Towards Responsible Prompt Design: a Framework for Secure and Ethical GenAI Outputs

Miriana Calvano[1], Antonio Curci[1,2], Giuseppe Desolda[1], Antonio Piccinno[1] and Alfonso Pio Pretorino[1]

[1]Department of Computer Science, University of Bari Aldo Moro, Via E. Orabona 4, 70125, Bari, Italy
[2]Department of Computer Science, University of Pisa, Largo B. Pontecorvo 3, 56197, Pisa, Italy

Abstract

Artificial Intelligence (AI) is spreading in many domains shaping the way in which humans carry out activities. AI is not a mere tool but can be considered a powerful instruments able to impact humans and the environment surrounding them. Among the technologies, Generative AI (GenAI) is receiving great attention in the current scenario due to its communication capabilities and being timely in responding to requests adapting to users' requests. As GenAI becomes deeply integrated into software development, the role of prompt engineering has transitioned from a simple interaction modality to a critical professional activity. It is important to define prompt that can lead to the generation of context-aligned results that do not represent a potential threat for humans but lead to the establishment of a human-AI symbiotic relationship in which both parties are enhanced. This article addresses the current lack of standardized guidelines for the definition of a prompt to create secure and legally compliant AI systems proposing a conceptual framework designed to guide users in creating prompts that matches these objectives. The proposed framework is at a preliminary stage and will be validated conducting real-world case studies to test its applicability and efficacy.

Keywords

Prompt Engineering, Large Language Models, Symbiotic Artificial Intelligence, Human-Centered Artificial Intelligence,

1. Introduction

As the use of Artificial Intelligence (AI) grows, safeguarding humans and considering them in all of their dimensions remains crucial. Among the various categories of AI, Generative Artificial Intelligence (GenAI) is the most widespread in modern society, being used by end users; it is defined as "AI that can create original content such as text, images, video, audio, or software code in response to a user's prompt or request" [1]. A substantial subset of GenAI consists of Large Language Models (LLMs), trained on large-scale textual datasets with the aim of generating coherent text and performing complex Natural Language Processing (NLP) tasks to carry out activities (e.g., script creation, dialog simulation) in numerous fields: from science to arts. In recent years, LLMs have demonstrated great potential in understanding cognitive concepts such as mathematical reasoning and problem-solving, gaining considerable attention for their ability to generate high-quality source code, thereby supporting software development by reducing development time and lowering the required level of expertise [2]. However, the training process carried out on open data can introduce significant security risks [3]. For example, a code generation tool is often trained on open-source repositories, which can contain code smells that can affect the quality of the generated code. Due to the stochasticity and non-determinism of GenAI systems, Prompt Engineering has emerged as a professional role, a distinct activity aimed at constraining model behavior, and a new modality of human-AI interaction that functions as a bridge between human

1st Workshop on Symbiotic AI for Cyber Social Security: Human–AI Collaboration for Social Resilience and Trustworthy Digital Ecosystems (SAI4CCS 2026), co-located with the 2nd Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09–13, 2026, Cagliari, Italy

miriana.calvano@uniba.it (M. Calvano); antonio.curci@uniba.it (A. Curci); giuseppe.desolda@uniba.it (G. Desolda); antonio.piccinno@uniba.it (A. Piccinno); a.pretorino@phd.uniba.it (A. P. Pretorino)
0000-0002-9507-9940 (M. Calvano); 0000-0001-6863-872X (A. Curci); 0000-0001-9894-2116 (G. Desolda); 0000-0003-1561-7073 (A. Piccinno); 0009-0000-8422-7808 (A. P. Pretorino)

© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-14

CEUR-WS.org/Vol-4329/paper15.pdf

intent and algorithmic execution. In fact, Prompt Engineering is the process of designing and optimizing natural language inputs to guide LLMs toward accurate and contextually relevant outputs [4].

The current literature lacks a standardized and in-depth exploration of the topic, thus bringing to the need for a guidance through indications, such as guidelines and frameworks, that can support users in creating effective prompts that generate ethically and legally compliant outputs, with particular attention to ensuring their safety and security. The need for such guidance arises for a variety of reasons, ranging from technological to scientific motivations. From the technological point of view, the integration of LLMs in real-world applications, such as healthcare, education, and customer service, has grown rapidly. However, poorly designed prompts can result in poor output quality, legal requirements being violated, or sensitive data being leaked. Synthesizing best practices and optimal prompt configurations into a structured conceptual framework can enhance security, ensure ethical and legal compliance, and mitigate risks by embedding preventive safety measures directly into the generative process.

