnaud Le Hors. 2024. The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence. *arXiv preprint arXiv:2403.13784*.

[97] Hari Subramonyam, Divy Thakkar, Jürgen Dieber, and Anoop Sinha. 2024. Content-Centric Prototyping of Generative AI Applications: Emerging Approaches and Challenges in Collaborative Software Teams. *arXiv preprint* arXiv:2402.17721.

[98] Krishna Ronanki, Beatriz Cabrero-Daniel, Jennifer Horkoff, and Christian Berger. 2023. Requirements Engineering using Generative AI: Prompts and Prompting Patterns. *arXiv preprint arXiv:2311.03832*.

[99] Farouq Sammour, Jia Xu, Xi Wang, Mo Hu, and Zhenyu Zhang. 2024. Responsible AI in Construction Safety: Systematic Evaluation of Large Language Models and Prompt Engineering. *arXiv preprint arXiv:2411.08320*.

[100] Yannan Li. 2024. Reduce AI Illusion Based on Data Science Technology and Prompt Engineering. *Applied and Computational Engineering*, 97, 152–156. https://doi.org/10.54254/2755-2721/97/20241463.

[101] S. M. Towhidul Islam Tonmoy, S. M. Mehedi Zaman, Vinija Jain, Anku Rani, Vipula Rawte, Aman Chadha, and Amitava Das. 2024. A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models. *arXiv preprint arXiv:2401.01313*. https://doi.org/10.48550/arXiv.2401.01313.

[102] Douglas Schuler and Aki Namioka (Eds.). 1993. *Participatory Design: Principles and Practices*. CRC Press, Boca Raton, FL.

[103] Clay Spinuzzi. 2005. The Methodology of Participatory Design. *Technical Communication*, 52(2), 163–174.

[104] Andy Stirling. 2008. "Opening Up" and "Closing Down": Power, Participation, and Pluralism in the Social Appraisal of Technology. *Science, Technology, & Human Values*, 33(2), 262–294. https://doi.org/10.1177/0162243907311265..

[105] Simon Thorne. 2024. Understanding and Evaluating Trust in Generative AI and Large Language Models for Spreadsheets. In *Proceedings of the European Spreadsheet Risks Interest Group (EuSpRiG)*, 65–78.

[106] Tomáš Ráčil, Petr Gallus, and Tomáš Šlajs. 2024. Efficiency Divide: Comparative Analysis of Human & Neural Network Algorithm Development. In *Proceedings of the 23rd European Conference on Cyber Warfare and Security (ECCWS 2024)*, 683–692.

[107] Mehdi Ben Amor, Michael Granitzer, and Jelena Mitrović. 2023. Impact of Position Bias on Language Models in Token Classification. In *Proceedings of the 2023 ACM International Conference on AI*. ACM, Article 3636126. DOI: https://doi.org/10.1145/3605098.3636126.

[108] Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity Bias in Preference Labeling by Large Language Models. *arXiv preprint arXiv: 2310.10076*.

[109] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and Prejudice: LLM Amplifies Self-Bias in Self-Refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics. DOI: 10.18653/v1/2024.acl-long.826.ACLAnthology+3

[110] Lance Eliot. 2024. Knowing About Temperature Settings When Using Generative AI Is Hot Stuff For Prompt Engineering. *Forbes* (July 29, 2024). Retrieved January 18, 2025, from https://www.forbes.com/sites/lanceeliot/2024/07/29/knowing-about-temperature-settings-when-using-generative-ai-is-hot-stuff-for-prompt-engineering/.

[111] Sander Schulhoff. 2025. Understanding Temperature, Top P, and Maximum Length in LLMs. *Learn Prompting* (January 2025). Retrieved January 22, 2025, from https://learnprompting.org/docs/intermediate/configuration_hyperparameters.

[112] Michelle Avery. 2024. From Benchmarks to Red-Teaming: Ensuring Robust and Responsible AI. *WillowTree* (2024). Retrieved from https://www.willowtreeapps.com/insights/genai-benchmarking-and-red-teaming.

