

Please quote as: Grau, M., Reinhard, P., Bauer-Hänsel, I. & Blohm, I. (2025).
Rethinking AI Design Practices: A Reference Architecture for Generative AI Systems.
European Conference on Information Systems (ECIS), Amman, Jordan: Aisel.

Association for Information Systems

AIS Electronic Library (AISeL)

ECIS 2025 Proceedings

European Conference on Information Systems
(ECIS)

June 2025

Rethinking AI Design Practices: A Reference Architecture for Generative AI Systems

Marc Christopher Grau

University of St. Gallen, marcchristopher.grau@unisg.ch

Philipp Reinhard

University of Kassel, philipp.reinhard@uni-kassel.de

Ingrid Bauer-Hänsel

Institute of Information Systems and Digital Business, ingrid.bauer-haensel@unisg.ch

Ivo Blohm

University of St. Gallen, ivo.blohm@unisg.ch

Follow this and additional works at: <https://aisel.aisnet.org/ecis2025>

Recommended Citation

Grau, Marc Christopher; Reinhard, Philipp; Bauer-Hänsel, Ingrid; and Blohm, Ivo, "Rethinking AI Design Practices: A Reference Architecture for Generative AI Systems" (2025). *ECIS 2025 Proceedings*. 6.
https://aisel.aisnet.org/ecis2025/des_research/des_research/6

This material is brought to you by the European Conference on Information Systems (ECIS) at AIS Electronic Library (AISeL). It has been accepted for inclusion in ECIS 2025 Proceedings by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

RETHINKING AI DESIGN PRACTICES: A REFERENCE ARCHITECTURE FOR GENERATIVE AI SYSTEMS

Completed Research Paper

Marc Grau, University of St. Gallen, St. Gallen, Switzerland, marcchristopher.grau@unisg.ch

Philipp Reinhard, University of Kassel, Kassel, Germany, philipp.reinhard@uni-kassel.de

Ingrid Bauer-Hänsel, University of St. Gallen, St. Gallen, ingrid.bauer-haensel@unisg.ch

Ivo Blohm, University of St. Gallen, St. Gallen, ivo.blohm@unisg.ch

Abstract

Generative AI (GenAI) holds the potential to transform various facets of society. While research and practice increasingly explore its diverse use cases and assess its impact, many companies continue to face significant challenges in designing and developing advanced GenAI systems, hindering their ability to derive value from these technologies. To address these challenges and offer guidance on emerging GenAI development practices, we propose a reference architecture that highlights key design decisions. Using a design science research approach, we developed the reference architecture informed by consortium workshops, literature reviews, and expert interviews. The reference architecture is grounded in eight key design requirements and serves as a strategic guide to streamline system design and assist organisational decision-making and communication. Additionally, the architecture contributes to the academic discourse by offering a structured framework for future research on the design peculiarities of GenAI systems, helping bridge the gap between practical implementations and theoretical insights.

Keywords: Generative Artificial Intelligence, Large Language Models, Foundation Models, Reference Architecture, Design Science Research

1 Introduction

Interest in Generative AI (GenAI) has surged since the release of ChatGPT in November 2022, as companies recognise its potential for operational efficiency and competitive advantage (Wessel et al., 2023). For example, Klarna's AI assistant, built on OpenAI's GPT model, autonomously handled nearly two-thirds of customer requests within its first month of operation (Klarna, 2024). However, while building prototypes with GenAI may be relatively fast, organisations often face challenges in scaling and integrating them effectively to achieve large-scale deployment and derive substantial benefits (Enholt et al., 2022). A primary barrier to scaling is the technical complexity of large-scale GenAI implementations and the technical changes that GenAI bring, which require a departure from traditional machine-learning architecture design approaches (Hagos et al., 2024).

GenAI introduces new paradigms for system design and development. For example, programming GenAI systems relies less on traditional code and more on prompt engineering, a critical consideration not considered in traditional machine-learning architectures (White et al., 2023). Models within a GenAI system can be orchestrated in a myriad of ways including chains (Wu et al., 2022) or multi-model systems (Hong et al., 2023). Furthermore, the "black-box" nature of GenAI models poses significant design and operational challenges, necessitating new experimental, iterative approaches that enable companies to better understand and reproduce model outputs (Schneider, 2024). This shift introduces novel requirements for model evaluation and adaptation, which must be reflected in architectural design of an overarching system. Additionally, GenAI's generative capability—producing new content such as language, images, and videos—results in non-deterministic outputs, complicating response verification

and the identification of errors like hallucinations (Ji et al., 2023). Consequently, testing systems including GenAI models demands innovative approaches to accommodate this non-determinism, as existing metrics from traditional machine learning contexts often fall short. Lastly, increasingly agentic workflows encompassing tool use, reflection, planning, or multi-agent collaboration, impacts system design (Chawla et al., 2024). This shift demands a rethinking of system architectures to integrate novel interaction approaches and handling of increased uncertainty, ensuring adaptable and robust GenAI system deployments. Despite its importance, little academic literature exists on GenAI system architecture design as of today. Some work has emerged from computer science, but it is often limited in scope, focusing on specific technical aspects without addressing broader organisational needs (Hassan et al., 2024). In the field of Information Systems (IS), research has primarily examined GenAI application areas, use cases, and ethical considerations (e.g., Baidoo-Anu & Ansah, 2023; Bouschery et al., 2023; Sigala et al., 2024) leaving a gap in guidance on the design and development of systems with GenAI components representing core functionality. Thus, a reference architecture can serve as an artefact for companies to better understand and communicate the development of systems with GenAI components, facilitating the effective implementation and integration of GenAI use cases in the corporate landscape (Feuerriegel et al., 2024). To address this gap and introduce an interdisciplinary IS perspective, this study explores the following research questions:

RQ1: *What design requirements does GenAI introduce for developing AI-enabled information systems?*

RQ2: *How can a reference architecture be designed to guide the development of GenAI systems?*

To identify design requirements and develop the reference architecture based on them, we applied a Design Science Research (DSR) approach, incorporating insights from consortium workshops, literature reviews, and expert interviews. This iterative process ensured both practical relevance and theoretical rigor, offering assistance for companies, but also an opportunity for design research on GenAI. The reference architecture, based on eight design requirements, provides a strategic framework for companies building GenAI systems across various business functions. It addresses key challenges, streamlines solution development, ensures adaptability, and fosters effective communication among stakeholders. For academics, this architecture highlights changes in GenAI system design compared to traditional AI, focusing on GenAI-specific challenges. It abstracts away general IT system design issues, adding to the discourse on GenAI system design and providing a common framework for discussing GenAI design and development. The remainder of this paper is organised as follows: We first introduce related work on the potential and challenges of GenAI, considering its unique characteristics, and review existing Machine Learning (ML) and GenAI developments to identify existing knowledge gaps. Next, we describe our methodology and present our findings. We then discuss these results in light of prior work and conclude by outlining limitations and suggesting directions for future research.