From a scientific perspective, it is valuable to investigate how specific linguistic instruction could shape AI behavior to avoid unexpected and harmful consequences. In this regard, it is important to consider the prompt not only as an aggregation of instructions, but also as an "interface" between humans and machines that allows a direct and continuous communication to reach a predefined objective. The motivation behind this lies in the incidents that have been occurring since the adoption of these tools; two significant and recent examples are reported in the following. The first one is the incident with Replit's coding assistant agent based on an LLM [1]. While working on a production database containing real data from companies and individuals, the AI agent dropped the entire database. This caused the operation to become irreversible, since the tables were replaced with empty ones and no backups or copies had been made before those actions [5]. The second and last incident dates back to November 2025, when Google released its new AI-powered IDE called Google Antigravity[2]. The user asked the agent to clear the server's cache and start the *Node.js* server. The command that was executed by the AI not only deleted the temporary files stored in the cache, but also erased all data on the drive without any confirmation from the user [6].

These concerns are also particularly relevant when employing LLMs in Software Engineering. In fact, in recent years, LLMs are also being employed to create software systems, build architectures, and design interaction mechanisms. Thus, ensuring that the artifacts generated by this technology are usable, safe, and secure is crucial to safeguard humans. This research lays its groundwork on these issues, proposing a conceptual framework to support humans in the software creation process by enabling them to write prompts that generate secure, ethical, and legally compliant outputs. The main objective is to create an instrument that guides practitioners throughout the entire creation process, from, for example, requirements definition to interface prototyping and code generation. It will also be worthwhile to investigate whether this framework can help prevent incidents like the Replit and Google Antigravity cases.

The article is structured as follows: section 2 reports the motivations and lays out the objectives of this research; section 3 illustrates the proposal of the conceptual framework that aims to fill the gap identified; section 4 draws the conclusions and presents the future work that will involve this research for refinement and validation.

2. Motivations and Objectives

Employing AI-based tools and, specifically, GenAI in software development is changing the way that practitioners are carrying out their tasks and activities. It is necessary to keep up with ever-evolving nature of this technology and understanding the state of the art to define solutions that can suit the current challenges. The background concepts and related work are reported below, followed by the objectives and the Research Question (RQ) that this work aims to answer.

[1] https://docs.replit.com/replitai/agent
[2] https://antigravity.google/

2.1. Related Work

LLMs like Gemini [7], Claude [8], GPT-4 [9], and LLaMA [10] have shaken the field of software development with their ability to generate software code starting from plain language text written by developers. Studies have demonstrated that about 62% of generated code contains vulnerabilities, highlighting the need for a security validation process before using it for production development [11]. In response to these challenges, research began to focus on whether and how such difficulties could be overcome using different prompting techniques.

Wang et al. proposes a framework for evaluating the quality of prompts, encompassing three dimensions: goal clarity, information completeness, and logical consistency [12]. Empirical experiments demonstrate that vulnerability rates increase monotonically as prompt normativity deteriorates (i.e., as the prompt becomes increasingly vague, incomplete, or contradictory). The authors found that using adequate prompting strategies, such as Chain-of-Thought (CoT) and "RE-ACT", mitigates these risks by up to $\approx 30\%$ for the most ambiguous prompts. The authors emphasize that careful prompt engineering represents a practical safeguard, redirecting LLM security research towards human-AI interaction rather than model-level flaws [12].

Chinnachamy present *PromptSecure*, which introduces a comprehensive, protocol-driven framework for prompt engineering. It addresses the issue of unsecured prompt interfaces and the absence of a standardized, secure approach to prompt engineering [13]. Unlike conventional prompt-tuning approaches, which focus on improving model performance, PromptSecure incorporates principles from cybersecurity, differential privacy, and software verification to create a robust prompt lifecycle. In particular, it introduces concrete mechanisms to control how prompts are created, checked, executed, and logged. The framework incorporates automated sanitization, static and runtime validation, and continuous monitoring to detect malicious or non-compliant prompts. By embedding cybersecurity controls and privacy safeguards and verification steps directly into the prompt pipeline, PromptSecure aims to prevent prompt injection, data leakage, and unauthorized behavior. Empirical evaluation shows that PromptSecure achieves 93% accuracy, compared to 78% for baseline systems. These results demonstrate that PromptSecure advances secure-by-design GenAI deployment by integrating cybersecurity principles and formal compliance scaffolds [13].

