Case study: validation of a policy chatbot

"Artificial intelligence will reach human levels by 2029".
Ray Kurzweil[127]

"I think we will have an AI that is smarter than any human being probably by the end of 2025".
Perplexity[128]

To illustrate the application of the LLM validation techniques described above, this section presents a case study of the validation of a company's internal policy chatbot.

Case definition

The company has developed a chatbot based on an open source LLM to answer questions and provide information about its internal policies. The main objective of this chatbot is to facilitate employee access to company policies.

The chatbot has been built using a cloud infrastructure and has been fed with all of the company's policies, which comprise approximately 1,000 pages of documentation. To improve its responsiveness, Retrieval-Augmented Generation (RAG) techniques have been applied, which allow the model to retrieve relevant information from its knowledge base before generating a response. Initially, the possibility of applying fine-tuning to the model was considered, but after initial testing it was concluded that the combination of the base LLM with RAG was sufficient to achieve satisfactory results.

Prior to its final implementation, the company has decided to conduct a thorough validation process to assess the chatbot's accuracy, security and suitability in the specific context of its intended use. This validation process aims to identify potential areas for improvement and to ensure that the chatbot meets the Company's quality standards and expectations.

Validation of the policy chatbot will be conducted using a combination of quantitative metrics and human evaluation techniques, following the multidimensional validation framework described in the previous section. The results of this process will be used to make informed decisions about the implementation of the chatbot and to establish a continuous improvement plan.

Design of the validation approach

In order to comprehensively validate the policy chatbot, a tailored validation approach was designed following the validation framework presented in the previous section, covering the key dimensions of the model lifecycle: data, design, evaluation, implementation, and usage. This approach combines quantitative metrics and human evaluation techniques, with the goal of obtaining a complete picture of the chatbot's performance and suitability in the company's specific context.

The tests and techniques selected for each dimension are summarized below:

Data

- Metrics: The Flesch-Kincaid scale will be used to evaluate the readability and complexity of the policies that feed the chatbot.
- Human evaluation: A representative sample of policies will be reviewed to identify possible inconsistencies, errors or ambiguities.

Model design

- Metrics: Specific elements of the LLM will be modified in the development code (e.g., the RAG technique and its hyperparameters, such as the size or the chunking strategy[129]) that may change its response performance, and the results will be compared against the original model.

[127] Ray Kurzweil (n. 1948). Director of Engineering at Google, computer scientist, inventor and futurist, known for the invention of OCR and for his contributions in AI.

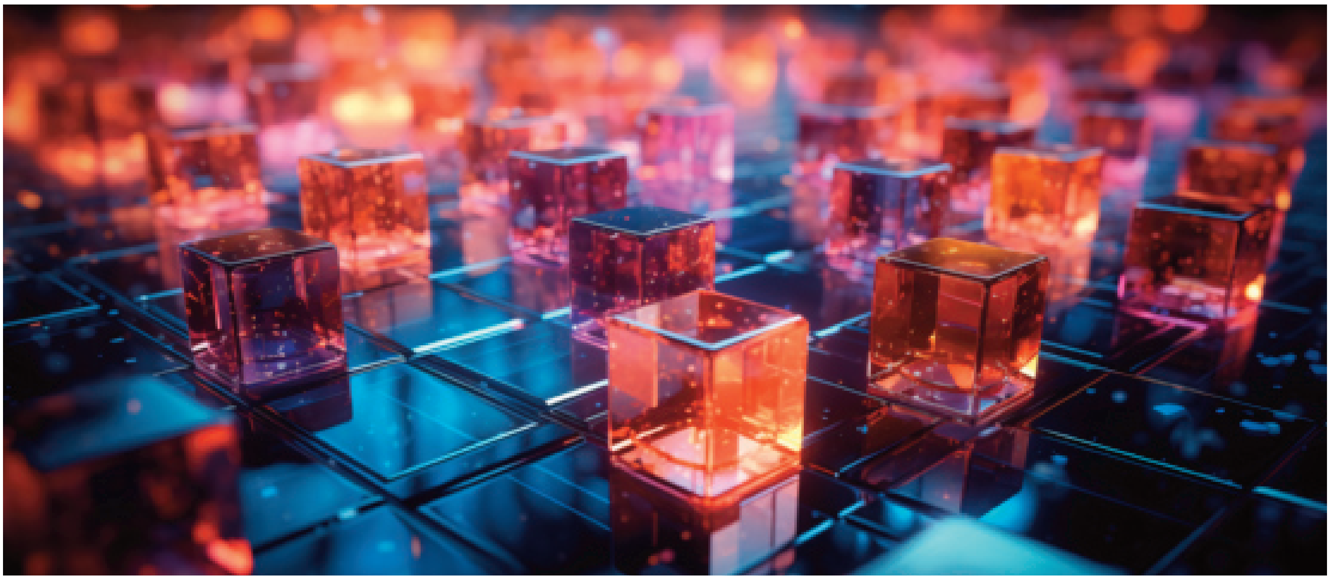
[128] Elon Musk (n. 1971), CEO of X, SpaceX, Tesla. South African-American entrepreneur, known for founding or co-founding companies such as Tesla, SpaceX and PayPal, owner of X (formerly Twitter), a social network that has its own LLM, called Grok.

