Speech-to-text and text-to-speech guardrails and evaluators

Marissa Coorengel | Hogeschool Utrecht | HCAIM BIP | Orq.ai

1 Introduction

In recent years, Text-to-Speech (TTS) and Speech-to-Text (STT) technologies have significantly advanced, enabling more natural and accessible human-computer interactions. However, ensuring the **reliability**, **accuracy**, and **safety** of these systems is crucial, particularly in sensitive applications such as healthcare, finance, and customer service. To address these challenges, guardrails and evaluators play a vital role in optimizing the performance and ethical integrity of these models.

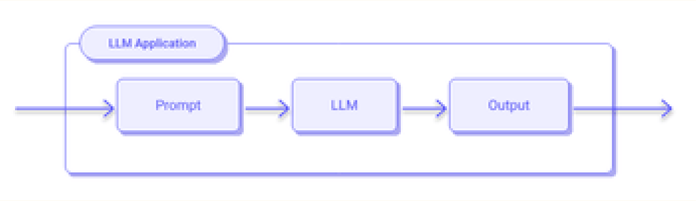
Guardrails act as preventive measures that ensure the generated or transcribed speech adheres to predefined rules, preventing harmful, biased, or misleading outputs. In TTS, guardrails help filter out inappropriate or toxic speech synthesis, while in STT, they can detect and flag misinterpretations that could lead to misinformation [3].

Evaluators are used to assess and monitor the quality of speech synthesis and recognition. These can be rule-based systems, machine learning models, or human evaluators who measure performance across accuracy, fluency, fairness, and robustness. Evaluators help refine models by identifying biases, reducing errors, and improving the overall user experience [4].

To research this topic, the following question will be investigated:

"How can guardrails and evaluators be designed and applied to make TTS and STT models safe and accurate for different use cases?"

Without guardrails



2 Literature review

There are already various guardrails that can be provided by TTS and STT systems. One of those is **rule-based string manipulation**. You can filter out bad-words or words you don't want your model to use. Another approach is a **LLM judge** that decides if an input is good or bad and for example decides if a user needs to be blocked [1]. An example of a LLM judge is NVIDIA's self-checking NeMo [2]. But you can also fine-tune your own LLM judge model.

Fig. 2 Figure that shows why it is important to have guardrails [5].

3 Research methodology

Literature review

Investigating existing guardrails and evaluators in TTS and STT technologies.

Stakeholder analysis

Identifying key users and their needs, including businesses, developers, and end-users.

Component test

Evaluating different guardrails and evaluators before integrating them into the system.

Ethical check

Ensuring that the solutions adhere to fairness, bias reduction, and inclusivity principles.

Acknowledgements. The HCAIM (the Human-Centred AI Master's Programme) Project is Co-Financed by the Connecting Europe Facility of the European Union Under Grant №CEF-TC-2020-1 Digital Skills 2020-EU-IA-0068. This Poster Was Created As Part of the Blended IP Event Organized Under Erasmus + Programme of the European Union.

With guardrails

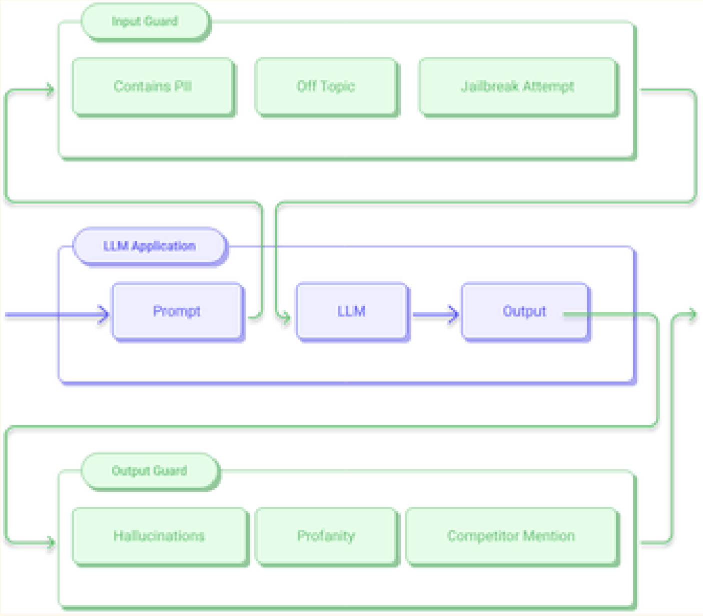


Fig. 1. Figure that explains what guardrails are [6].

5 Conclusion

The development of reliable and ethical TTS and STT systems is crucial as these technologies become more integrated into daily life. Guardrails and evaluators provide a structured approach to **mitigating risks**, improving performance, and ensuring user trust. By conducting thorough research, testing, and ethical assessments, we can develop AI-driven speech technologies that are both **effective** and **responsible**. Also, continuous refinement and stakeholder collaboration will be essential in advancing these systems for broader and safer adoption.

References.
[1]. Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., Cohen, J., & NVIDIA. (2023). NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. arXiv:2310.10501v1 [cs.CL] 16 Oct 2023. https://arxiv.org/pdf/2310.10501
[2]. Luden, I. (2024, November 21). The landscape of LLM guardrails: intervention levels and techniques. Medium. Start your day by listing your most important tasks. Use tools time-blocking to prioritize and manage your workload effectively.
[3]. What are AI guardrails? (2024, November 14). McKinsey & Company. https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-are-ai-guardrails
[4]. Matuszewska, J., & Miquido. (2025, January 23). AI Evaluation Definition. Miquido. https://www.miquido.com/ai-glossary/ai-model-evaluation/
[5]. Omroep Zeeland. (2022, November 23). Auto bovenop vangrail bij Kamperland. Omroep Zeeland. https://www.omroepzeeland.nl/nieuws/15144053/auto-bovenop-vangrail-bij-kamperland
[6]. LLMs in production with guardrails. (n.d.). https://www.ionio.ai/blog/llms-in-production-with-guardrails

4 Preliminary considerations

The findings of this study will serve as a foundation for future advancements in TTS and STT technologies. By demonstrating **effective implementation** of guardrails and evaluators, this research will encourage further exploration into **save** and **ethical** AI-driven speech systems.