In another study, Nazzal et al. investigate a dual-objective prompt-optimization framework to address the critical vulnerability code generated by LLMs [14]. This code inherits insecure patterns from its massive open-source training corpora and consequently produces code with common weaknesses (e.g., SQL injection). The framework proposed by the authors, named *PromSec*, combines a Generative Adversarial Graph Neural Network (gGAN) with a static code analyzer. The main objective of gGAN is to clean the code by reducing the number of vulnerabilities, and the resulting secure code drives the LLM to generate improved prompts for the next iteration. The empirical evaluation, performed across four state-of-the-art LLMs, shows that PromSec eliminates all identified vulnerabilities, and the prompts are transferable to other LLMs, generalizable in programming language, and to unseen vulnerabilities in training[14].

These approaches and frameworks focus on the security of the generated output, lacking in effective validation of the ethical and legal factors that could affect the output [12, 13, 14]. When humans interact with an AI system, it is essential that the interaction is secure and that the information provided by the system is reliable. This field is known as Human Centered Artificial Intelligence (HCAI) [15, 16] and it aims to design, develop, and evaluate AI systems that are safe, reliable, and trustworthy [15]. HCAI is rooted in Human-Centered Design (HCD), which is reported in the ISO 9241:210 and it indicates the processes necessary to involve humans in all the phases of creating systems, products, or services [17]. In the context of HCAI, the traditional approach of building AI systems that merely focused on complex mathematical structures and sophisticated algorithms is accompanied by other aspects that belong to the field of Human-Computer Interaction (HCI) with interaction design and psychology [18].

Within this perspective, a step forward in the context of human-centered systems is Symbiotic Artificial Intelligence (SAI), which is represents a specialization of HCAI, in which humans and AI engage in a continuous, mutual collaboration aimed at augmenting human capabilities rather than

replacing them, fostering a bidirectional relationship in which both parties improve over time [19, 20]. This means that humans and AI collaborate in a symbiotic way, where the limitations of one are compensated by the strengths of the other. The approach necessary to build these systems inevitably multidisciplinary, touching on various fields that range from the technicalities of *Artificial Intelligence* and *Software Engineering* to the more ethical and human issues tackled by *Law & Ethics* [18], keeping HCI at the core of the scenario. Thus, in SAI systems there must be interaction mechanisms that enhance humans and enable them to fully understand the processes, outputs, and structure of AI reasoning, for example, by putting in place proper *Transparency* and *Explainability* techniques that ensure that humans can be *in-the-loop*. This ensures that humans remain in control of the decision-making process, avoiding negative consequences on themselves or their environment.

In this regard, legal and governmental bodies are also showing interest in regulating and standardizing the practices of designing, developing, and deploying AI systems in order to ensure that humans are safeguarded and protected. An exemplary case is the *Artificial Intelligence Act*, which is a legal framework released by the European Union (EU) in 2024. It defines four categories of risk that dictate requirements and obligations that practitioners must follow in order to being authorized to release the systems they build [21]. An example of how this regulation and human-AI symbiosis can be merged to reach a common objective is the framework *FrameSAI*, which is a three-layer framework that aims to support practitioners in building legally compliant SAI systems through practical indications: principles, guidelines, and success criteria [22].

The efforts in operationalizing the regulations concerning AI are yet to be fully investigated; due to the novelty of the field, there is limited research available or related literature available on defining a conceptual framework or providing users with guidelines to help them create prompts that generate secure, ethical, and legally compliant outputs.

2.2. Objectives and Research Questions

Based on the topics discussed above, the overall objective of this research is to define a conceptual framework for writing effective prompts that generate outputs ethically and legally compliant, with particular attention to ensuring the safety and security. In order to reach this goal, it is necessary to analyze the mechanisms that are commonly employed by users when interacting with a LLMs and to investigate whether or not legal and ethical factors are influencing the output produced by the designed prompt. Thus, it is necessary to improve the security, safety, trust, fairness, and transparency of the generated output.

