

# Design Features for Explainable Generative AI (GenXAI) Systems in Knowledge-Intensive Service Work

Philipp Reinhard  
University of Kassel  
[philipp.reinhard@uni-kassel.de](mailto:philipp.reinhard@uni-kassel.de)

Mahei Manhai Li  
University of St.Gallen  
[mahei.li@unisg.ch](mailto:mahei.li@unisg.ch)

Matteo Fina  
Goethe University Frankfurt  
[matteo.fina@its.uni-frankfurt.de](mailto:matteo.fina@its.uni-frankfurt.de)

Christoph Peters  
University of the Bundeswehr Munich  
[christoph.peters@unibw.de](mailto:christoph.peters@unibw.de)

Jan Marco Leimeister  
University of St.Gallen  
[janmarco.leimeister@unisg.ch](mailto:janmarco.leimeister@unisg.ch)

## Abstract

*The use of generative AI (GenAI) and large language models (LLMs) in knowledge-intensive fields like customer support is rapidly growing. While GenAI responses often appear persuasive, they carry the risk of inaccuracies and hallucinations. Hence, users must critically evaluate responses to reach appropriate reliance and knowledge utilization. Despite technological advancements, design knowledge for enhancing human-GenAI interaction from an explainable AI (XAI) perspective remains lacking. Thus, this study applies the design science research (DSR) approach to develop explanations that aid human interaction with GenAI systems. Drawing from XAI literature and human reasoning theories, we built and evaluated seven design features and instantiated a prototype that contributes to the development of reliable explainable GenAI (GenXAI).*

**Keywords:** Generative AI, XAI, Large Language Models, Design Science

## 1. Introduction

Chatbots, question-answering tools, and other natural language services leverage the generative power of AI, particularly LLMs, which enhance both language understanding and generation (Feuerriegel et al., 2024). LLMs, for example, can be applied to enable self-service systems for customers or to facilitate knowledge use by customer support agents (Brynjolfsson et al., 2025). Despite its potential for knowledge management (Storey, 2025), human interaction with GenAI systems presents significant challenges in establishing an appropriate level of reliance (Kim et al., 2025; Pafla et al., 2024). While LLM responses might sound convincing and coherent, these large-scale systems inherently carry the risk of hallucinations (Azamfirei et al., 2023). Hallucinations constitute responses that do not stay true to the supplied source input and facts, despite appearing

fluent, natural, persuasive, and consistent (Ji et al., 2023). Technological advancements have been made to reduce non-factual responses (Tonmoy et al., 2024). For instance, retrieval-augmented generation (RAG) minimizes inaccuracies by providing LLMs with relevant context retrieved from knowledge sources (Lewis et al., 2020).

However, research on supporting users in detecting incorrect responses and fostering appropriate reliance remains in its early stages (e.g., Leiser et al., 2024; Nahar et al., 2024; Pafla et al., 2024). Current *explainable GenAI (GenXAI)* designs are fragmented and insufficiently grounded in established XAI research. Thus, we advocate for drawing upon established work on explainability and transparency in AI systems (e.g., Bauer et al., 2023) and complementing them with theories of human reasoning and decision-making (Johnson-Laird, 2010). We aim to advance the design of GenXAI systems (Schneider, 2024) that support human-GenAI interaction and knowledge utilization through explanation, asking: *How can GenXAI systems be designed based on XAI and human reasoning theory?*

We follow a DSR approach (Peffer et al., 2007) to design a GenXAI interface for the knowledge-intensive field of customer service, incorporating multiple explanation mechanisms, and to contribute to the research on appropriate reliance (Jussupow et al., 2024; Schemmer et al., 2023). While explainability in XAI encompasses algorithmic, data-driven, and outcome-oriented perspectives, our research specifically examines the latter - investigating how end users perceive and depend upon post-hoc generated outputs in practical service environments.

## 2. Conceptual background

### 2.1. LLMs in customer support

LLMs are deep learning-based foundation models that process text and can be fine-tuned for tasks like

summarization and sentiment analysis (Feuerriegel et al., 2024). LLMs excel at question-answering by generating responses rather than merely retrieving knowledge, unlike traditional information retrieval systems (Durcikova et al., 2011). Customer support departments see the greatest potential in question-answering over knowledge bases (Brynjolfsson et al., 2025; Reinhard et al., 2024), as it helps make information more accessible to customers and knowledge workers (Durcikova et al., 2011; Ko & Dennis, 2011). In the realm of LLMs, research and practical applications have thus shifted toward the utilization of RAG (Lewis et al., 2020; Peng et al., 2023). Thereby, LLMs are infused with retrieved knowledge sources such as customer service tickets (Xu et al., 2024). Despite their potential in various language tasks, LLMs and RAG systems face significant challenges, notably the issue of hallucinations (Zhang et al., 2023). Given the definition of hallucination in psychology, that hallucination is an unreal perception that feels real, we refer to hallucinations as AI responses that seem fluent, natural, persuasive, and consistent but deviate from the supplied source input (Ji et al., 2023). Researchers from the service domain have already contributed approaches to detect and mitigate such fabricated answers (Martino et al., 2023). Techniques such as RAG can lower the risk of generating inaccuracies or fabrications, but cannot guarantee fully factual responses (Lewis et al., 2020).