[113] Angelina Wang, Aaron Hertzmann, and Olga Russakovsky. 2024. Benchmark Suites Instead of Leaderboards for Evaluating AI Fairness. *Patterns* 5, 11 (2024), 101080. DOI: https://doi.org/10.1016/j.patter.2024.101080.

[114] Restack. 2025. Benchmarking AI in Sustainability. *Restack* (2025). Retrieved May 21, 2025, from https://www.restack.io/p/sustainable-ai-answer-benchmarking-ai-sustainability-cat-ai.

[115] Lee Boonstra. 2024. Documenting Your Prompts: A Best Practice for Success. *Medium* (2024). Retrieved from https://medium.com/google-cloud/documenting-your-prompts-a-best-practice-for-success-1278f2c0344e.

[116] S. Kingson. 2024. How Technical Writers Can Master Prompt Engineering. *Document360* (2024). Retrieved from https://document360.com/blog/prompt-engineering-for-technical-writers/.

[117] Thalia Khan, Albert Tanjaya, Jacob Pratt, and John Howell. 2025. Transparency Through Documentation: A Pathway to Safer AI. *Partnership on AI* (2025). Retrieved May 21, 2025, from https://partnershiponai.org/transparency-through-documentation-a-pathway-to-safer-ai/.

[118] Riikka Koulu and Jörg Pohle. 2024. Legal Design Patterns: New Tools for Analysis and Translations Between Law and Technology. *Digital Society* 3, 2 (2024), 1–13. DOI: https://doi.org/10.1007/s44206-024-00109-y.

[119] PromptHero. 2025. Search Prompts for Stable Diffusion, ChatGPT & Midjourney. (January 2025). Retrieved January 17, 2025, from https://prompthero.com/.

[120] The Prompt Index. 2025. AI Prompt Database. *The Prompt Index* (2025). Retrieved January 18, 2025, from https://www.thepromptindex.com/prompt-database.

[121] Promptsty.com. 2024. Prompts for Artificial Intelligence Ethics: Essential Guide. *Promptsty* (2024). Retrieved January 17, 2025, from https://promptsty.com/prompts-for-artificial-intelligence-ethics/.

[122] Maastricht University. 2025. AI Prompt Library. *Maastricht University* (January 2025). Retrieved January 17, 2025, from https://www.maastrichtuniversity.nl/about-um/education-at-um/edlab/ai-education-maastricht-university/ai-prompt-library.

[123] Tian Y. Liu, Stefano Soatto, Matteo Marchi, Pratik Chaudhari, and Paulo Tabuada. 2024. Meanings and Feelings of Large Language Models: Observability of Latent States in Generative AI. *arXiv preprint arXiv:2401.12345*.

[124] Anthropic. 2025. *System Prompts: Claude 3.5 Sonnet*. Retrieved January 17, 2025, from https://docs.anthropic.com/en/release-notes/system-prompts.

[125] GenAIScript. 2025. *System Prompts*. Retrieved January 17, 2025, from https://microsoft.github.io/genaiscript/reference/scripts/system/.

[126] Bachaalany, Elias. 2025. *TheBigPromptLibrary: A Collection of Prompts, System Prompts, and LLM Instructions*. Retrieved January 17, 2025, from https://github.com/0xeb/TheBigPromptLibrary.

[127] Alex, Vlad. 2025. *ChatGPT-System-Prompts*. Retrieved January 17, 2025, from https://github.com/mustvlad/ChatGPT-System-Prompts.

[128] Korzyński, Paweł, Mazurek, Grzegorz, Krzypkowska, Pamela, and Kurasiński, Artur. 2023. *Artificial Intelligence Prompt Engineering as a New Digital Competence: Analysis of Generative AI Technologies Such as ChatGPT. Entrepreneurial Business and Economics Review* 11, 3, 25–38..

[129] Lo, Leo S. 2023. *The Art and Science of Prompt Engineering: A New Literacy in the Information Age. Internet Reference Services Quarterly* 27, 4, 203–210..

[130] Gonzaga University School of Law. 2024. *Generative AI & Legal Research: A Guide for Students and Faculty on Using Generative AI as a Tool for Legal Research and Writing*. Retrieved January 17, 2025, from https://libguides.law.gonzaga.edu/c.php?g$=$1374374.