Master's Thesis: Iterative Evaluation and Further Development of AI Skills and Agents

Use Case: A Persona Skill for Marketing on the AI Platform Helikon

Master's thesis proposal in cooperation with Thalia
Departments: AI & Data Platforms and Marketing

Summary

In one sentence. This thesis develops and tests, at the book retailer Thalia, a largely automated process for measuring and improving the quality of AI assistants, using the marketing department's persona assistant as the pilot case.

In one paragraph. Thalia runs an internal AI platform called Helikon, on which departments build so-called skills: AI assistants tailored to one specific task. Building a skill is fast, but checking whether it actually works well is slow, manual and unmeasured. This thesis designs, builds and evaluates a reusable quality loop that combines fixed test cases, automated grading by a second AI model and targeted human spot checks, so that business experts spend their scarce time only where their judgement cannot be replaced. The pilot case is the marketing department's persona skill, which lets employees assess drafts and ask questions from the perspective of defined customer personas. The thesis measures how reliable the automated grading is compared with human experts, how much the loop reduces effort and iteration time, and which ownership model is needed for it to work in day-to-day operations. The result is a reference process that can be transferred to other skills on the platform and that provides the evidence base required for internal AI governance and regulatory quality requirements.

1. Background

Thalia Bücher GmbH, one of the largest book retailers in German-speaking Europe, runs its own internal AI platform called Helikon. The platform reaches around 6,000 employees across the group. On it, business departments increasingly build their own skills and agents. A skill is a packaged AI capability for one specific task: it combines its own instructions to the language model (the prompt), the background information the model is given (the context), its own data sources and, in some cases, access to software tools. An agent is a skill that can work through multi-step tasks on its own. Building such a skill is fast. Assuring its quality is not.

Today, quality emerges mostly implicitly. A skill is built, the requesting department tries it out, comments are passed on informally, and the skill is adjusted. This causes three problems. First, quality is not measurable: whether a change made the skill better or worse remains an assertion. Second, every round of revision creates manual work on both sides, in the department that tests and comments and in the developing team that collects, interprets and implements the feedback. Neither scales alongside day-to-day business. Third, there is no protection against regressions: a change to the prompt, the underlying model or the context can silently break cases that previously worked.

Master's Thesis: Iterative Evaluation and Further Development of AI Skills and Agents

The marketing department's persona skill illustrates this bottleneck. An alpha version is running. The next step is not new development but ongoing refinement: assessment by the business departments, revision and intensive testing. For exactly this step, no workable process exists yet.

2. Problem Statement

What is needed is an evaluation and iteration process for AI skills and agents that meets two opposing requirements at once: the assessment must be meaningful, and the effort for everyone involved must be minimal. The scarce resource is expert judgement from the business side. It should be used only where it cannot be replaced, and everywhere else the assessment should be carried by automated methods.

This raises open questions about how reliable automated assessment is. One established approach is known as LLM-as-a-Judge: a second large language model grades the outputs of the skill against defined criteria. Such judges produce assessments at scale, but they have known biases. They may favour answers that appear in a particular position or that are simply longer, they may prefer outputs that resemble their own style, and their reading of the grading criteria can drift over time. Using them is only justifiable if their agreement with human expert judgement is known and continuously checked.

3. Objective

The aim of the thesis is to design, prototype and evaluate a reusable process for the iterative quality assurance and further development of skills and agents on Helikon.

Guiding research questions:

1. How can an evaluation loop for AI skills be designed so that human expert assessment is needed only at selected points and with little effort, while the bulk of the assessment runs automatically?
2. Which combination of defined test sets, LLM-as-a-Judge and human spot checks achieves sufficient agreement with the quality judgement of business experts?
3. How can feedback signals be captured and processed so that they systematically turn into test cases and improvements to the skill?
4. What effect does the process have on output quality, iteration time and effort in the business department and in the developing team?
5. Which maintenance and ownership model is needed for the loop to actually run in everyday operations?