## 2.2. Explainable generative AI (GenXAI)

XAI aims to provide humans with understandable explanations for AI's output (Bauer et al., 2023) and procedures that not only enhance human understanding and ability to validate outcomes but also engage them in critical appraisal of the outputs generated by AI (Miller, 2019). Although prior research offers valuable insights into the interplay between XAI and user behavior (Förster et al., 2020), few studies examine how explanations shape decision-making in GenAI-based knowledge systems (Kim et al., 2025; Reinhard et al., 2025). With GenAI, XAI is evolving in two ways: (1) generating novel explanations and translating technical information into natural language (e.g., Cambria et al., 2024; He et al., 2025) - also known as conversational XAI; and (2) demanding new approaches to explainability, especially when it comes to mitigating challenges such as hallucinations (e.g., Kim et al., 2025; Leiser et al., 2024; Pafla et al., 2024). Hence, Schneider et al. (2024) introduced GenXAI as an emerging field. However, despite the ability of LLMs to explain their reasoning (self-explanatory) (Nicula et al., 2023),

research on XAI in GenAI-based knowledge systems remains limited (Danry et al., 2023). To close this gap, we integrate established explanation types into GenAI agents to enhance human-AI interaction and systematize design knowledge, focusing on post-hoc, result-oriented explanations where human-GenAI interaction occurs, rather than on deeper and more technical data- or model-level transparency.

## 3. Methodology

We adopt a design science research (DSR) approach by Peffers et al. (2007) comprising problem identification and relevance, design and development of an artifact, and demonstration and evaluation. This approach is suitable as it allows us to both conceptualize explanation mechanisms and evaluate design features (DFs) that integrate XAI with human reasoning principles. First, to ensure practical relevance (Hevner, 2007), we hold two workshops with customer service experts ( $n_1=13$ ;  $n_2=12$ ). In the *first workshop* ( $n=13$ , IT support agents & managers), participants validated ChatGPT (GPT-3) responses to technical issues. They intuitively verified answers by testing suggested steps and identifying multiple errors. For example, they noted that solution steps vary by system type (e.g., browser, app, mobile) and flagged issues like repetitive phrasing, overly specific answers, or anomalous words, concluding that ChatGPT's knowledge remains superficial. In the *second workshop* ( $n=12$ ), we introduced additional information to the GenAI response. Participants found added details, such as sources and alternative responses, useful, but emphasized the need for better guidance in prompting GenAI.

**Table 1. Excerpt of related GenXAI work**

Authors	Features	Key findings
Leiser et al., 2024	Source, confidence score, source quality, etc.	Source links and metric confidence scores were crucial to users.
Schelhorn et al., 2025	Chain-of-thought (CoT) prompting	CoT prompting indicates an indirect positive effect on task effectiveness.
Do et al., 2024	Factuality score and source attribution	Participants preferred color-coded phrases based on factuality scores.
Cheng et al., 2024	Textual counterfactuals	Interactive tool for generating and analyzing textual counterfactuals.
Kim et al., 2025	Sources, explanation, and sources	Explanations increase reliance on all responses, but sources and inconsistencies reduce reliance on incorrect ones.
Steyvers et al., 2025	Model confidence and explanation	Aligning LLM explanations with model confidence improved accuracy perception.
Pafla et al., 2024	Saliency-based explanation, self-explanation	Participants found human saliency maps more helpful than machine ones.

In terms of rigor, we extend these insights by systematically reviewing 107 experimental studies on XAI (2000-2025), spanning literature on appropriate reliance on GenAI advice (Kim et al., 2025) (Table 1). Findings reveal that research on explainability and reliance is fragmented, lacking a strong theoretical foundation, and aggregated design knowledge. To fill this gap, we developed a framework that integrates explanation types with human reasoning styles, deriving seven XAI-informed DFs from a literature review and instantiating them in an RAG-based prototype tested on telecom data. To evaluate the system, we conducted a *think-aloud study* with GenAI users and experts (n=8) (Jaspers et al., 2004), allowing participants to explore explanations using random dataset examples. We applied qualitative content analysis, inductively deriving codes and grouping them under reasoning types to reveal recurring patterns in participant feedback.

#### 4. Design of the GenXAI system

The psychology of reasoning literature offers valuable insights for designing more human-like and effective explanations in XAI. Wang et al. (2019) provide a conceptual framework that synthesizes key reasoning theories, while Hoffman and Klein (2017) explore how individuals formulate and accept explanations, emphasizing logical structures, interpretation processes, and counterexamples. Reasoning is shaped by content, prior knowledge, and individual strategies rather than strict logical manipulation (Johnson-Laird, 2010). These insights underscore the need for explanations that align with formal logic while remaining accessible to diverse users. Building on this foundation and the systematization of XAI literature via our systematic literature review, we identified seven reasoning types that inform the design of explanations in LLM applications (Table 2). We selected these types because they represent foundational modes of human reasoning that directly shape how individuals seek, interpret, and evaluate explanations. Each type draws on cognitive science and XAI research, keeping our DFs both robust and cognitively plausible.

**Factual reasoning.** Factual reasoning manifests in various forms of explanations that explain why a certain decision was made. Explanations that reference underlying domain knowledge, or heuristics, are effective in increasing human confidence in AI outcomes while reducing subjective workload (Gregor & Benbasat, 1999; Lai et al., 2022).

Trace explanations, such as displaying sources and references (Leiser et al., 2024; Sharma et al., 2024), are anticipated to augment trust in the system

(Mao & Benbasat, 2001). In RAG-based LLMs, this is done by referencing the top-k sources. In addition, users with contextualized access to sources (e.g., in the form of referenced paragraphs) (*DF1 References and knowledge sources*) (Kim et al., 2025; Leiser et al., 2024) will be more willing to look for the underlying roots for the correct answer, and request deep explanations (Mao & Benbasat, 2001). In customer support services, tracing responses to verifiable sources plays a crucial role in building trust and delivering consistent, evidence-based assistance. Explanations that offer content-backed justification benefit agents and customers alike by promoting deeper engagement with knowledge (Mao & Benbasat, 2001).