2 Related Work

2.1 Generative AI and Large Language Models

The field of AI has recently raised the potential of GenAI, which is capable of generating new data from patterns in its large training data and reaching human-like capabilities in terms of processing and creating text, images, video, and audio (Feuerriegel et al., 2024; Schneider et al., 2024). The basis of these types of technologies refers to foundation models representing "*large-scale AI models that are pre-trained on vast amounts of data and that can be adapted to for downstream application*" (Schneider et al., 2024). Large language models (LLM) constitute a subclass of foundation models that show new capabilities regarding their natural language understanding and language generation (Feuerriegel et al., 2024). For example, ChatGPT is based on a GPT language model and is specifically fine-tuned for the task of chatting with humans. Through an approach of self-supervised and in-context learning, LLMs possess extensive world knowledge, however, require finetuning and adaptation to domain-specific context and data. In enterprises, the potential of generative AI lies in linking internal knowledge bases to large language models, as pre-trained models are not domain specific. Hence, research and practice

focus on retrieval-augmented generation (RAG) systems, which search enterprise databases and provide retrieved knowledge as context for the language model (Lewis et al., 2020).

GenAI's unique technical characteristics present challenges to businesses and disrupt current architectural and procedural approaches within the AI development life cycle (Schneider et al., 2024). GenAI shifts from traditional discriminative tasks like predictions to generative tasks that can produce complex, natural language-based content. While the responses might sound convincing and coherent, the nature of the large-scale systems inherit the risk of hallucinations (Azamfirei et al., 2023). Hence, implementing such technologies in enterprises requires measures such as increased explainability to mitigate or reduce the risk of erroneous responses to keep overall performance high (Alkaissi & McFarlane, 2023, Sallam, 2023). Furthermore, the interactive nature of GenAI introduced prompt engineering, which allows users to guide and influence AI outputs using natural language. This shift moves beyond programming and model training to enhance user control and interactivity (White et al., 2023, Liu & Chilton, 2022) while challenging the controllability and reproducibility of GenAI behaviour. Organisations are tasked with selecting suitable generative AI models out of a large set of pre-trained models (Feuerriegel et al., 2024, Kolides et al., 2023). Moreover, there is an emerging need for data-centric AI management spanning training data development, inference data development, and data maintenance (Hamid, 2022, Zha et al., 2023). Given these changes and the associated challenges, existing best practices and knowledge in AI system development are being put to the test. The unique characteristics of this broadly applicable technology demand a critical reevaluation of established approaches to designing AI systems.

2.2 Designing AI Systems

Given the widespread use of AI technology across all industries, scholars started to investigate the peculiarities of building AI-based systems along the AI life cycle (Ozmen Garibay et al., 2023, Berente et al., 2021). From a management perspective, AI presents challenges such as efficient human-AI interactions (i.e., trust, aversion, overreliance, etc.) or ethics and accountability (i.e., bias, fairness, responsibility) (Berente et al., 2021). In terms of software engineering and development, AI requires innovative approaches to governing data, training models, evaluating performance and consequences, and studying its effective use (Zolbanin & Aubert, 2025). Designing AI systems typically follows overarching phases including (1) data collection and curation, (2) model design and experimentation, and (3) deployment and use (Ozmen Garibay et al., 2023). Prior research in the domain of AI, has not only contributed to a better understanding of the AI lifecycle (e.g., Amershi et al., 2019, Zolbanin & Aubert, 2025) but also derived guidelines and approaches to building better, more responsible and human-centered systems (e.g., Abedin et al., 2022, Dennehy et al., 2023). Prior studies highlighted the importance of ensuring data quality for subsequent model training including data collection, data cleaning, and labeling (Amershi et al., 2014). In traditional machine learning projects, domain experts are involved in labeling the data suitable for training the AI model. This perspective of human-in-the-loop was further extended to involving humans in auditing and altering the AI systems in both model development and model use phases (Grønsund & Aanestad, 2020). During the phase of model design and experimentation, AI developers perform feature engineering, select, train, and evaluate models (Amershi et al., 2019). The main leverages for improving model performance relate to extracting the most valuable features (Turner, 1999) and tuning the hyperparameters of machine learning algorithms (Probst et al., 2019). Subsequently, the models are deployed and used in real-world settings. In the realm of DSR, scholars aggregated design knowledge for building useful AI applications in various domains (e.g., Meske & Bunde, 2022) and improving the interaction between humans and AI (e.g., Storey et al., 2024, Cao et al., 2024). Finally, AI research discussed approaches for continuous improvement and operations and introduced the concept of machine learning operations (MLOps) (e.g., Steidl et al., 2023, Kreuzberger et al., 2023). The shift from traditional machine learning tasks toward generative, non-deterministic LLM-based tasks significantly increases complexity for companies. For instance, designing GenAI systems requires novel considerations such as prompt engineering strategies (White et al., 2023), new approaches to testing and monitoring, and addressing “hallucinations” that may surface from opaque language models (Ji et al., 2023). Induced heightened uncertainty and adjusting to new

design paradigms potentially hinder the scaling of GenAI applications in practice. Prior studies on GenAI focus largely on use cases (Sigala et al., 2024), ethical and governance challenges (Dennehy et al., 2023), or open questions about trust, bias, and accountability (Fui-Hoon Nah et al., 2023). However, the fundamental design of a GenAI system such as its architecture and key development processes are sparsely addressed. While unique interaction capabilities such as chain-of-thought or AI chains are increasingly investigated (e.g., Wu et al., 2022), few studies consider overarching design and implementation challenges concerning the development of a scalable architecture that enables to build standardized implementations of AI use cases that researchers and practitioners face. To address this, we engage GenAI experts to develop a reference architecture guiding the design, development, and improvement of GenAI systems.

3 Methodology

To design and evaluate a reference architecture for GenAI systems, we employed DSR methods. Originating in computer and engineering sciences, DSR aims to address practical problems through developing innovative artifacts (Peffer et al., 2007), such as concepts, models, or instantiations (Hevner et al., 2004). Our problem-centered DSR approach included two iterative design and evaluation cycles (Peffer et al., 2007), combining consortium workshops, literature and expert interviews.

Problem-Centered DSR Initiation: Our study was motivated by challenges observed in a research consortium involving quarterly meetings of researchers and practitioners focused on GenAI. Established in November 2022 with four participating companies, initial discussions covered GenAI use cases and governance. However, as companies progressed in GenAI adoption, attention shifted to technical setups and architecture design. By March 2024, a dedicated workshop brought together IT architects from the consortium to share architectural proposals and highlight development and design challenges, helping us identify key pain points and core issues. Insights gathered on the abstract problem and requirements for a GenAI adapted architecture were then analysed and supplemented with existing academic research and grey literature (i.e. publicly available descriptions of architectures deployed by businesses) on GenAI architectures (Lee et al., 2011). The combined analysis led to seven key requirements, designed to be abstract enough to generalise across different solutions, ensuring broad applicability. Thus, the results of this DSR phase include a validated understanding of the problem and a first specification of design requirements.

1st Iteration: In the following consortium workshop in June 2024, we again de-abstracted to the practical challenges and presented and discussed the identified requirements for a GenAI architecture to the participating practitioners. Feedback to the presented design requirements was collected both verbally and in written format on feedback cards. Based on the collected feedback from the consortium, we iterated and refined the design requirements and then drafted an initial version for a GenAI reference architecture by translating requirements to architectural components.