[129] The chunking strategy refers to the process of dividing the input text to an LLM into smaller, more manageable units ("chunks") during use or implementation.

- Human evaluation: A thorough review of the chatbot components will be performed, including RAG configuration, input and output filters, prompt definition, and hyperparameter optimization. In addition, A/B testing will be conducted to compare the chatbot's performance with other LLMs available in the market.

Evaluation of the model

- Privacy and security
  - Metrics: K-anonymity tests will be applied to evaluate the protection of personal data in chatbot responses, and PII (Personal Identifiable Information) tests will be applied to identify sensitive attributes in the data, using PIIfilter.
  - Human assessment: Ethical hacking tests will be performed to identify potential vulnerabilities and detailed logs of chatbot interactions will be maintained.
- Accuracy
  - Metrics: Word Error Rate (WER) and ROUGE metrics will be used to assess the accuracy of chatbot responses compared to the original policies. Domain-specific benchmarks, such as a set of questions and answers designed by the company's policy experts, will also be used.
  - Human evaluation: A case-by-case review of a representative sample of chatbot interactions will be performed to identify possible errors or inaccuracies.
- Consistency
  - Metrics: Cosine Similarity and Jaccard Index will be used to assess the consistency of chatbot responses to similar queries.
  - Human evaluation: A/B tests will be conducted to compare chatbot responses in different scenarios and a case-by-case review will be performed to identify possible inconsistencies.
- Robustness
  - Metrics: Tools such as TextFooler will be used to generate adversarial text and evaluate the chatbot's resilience to misleading information. In addition, the number of chatbot rejections to malicious prompts will be counted.
  - Human evaluation: Ethical hacking tests and mock incidents will be conducted to evaluate the chatbot's ability to handle adverse situations.
- Adaptability
  - Metrics: The chatbot’s performance will be evaluated against new policies or updates using few-shot learning techniques. The chatbot's response to languages not used in the policies or requests for translations into languages not included in the RAG (e.g., Polish) will be evaluated.
  - Human evaluation: A/B testing and case-by-case reviews will be conducted to evaluate the chatbot's ability to adapt to new scenarios.
- Explainability
  - Metrics: Explainability techniques, such as SHAP, will be used to understand the chatbot's decision-making process. The chatbot's intrinsic interpretability module, which provides an explanation of the origin of the information in the response to the user, will be evaluated.
  - Human evaluation: The user experience (UX) will be monitored and a focus group will be conducted to evaluate users' perceptions of the chatbot's transparency and explainability.
- Biases and fairness
  - Metrics: The AI Fairness 360 toolkit will be used to assess potential demographic bias in chatbot responses. Specific benchmarks, such as the Bias Benchmark for QA (BBQ), will also be used to measure fairness in the context of company policies.
  - Human evaluation: Ethical hacking tests and a focus group will be conducted to identify potential bias or discrimination in the chatbot's responses.
- Toxicity
  - Metrics: Perspective API and Hatebase API tools will be used to assess the presence of toxic or inappropriate language in chatbot responses. In addition, specific benchmarks, such as RealToxicityPrompts, will be used to measure toxicity in the context of corporate policy.
  - Human evaluation: Ethical hacking tests will be conducted to identify potential instances of offensive or inappropriate language in chatbot interactions.



Implementation and use

- Scalability
  - Metrics: System stress tests will be performed using Apache JMeter to evaluate the chatbot's performance under heavy workloads.
  - Human evaluation: Simulations will be conducted to evaluate the chatbot's ability to handle an unforeseen increase in the number of users or queries.
- Efficiency
  - Metrics: Response time (Time-to-First-Byte, TTFB), resource usage (GPU/CPU, memory) and latency will be measured to evaluate chatbot efficiency.
- User acceptance
  - Metrics: A checklist of user requirements will be created and user satisfaction will be measured using indicators such as Net Promoter Score (NPS) and Customer Satisfaction Score (CSAT).
  - Human evaluation: User experience (UX) tracking will be conducted to evaluate user acceptance and satisfaction with the chatbot.

This customized validation approach will enable the company to obtain a comprehensive evaluation of the policy chatbot, identify areas for improvement and ensure its suitability for its intended use. The results of these tests and evaluations will be used to make informed decisions about the implementation and the chatbot's ongoing refinement.

Results

After applying the customized validation approach to the policy chatbot, promising results were obtained, demonstrating its overall suitability for the company's intended use (Figure 13). The chatbot achieved satisfactory performance in most evaluated dimensions, meeting quality standards and established expectations.

With respect to the quality of input data, the policies that fed the chatbot were generally found to be of sufficient readability and complexity to be understood by users. In addition, the human review did not identify any significant inconsistencies or errors in the content of the policies.

The model design also proved appropriate for the use case, with optimal configuration of the chatbot components and superior performance compared to other LLMs available on the market.