**Table 2. Derivation of design features for GenXAI from literature and human reasoning**

Reasoning Type	Feature	Relevance
<b>Factual</b> - Drawing conclusions based on verifiable facts and evidence ( <i>Why</i> ) (e.g., Kim et al., 2025; Leiser et al., 2024; Mao & Benbasat, 2001).	<b>DF1:</b> References and knowledge sources	<i>Agents can access relevant knowledge from the knowledge base.</i>
<b>Causal</b> - Identifying relationships between causes and effects (e.g., Ashktorab et al., 2019; Bo et al., 2025; Pafla et al., 2024; Papenmeier et al., 2022)	<b>DF2:</b> Source attribution	<i>Agents are directed to the most relevant phrases and efficiently assess the response.</i>
<b>Analogical</b> - Identifying relational similarities ( <i>What-else</i> ) (e.g., Leichtmann et al., 2024; H. Lu et al., 2023)	<b>DF3:</b> Similar cases from the underlying database	<i>Agents can validate the response by comparing it with alternatives.</i>
<b>Transfactual</b> - Demonstrating how varying inputs can influence the outcome ( <i>What-if</i> ) (e.g., Hoffman & Klein, 2017)	<b>DF4:</b> Recommended prompts	<i>Agents can explore the existing knowledge space.</i>
<b>Counterfactual</b> - Considering hypothetical changes that are needed for a different outcome (e.g., Cheng et al., 2024; Naiseh et al., 2023; Silva et al., 2023)	<b>DF5:</b> Critical information and counterfactuals	<i>Agents are prompted to add missing information and to clarify.</i>
<b>Probabilistic</b> - Making decisions based on the likelihood, probability or uncertainty of outcomes (e.g., Cramer et al., 2008; Steyvers et al., 2025)	<b>DF6:</b> Confidence Score	<i>Agents use it to heuristically assess knowledge relevance.</i>
<b>Deductive</b> - Logical inference that begins with general premises and applies structured reasoning to arrive at a certain conclusion (e.g., Wei et al., 2022; Chu et al., 2023)	<b>DF7:</b> Chain-of-thought	<i>Agents can understand how the AI came up with the response and how knowledge was prepared.</i>

**Causal reasoning.** Causal reasoning refers to the cognitive process of identifying and mapping

relationships between causes and effects (Kuhn, 2012). In the context of XAI, it is closely related to factual reasoning, but distinguishes itself by focusing on why a particular prediction occurs, based on the influence of specific input features. To operationalize this logic, feature-based methods, commonly known as feature attribution methods, are widely used within the field of explainable AI (XAI) (Bauer et al., 2023; He et al., 2023). These methods focus on determining the relevant details influencing an AI model's decision-making process, clarifying the factors crucial for its outcomes. Attribution-based explanations link the input with the prediction by demonstrating the importance of certain inputs to the system's advice (Bauer et al., 2023).

Highlighting important keywords and phrases in knowledge references influencing the prediction has been applied to multiple other studies on XAI in text-based decision support systems (e.g., Ashktorab et al., 2019; Lai et al., 2022). Thus, in **DF2 Source attribution**, phrases are color-coded based on their importance, ranging from non-important to highly important, aligning with the previous studies on hallucinations (Bo et al., 2025; Do et al., 2024; Pafila et al., 2024). In customer service contexts, such visual explanations play a critical role in ensuring reliable knowledge utilization. By transparently showing which parts of the customer query or knowledge base response were most influential, DF2 helps service agents verify the basis of the system's advice and reduces reliance on unsupported outputs.

**Analogical reasoning.** The idea behind analogical reasoning lies in providing similar cases that help users compare and relate GenAI responses to familiar data points. Such explanations are also called example-based, case-based, similarity-based, or evidence-based explanations (Dodge et al., 2019). Because these *What-else* explanations provide additional information (Jiang et al., 2022), they require increased mental effort to understand (Miller, 2019) - especially in cases of high uncertainty.

Motivated by traditional machine learning systems, what-else explanations show similar input instances that result in similar model outputs (**DF3: Similar cases**), typically drawn from the training data via nearest-neighbor search (Dodge et al., 2019). For GenAI systems, we instantiate DF3 by providing the nearest retrieved cases from our underlying database.

**Transfactual reasoning.** Transfactual reasoning refers to exploring how different inputs or conditions could change an outcome, often through what-if scenarios (Hoffman & Klein, 2017). It entails looking ahead and examining not only anticipated events but also possible shifts in context, fundamental assumptions, or critical factors. This type of reasoning

supports anticipatory thinking by prompting questions like, "What will happen if things change?" (Gary et al., 2011).

To implement transfactual reasoning, **DF4 Recommended prompts** offers users three related prompts that suggest alternative or additional questions they could ask to explore how the response might change. In the context of customer service, DF5 empowers both customers and service agents to actively navigate the underlying knowledge space. This facilitates a more exploratory and dialogic form of interaction, where users are encouraged not only to refine their initial query but also to consider related issues or surface overlooked aspects of a problem.

**Counterfactual reasoning.** Counterfactual reasoning focuses on pivotal moments where alternative outcomes could have happened, aiming to excuse and explain poor outcomes, which can improve future results (Kahneman & Tversky, 1982). Utilizing hypothetical scenarios, counterfactual explanations illustrate how altering inputs could affect predictions, facilitating informed decision-making within human-AI interactions (Naiseh et al., 2023; Silva et al., 2023). Counterfactual explanations often involve the addition of new information rather than the removal of existing data (Cheng et al., 2024). The rationale behind that is that identifying missing information is crucial for understanding predictions.

