Pilot Study: Evaluating Knowledge-Management Chatbots in the Third Sector

Prepared by the Hertie School Data Science Lab for the Bertelsmann Stiftung

Background and Motivation

AI systems such as retrieval-augmented generation (RAG) chatbots are increasingly used by third-sector and public-interest organisations to make institutional knowledge more accessible. These tools promise to improve efficiency, support evidence-based decision-making, and strengthen knowledge-sharing, but they also introduce questions of accuracy, reliability, and fairness.

Bertelsmann Stiftung has been an early adopter of such tools, developing two internal chatbots:

- **RichtlinienRadar**, which helps staff navigate internal policies and agreements.
- **PublikationsBot**, which enables search, comparison, and synthesis across roughly 3,000 grey publications.

While these systems have demonstrated potential, systematic evaluation is needed to understand their reliability and limitations, ensure compliance with data governance standards, and build confidence among users.

This collaborative pilot project between the **Hertie School** and **Bertelsmann Stiftung** will develop and test a scientifically grounded, **reproducable evaluation framework** for measuring chatbot response quality, robustness, and fairness. The findings will support internal improvement as well as contribute to best practices for responsible AI use in the wider NGO and public sectors.

Purpose and Scope

The pilot aims to create an initial evaluation framework that can later scale into a broader study.

This initial study will:

1. Develop and test a lightweight evaluation approach to measure response quality, robustness, and fairness for one or two selected chatbot domains (e.g., RichtlinienRadar).
2. Design/adapt and validate a evaluation interface for internal reviewers to assess chatbot outputs using predefined quality and relevance criteria.
3. Produce an initial findings report and roadmap for a potential continuation project.

This smaller study intentionally limits the scope of system testing and dataset size to ensure feasibility within a 6-month timeframe and budget.

Objectives

1. Response Quality

- Build a modular evaluation workflow (based on transformer-based models and expert validation) to measure quality, i.e., accuracy, grounding, and relevance of chatbot responses.
- Depending on what data is available, develop or adapt a simple evaluation interface to facilitate internal staff review and annotation.

2. Robustness and Fairness

- Conduct a targeted robustness test using rephrased and multilingual prompts to check for consistency.
- Identify and analyse any systematic inconsistencies, biased output or demographic/gender-coded disparities.

3. Synthesis and Roadmap

- Combine findings and insights from the previous two objectives to identify improvement areas.
- Produce a short summary report and draft best practice guidelines outlining lessons learned and recommendations for a possible follow-up study.

Methodology

The study plans to integrate a diverse set of analytical techniques, with the exact setup depending on data availability. Potential directions include machine-learning–based benchmarking, human-in-the-loop validation, and robustness testing. This approach aims to ensure both technical rigor and contextual relevance. The study's design follows internationally recognised principles of trustworthy AI (e.g. OECD AI Principles, EU AI Act risk management framework).

Analytical Approach

1. Response Quality Evaluation

- Develop a structured validation protocol and a lightweight supporting annotation interface enabling Bertelsmann's domain experts to label and enrich chatbot query-response pairs along pre-defined criteria-assessing correctness, relevance, and contextual accuracy, and providing expert reference answers or corrective annotations where appropriate. This will yield in a high-quality **query–response dataset** representing typical policy and publication queries.
- Use **transfer learning and transformer-based models** for retrieval and response assessment:
    - *(Multilabel) text classification* (e.g.,based on fine-tuned DeBERTa) to detect correctness and topical fit.

    - Combining *sentence similarity* (e.g., based on fine-tuned or off-the-shelf SBERT) and *reranking models* (e.g., based on fine-tuned or off-the-shelf cross-encoders) to measure semantic grounding and contextual accuracy.

2. Robustness and Fairness Evaluation

- Conduct **robustness testing** applying models from 1. to rephrased and multilingual variants of the original queries to measure stability of outputs.
- Assess **fairness** by introducing gender-coded and demographically varied prompts and statistically analysing parity in responses.
- Summarise deviations or inconsistencies in structured reports.

3. Synthesis and Benchmarking

- Aggregate all results into a reporting template.
- Compare findings against baseline metrics (retrieval recall, semantic relevance, etc.) to guide iterative system improvement.

Tools and Techniques

- Python-based evaluation pipeline using **Hugging Face Transformers** and **OpenAI Eval** frameworks.
- Custom scripts for classification, similarity, and robustness scoring.
- Secure Jupyter-based environment hosted at Hertie School for joint analysis and documentation.

Data Governance and Ethics

- Evaluation will intend to use internal policy and publication documents only—no personal data, external connectivity, or sensitive information.
- All work will comply with GDPR and Bertelsmann's internal governance standards.
- Ethical and privacy risks are therefore considered minimal.

Collaboration Structure

Hertie School

- Lead research design, conceptual development, methodology, and analysis.
- Develop the prototype evaluation workflow and run core tests.
- Coordinate reporting and final synthesis.

Bertelsmann Stiftung

a) Provide domain ground-truthing and validate chatbot outputs against internal sources.
b) Co-design and test the validation protocol and evaluation interface for staff feedback.
c) Participate in coding and interpretation of results to ensure contextual accuracy and organisational relevance.

Joint Activities

- Define scoring rubrics and data samples.
- Conduct internal review sessions at mid-point and end.
- Co-author the final pilot report and outline for a possible continuation study.

Work Plan and Timeline

Duration: 6 months (flexible start in January 2026)

Phase	Months	Focus	Key Outputs
1. Setup & Design	Jan–Feb 2026	Define metrics, dataset sampling, interface prototype	Evaluation framework draft
2. Evaluation & Testing	Mar–May 2026	Run small-scale benchmarking, robustness checks, expert reviews	Preliminary results
3. Analysis & Reporting	Jun 2026	Synthesis, lessons learned, recommendations	Pilot report & roadmap

Deliverables

1. **Prototype evaluation workflow** (scripts, rubrics, test queries).
2. **Evaluation interface** and example annotation outputs.
3. **Preliminary findings report** summarising accuracy, robustness, and fairness results.
4. **Short best-practice brief** outlining methodology and recommendations for next steps.

Preliminary Budget

Category	Description	Costs (€)
Personnel	• Faculty Lead (Sascha Goebel, 0.2 FTE, 6 months, Drew Dimmery, 0.2 FTE, 3 months), • Data Scientist (0.5 FTE, 6 months), • Research Assistant (0.25 FTE, 6 months), • coordination (0.3 FTE, 6 months)	50 000
Technical Resources	Compute, storage, annotation tools (via Azure/Hertie resources)	2 000
Travel	• 2 trips for the Faculty Lead	1 500
Total (Net)		**53 5000**
+ 7% VAT		3745
Total		**57245**