In terms of model evaluation, the chatbot achieved positive results in most of the metrics and tests applied. The high accuracy of the responses, the consistency in handling similar queries and the ability to adapt to new scenarios stand out. However, some areas for improvement were identified in aspects such as explainability, bias detection, and the response to very specific questions where further model refinement of the model is required. In the area of cybersecurity, a more detailed analysis of the specific vulnerabilities of the open-source LLMs used is required to mitigate this risk in production.

In terms of implementation and use, the chatbot demonstrated good scalability and efficiency in handling high workloads. In addition, user satisfaction was high, indicating a good acceptance of the tool in the company context.

Figure 13. Summary of results of policy chatbot human evaluation metrics and techniques.

Dimension	Test	Result	Interpretation
Datas	**Flesch-Kincaid**	Adequate legibility (grade 8)	The policies are understandable to most users.
	Human Review	No significant inconsistencies	The policies are consistent and free of material misstatement.
Model design	**Challenger models**	Parameter improvements identified	Adapting RAG parameters to the policy context (i.e., chunk size) is required to improve information capture on very specific questions.
	Component overhaul	Optimum configuration	Chatbot design is appropriate for the use case.
	A/B testing	Superior performance compared to other LLMs	Chatbot outperforms other models available on the market
Model Evaluation	**K-anonimato**	Adequate protection of personal data	Chatbot does not reveal sensitive information in its responses.
	Ethical hacking	Identified minor vulnerabilities	Adjustments required to strengthen chatbot security
	Word Error Rate (WER)	WER < 5%	Chatbot responses are highly accurate
	ROUGE	ROUGE-L > 0.8	Chatbot responses adequately capture the content of the policies
	Cosine similarity / Jaccard index	Similarity > 0.9	Chatbot provides consistent responses to similar queries
	TextFooler	Resiliencia moderada ante texto adversario	Chatbot is moderately robust to misleading information
	Few-shot learning	Satisfactory adaptability	Chatbot can adapt to new policies or updates with minimal training, but it is required to monitor and add those new documents to the RAG periodically.
	SHAP	Satisfactory adaptability	Improvements are required in the chatbot's ability to explain its decisions , although the RAG component has been built in such a way that the LLM gives a self-explanatory answer.
	AI Fairness 360 / BBQ	Identified minor demographic biases	The chatbot presents some biases that need to be mitigated
	Perspective API / RealToxicityPrompts	Low toxicity (< 5%)	Chatbot responses rarely contain toxic or inappropriate language
Implementation and use	**Apache JMeter**	Satisfactory scalability (up to 1000 users)	Chatbot can handle high workloads without significant performance degradation
	TTFB / / Resource usag / Latencia	Adequate efficiency (TTFB < 1s, moderate use)	Chatbot responds quickly and uses resources efficiently
	NPS / CSAT	High satisfaction (NPS > 60, CSAT > 80%)	Users are highly satisfied with the chatbot and would recommend it to others

These results indicate that the policy chatbot is well on its way to being implemented in the company, although some specific areas were identified that require further improvement. The following section presents the main conclusions and recommendations derived from this validation process.

Main conclusions

The policy chatbot validation process has shown that this LLM-based system can be a valuable tool for facilitating employee access to relevant corporate information. The results of the various tests and evaluations indicate that the chatbot largely meets the quality, security and efficiency requirements set by the organization.

Strengths identified included the accuracy and consistency of the chatbot's responses, its ability to adapt to new scenarios, and its scalability to handle large workloads. In addition, user satisfaction with the tool is high, indicating good acceptance and adoption by employees.

However, the validation process has also revealed some areas for improvement that need to be addressed before the final implementation of the chatbot. In particular, the following recommendations are made:

1. **Improve the explainability of the model:** It is necessary to develop more advanced techniques so that the chatbot can provide clear and understandable explanations of its decision-making process. This will increase transparency and user confidence in the tool. While the RAG component has been built in such a way that the LLM gives a self-explanatory answer and refers to the corresponding policy, this explanation is not entirely clear for very specific questions.

2. **Mitigate identified biases:** Although the identified biases are small, it is advisable to apply debiasing techniques to ensure that chatbot responses are fair and non-discriminatory. Periodic review of biases and implementation of corrective measures where necessary is suggested.

3. **Strengthen security and privacy:** While the chatbot meets basic personal data protection standards, additional and recurring ethical hacking tests and more robust security measures are recommended to prevent potential vulnerabilities

4. **Establish a monitoring and continuous improvement plan:** It is essential to define a process for regularly monitoring and evaluating the chatbot's performance in order to identify opportunities for improvement and ensure its optimal performance in the long term. This plan should include collecting feedback from users, regularly updating policies and including them in the chatbot database, monitoring to improve the parameters used in the RAG and updating them, and incorporating new techniques and technologies as they become available.

In conclusion, the policy chatbot has shown potential to improve the efficiency and accessibility of information in the company. With the implementation of the suggested improvements and a focus on continuous improvement, this LLM-based system can become a strategic tool for organizational success. The final recommendation has been to proceed with the implementation of the chatbot, taking into account the observations and recommendations derived from this validation process.