Thus, our fourth feature is instantiated by finding related cases to the given prompt and extracting the most critical information that changes the GenAI response (**DF5 Critical information and counterfactuals**). In customer service, integrating system-level (e.g., software versions) and contextual information (e.g., updates) is key to finding relevant knowledge. DF5 supports this by improving intent disambiguation and issue clarification for more accurate retrieval.

**Probabilistic reasoning.** Information on the local performance, also phrased as certainty ratings (Cramer et al., 2008) or confidence scores (Jiang et al., 2022), can act as a probabilistic approach to trusting GenAI advice. For example, prediction confidence scores can explain the extent to which the prediction might be correct or incorrect. A high score conveys trustworthiness, while a low score will make the system look incompetent and less credible (Jiang et al., 2022). Local performance indicators represent the credibility of the GenAI system but elicit a low degree of additional information and arguments.

(**DF6 Confidence score**) We instantiate performance explanations as confidence scores in providing a correct answer to the given request (Do et al., 2024; Leiser et al., 2024). The confidence score (0–1), derived from cosine similarity in the RAG system,

indicates response accuracy and source relevance, giving customers and agents a quick heuristic to judge answer appropriateness.

**Deductive reasoning.** Deductive reasoning is a logical process where a conclusion necessarily follows from given premises, ensuring its truth if the premises are true (Johnson-Laird, 1999). Similarly, CoT reasoning is a modern approach aimed at enhancing and clarifying LLM’s responses in the era of GenAI (Chu et al., 2023). CoT enables structured problem-solving through a clear and logical sequence of reasoned deductions. The approach has been applied to multiple domains in natural language processing such as question-answering (P. Lu et al., 2022).

In the case of RAG systems, CoT reasoning provides information on “how” the system derived a response (Schelhorn et al., 2025), including searching databases, retrieving documents, and formulating answers (**DF7 Chain-of-thought**). In customer support, this transparency helps agents follow the AI’s reasoning, verify steps, and validate conclusions. From a knowledge management perspective, CoT fosters structured reuse and traceability of knowledge, making problem-solving more interpretable and auditable.

## 5. Demonstration and evaluation

Afterward, we translated the design features into an instantiated artifact. We developed a GPT4-based GenXAI system that provides human support agents with responses and corresponding explanation features to given customer requests. Motivated by the telecommunications domain, we scraped publicly available FAQs, knowledge base articles, and forum posts from a major mobile provider. The system was trained on this 191-entry knowledge base using a state-of-the-art RAG approach (Lewis et al., 2020). The RAG system was tested with a RAGAS evaluation framework (Es et al., 2023) to ensure the quality of the AI responses and depict a realistic case. The resulting GenXAI system comprises an established chat interface that allows users to prompt the LLM system and retrieve numerous forms of explanations. Figure 1 depicts the instantiated prototype that provides an overview of XAI-driven features for LLM systems.

To evaluate each DF and the instantiated GenXAI system, we utilized the think-aloud method (Jaspers et al., 2004). We opted for the think-aloud method as it captures participants’ immediate reasoning and sensemaking processes (Payne, 1994; Vitalari, 1985). Despite the large number of DFs, this yields rich qualitative insights into each feature. We approached experienced GenAI users (n=8; 4 female; 4 male;

The screenshot displays the GenXAI prototype interface with several explainability features (DFs) overlaid:

- DF7 - Deductive:** A box explaining that to answer the question, the system needed to know what could cause EDGE reception on an iPhone 11 and to retrieve similar cases from the database to find troubleshooting steps.
- DF4 - Transfactual:** A box with prompts: "What can I do if my iPhone 11 is stuck on EDGE and won't connect to LTE?", "What could be causing my iPhone 11 to show only EDGE reception while other devices in the same area have LTE?", and "My iPhone 11 has been stuck on EDGE for weeks while my other phones on a different plan has LTE everywhere. What can I do?"
- DF6 - Probabilistic:** A box showing a similarity score of 92% (highly confident) based on facts from the underlying data base.
- DF1 - Factual:** A box showing source information: "Source Title: Everywhere EDGE Solutions", "Source Type: Forum discussion (internal data)", and "Source Link: http://telco.dv.tst/everywhere-only-EDGE/".
- DF2 - Causal:** A box showing text passages that were particularly important for the answer, categorized by importance (very important, mildly important, not important).
- DF3 - Analogical:** A box showing summaries of similar cases from the database, including a top-1 similar case and a top-2 similar case.
- DF5 - Counterfactual:** A box showing a counterfactual prompt: "Hello! For a few weeks now I have had the problem on my iPhone 11 that I only have EDGE reception. I have tested this issue in several locations, including downtown and the airport, while my private mobile phone with a..."

Figure 1. Final initiated GenXAI prototype

$M_{age}=29.9$  years) who participated voluntarily to use the artifact and evaluate it. We provided a naturalistic setting (Venable et al., 2016), assuming that the participants were solving technical issues concerning their mobile phones. They were provided with support by our GenXAI system and were asked to elaborate on the quality and reliability of the response, given the designed explanation features. The sessions took place via Microsoft Teams and were recorded and transcribed. The sessions were on average 45 minutes long. The participants were provided with an example prompt and response at the beginning. Then they were randomly shown each DF for a different and again randomized prompt-response pair. At the end, we showed the whole instantiated prototype (Figure 1).