2nd Iteration: The refined design requirements and the subsequently designed GenAI reference architecture were then evaluated through in-depth semi structured interviews (Myers & Newman, 2007) involving two consortium members and predominantly additional external experts from the field to broaden perspectives and enhance validity (Myers & Newman, 2007). Table 1 below provides a more detailed overview about our interview participants. We tried to achieve a broad diversity of experience and roles including IT architects and CIOs providing higher-level perspectives, whereas ML engineers and consultants provide detailed implementation insights on things such as prompt engineering or orchestration. The interview guide was co-developed by the research team, and each interview involved two researchers. Interviews, averaging 53 minutes, were conducted on Microsoft Teams, transcribed using AI tools including manual processing, and analysed with the qualitative analysis tool Atlas.ti. An open-coding approach was applied, and findings were discussed among the research team (Saldaña, 2009). To minimise errors and ensure inter-coder-reliability and internal validity, the researchers cross-checked the codes. As a result, an additional design requirement was derived by splitting an overly broad existing requirement, while adjustments to the GenAI reference architecture were made at the component level, keeping the overall dimensions and layers intact. We revisited each design requirement

to confirm its alignment with the practical challenges described by practitioners and its appropriately granular representation in the architecture. This iterative development and evaluation cycle anchored the GenAI reference architecture in practical experience and expertise. Our results include a practically grounded and validated problem understanding, a validated set of design requirements, and a validated model for a GenAI reference architecture. This publication reports our findings, thereby contributing to the existing knowledge base and discourse on designing GenAI systems (Hevner, 2007).

ID	Background & Position	Domain	Consortium	Experience	Min.
I01	AI Researcher	Academia	No	5 years	58
I02	IT Architect	Financial Services	Yes	8 years	57
I03	Senior ML Engineer	Software Dev.	No	3 years	54
I04	Manager AI & Data Engineering	Consulting	No	14 years	49
I05	Lead IT-Architect	Aviation	No	21 years	54
I06	AI Researcher	Academia	No	3.5 years	47
I07	Senior Consultant Data & AI	Consulting	No	5 years	44
I08	Lead Data Science & AI	Industrial	No	13 years	54
I09	Business Lead Data & AI	Consulting	No	6.5 years	46
I10	Head of AI Projects	Media	No	8 years	65
I11	ML Engineer	Financial Services	Yes	6.5 years	53
I12	Head of Data & AI	Consulting	No	6.5 years	59
I13	Chief Technology Officer	Life Sciences	No	6 years	47

Table 1. Overview of interview participants.

4 Results

4.1 Design Requirements for Generative AI Systems

Following we report the Design Requirements (DR) derived from the consortium workshops and validated through the interviews. Table 2 summarises the final, validated list of design requirements for a GenAI reference architecture and gives a short description for each.

Design Requirement	Description
<i>DR 1: Data Availability and Representation</i>	The GenAI system requires structured, high-quality, context-specific data from internal and external sources, formatted to suit specific use cases.
<i>DR 2: Interoperability</i>	Internal systems, external systems, databases, models, and tools must communicate through standardised methods to enhance modularisation, replaceability and interaction possibilities across systems.
<i>DR 3: Orchestration</i>	Models must be systematically orchestrated to enable efficient data access, model interactions, and tool integration within flexible workflows, whether structured as agentic systems, multi-agent frameworks, sequential chains, or non-AI systems.
<i>DR 4: Model Evaluation and Adoption</i>	Organisations need to establish robust evaluation frameworks and selection processes to continuously assess and deploy optimal models for each use case.
<i>DR 5: Prompt Engineering</i>	The system requires versioned and iteratively refined prompting strategies tailored to specific use cases and shared across the organisation.
<i>DR 6: Guardrails</i>	The system must establish robust guardrails to ensure compliance with corporate, ethical, and legal standards across all system interfaces.
<i>DR 7: Continuous Oversight and Iterative Improvement</i>	The system must implement continuous monitoring and testing at both model and system levels to enable iterative improvements and experimentation, ensuring optimal performance.

DR 8: <i>Human-in-the-Loop</i>	The system requires human involvement for supervision, feedback, annotation, and decision-making to ensure process accuracy and ongoing system improvement.
--------------------------------	---

Table 2. Summary of the Design Requirements for a GenAI Reference Architecture.

DR 1 - Data Availability and Representation: The availability of an abundance of data is crucial for training generative AI models such as transformer-based language models (Hoffmann et al., 2024). While this is predominantly the case for pretraining foundational models and finetuning models (Min et al., 2023), GenAI applications increasingly rely on high-quality, context-specific data at runtime to improve output quality exemplified in RAG use cases (Lewis et al., 2020). I12 solidifies this requirement: *"I would say the most crucial part is the data availability and the data access. And the more or the easier it is to handle the data, the easier you can do rapid prototyping and build use cases and try to implement actual AI solutions or Gen AI solutions for your company"*. Whether it is GenAI or AI, issues at companies remain that the legacy data architecture constrains storing and processing data from various sources in an integrated fashion (Cao & Iansiti, 2022), reflected by I13: *"First, it's the same as with all data science and data-related technologies: companies typically just don't have their data in order. It's not maintained, it's not clean. There are different versions of documents, and so on"*. Given available data, the challenge for GenAI system remains in modeling data sources in a suitable structure for the underlying use case such as vector databases (Wang et al., 2024) or knowledge graphs (Xu et al., 2024). As I12 put it into words: *"Because in theory, you always have the data, but it's just modeling the data that is always the most difficult part"*. Our interviewed experts put emphasis on how structuring available data - whether it is coming from internal or external source - can add significant value to the overall system: *"I think the enthusiasm around GenAI was so great that you would hear everywhere. Yes, you can just feed it anything, and it will already know how to respond. But during the project and testing phases, we realised that it's not as simple as one might imagine"* (I11). Thus, access to high-quality, domain-specific, proprietary data is essential for developing effective generative AI systems. This data should be structured and adapted to the context, enriching the system with additional information from both internal and external sources.

DR 2 – Interoperability: Interoperability between different modules is a requirement to information systems in general and key to enabling their potential (Hodapp & Hanelt, 2022). I09 addresses the importance of well-defined APIs within GenAI workflows: *"When I think about agent systems, I need APIs to trigger specific applications, certain automations, or data workflows that then provide me with additional information. In this context, I believe APIs are extremely important"*. However, the highly volatile environment in GenAI further stresses the importance of interoperability: *"One must be particularly careful to ensure that integration with LLMs remains flexible and standardised, as the world of LLMs is highly volatile"* (I02). Access to other systems internal or external to the organisation improve the usability of the GenAI system in providing access to extended functionality and access to data sources: *"In the end, it also needs to have an API or something similar. That means, whatever I develop, regardless of the system or platform, only then can I ensure that in the future we can combine these elements or communicate with agents, and so on"* (I08). Adapting and integrating legacy systems is important to ensure seamless operation within the existing system landscape: *"There are quite a few legacy systems, some of which are very central, and we can't just disconnect them directly. That's why we always have an abstraction layer with an API, and we set this as a requirement for our customer"* (I03). Furthermore, standardisation for better interoperability and communication between language models should be considered. Communication formats in chained or agentic systems are not necessarily best handled in natural language as format suitability varies with the underlying task (Chen et al., 2024). Hence, careful considerations of interoperability are required on both between-model and between-system communication.

DR 3 – Orchestration: Prompting strategies such as Chain-of-Thought (Wei et al., 2024) or the ReAct paradigm (Yao et al., 2023) naturally involve repeated calls to an LLM. Multi-agent systems rely on communication structures involving multiple language models with different capabilities, instructions and knowledge bases (Guo et al., 2024). I12 differentiates: *"An LLM call is rarely an LLM call."*

Especially if you are looking at agents, an LLM call is actually ten LLM calls. Then one of them, you could have humans calling the RAG, but in most cases, you're going to have some LLM involvement before you go to the RAG world, if only to restructure the input". Further emphasised by I13: *"In the background, various prompt templates dynamically incorporate the relevant information at the right time and place to generate partial increments. And all of this builds upon each other".* Thus, a GenAI system requires careful consideration of model selection based on capabilities, along with efficient model routing, chain construction or general workflow orchestration, including the integration and communication with other systems. Workflow orchestration depends on multiple things such as applied prompting strategies or the autonomy of models in task execution.