Starting from this goal, the following RQ was formulated:

RQ: How to structure and script prompts that foster the generation of human-centered, secure, and legally compliant ?

The RQ aims to investigate how GenAI can be leveraged to create outputs that prioritize ethics while ensuring safety and compliance with existing legal and regulatory frameworks. As anticipated in section 1, the need for guidance is becoming more relevant due to the rapid adoption of GenAI technologies in the software development field, with the consecutive growth of concerns related to trust, data protection, and ethical use. The research work will examine design methodologies and security mechanisms to identify best practices to build prompts that lead to secure, ethical, and legally compliant outputs.

3. Towards Secure Prompt Engineering

The RQ guided the definition of an approach that aims at laying the groundwork for the conceptual framework that combines different elements of prompt design with respect to the aspects of interest of the research (e.g., components, security, ethics). Thus, the proposal aims at representing a step forward towards secure prompt engineering and the operationalization of its best practices that satisfy these requirements. They are described and illustrated below.

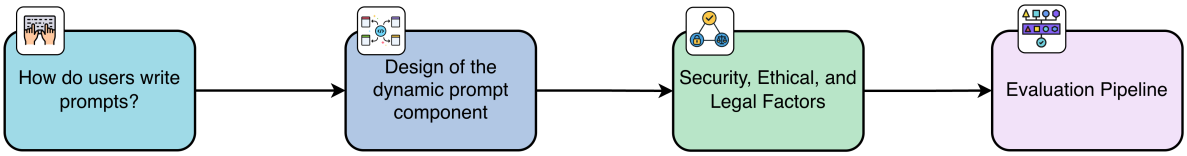


Figure 1: Visual representation of four phases of the proposed approach

3.1. Proposed Approach

The approach is composed into four main phases with the final aim of investigating and optimizing prompt engineering for generating secure, ethical, and legally compliant output. The four phases are described below and illustrated in fig. 1.

P1: it investigates the mechanisms employed by users when writing prompts. This involves examining how they structure prompts for different objectives, such as interface design or coding a method, in order to identify common strategies and mistakes and extract generalizable principles to guide more effective prompt design.

P2: it consists of the design of a dynamic prompt component that can adaptively incorporate specific elements based on the user's objective. For instance, if the objective is vulnerability remediation (e.g., CVE-based tasks), the component will ensure that the prompt incorporates known vulnerabilities and their context.

P3: it considers ethical and legal factors, together with best practices for security. They could be potentially derived from existing literature while ensuring compliance with regulations such as General Data Protection Regulation (GDPR)[23] and EU AI Act. For example, by possibly translating the obligations set by these regulations into a set of best practices for writing prompts and using models that comply with their privacy requirements and risk-based classification.

P4: it refers to the testing pipeline, which is developed to evaluate the framework throughout the software creation process, including the design and evaluation phases.

3.2. Conceptual Framework

Based on these four phases, a conceptual framework is being defined to provide the necessary elements to construct optimal prompts that generate secure, ethically and legally compliant outputs. The proposed framework, shown in fig. 2, is described below. It provides details about the components and the two actors (i.e., *AI Providers* and *Human*) along with their role in the whole process.

(C1) Input: it consists of both a prompt and a set of settings. The prompt contains instructions and information about the objective that the human wants to achieve. In some specific cases, the AI system allows the human to adjust the model's parameters (referred to as "Settings" in Figure fig. 2) that control how the model generates its responses (output). An example is the *temperature* parameter, where a low value (e.g., 0.3) makes the responses more deterministic, while a high value introduces more variability in the produced output. Other key parameters include *max tokens*, which limits the output's length; *top-p* (nucleus sampling), which restricts token choices to the most probable ones.

(C2) AI system: in this scenario, it can be described as a man in the middle, which is receiving the humans' input and transmitting it to the AI model. Once the model finishes with its elaboration, the output is sent back to the human by leveraging the AI system's User Interface (UI) to present the output.

(C3) AI Act: it is a regulatory framework enacted by the EU to govern the development, deployment, and use of AI systems. It adopts a risk-based approach, classifying AI systems into unacceptable high, limited, and minimal risk categories. The regulation aims to guarantee that safety and fundamental rights are ensured in AI systems [21]. In this workflow, its role is to categorize and regulate the AI system by guaranteeing its compliance with the enacted law.