**GenAI response.** Despite providing users with excessive additional information as part of our GenXAI system, participants still focused on the actual responses (I01). GenAI responses should include more step-by-step instructions and could be more precise, referring to how knowledge is depicted (II01, I02, I06, I07). For example, the phrase “for some users” was utilized as a heuristic to trust the system (I03). In addition, domain-specific terms facilitated trust (I05). Finally, the decision importance and the consequences of incorrect AI advice are mentioned as being important for assessing the appropriateness of the extracted knowledge (I07, I08).

**Explanations.** We provide an overview of representative statements for each explanation in Figure 2. Our evaluation revealed two layers of feedback: usability-related concerns with interface design (e.g., text overload, vague labeling, limited interactivity) and conceptual concerns regarding the clarity and utility of the reasoning types. We highlight the latter in bold to distinguish feedback on the reasoning types and their underlying logic. Some users requested “explanations for the explanations”, highlighting the need for improved usability (I01). While explanations are based on human reasoning principles, some features, especially DF3, were found non-intuitive.

For *factual explanations* (DF1), users emphasize the importance of clickable source links for verifying the knowledge (I03, I04, I07) and suggest clearer source types, distinguishing between internal and external links (I01, I02, I03). Concerns about source reliability, especially for forum-based sources, lead to requests for additional context, such as text snippets (I06). Some users prefer reviewing the source before confirming validity (I05, I08).

*Causal explanations* (DF2) benefit from color coding, but users seek label clarifications via a clearer description or an info button (I01, I04). They request better explanations for why certain parts are deemed

irrelevant and clearer indications of knowledge sources (I02, I03). Users also want insights into the technical workings of the knowledge management system and the explanations, including similarity scores and concerns about processing speed (I06).

*Analogical explanations* (DF3) are seen as cluttered and poorly structured, prompting calls for collapsible sections, icons, and a clearer layout (I01). Some experts stress the need for consistent responses better aligned with the problem, noting misleading titles (I02, I05, I06). Responses are often too long, making assessment difficult, and users call for concise information (I04, I06). While a button to request similar cases is appreciated, concerns remain about interface overload (I03).

*Transfactual explanations* (DF4) (similar prompts) are often too similar to the original input, making them less useful. Users suggest focusing on next steps instead (I02, I06). Prompts should be relevant but not redundant, as some users feel they do not refine knowledge inquiries effectively (I03, I04). The use of transfactual explanations is seen as effort-intensive, with a preference for summarizing related cases instead (I05). Users trust repeatable answers but feel that simply presenting examples lacks clear intent (I08).

*Counterfactual explanations* (DF5) are often non-intuitive, with users suggesting a more interactive approach to allow refinement or further prompting (I02). Responses feel repetitive and case-dependent, leading to redundancy (I03). Trust is an issue, as identical responses seem probabilistic rather than explanatory (I04, I06). Excessive text also makes these explanations cognitively demanding, leaving users unsure whether the information influences their decisions or is just additional context (I07, I08).

For *probabilistic reasoning* (DF6), users find confidence estimates overly high. Some users worry that high confidence values encourage blind trust in AI outputs (I03). They suggest displaying confidence only when AI is uncertain (I01, I02, I06). Many find the slider interface unclear, proposing alternatives like a score bar or integrated color indicators (I01, I03, I05). Users also question how confidence scores are generated and request clearer explanations of their reliability (I04).

For *deductive reasoning* (DF7), while the structured reasoning process is appreciated, responses are often too vague, prompting calls for more concrete problem-solving steps (I02, I05). Users also stress the need to avoid redundancy (I02). Although the reasoning process builds trust, users want clearer visibility into retrieved sources (I04, I06).

<b>DF1: References and knowledge source</b>	"makes it valid" (I01)   "link to source should always be shown" (I02)   " <b>How trustworthy is the forum?</b> " (I06)
<b>DF2: Source attribution</b>	"additional context is great" (I02)   "how does this work technically?" (I06)   "explain what 'very important' and 'mildly important' actually mean" (I01)   "lack of time to check the details" (I02)   " <b>which snippets were included in the response with what similarity?</b> " (I06)
<b>DF3: Similar cases from the underlying database</b>	"Now we have a lot of text" (I06)   " <b>add additional cases on demand</b> " (I01)   "what are similar cases?" (I02)   "No time to look at this" (I02)
<b>DF4: Recommended Prompts</b>	"If the same answer appears often, I tend to trust it." (I08)   "Too effort-intensive; summarizing other cases would be more important." (I05)   " <b>Prompts are too similar</b> " (I02)
<b>DF5: Critical information and counterfactuals</b>	"Too much repetition" (I03)   " <b>If it is always the same, I would assume it is correct</b> " (I06)   " <b>I would rather ask again, I would prompt it myself</b> " (I02)
<b>DF6: Confidence score</b>	"Confidence is always high" (I01)   " <b>100% is not meaningful – rather, it makes me skeptical.</b> " (I06)   " <b>Facilitates overreliance!</b> " (I03)
<b>DF7: Chain-Of-Thought</b>	„Helpful, but the concrete steps for problem-solving are missing" (I05)   "Alarm bell ... <b>I can't see what it retrieved; the source is missing</b> " (I06)   "Breaking down the process is good, but the response is too vague and it should not be repetitive." (I02)

**Figure 2. Exemplary statements from the think-aloud study**

**Overall GenXAI System.** Users highlight concerns about information overload, emphasizing the need to balance factual depth with conciseness. Excessive text makes responses difficult to process and reduces actual engagement with provided explanations (I05). Additionally, some explanations are generated by the system itself (self-explanations), raising concerns about hallucinated content and trustworthiness - users question whom they can actually trust (I01, I03). To improve usability, the system should be more interactive, with features available on demand via buttons rather than displayed by default. Users also demonstrate an interesting heuristic when evaluating responses: If an alternative response provides counterfactual examples, they assume the original might be incorrect, whereas a lack of variance reinforces belief in correctness (I07, I08). These findings underscore the need for interactive and adaptive DFs to enhance trust and comprehension. Specifically, adapting the GenXAI interface to task characteristics (e.g., information retrieval, problem-solving) and user characteristics (e.g., domain expertise, AI literacy) is expected to enhance knowledge utilization.