DR 4 - Model Evaluation and Adaptation: To achieve an optimal task-model fit, model choice plays an essential role in building GenAI systems. Model selection implies that criteria for model evaluation must be present. Experts acknowledge the need for choosing suitable models for their use cases but iterate on difficulties in expressing model capabilities and how to evaluate them. I02 stresses the dynamic field requiring flexibility: *"You need to be able to quickly and flexibly access new LLMs when they become available, whether it is running them on-premises because they are open source or accessing another version like a newer GPT model. That's quite a challenge. There's also the decision of which LLM to use in different scenarios, as there are various reasons for choosing one [...], but in just a few days there could be completely different LLMs. So, you need to build everything flexibly and encapsulated to adapt to these changes".* Chang et al. (2024) distinguish between ‘what’, ‘where’, and ‘how’ in evaluating LLMs, which includes evaluation on task levels such as summarisation but also more abstract constructs like trustworthiness and robustness. These are however not always applicable in practice: *"We use simple accuracy scores and similar metrics. But things like perplexity and all those benchmarks are almost uninterpretable for us when it comes to specific use cases" (I07),* as there is still uncertainty about exact measurements: *"And that's the challenging part—how do you represent the delta in numbers? In text, you might be able to explain it well, but what metrics do you use to convey the difference? We have metrics like Completeness, Helpfulness, Actuality, I think, and also Hallucination" (I03).* In general, traditional metrics (ROUGE, BLEU, METEOR) have poor correlation with human judgement of hallucination (Ji et al., 2023) and specialised metrics for hallucination remain scarce. With growing reliance on Language-Models-as-a-Service, difficulties in reproducibility and reliability are further increased due to a lack of information from providers (La Malfa et al., 2024).

DR 5 - Prompt Engineering: Effective prompting significantly alters the output of LLM-based AI systems (White et al., 2023) and closely relates to individual model orchestration as I13 describes their experimental approach: *"I think it also depends on the model. Prompt engineering varies a bit per model. Typically, things like Chain-of-Thought are very powerful, and context is important. But then it depends—sometimes things don't work as expected, and you have to do a lot of experimentation. So, the best practices are the classic approaches, in my opinion, and as for prompt libraries—yes, we're basically selling our prompt libraries at the moment".* Prompting requires skill and training (Zamfirescu-Pereira et al., 2023) and both programmers and business users within an organisation could profit from standardised prompt sharing similar to prompt patterns as established by White et al. (2023). I04 evaluates prompt development and sharing as follows: *"The thing is that business users—and it's understandable, as it is not their role—sometimes think prompt engineering is just about phrasing the question clearly. But it's also about strategies like one-shot prompting, few-shot prompting, or realising that you need to include examples in the prompt. All these techniques, these real prompting strategies—yes, there is value in sharing them. I've seen interest in this. However, I haven't seen a prompt library, like an enterprise-level solution; I haven't come across one".* Nonetheless, output evaluation should not be restricted to a single model call but evaluated on the overall system including communication between different modules. This aligns with more recent developments such as engineering prompts using varieties of RAG-patterns (Lewis et al., 2020), where prompts follow engineered schemata or a higher reliance on model-based reasoning (Yao et al., 2023).

DR 6 – Guardrails: Guardrails are predominantly required to block content categorised as being harmful, incorrect, system misuse or threat to privacy and security (Fui-Hoon Nah et al., 2023). In a

larger scope a responsible use of AI encompasses addressing the effects of using AI for decision-making processes, chat settings, recommender systems or classification tasks (Dennehy et al., 2023). Issues will become more prevalent with increasingly agentic systems not being fully under human control (Chan et al., 2023). I13 states the following considerations: *"Traditionally, we have these three points—Corporate, Ethical, and Legal—which should go into our check module, so it's uploadable and dynamically checks content to see if anything is inappropriate"*. Companies differ between applications that are for internal or external use. Customer-facing GenAI systems need to be handled with extreme care as there is higher exposure to adversarial attacks or larger damage induced by generating undesired content (Kenthapadi et al., 2023). I05 relates the issue to the black-box nature of models: *"It becomes more critical with the models that actually interact externally. The main challenge is to ensure that they don't convey incorrect information, that they communicate with the right tone, and that everything is properly tested to get a handle on the black-box nature of these systems"*. Many companies rely on the guardrails already provided by model vendors, acknowledging that it would be challenging to implement their own, even though it is necessary: *"Yeah, with smaller domain-specific models, you need to know how to effectively implement guardrails at scale in a repeatable way on top of your raw LLM. Because if you just take a model from HuggingFace and ask it to tell a racist joke, it probably will, whereas with OpenAI's ChatGPT at this point, it wouldn't respond that way you would have to put in effort and try to trick the model into doing so"* (I04).

DR 7 – Continuous Oversight and Iterative Improvement: The development of GenAI systems requires an experimental mindset. The generative variability of model outputs (Weisz et al., 2024) or unexpected deterioration of model performance requires careful monitoring and testing of the system. I04 describes the process as: *"It's like, I have version one. I evaluate, I track, I monitor, I refine, and maybe I release a version two of the system with a tweaked prompting strategy. I track versions and I have experiments and I can see like, yeah, this version of the whole system, not the model, it's maybe better with different chunking, for instance"*. All modules within a generative AI system, such as information retrieval in a RAG system or adjustments to prompting strategies, can influence subsequent steps in unpredictable ways. Therefore, pipelines must be designed for flexibility to support rapid iteration. Additionally, providing sandboxes for quick testing cycles enhances the efficiency of experimental setups: *"Because, in general, when you set something up, it should be possible to quickly adjust, retrain, or make other changes as needed. The pipeline should be functional and easy to implement for these adjustments"* (I11). Nonetheless, monitoring and tracking systems postulates the presence of a testing suite for evaluation on system, model and data level, which remains a challenge as described by I05: *"My gut feeling is that we don't yet have a fully developed testing framework for this. I'm not sure if anyone in the world even has one"*. While system monitoring is essential, it is equally important to stay informed about market developments. The frequent release of new open-source models and proprietary models accessible via APIs can present opportunities for cost savings and performance enhancements. However, achieving this requires a system architecture that facilitates seamless module replacements and upgrades: *"The market is, of course, evolving rapidly. New models are coming out every few weeks or months. You have to keep an eye on that and be willing to replace the model you're currently using with a newer one if it's cheaper, offers better quality, or for other reasons. This will be important in the future as well"* (I08).