(C4) GDPR: it is a law enacted by the EU to protect the personal data of EU residents. It defines how personal data can be processed and transferred, how organizations must protect it, and establishes EU residents' right over personal data possession, use, and collection of their personal data. The GDPR enforces principles such as lawfulness, fairness, data minimization, accuracy, and accountability. It is important to note that the GDPR applies to organizations outside EU both territorially and legally. Any organization that accesses or processes personal data of EU residents must comply with the GDPR [23]. In this scenario, the GDPR regulates how human's data is handled by the AI system and by the model.

(C5) Output of GenAI systems: it can be returned in different forms, and it should contain information coherent with the content of the inputs provided by the human.

The actors, with their respective roles in the interaction, are reported below:

(A1) The AI providers (e.g., Google, Meta, OpenAI) offer their services for external use through AI systems. During the models' design phase, such companies must develop solutions in accordance with the EU'sAI Act and the GDPR.

(A2) The human interacts with the AI model through the User Interfaces (UIs) of the AI system, which receives the input on one side and directly communicates with the model on the other. The human's input consists of both a prompt and a set of settings. Once the human finishes crafting the input, it is then sent to the AI systems, which, based on the prompt and settings, generate an output returning it to the human.

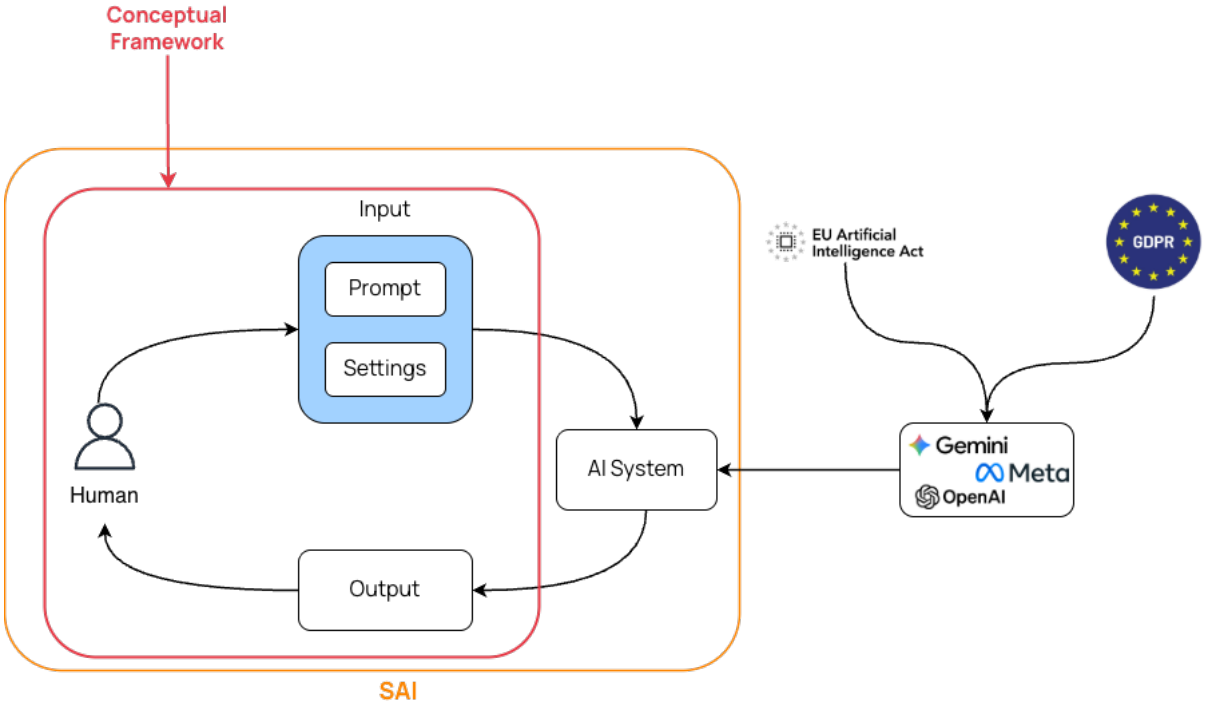


Figure 2: Conceptual framework for constructing an optimal prompt

The human-AI interaction raises several key challenges: ethical and privacy concerns about how data is handled; the interaction lacks close collaboration between humans and AI; finally, the quality and accuracy of data can be poor, and critical decisions often lack sufficient human judgment. In the proposed approach, the connection between actors and components creates a loop through which the human is able to control the interaction with the AI models. This underscores that SAI helps ensure the security, ethical, and legal compliance of the generated output. These issues can be partially solved by SAI, as motivated in section 2, but in cases where these symbiotic elements are not integrated in AI systems, humans can still take measures to minimize potential risks by using the conceptual framework (in red).

