Beyond Chatbots: Designing Reliable and Transparent Interactive Service Agents Through Hierarchical Planning

Human-Centered Design and Evaluation for AI-Driven Customer Support

Name:	*Anonymized*
Matrikelnummer:	*Anonymized*
Department:	Computer Science Department
Supervisor:	Prof. Dr. Iryna Gurevych
Responsible Staff:	Dr. Thomas Arnold

Date of submission: February 28, 2026
Darmstadt

TECHNISCHE UNIVERSITÄT DARMSTADT

Computer Science Department

Ubiquitous Knowledge Processing Lab

1 Motivation

As digital service expectations rise, the gap between the demand for immediate, 24/7 support and the limited operational capacity of human-only service desks has reached a critical threshold. Interactive Service Agents (ISAs) are rapidly transforming this landscape by enabling scalable, efficient, and natural interactions that overcome the deterministic constraints of traditional chatbots [1].

As of 2025, advancements in Large Language Models (LLMs) have pushed ISAs across finance, retail, and IT sectors toward increasing autonomy, allowing for end-to-end task execution and human-mimetic dialogue [2, 3, 4]. These developments offer organizations the opportunity to enhance customer experience and operational value, but they also raise crucial questions about workflow redesign, human-AI collaboration, and the future of customer service work.

However, this technological acceleration introduces significant friction. While ISAs can improve first-contact resolution and reduce operating costs, existing solutions frequently struggle with reliability, transparency, and the preservation of human empathy in mission-critical scenarios [5]. Complex issues still require human judgment, as AI development is not yet advanced enough to handle high-stakes nuances autonomously. Consequently, most deployments remain isolated pilot projects [2]. Research has historically focused on technical benchmarks, leaving the critical dimensions of user perception, trust, and human-AI augmentation underexplored [6].

Central Research Question (CRQ):

> *How can the operational reliability and transparency provided by a hierarchical agent framework establish the user trust necessary for effective human-AI collaboration in professional customer support?*

The objective is to develop a framework for human-centered ISA deployment, leveraging AI to automate routine tasks and augment human agents, ensuring reliability and proactive collaboration [4, 6].

2 Approach

This chapter outlines the methodological foundation of the thesis, situating the work within the rapid transformation of customer service from rigid rule-based automation to **Interactive Service Agents (ISAs)** powered by Large Language Models (LLMs). Despite recent advances, ISAs still face significant challenges in grounding, reliability, and integration [1, 6, 4, 2].

Customer service automation has evolved through three distinct technological phases: Early **Rule-Based Systems** (Phase 1) relied on deterministic flows with limited Natural Language Understanding (NLU), offering stability but poor adaptability [1]. The subsequent phase of **Retrieval-Augmented Chatbots** (Phase 2) introduced semantic search to improve informational accuracy, yet these systems still lacked flexible reasoning capabilities [5].

The current era of **Generative ISAs** (Phase 3), exemplified by LLM-based agents like ChatGPT [7] and Gemini [8], supports complex multi-turn dialogue and tool use. However, despite their reasoning capabilities, these systems remain prone to hallucinations and brittle interactions with enterprise systems [4, 3]. This thesis builds specifically on this **third phase**, addressing the critical gap between strong linguistic reasoning and the need for consistent, safe action execution in enterprise environments.

While generative ISAs exhibit advanced reasoning, they often fail in **operational reliability**, frequently producing malformed API calls or inconsistent tool usage [3, 6]. To resolve these issues, this thesis introduces the **Hierarchical Service Agent Framework (HSAF)**, a design methodology for constructing actionable and verifiable ISAs. The framework contributes two core innovations: a **structured planning pipeline** that separates high-level reasoning from surface-level generation to increase controllability [5, 6], and an **Action-Schema Validator** that enforces strict adherence to enterprise API specifications. Together, these mechanisms enable the deployment of ISAs that are both reasoning-capable and execution-safe.

The analysis is grounded in three complementary theoretical perspectives adapted from [1] and integrated with contemporary research on human-AI collaboration [4]. These frameworks provide a multi-dimensional lens to evaluate technical quality, user acceptance, and service value.

1. Human-Computer Interaction (HCI) & Conversational Agency

This perspective guides the evaluation of **SQ1 (Interaction Quality)**. It focuses on the system's ability to maintain *dialog coherence* and exhibit *perceived intelligence* during multi-turn exchanges. Within this framework, the ISA is analyzed not merely as a tool, but as a "communicative actor." Key metrics derived from this view include responsiveness, navigational clarity, and the mitigation of "uncanny valley" effects in anthropomorphic (humanlike) dialogue [1].

2. Technology Acceptance Model (TAM)

To address **SQ2 (User Acceptance)**, the study employs constructs from the Technology Acceptance Model. Beyond the traditional metrics of *Perceived Usefulness (PU)* and *Perceived Ease of Use (PEOU)*, this thesis incorporates ISA-specific determinants such as **trust** and **perceived risk** [6]. Given the "black-box" nature of generative AI, user trust is a critical variable for adoption in professional settings [2].

3. Service-Dominant Logic (S-DL)

The **SQ3 (Comparative Performance)** analysis is informed by Service-Dominant Logic, which emphasizes *value-in-use* and the co-creation of value between the service provider and the customer. This framework is essential for defining the **Human-AI Boundary**: it helps identify complex or empathetic scenarios where the "value" is maximized by escalating to a human agent rather than persisting with automation [5, 4].