DR 8 - Human-in-the-Loop: Human-in-the-loop or human-AI collaboration has already been devoted a lot of research in IS literature (Dennis et al., 2023). While we do not engage in discussion about how to structure these interactions, we identify three human tasks essential to GenAI systems development. First, human supervision serves as an organisational guardrail in ensuring correct system output and throughput: *"Yes, we actually have that at every step. Everywhere something is generated by the LLM, we have defined processes where humans review it"* (I03). *"We always have an additional human layer in between to check if it fits, makes sense, or if any adjustments are needed. This way, we ensure that even when using GenAI, we can maintain the quality standards"* (I06). Hence, integration of steps involving human judgement and their optimal integration within AI workflows becomes essential. Second, humans are required in organising and annotating data sources ensuring high-quality standards. Companies for example build dedicated knowledge bases as described by I03: *"The entire knowledge*

database is also under human authority and is not automatically assembled from external data sources. It always requires a person, an expert within the organisation, to review and set it up. So, implicitly, this approach also includes a quality control aspect". Third, human feedback is required for improving a running system. Feedback collected at runtime can serve as quality control as well as source of training data as elaborated by I10: *"A feedback system, where each user can give a thumbs-up or thumbs-down, along with a text entry explaining what they liked or disliked. We then use this feedback for generating training data or even as a basis for creating additional synthetic data."*

4.2 Reference Architecture for Generative AI Systems

The proposed reference architecture for GenAI systems (Figure 1) was developed based on the design requirements and validated through the expert interviews. It is structured into four dimensions, each comprising various layers and components. At its core is the GenAI dimension, which integrates diverse components to define its functionality. Preceding this are the human and interface dimensions, which outline access mechanisms, while the enablers form the more technical foundation. Surrounding certain layers are dotted boxes representing enterprise settings such as corporate policies, evaluation metrics, model, and data governance, highlighting the need for both technological and organisational considerations. The architecture serves as a flexible reference, not necessarily requiring in-house development or complete implementation. Organisations can adapt, exclude, or outsource modules based on their specific use cases and strategic priorities. Each dimension is detailed with references to the design requirements below.

Human: This dimension addresses the human role involved in the functioning of the overall system. Humans supervise actions or verify system outputs. Additionally, through both feedback and annotation systems can be improved from a data perspective (DR8). Conducted experiments – preferably within testing sandboxes – assist in achieving system improvements (DR7). Changes in models, data sources, prompting strategies or orchestration could potentially impact the overall system and require carefully designed experiments.

Interfaces: Interactions with generative models are increasingly multimodal (Gao et al., 2024) and interfaces could even be generated dynamically based on a users input (Okopnyi et al., 2024). Better visual understanding of documents (Wang et al., 2024) or easy translation of audio (Radford et al., 2022) improves interoperability and orchestration within the system and facilitates the interaction with humans due to lower data processing restrictions (DR2, DR3).

GenAI System: At the core of the reference architecture, the GenAI system is divided into four interacting layers. Encompassing the whole system is the module for continuous testing and monitoring accounting for the probabilistic uncertainty inherent to generative AI and reflecting the experimental approach required during development (DR4, DR7). The *security layer* serves as protection against attacks and helps steer the behavior of the model. Checking inputs and outputs, the system ensures that harmful requests do not reach the model and that the model operates according to company policies and ethical guidelines. This layer also addresses generative AI-specific issues such as hallucinations by incorporating features like fact-checking, providing sources, or offering confidence scores. Additionally, IAM roles need to be defined such that different users have access to models and data based on their roles within the organisation (DR6). Within the *application layer*, we distinguish between three GenAI-specific application modules. A single system may incorporate multiple agentic application logics, represented by the layered structure. Model orchestration, prompting strategies (DR5), and tool access are closely interconnected, where changes in one module impact the others (DR3). Therefore, these components need to be tested in combination using custom strategies and metrics tailored to specific use cases. This approach presupposes interoperable modules (DR2), and the choice of model also influences how these interactions unfold (DR4). The *connection layer* incorporates API representations, which address not only how GenAI systems are triggered or prompted but also extend beyond this scope. It includes ensuring access to internal and external systems, such as ERPs and CRMs, to expand data availability (DR1) and application capabilities, particularly in organisations with legacy systems (DR2). Additionally, the use of prompting strategies and tool calling may require model-

generated code (Liu et al., 2023) to be executed within secure application spaces (Wang et al., 2024). The *model layer* differentiates between general-purpose models, such as GPT-4 or Claude from Anthropic, and custom trained models. General-purpose models offer a broad range of capabilities, while customised versions are tailored to specific company use cases or trained to excel in particular tasks. These models can be sourced from third parties or developed and hosted by the organisation itself. Model ownership is closely related to security considerations, as some queries may need to be processed on-premises to ensure data privacy. Training pipelines and corresponding data sources need to be preestablished to ensure quick and iterative development cycles (DR1). This postulates clear definition of model evaluation metrics and detection of divergences in performance (DR4).

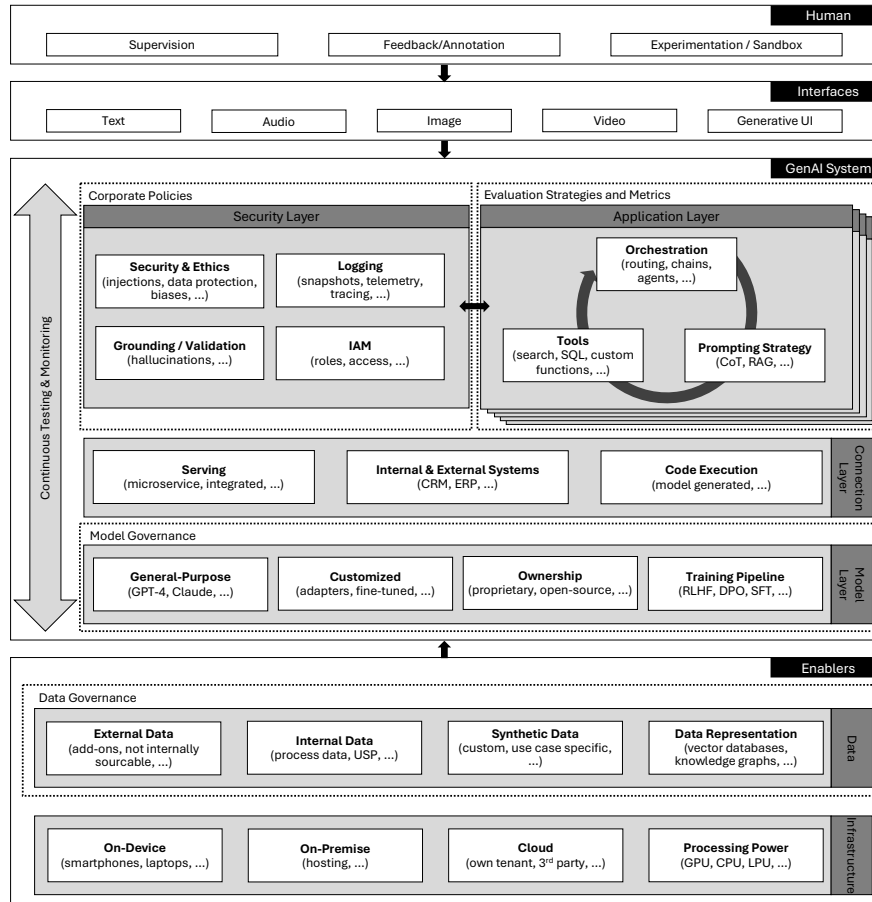


Figure 1. A Reference Architecture for Generative AI Systems.

Enablers: The enablers dimension encompasses data and infrastructure, which serve as foundational elements for the GenAI system. Data can be sourced and represented in various ways (DR1), including synthetically created to tailor to specific use cases (Li et al., 2023). It is a critical asset required across the entire architecture, whether for adding context at runtime or for training organisation-specific models. High-quality data acts as a key differentiator in the performance of AI systems (Rangan & Yin, 2024). Infrastructure is driven by the system's use cases and requirements. Models can be deployed in different environments: directly on-device, on-premises for enhanced data security, or in the cloud. These deployment choices influence the hardware used, leading to decisions between GPUs, CPUs, or LPUs, depending on the specific processing power needs of the system.