4. Conclusions

The widespread adoption of AI is changing activities and processes in many fields; when it comes to software engineering, LLMs have been introduced in different phases of software's life-cycle, being used for multiple purposes in design, development, deployment, and testing. As this technology continues to advance and change, it is important to keep up with the ever-evolving requirements and characteristics to ensure a smooth and appropriate integration of these tools in software engineering practices. Facing these challenges requires a multidisciplinary approach which is based on foundational ethical guidelines, regulatory frameworks, and the human-centered approach. More specifically, designers and developers must create systems that reinforce human control and oversight, protect them from harm and threats, and accurately fulfill the tasks they are built for. Ensuring that such requirements are satisfied, can allow humans and AI to work symbiotically together, enabling them to support each other and improve over time.

This research work proposes an approach and a conceptual framework to optimize prompt engineering for a secure, ethical, and lawful output. The proposed approach unfolds of four components and lays the groundwork for a conceptual framework that aims at operationalizing the above-mentioned goal. Since this investigation is in its preliminary stages, its next steps are necessary to refine the framework and translate it from a conceptual into an operational one. It is planned to perform a systematic literature review concerning the use of LLMs in the context of software engineering in order to have a clear overview of the practices that are currently being employed in this context. Then, a refined version of the framework will be delineated, which will be validated through case studies and experiments, for example, by employing it during software creation.

Acknowledgments

This research is supported by the following projects: SERICS - "Security and Rights In the CyberSpace - SERICS" (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU; SMART: System for Matching and Ranking Talents, Avviso "Reti - Sostegno alla ricerca collaborativa" - Codice Progetto CGGUBN6

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] C. Stryker, M. Scapicchio, What is generative ai?, IBM Think, https://www.ibm.com/think/topics/generative-ai, 2025. URL: https://www.ibm.com/think/topics/generative-ai, accessed: 2025-11-15.

[2] A. Ziegler, E. Kalliamvakou, X. A. Li, A. Rice, D. Rifkin, S. Simister, G. Sittampalam, E. Aftandilian, Productivity assessment of neural code completion, in: Proceedings of the 6th ACM SIGPLAN

International Symposium on Machine Programming, MAPS 2022, Association for Computing Machinery, 2022, pp. 21–29. URL: https://dl.acm.org/doi/10.1145/3520312.3534864. doi:10.1145/3520312.3534864.

[3] M. L. Siddiq, S. H. Majumder, M. R. Mim, S. Jajodia, J. C. S. Santos, An empirical study of code smells in transformer-based code generation techniques, in: 2022 IEEE 22nd International Working Conference on Source Code Analysis and Manipulation (SCAM), 2022, pp. 71–82. URL: https://ieeexplore.ieee.org/abstract/document/10006873. doi:10.1109/SCAM55253.2022.00014, ISSN: 2470-6892.

[4] J. Berryman, A. Ziegler, Prompt Engineering for LLMs: The Art and Science of Building Large Language Model-Based Applications, O'Reilly Media, 2025.

[5] I. Kovvuru, Replit AI Deletes Production Database: 2025 DevOps Security Lessons for AWS Engineers, 2025. URL: https://medium.com/@ismailkovvuru/replit-ai-deletes-production-database-2025-devops-security-lessons-for-aws-engineers-4984c6e7a73d.

[6] r/google_antigravity community, Google antigravity just deleted the contents of their website, https://www.reddit.com/r/google_antigravity/comments/1p82or6/google_antigravity_just_deleted_the_contents_of/, 2014. Accessed: 2025-12-09.

[7] T. Gemini, Gemini: A Family of Highly Capable Multimodal Models, 2025. URL: http://arxiv.org/abs/2312.11805. doi:10.48550/arXiv.2312.11805.