Framework	Application in this Thesis
HCI & Agency	Criteria for evaluating linguistic capabilities, turn-taking, and error recovery strategies.
TAM (Extended)	Measurement of trust, transparency, and willingness to use ISAs for critical tasks.
S-DL	Definition of escalation protocols and determining the functional limits of automation.

Table 2.1: Alignment of theoretical frameworks with research objectives

The research follows a **Design Science Research (DSR)** approach, combining system development with iterative empirical evaluation. The HSAF prototype will be implemented in Python using modern LLMs (e.g., Llama 3) and connected to a simulated CRM environment via a function-calling interface. The evaluation utilizes a mixed-methods strategy: Quantitative Benchmarking using the custom **CSD-TS Dataset** to measure success rates and resolution times, and a Qualitative User Study.

To specifically address human-AI collaboration, the user study will involve $N \approx 50$ **professional customer service representatives**. This target group is chosen strictly because their professional expertise allows for a nuanced assessment of how ISAs assist or hinder complex workflows, providing insights into operational trust and transparency that general end-users cannot offer.

Integrating ISAs into enterprise workflows involves inherent technical and methodological hurdles. Implementing the HSAF requires navigating **complex data annotation** for the CSD-TS dataset, which must be meticulously labeled to ensure a valid ground truth. Furthermore, while the **simulated CRM environment** allows for controlled testing, it may not fully capture the latency, "noisy" data, or authentication complexities of live enterprise APIs. There is a risk that findings on perceived safety may be *biased by this simulation*; however, this is mitigated by focusing the evaluation on the *logic* of the Action-Schema Validator rather than network performance. Finally, the CSD-TS dataset, while diverse, may face generalizability limits across different industries, a factor that will be addressed in the final discussion of the results.

3 Schedule

The following schedule outlines the 6-month workflow, structured into five phases from foundation to final submission of the thesis. Starting from KW45 of 2025 (start of the module EiwA)

Phase 1: Orientation & Foundation (Month 1)		
Week 1–2	**Literature & Scoping**	Systematic review of Generative ISAs and reliability issues. Refinement of research gap and theoretical framework (HCI, TAM, S-DL).
Week 3–4	**Framework Design**	Conceptual design of the Hierarchical Service Agent Framework (HSAF). Definition of action schema, planning logic, and system requirements.
Phase 2: System Development (Months 2–3)		
Week 5–8	**Prototype Implementation**	Development of HSAF prototype in Python (e.g., Llama 3). Implementation of the structured planning pipeline and function-calling interface.
Week 9–12	**Validator & Simulation**	Implementation of Action-Schema Validator (JSON Schema/Pydantic). Construction of simulated CRM environment and initial functional testing.
Phase 3: Data & Evaluation Setup (Month 4)		
Week 13–14	**Dataset Creation**	Creation of the CSD-TS Dataset covering routine and edge-case scenarios. Annotation of ground truth actions for benchmarking.
Week 15–16	**Study Design**	Finalization of study protocol and questionnaires (SUS, TAM constructs). Recruitment of $N \approx 50$ professional customer service representatives.
Phase 4: Empirical Evaluation (Month 5)		
Week 17–18	**Quantitative Benchmarking**	Execution of automated benchmarks (SQ3). Measurement of Action Success Rate, resolution speed, and error analysis vs. baseline.
Week 19–20	**User Study Execution**	Conducting user study (SQ1–2). Collection of interaction logs and survey responses regarding trust, empathy, and usability.
Phase 5: Writing & Finalization (Month 6)		
Week 21–23	**Analysis & Writing**	Statistical analysis of results. Drafting Evaluation, Discussion, and Conclusion chapters. Refinement of Approach and Implementation sections.
Week 24	**Review & Submission**	Final proofreading, formatting check (bibliography, layout), and submission of thesis and artifacts.

Bibliography

[1] Rejwan Bin Sulaiman. “AI-Based Chatbot: An Approach of Utilizing on Customer Service Assistance”. In: *arXiv* (2019). URL: https://arxiv.org/pdf/2207.10573.

[2] McKinsey & Company. *The State of AI in 2025: Agents, Innovation, and Transformation*. 2025. URL: https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai.

[3] Capgemini Research Institute. *Unleashing the Value of Customer Service: The Transformative Impact of Gen AI and Agentic AI*. 2025. URL: https://www.capgemini.com/news/press-releases/generative-and-agentic-ai-set-to-transform-customer-service-into-a-strategic-value-driver-for-businesses/.

[4] M. Costa and S. Ghosh. *Human-AI Collaboration in Service Systems*. 2023. URL: https://www.researchgate.net/publication/389762161_Empowering_customer_service_with_generative_AI_enhancing_agent_performance_while_navigating_challenges.

[5] R. Smith, E. McKeon, and M. C. Gonzalez. *The AI Revolution in Customer Service and Support: A Practical Guide*. O’Reilly Media, 2024. URL: https://www.oreilly.com/.

[6] P. Reinhard et al. “Generative AI in Customer Support Services: A Framework for Augmenting the Routines of Frontline Service Employees”. In: *HICSS*. 2024. URL: https://pubs.wi-kassel.de/wp-content/uploads/2023/12/JML_956.pdf.

[7] OpenAI. *ChatGPT*. 2023. URL: https://chat.openai.com/.

[8] Google. *Google Gemini*. 2024. URL: https://gemini.google.com/.

Declaration on the Use of AI Tools:

The present work was created independently in terms of content and concept. The tools Grammarly and ChatGPT (GPT-4) / Google (Gemini 3), were used for linguistic review and formulation assistance (correction of spelling and grammar). Responsibility for the content and scientific accuracy rests solely with the author and the sources at hand.