## 6. Discussion and conclusion

In this study, we designed and evaluated a GenXAI system for facilitating appropriate knowledge utilization in LLM-based knowledge management systems. Drawing from (1) contemporary GenXAI research, (2) extensive literature on XAI, (3) foundational theories of human reasoning, and (4) practical insights from customer support services, we

identified seven DFs to enhance human-GenAI interaction for knowledge-intensive service work. In doing so, we integrated established design knowledge from the XAI literature with emerging LLM technologies, thereby creating a conceptual bridge to knowledge management. We implemented the DFs in our final GenXAI system and evaluated them through a think-aloud study with GenAI experts.

We contribute design knowledge to the emerging class of GenXAI systems (Schneider, 2024). In contrast to prior approaches to facilitating appropriate knowledge utilization in the LLM-based systems (Leiser et al., 2024; Pafla et al., 2024), we build upon the theoretical foundations of XAI and human reasoning to lay a foundation for further research on reliance on GenAI advice, particularly concerning hallucinations. Our study extends the existing XAI literature by examining several types of human reasoning as a foundation for explanation features in LLM-enabled knowledge management systems. Hence, this study contributes to the literature on decision-making in AI advice (Jussupow et al., 2021) and XAI (Bauer et al., 2023) by conceptualizing explanations for GenAI systems. Unlike existing XAI taxonomies that focus on technical methods, our design emphasizes the cognitive functions explanations serve, mirroring how humans think and reason. By projecting these reasoning modes onto GenAI agents, we bridge technical explanation methods with psychological theory. By integrating human reasoning theory with design knowledge from XAI, our study provides an integrative perspective that complements existing work on explanation types and trust calibration, while addressing a theoretical gap in how explanations are cognitively anchored and systematically connected to user reliance in GenAI systems.

Our results imply important consequences for organizations applying GenAI-based systems. The analysis supports practitioners in effectively designing GenAI-based retrieval tools such as question-answering systems and conversational agents in realms with high expectations regarding compliance and factuality, such as customer support services. We especially identify two crucial design challenges. First, several participants requested *explanations for explanations*, underscoring the challenge of reducing complexity, simplifying explanation logics, and the potential cognitive load imposed by multi-layered explanations. This highlights the need for adaptive explanation strategies that align depth with user expertise. Second, the tendency to heuristically over-trust high confidence scores illustrates the risk of miscalibrated trust, suggesting that confidence cues must be carefully designed to prevent blind reliance.

## 6.1. Limitations and future research

Our study has several limitations. *First*, our set of explanations is not exhaustive; rather, we focused on identifying representative explanation types based on prior XAI literature in traditional machine learning. *Second*, our evaluation was limited to a sample of eight GenAI experts, excluding domain professionals such as support agents or managers who directly interact with these systems. Assessing the design in real-world applications across different domains, such as education and healthcare, would provide deeper insights into its practical effectiveness. *Third*, we showed participants only static explanations. Future iterations could allow users to specify what they want to understand from the AI and how they might re-prompt their questions. *Fourth*, due to the large number of explanatory features, we did not conduct a quantitative analysis of their effectiveness. Future studies could explore structured evaluation methods, such as best-worst scaling or benchmarking, to compare the design with existing solutions, for example, with regard to hallucination detection. *Fifth*, we acknowledge that explainability spans multiple layers, from algorithm design to training data and generated outputs. Our contribution is situated at the post-hoc layer, only providing an integrative perspective that bridges technical explanation methods with human reasoning theory to improve how service professionals interpret and act on GenAI outputs. Finally, a deeper theoretical understanding is needed to explore how and why different reasoning types influence users' trust in GenAI. Investigating the cognitive mechanisms underlying trust formation could further refine explanation strategies and improve system transparency.

Our study lays the groundwork for explainable GenAI (GenXAI) systems, and it opens up several exciting avenues for future research. First, additional design options could enhance explainability, such as hallucination detectors and warning mechanisms to mitigate various types of GenAI inaccuracies, including ambiguities, irrelevant sources, factual inaccuracies, and hallucinations. Exploring how different types of hallucinations impact user trust and decision-making would further refine system reliability. Second, future research could explore user-adaptive system designs that tailor explanations to individual preferences and cognitive styles. Personalizing explanations could improve user engagement and comprehension, making human-GenAI interactions more effective. Third, integrating explanations directly into the response rather than as separate elements could enhance usability. As our evaluation highlighted, users emphasize the AI's

response itself, suggesting that analyzing response features influencing reliance would provide complementary insights into explanation design in LLM systems (Jakesch et al., 2023). Finally, our design lays a foundation for advancing XAI technologies, through enhanced RAG, knowledge graph integration, and neurosymbolic computing, to move beyond purely post-hoc, prediction-based explanations.