5 Discussion

Employing a design science research approach (Peppers et al., 2007) we iteratively defined eight design requirements for GenAI systems, which served as the foundation for drafting and evaluating a reference

architecture and delineates the unique problem space of designing and developing GenAI systems (Maedche et al., 2019). By bridging the gap between technical system design and organisational needs, our research extends the existing body of knowledge on the design and development of AI systems to the domain of GenAI, where new paradigms of prompt-driven interactions, non-deterministic behavior, and novel interaction mechanisms require different design and development considerations. Our reference architecture takes a holistic perspective by addressing not only technical aspects but also business and governance concerns. From a technical perspective, it ensures robust and considerate data handling, model orchestration, and interoperability. Business considerations are integrated through the focus on modularisation, flexibility and continuous improvement. The modularisation and layered structure allow organisations to evaluate individual system modules according to their needs and assists in sourcing decision-making processes. Finally, governance aspects are emphasised through the implementation of guardrails and humans-in-the-loop to comply with ethical, legal, and corporate standards, which are essential given the increasing regulatory scrutiny around GenAI systems. By deriving a reference architecture as an artifact of a nascent design theory (Gregor & Hevner, 2013), we contribute prescriptive knowledge to the literature on AI development practices by identifying novel approaches necessary for implementing GenAI systems, while still addressing the human role within the system and AI lifecycle (Ozmen Garibay et al., 2023). Our results highlight the need for further research into the challenges of designing GenAI systems, emphasising that GenAI's distinct characteristics necessitate adaptations. For instance, the non-deterministic nature and reliance on prompt-based interactions call for a rigorous approach to system monitoring, iterative experimentation and novel evaluation criteria. Moreover, we contribute to the literature on AI governance by deepening the understanding of designing increasingly agentic, autonomous and less controllable GenAI systems (Berente et al., 2021). Stating dynamic mechanisms of model orchestration or human involvement within system processes integrated into such a sociotechnical system, our research addresses necessary design decisions. Finally, we extend the existing knowledge on GenAI systems (Feuerriegel et al., 2024; Kolides et al., 2023; Schneider et al., 2024) by adapting a design-centric perspective. Drawing from field insights, we explore how emerging GenAI use cases can be systematically implemented and efficiently designed, highlighting key trade-offs involved in the process. Building on the work of Banh & Strobel (2023), we provide a comprehensive view of information systems with GenAI components. Our contribution includes a deeper analysis on the design of the application layer and the addition of critical considerations such as security and human roles. From a practical perspective, the reference architecture developed in this study offers value for organisations integrating GenAI systems. It serves as a strategic blueprint, helping businesses streamline system design and development by providing a clear framework that aligns with best practices. For IT architects and system designers, this architecture facilitates scalable, modular implementations, reducing the time and cost associated with building GenAI solutions from scratch. Beyond guiding development, the architecture supports strategic decision-making. CTOs and IT project managers can use it to evaluate trade-offs, such as the choice between cloud-based and on-premises deployments or proprietary versus open-source models. By clarifying complexities involved in GenAI integration, it helps prioritise investments and make informed buy-vs-make decisions, ensuring efficient resource allocation. Furthermore, the architecture promotes standardisation within a company's GenAI system landscape, allowing to reuse modules across projects. This standardisation not only reduces duplication but also fosters organisational learning. Domain experts and IT teams can leverage this shared framework to enhance cross-functional collaboration and communication, ensuring that business goals align with technical capabilities.

6 Conclusion and Future Research

GenAI holds significant potential to enhance efficiency and transform business processes (Wessel et al., 2023). Yet, it also presents new complexities – from prompt-driven interactions to unpredictable outputs – that do not readily fit into conventional AI development processes. Thus, scaling its benefits requires advancements and a rethinking of system design and integration. To support researchers and practitioners to make value out of GenAI technologies, this study introduces a GenAI reference architecture grounded in eight design requirements encompassing layers and components necessary for

effective system development, providing a foundational framework for both academic research and practical applications. Practitioners could use this reference architecture for implementation guidance and to gain a holistic understanding helping to incorporate GenAI systems into the overall IT landscape. Academics could orientate their research more precisely within the individual components and implement improvements considering the overall system. Nevertheless, our study has limitations that open avenues for future research. First, while adjustments were made based on interview feedback, the architecture was not subjected to a secondary validation. A secondary evaluation could be conducted in the form of an action design research where the reference architecture is validated in a live environment. Second, further evaluation should include additional experts outside the technical field such legal experts and engage in more in-depth discussion with experts in specialised fields like prompt engineering. Third, as GenAI technologies rapidly evolve, our architecture reflects the current state of the field. Continuous adaptation will be necessary to incorporate emerging features and maintain its relevance. Moreover, future research could focus on testing the architecture in real-world settings to refine its applicability further. Beyond these limitations, there remain unexplored areas that warrant deeper investigation. As organisations face crucial build-or-buy decisions, limited in-house expertise often drives reliance on third-party solutions, introducing complexities like vendor dependencies and data privacy issues. Investigating these impacts on the development of GenAI systems is a promising area for future research. Furthermore, while our study addressed the components of GenAI system design, it did not fully explore the potential and implications of increasingly agentic systems. Further research is needed to understand the design challenges in terms of multi-agent configurations or interaction orchestration.

7 Acknowledgements

This work was supported by Innosuisse and is part of the project 104.954 IP-SBM

References

- Abedin, B., Meske, C., Junglas, I., Rabhi, F., & Motahari-Nezhad, H. R. (2022). Designing and Managing Human-AI Interactions. *Information Systems Frontiers*, 24(3), 691–697. <https://doi.org/10.1007/s10796-022-10313-1>
- Alkaissi, H., & McFarlane, S. I. (2023). Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*, 15(2), e35179. <https://doi.org/10.7759/cureus.35179>
- Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., & Zimmermann, T. (2019). Software engineering for machine learning: A case study. *Proceedings of the 41st International Conference on Software Engineering: Software Engineering in Practice*, 291–300. <https://doi.org/10.1109/ICSE-SEIP.2019.00042>
- Amershi, S., Cakmak, M., Knox, W. B., & Kulesza, T. (2014). Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4), Article 4. <https://doi.org/10.1609/aimag.v35i4.2513>
- Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care (London, England)*, 27(1), 120. <https://doi.org/10.1186/s13054-023-04393-x>
- Baidoo-Anu, D., & Ansah, L. O. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, 7(1), 52–62.
- Banh, L., & Strobel, G. (2023). Generative artificial intelligence. *Electronic Markets*, 33(1), 63. <https://doi.org/10.1007/s12525-023-00680-1>
- Berente, N., Gu, B., Recker, J., & Santhanam, R. (2021). Special Issue Editor's Comments: Managing Artificial Intelligence. *Management Information Systems Quarterly*, 45(3), 1433–1450.
- Bouschery, S. G., Blazeovic, V., & Piller, F. T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. *Journal of Product Innovation Management*, 40(2), 139–153. <https://doi.org/10.1111/jpim.12656>