4. Use Case: The Marketing Persona Skill

In marketing, personas are fictional but evidence-based profiles of typical customer groups. They describe a group's motivations, habits, reservations and language, and they help teams

Master's Thesis: Iterative Evaluation and Further Development of AI Skills and Agents

keep the customer's perspective in view when designing campaigns and products. At Thalia, a set of approved persona descriptions is documented in a central knowledge base.

The persona skill makes this customer perspective available in everyday work. It draws exclusively on the approved persona descriptions from the knowledge base and does not invent target-group characteristics. If the knowledge base offers no basis for a question, the skill says so openly rather than speculating.

Three usage scenarios are in focus:

Assessment from the personas' perspective. Landing pages, newsletters, visual motifs or feature concepts are assessed from the viewpoint of several personas. Each assessment comes with a justification and a reference to the specific persona characteristics it is based on.

Chat with a persona. Teams in UX, marketing, brand and communications can question a persona about its motivations, reservations and language. The answers are explicitly labelled as the persona's view, not as a statement by a real customer.

Automated evaluation. A judge model checks the skill's outputs against fixed criteria, in particular fidelity to the persona, quality of reasoning, traceability to the source descriptions and freedom from hallucination, that is, from invented facts. Defined and interchangeable test sets make the effect of every change reproducibly measurable. Human reviewers remain in the loop to calibrate the judge and to check ongoing samples.

The case is well suited as a research object because the quality of the outputs depends heavily on context and tone and cannot be reduced to simple right-or-wrong checks. This is precisely where it becomes clear whether automated assessment holds up. The process to be developed is not meant to be tailored to this one case but to be transferable to other Helikon skills.

Scope: The skill is not a substitute for real market research. It serves as a preliminary check and a generator of hypotheses.

5. Approach

The thesis follows a design-oriented research approach, with methodological and implementation work given equal weight.

Analysis. Documentation of the current process around the persona skill, elicitation of the relevant quality dimensions together with the business departments involved, and positioning of the work within the current state of research and practice on evaluating applications of large language models and agentic systems.

Design. Design of the evaluation loop, including grading rubrics; construction of interchangeable test sets from real user requests; definition of the assessment tiers, from deterministic checks through LLM-as-a-Judge to human spot checks; design of low-effort

Master's Thesis: Iterative Evaluation and Further Development of AI Skills and Agents

feedback mechanisms embedded in the context of use; and definition of how findings feed back into changes to the skill and into regression tests.

Implementation. Prototype implementation of the loop in the Helikon environment, including management of test cases, automated evaluation runs, analysis of results and traceability of skill versions.

Evaluation. Calibration of the judge model against expert judgements, several iteration cycles on the persona skill, and assessment of the process in terms of output quality, agreement measures between judge and experts, iteration time and time spent in the business department and in development. The thesis concludes with an assessment of how transferable the process is to other skills.

6. Hypotheses to Be Tested

- The customer perspective enters the work process earlier, and the number of iterations until sign-off decreases.
- Feedback loops run in minutes rather than through meetings and coordination rounds.
- Personas are used in everyday work rather than being retrieved from a document only at the start of a campaign.
- Quality becomes measurable instead of a matter of gut feeling, and changes can be released without a complete manual review.
- The judge model and the expert spot checks correlate strongly enough to reduce manual effort substantially.
- The test scenario and the judge approach are transferable to other Helikon skills.

7. Expected Results

- A reference process for the iterative evaluation and further development of skills and agents
- A working implementation of the evaluation loop on Helikon
- Calibrated grading rubrics and interchangeable test sets for the persona skill
- Empirical evidence on the reliability of automated assessment in the context studied
- A maintenance and ownership model, together with recommendations for scaling the process to the full skill portfolio

At the same time, the process provides the documented evidence that internal AI governance and regulatory requirements for the quality management and monitoring of AI systems demand.