## 7. References

- Ashktorab, Z., Jain, M., Liao, Q. V., & Weisz, J. D. (2019). Resilient Chatbots: Repair Strategy Preferences for Conversational Breakdowns. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. ACM.
- Azamfirei, R., Kudchadkar, S. R., & Fackler, J. (2023). Large language models and the perils of their hallucinations. *Critical Care*, 27(1), 1–2.
- Bauer, K., Zahn, M. von, & Hinz, O. (2023). Expl (AI) ned: The impact of explainable artificial intelligence on users' information processing. *Information Systems Research*, 34(4), 1582–1602.
- Bo, J. Y., Wan, S., & Anderson, A. (2025). To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on Large Language Models. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. ACM.
- Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. *The Quarterly Journal of Economics*.
- Cambria, E., Malandri, L., Mercorio, F., Nobani, N., & Seveso, A. (2024). Xai meets llms: A survey of the relation between explainable ai and large language models. *ArXiv Preprint ArXiv:2407.15248*.
- Cheng, F., Zouhar, V., Chan, R. S. M., Fürst, D., Strobel, H., & El-Assady, M. (2024). Interactive analysis of LLMs using meaningful counterfactuals. *ArXiv Preprint ArXiv:2405.00708*.
- Chu, Z., Chen, J., Chen, Q., Yu, W., He, T., Wang, H., Peng, W., Liu, M., Qin, B., & Liu, T. (2023). A survey of chain of thought reasoning: Advances, frontiers and future. *ArXiv Preprint ArXiv:2309.15402*.
- Cramer, H., Evers, V., Ramlal, S., van Someren, M., Rutledge, L., Stash, N., Aroyo, L., & Wielinga, B. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5), 455–496.
- Danry, V., Pataranutaporn, P., Mao, Y., & Maes, P. (2023). Don't Just Tell Me, Ask Me: AI Systems that Intelligently Frame Explanations as Questions Improve Human Logical Discernment Accuracy over Causal AI explanations. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.
- Do, H. J., Ostrand, R., Weisz, J. D., Dugan, C., Sattigeri, P., Wei, D., Murugesan, K., & Geyer, W. (2024). Facilitating human-LLM collaboration through

- factuality scores and source attributions. *ArXiv Preprint ArXiv:2405.20434*.
- Dodge, J., Liao, Q. V., Zhang, Y [Yunfeng], Bellamy, R. K. E., & Dugan, C. (2019). Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM.
- Durcikova, A., Fadel, K. J., Butler, B. S., & Galletta, D. F. (2011). Research note—knowledge exploration and exploitation: the impacts of psychological climate and knowledge management system access. *Information Systems Research*, 22(4), 855–866.
- Es, S., James, J., Espinosa-Anke, L., & Schockaert, S. (2023). Ragas: Automated evaluation of retrieval augmented generation. *ArXiv Preprint ArXiv:2309.15217*.
- Feuerriegel, S., Hartmann, J., Janiesch, C., & Zschech, P. (2024). Generative ai. *Business & Information Systems Engineering*, 66(1), 111–126.
- Förster, M., Klier, M., Kluge, K., & Sigler, I. (2020). Evaluating Explainable Artificial Intelligence – What Users Really Appreciate. In *European Conference on Information Systems (ECIS)*. Aisel.
- Gary, K., David, S., & Chew Lock, P. (2011). Anticipatory Thinking. In *Informed by Knowledge* (pp. 249–260). Psychology Press. <https://doi.org/10.4324/9780203847985-23>
- Gregor, S., & Benbasat, I. (1999). Explanations from intelligent systems: Theoretical foundations and implications for practice. *MIS Quarterly*, 497–530.
- He, G., Aishwarya, N., & Gadiraju, U. (2025). Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*. ACM.
- He, G., Kuiper, L., & Gadiraju, U. (2023). Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. ACM.
- Hoffman, R. R., & Klein, G. (2017). Explaining explanation, part 1: theoretical foundations. *IEEE Intelligent Systems*, 32(3), 68–73.
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, 120(11).
- Jaspers, M. W. M., Steen, T., van den Bos, C., & Geenen, M. (2004). The think aloud method: a guide to user interface design. *International Journal of Medical Informatics*, 73(11-12), 781–795.
- 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 Computing Surveys*, 55(12), 1–38.
- Jiang, J., Kahai, S., & Yang, M. (2022). Who needs explanation and when? Juggling explainable AI and user epistemic uncertainty. *International Journal of Human-Computer Studies*, 165, 102839.
- Johnson-Laird, P. N. (1999). Deductive reasoning. *Annual Review of Psychology*, 50(Volume 50, 1999), 109–135.
- Johnson-Laird, P. N. (2010). Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43), 18243–18250.
- Jussupow, E., Benbasat, I., & Heinzl, A. (2024). An integrative perspective on algorithm aversion and appreciation in decision-making. *MIS Quarterly*, 48(4).
- Jussupow, E., Spohrer, K., Heinzl, A., & Gawlitza, J. (2021). Augmenting medical diagnosis decisions? An investigation into physicians’ decision-making process with artificial intelligence. *Information Systems Research*, 32(3), 713–735.
- Kahneman, D., & Tversky, A. (1982). The psychology of preferences. *Scientific American*, 246(1), 160–173.
- Kim, S. S. Y., Vaughan, J. W., Liao, Q. V., Lombrozo, T., & Russakovsky, O. (2025). Fostering Appropriate Reliance on Large Language Models: The Role of Explanations, Sources, and Inconsistencies. *ArXiv Preprint ArXiv:2502.08554*.
- Ko, D.-G., & Dennis, A. R. (2011). Profiting from knowledge management: The impact of time and experience. *Information Systems Research*, 22(1), 134–152.
- Kuhn, D. (2012). The development of causal reasoning. *Wiley Interdisciplinary Reviews: Cognitive Science*, 3(3), 327–335.
- Lai, V., Carton, S., Bhatnagar, R., Liao, Q. V., Zhang, Y [Yunfeng], & Tan, C. (2022). Human-AI Collaboration via Conditional Delegation: A Case Study of Content Moderation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. ACM.
- Leichtmann, B., Hinterreiter, A., Humer, C., Streit, M., & Mara, M. (2024). Explainable Artificial Intelligence Improves Human Decision-Making: Results from a Mushroom Picking Experiment at a Public Art Festival. *International Journal of Human-Computer Interaction*, 40(17), 1–18.
- Leiser, F., Eckhardt, S., Leuthe, V., Knaeble, M., Mädche, A., Schwabe, G., & Sunyaev, A. (2024). HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM.
- Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., & others (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.
- Lu, H., Ma, W., Wang, Y., Zhang, M., Wang, X [Xiang], Liu, Y., Chua, T.-S., & Ma, S. (2023). User Perception of Recommendation Explanation: Are Your Explanations What Users Need? *ACM Transactions on Information Systems*, 41(2), 1–31.
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafford, O., Clark, P., & Kalyan, A. (2022). Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in*