- Cao, L., Lindman, J., Cronholm, S., & Göbel, H. (2024). Design Principles for Hybrid Decision Support Systems in IT Service Management. *ECIS 2024 Proceedings*.
- Cao, R., & Iansiti, M. (2022). Digital transformation, data architecture, and legacy systems. *Journal of Digital Economy*, 1(1), 1–19. <https://doi.org/10.1016/j.jdec.2022.07.001>
- Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., ... Maharaj, T. (2023). Harms from Increasingly Agentic Algorithmic Systems. *Proceedings of the 2023 ACM Conference on FAccT*, 651–666. <https://doi.org/10.1145/3593013.3594033>
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.*, 15(3), 39:1-39:45. <https://doi.org/10.1145/3641289>
- Chawla, C., Chatterjee, S., Gadadinni, S. S., Verma, P., & Banerjee, S. (2024). Agentic AI: The building blocks of sophisticated AI business applications. *Journal of AI, Robotics & Workplace Automation*.
- Chen, L., Zaharia, M., & Zou, J. (2023). *How is ChatGPT's behavior changing over time?* (arXiv:2307.09009). arXiv. <https://doi.org/10.48550/arXiv.2307.09009>
- Chen, W., Yuan, C., Yuan, J., Su, Y., Qian, C., Yang, C., Xie, R., Liu, Z., & Sun, M. (2024). Beyond Natural Language: LLMs Leveraging Alternative Formats for Enhanced Reasoning and Communication. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 10626–10641). Association for Computational Linguistics. <https://aclanthology.org/2024.findings-emnlp.623>
- Dennehy, D., Griva, A., Pouloudi, N., Dwivedi, Y. K., Mäntymäki, M., & Pappas, I. O. (2023). Artificial Intelligence (AI) and Information Systems: Perspectives to Responsible AI. *Information Systems Frontiers*, 25(1), 1–7. <https://doi.org/10.1007/s10796-022-10365-3>
- Dennis, A. R., Lakhiwal, A., & Sachdeva, A. (2023). AI Agents as Team Members: Effects on Satisfaction, Conflict, Trustworthiness, and Willingness to Work With. *Journal of Management Information Systems*, 40(2), 307–337. Scopus. <https://doi.org/10.1080/07421222.2023.2196773>
- Enholm, I. M., Papagiannidis, E., Mikalef, P., & Krogstie, J. (2022). Artificial Intelligence and Business Value: A Literature Review. *Information Systems Frontiers*, 24(5), 1709–1734. <https://doi.org/10.1007/s10796-021-10186-w>
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative AI. *BISE*, 66(1), 111–126. <https://doi.org/10.1007/s12599-023-00834-7>
- Fügener, A., Grahl, J., Gupta, A., & Ketter, W. (2022). Cognitive Challenges in Human–Artificial Intelligence Collaboration: Investigating the Path Toward Productive Delegation. *Information Systems Research*, 33(2), 678–696. <https://doi.org/10.1287/isre.2021.1079>
- Fui-Hoon Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information Technology Case and Application Research*, 25(3), 277–304. <https://doi.org/10.1080/15228053.2023.2233814>
- Gao, D., Hu, S., Lin, Q., & Shou, M. Z. (2024). AssistGPT: Towards Multi-modal Agent for Human-Centric AI Assistant. *Proceedings of the 5th International Workshop on Human-Centric Multimedia Analysis*, 3–5. <https://doi.org/10.1145/3688865.3689482>
- Gregor, S., & Hevner, A. R. (2013). Positioning and presenting design science research for maximum impact. *MIS Q.*, 37(2), 337–356. <https://doi.org/10.25300/MISQ/2013/37.2.01>
- Grønsund, T., & Aanestad, M. (2020). Augmenting the algorithm: Emerging human-in-the-loop work configurations. *Journal of Strategic Information Systems*, 29(2). Scopus. <https://doi.org/10.1016/j.jsis.2020.101614>
- Guo, T., Chen, X., Wang, Y., Chang, R., Pei, S., Chawla, N. V., Wiest, O., & Zhang, X. (2024). *Large Language Model Based Multi-agents: A Survey of Progress and Challenges*. 9, 8048–8057. <https://doi.org/10.24963/ijcai.2024/890>
- Hagos, D. H., Battle, R., & Rawat, D. B. (2024). Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. *IEEE Transactions on Artificial*

- Intelligence*, 1–21. IEEE Transactions on Artificial Intelligence. <https://doi.org/10.1109/TAI.2024.3444742>
- Hamid, O. H. (2022). From Model-Centric to Data-Centric AI: A Paradigm Shift or Rather a Complementary Approach? *2022 8th International Conference on Information Technology Trends (ITT)*, 196–199. <https://doi.org/10.1109/ITT56123.2022.9863935>
- Hassan, A. E., Lin, D., Rajbahadur, G. K., Gallaba, K., Cogo, F. R., Chen, B., Zhang, H., Thangarajah, K., Oliva, G., Lin, J. (Justina), Abdullah, W. M., & Jiang, Z. M. (Jack). (2024). Rethinking Software Engineering in the Era of Foundation Models: A Curated Catalogue of Challenges in the Development of Trustworthy FMware. *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 294–305. <https://doi.org/10.1145/3663529.3663849>
- Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 19(2). <https://aisel.aisnet.org/sjis/vol19/iss2/4>
- Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Q.*, 28(1), 75–105.
- Hodapp, D., & Hanelt, A. (2022). Interoperability in the era of digital innovation: An information systems research agenda. *Journal of Information Technology*, 37(4), 407–427. <https://doi.org/10.1177/02683962211064304>
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., ... Sifre, L. (2024). Training compute-optimal large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030.
- Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., Zhang, C., Wang, Z., Yau, S. K. S., Lin, Z., Zhou, L., Ran, C., Xiao, L., Wu, C., & Schmidhuber, J. (2023, October 13). *MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework*. The Twelfth International Conference on Learning Representations.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, 55(12), 248:1–248:38. <https://doi.org/10.1145/3571730>
- Kenthapadi, K., Lakkaraju, H., & Rajani, N. (2023). Generative AI meets Responsible AI: Practical Challenges and Opportunities. *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 5805–5806. <https://doi.org/10.1145/3580305.3599557>
- Klarna. (2024, February 27). *Klarna AI assistant handles two-thirds of customer service chats in its first month | Klarna International*. <https://www.klarna.com/international/press/klarna-ai-assistant-handles-two-thirds-of-customer-service-chats-in-its-first-month/>
- Kolides, A., Nawaz, A., Rathor, A., Beeman, D., Hashmi, M., Fatima, S., Berdik, D., Al-Ayyoub, M., & Jararweh, Y. (2023). Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory*, 126, 102754. <https://doi.org/10.1016/j.simpat.2023.102754>
- Kreuzberger, D., Kühl, N., & Hirschl, S. (2023). Machine Learning Operations (MLOps): Overview, Definition, and Architecture. *IEEE Access*, 11, 31866–31879. IEEE Access. <https://doi.org/10.1109/ACCESS.2023.3262138>
- La Malfa, E., Petrov, A., Frieder, S., Weinhuber, C., Burnell, R., Nazar, R., Cohn, A., Shadbolt, N., & Wooldridge, M. (2024). Language-Models-as-a-Service: Overview of a New Paradigm and its Challenges. *J. Artif. Int. Res.*, 80. <https://doi.org/10.1613/jair.1.15865>
- Lee, J. S., Pries-Heje, J., & Baskerville, R. (2011). Theorizing in design science research. *Proceedings of the 6th International Conference on Service-Oriented Perspectives in DSR*, 1–16.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 9459–9474.