[8] Anthropic, Claude sonnet 4.5, https://www.anthropic.com/news/claude-sonnet-4-5, 2025. Accessed: 2025-12-05.

[9] OpenAI, GPT-4 Technical Report, 2024. URL: http://arxiv.org/abs/2303.08774. doi:10.48550/arXiv.2303.08774.

[10] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. URL: http://arxiv.org/abs/2302.13971. doi:10.48550/arXiv.2302.13971.

[11] N. Tihanyi, T. Bisztray, M. A. Ferrag, R. Jain, L. C. Cordeiro, How secure is AI-generated code: a large-scale comparison of large language models, Empirical Software Engineering 30 (2025) 47. URL: https://link.springer.com/10.1007/s10664-024-10590-1. doi:10.1007/s10664-024-10590-1.

[12] B. Wang, Y. Zhong, M. Wan, W. Yu, Y. Ouyang, Y. Huang, H. Li, Is Your Prompt Poisoning Code? Defect Induction Rates and Security Mitigation Strategies, 2025. URL: http://arxiv.org/abs/2510.22944.

[13] T. Chinnachamy, Promptsecure: Secure Prompt Engineering Protocols for Regulated Genai Environments, International Journal of Innovative Science and Research Technology (2025) 3021–3029. URL: https://www.ijisrt.com/promptsecure-secure-prompt-engineering-protocols-for-regulated-genai-environments. doi:10.38124/ijisrt/25jul1787.

[14] M. Nazzal, I. Khalil, A. Khreishah, N. Phan, PromSec: Prompt Optimization for Secure Generation of Functional Source Code with Large Language Models (LLMs), in: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, CCS '24, Association for Computing Machinery, New York, NY, USA, 2024, pp. 2266–2280. URL: https://dl.acm.org/doi/10.1145/3658644.3690298. doi:10.1145/3658644.3690298.

[15] B. Shneiderman, Human-Centered AI, Oxford University Press, 2022. URL: https://doi.org/10.1093/oso/9780192845290.001.0001. doi:10.1093/oso/9780192845290.001.0001.

[16] W. Xu, Toward human-centered AI: a perspective from human-computer interaction, Interactions 26 (2019) 42–46. URL: https://doi.org/10.1145/3328485. doi:10.1145/3328485.

[17] I. O. for Standardization, Iso 9241:210 - ergonomics of human-system interaction: Human-centred design for interactive systems, 2019. URL: https://www.iso.org/standard/77520.html.

[18] G. Desolda, G. Dimauro, A. Esposito, R. Lanzilotti, M. Matera, M. Zancanaro, A human–AI interaction paradigm and its application to rhinocytology, Artificial Intelligence in Medicine 155 (2024) 102933. URL: https://www.sciencedirect.com/science/article/pii/S0933365724001751. doi:10.1016/j.artmed.2024.102933.

[19] D. Szafran, R. L. Bach, "the human must remain the central focus": Subjective fairness perceptions in automated decision-making, Minds and Machines 34 (2024) 24. URL: https://doi.org/10.1007/s11023-024-09684-y. doi:10.1007/s11023-024-09684-y.

[20] M. Calvano, A. Curci, G. Desolda, A. Esposito, R. Lanzilotti, A. Piccinno, Building Symbiotic Artificial Intelligence: Reviewing the AI Act for a Human-Centred, Principle-Based Framework, Minds and Machines 36 (2025) 1. URL: https://link.springer.com/10.1007/s11023-025-09753-w. doi:10.1007/s11023-025-09753-w.

[21] European Parliament, Council of the European Union, Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), 2024.

[22] M. Calvano, A. Curci, R. Lanzilotti, A. Piccinno, FrameSAI: A Three-Layer Framework to Create Symbiotic AI Systems, in: Joint Proceedings of the Thematic Workshops at Ital-IA 2025 Colocated with the 5th National Conference on Artificial Intelligence, Organized by CINI (Ital-IA 2025), volume 4121, CEUR Workshop Proceedings, 2025. URL: https://ceur-ws.org/Vol-4121/Ital-IA_2025_paper_9.pdf.

[23] European Parliament, Council of the European Union, Regulation (EU) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/EC (general data protection regulation) (text with EEA relevance), 2016. URL: http://data.europa.eu/eli/reg/2016/679/oj, legislative Body: EP, CONSIL.