- Neural Information Processing Systems*, 35, 2507–2521.
- Mao, J.-Y., & Benbasat, I. (2001). The effects of contextualized access to knowledge on judgement. *International Journal of Human-Computer Studies*, 55(5), 787–814.
- Martino, A., Iannelli, M., & Truong, C. (2023). Knowledge injection to counter large language model (LLM) hallucination. In *European Semantic Web Conference*. Symposium conducted at the meeting of Springer.
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38.
- Nahar, M., Seo, H., Lee, E.-J., Xiong, A., & Lee, D. (2024). *Fakes of Varying Shades: How Warning Affects Human Perception and Engagement Regarding LLM Hallucinations*.
- Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2023). How the different explanation classes impact trust calibration: The case of clinical decision support systems. *International Journal of Human-Computer Studies*, 169, 102941.
- Nicula, B., Dascalu, M., Arner, T., Balyan, R., & McNamara, D. S. (2023). Automated assessment of comprehension strategies from self-explanations using LLMs. *Information*, 14(10), 567.
- Pafla, M., Larson, K., & Hancock, M. (2024). Unraveling the Dilemma of AI Errors: Exploring the Effectiveness of Human and Machine Explanations for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM.
- Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. *ACM Transactions on Computer-Human Interaction*, 29(4), 1–33.
- Payne, J. W. (1994). Thinking aloud: Insights into information processing. *Psychological Science*, 5(5), 241–248.
- Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, 24(3), 45–77.
- Peng, B., Galley, M., He, P., Cheng, H., Xie, Y., Hu, Y., Huang, Q., Liden, L., Yu, Z., Chen, W., & Gao, J. (2023). *Check Your Facts and Try Again: Improving Large Language Models with External Knowledge and Automated Feedback*.
- Reinhard, P., Li, M. M., Fina, M., & Leimeister, J. M. (2025). Fact or Fiction? Exploring Explanations to Identify Factual Confabulations in RAG-Based LLM Systems. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*.
- Reinhard, P., Li, M., Peters, C., & Leimeister, J. M. (2024). Generative AI in customer support services: a framework for augmenting the routines of frontline service employees. In *Proceedings of the 57th Hawaii International Conference on System Sciences (HICSS)*.
- Schelhorn, T., Gnewuch, U., & Maedche, A. (2025). The Impact of Chain-of-Thought Display on the Effective Use of LLM-based Analytics Agents.
- Schemmer, M., Kuehl, N., Benz, C., Bartos, A., & Satzger, G. (2023). Appropriate reliance on AI advice: Conceptualization and the effect of explanations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*.
- Schneider, J. (2024). Explainable generative ai (genxai): A survey, conceptualization, and research agenda. *Artificial Intelligence Review*, 57(11), 289.
- Sharma, N., Liao, Q. V., & Xiao, Z. (2024). Generative echo chamber? effect of llm-powered search systems on diverse information seeking. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (pp. 1–17). ACM.
- Silva, A., Schrum, M., Hedlund-Botti, E., Gopalan, N., & Gombolay, M. (2023). Explainable Artificial Intelligence: Evaluating the Objective and Subjective Impacts of xAI on Human-Agent Interaction. *International Journal of Human-Computer Interaction*, 39(7), 1390–1404.
- Steyvers, M., Tejada, H., Kumar, A., Belem, C., Karny, S., Hu, X., Mayer, L. W., & Smyth, P. (2025). What large language models know and what people think they know. *Nature Machine Intelligence*, 1–11.
- Storey, V. C. (2025). Knowledge management in a world of generative AI: Impact and Implications. *ACM Transactions on Management Information Systems*.
- Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *ArXiv Preprint ArXiv:2401.01313*.
- Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: a framework for evaluation in design science research. *European Journal of Information Systems*, 25(1), 77–89.
- Vitalari, N. P. (1985). Knowledge as a basis for expertise in systems analysis: An empirical study. *MIS Quarterly*, 221–241.
- Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15). ACM.
- Wei, J., Wang, X [Xuezhi], Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., & others (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35, 24824–24837.
- Xu, Z., Cruz, M. J., Guevara, M., Wang, T., Deshpande, M., Wang, X [Xiaofeng], & Li, Z. (2024). Retrieval-augmented generation with knowledge graphs for customer service question answering. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*.
- Zhang, Y [Yue], Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y [Yu], Chen, Y., & others (2023). Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv Preprint ArXiv:2309.01219*, 2(5).