- Li, Z., Zhu, H., Lu, Z., & Yin, M. (2023). Synthetic Data Generation with Large Language Models for Text Classification: Potential and Limitations. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on EMNLP*. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.emnlp-main.647>
- Liu, C., Lu, S., Chen, W., Jiang, D., Svyatkovskiy, A., Fu, S., Sundaresan, N., & Duan, N. (2023). Code Execution with Pre-trained Language Models. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the ACL 2023* (pp. 4984–4999). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-acl.308>
- Liu, V., & Chilton, L. B. (2022). Design Guidelines for Prompt Engineering Text-to-Image Generative Models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–23. <https://doi.org/10.1145/3491102.3501825>
- Lu, Q., Zhu, L., Xu, X., Whittle, J., Zowghi, D., & Jacquet, A. (2024). Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. *ACM Comput. Surv.*, 56(7), 173:1-173:35. <https://doi.org/10.1145/3626234>
- Maedche, A., Gregor, S., Morana, S., & Feine, J. (2019). Conceptualization of the Problem Space in Design Science Research. In B. Tulu, S. Djamasbi, & G. Leroy (Eds.), *Extending the Boundaries of Design Science Theory and Practice* (pp. 18–31). Springer International Publishing. https://doi.org/10.1007/978-3-030-19504-5_2
- Meske, C., & Bunde, E. (2022). Design Principles for User Interfaces in AI-Based Decision Support Systems: The Case of Explainable Hate Speech Detection. *Information Systems Frontiers*, 25(2), 743–773. <https://doi.org/10.1007/s10796-021-10234-5>
- Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., & Roth, D. (2023). Recent Advances in Natural Language Processing via Large Pre-trained Language Models: A Survey. *ACM Comput. Surv.*, 56(2), 30:1-30:40. <https://doi.org/10.1145/3605943>
- Myers, M. D., & Newman, M. (2007). The qualitative interview in IS research: Examining the craft. *Information and Organization*, 17(1), 2–26. <https://doi.org/10.1016/j.infoandorg.2006.11.001>
- Okopnyi, P., Nordberg, O. E., & Guribye, F. (2024). Against Generative UI. *Proceedings of the Halfway to the Future Symposium*, 1–4. <https://doi.org/10.1145/3686169.3686184>
- Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotko, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., ... Xu, W. (2023). Six Human-Centered Artificial Intelligence Grand Challenges. *International Journal of HCI*, 39(3), 391–437. <https://doi.org/10.1080/10447318.2022.2153320>
- Peffer, K., Tuunanen, T., Rothenberger, M., & Chatterjee, S. (2007). A Design Science Research Methodology for Information Systems Research. *J. Manage. Inf. Syst.*, 24(3), 45–77. <https://doi.org/10.2753/MIS0742-1222240302>
- Probst, P., Boulesteix, A.-L., & Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *J. Mach. Learn. Res.*, 20(1), 1934–1965.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2022). *Robust Speech Recognition via Large-Scale Weak Supervision* (arXiv:2212.04356). arXiv. <https://doi.org/10.48550/arXiv.2212.04356>
- Rangan, K., & Yin, Y. (2024). A fine-tuning enhanced RAG system with quantized influence measure as AI judge. *Scientific Reports*, 14(1), 27446. <https://doi.org/10.1038/s41598-024-79110-x>
- Saldaña, J. (2009). *The coding manual for qualitative researchers* (pp. xi, 223). Sage Publications Ltd.
- Sallam, M. (2023). ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare (Basel, Switzerland)*, 11(6), 887. <https://doi.org/10.3390/healthcare11060887>
- Schneider, J. (2024). Explainable Generative AI (GenXAI): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11). <https://doi.org/10.1007/s10462-024-10916-x>
- Schneider, J., Meske, C., & Kuss, P. (2024). Foundation Models. *Business & Information Systems Engineering*, 66(2), 221–231. <https://doi.org/10.1007/s12599-024-00851-0>
- Sigala, M., Ooi, K.-B., Tan, G. W.-H., Aw, E. C.-X., Cham, T.-H., Dwivedi, Y. K., Kunz, W. H., Lethere, K., Mishra, A., Russell-Bennett, R., & Wirtz, J. (2024). ChatGPT and service:

- Opportunities, challenges, and research directions. *Journal of Service Theory and Practice*, 34(5), 726–737. <https://doi.org/10.1108/JSTP-11-2023-0292>
- Steidl, M., Felderer, M., & Ramler, R. (2023). The pipeline for the continuous development of artificial intelligence models—Current state of research and practice. *Journal of Systems and Software*, 199, 111615. <https://doi.org/10.1016/j.jss.2023.111615>
- Storey, V. C., Hevner, A. R., & Yoon, V. Y. (2024). The design of human-artificial intelligence systems in decision sciences: A look back and directions forward. *Decision Support Systems*, 182, 114230. <https://doi.org/10.1016/j.dss.2024.114230>
- Turner, H. (1999). A logic of universal causation. *Artificial Intelligence*, 113(1), 87–123. [https://doi.org/10.1016/S0004-3702\(99\)00058-2](https://doi.org/10.1016/S0004-3702(99)00058-2)
- Wang, D., Raman, N., Sibue, M., Ma, Z., Babkin, P., Kaur, S., Pei, Y., Nourbakhsh, A., & Liu, X. (2024). DocLLM: A Layout-Aware Generative Language Model for Multimodal Document Understanding. In L.-W. Ku, A. Martins, & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (pp. 8529–8548). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.acl-long.463>
- Wang, J., Luo, X., Cao, L., He, H., Huang, H., Xie, J., Jatowt, A., & Cai, Y. (2024). *Is Your AI-Generated Code Really Safe? Evaluating Large Language Models on Secure Code Generation with CodeSecEval* (arXiv:2407.02395). arXiv. <https://doi.org/10.48550/arXiv.2407.02395>
- Wang, X., Wang, Z., Gao, X., Zhang, F., Wu, Y., Xu, Z., Shi, T., Wang, Z., Li, S., Qian, Q., Yin, R., Lv, C., Zheng, X., & Huang, X. (2024). Searching for Best Practices in Retrieval-Augmented Generation. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17716–17736). Association for Computational Linguistics. <https://aclanthology.org/2024.emnlp-main.981>
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2024). Chain-of-thought prompting elicits reasoning in large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 24824–24837.
- Weisz, J. D., He, J., Muller, M., Hoefer, G., Miles, R., & Geyer, W. (2024). Design Principles for Generative AI Applications. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–22. <https://doi.org/10.1145/3613904.3642466>
- Wessel, M., Adam, M., Benlian, A., & Thies, F. (2023). Generative AI and its transformative value for digital platforms. *Journal of Management Information Systems*.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT* (arXiv:2302.11382). arXiv. <https://doi.org/10.48550/arXiv.2302.11382>
- Wu, T., Terry, M., & Cai, C. J. (2022). AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model Prompts. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 1–22. <https://doi.org/10.1145/3491102.3517582>
- Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X., & Li, Z. (2024). Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2905–2909. <https://doi.org/10.1145/3626772.3661370>
- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. *The Eleventh International Conference on Learning Representations*.
- Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. <https://doi.org/10.1145/3544548.3581388>
- Zha, D., Bhat, Z. P., Lai, K.-H., Yang, F., Jiang, Z., Zhong, S., & Hu, X. (2023). *Data-centric Artificial Intelligence: A Survey* (arXiv:2303.10158). arXiv. <https://doi.org/10.48550/arXiv.2303.10158>
- Zolbanin, H., & Aubert, B. (2025). A process model for design-oriented machine learning research in information systems. *The Journal of Strategic Information Systems*, 34(1), 101868. <https://doi.org/10.1016/j.jsis.2024.101868>