1.4.1. KLM - Next-Gen XR/AI Integration for Cabin Crew Training

On-Premise Conversational AI for Realistic and Emotional Dialogue

IFM4040: Joint Interdisciplinary Project

TU Delft KLM

1.4.1. KLM - Next-Gen XR/AI Integration for Cabin Crew Training

On-Premise Conversational AI for Realistic and Emotional Dialogue

by

Student Name	Student Number	Email
Winnie Cheng	5280532	W.V.Cheng@student.tudelft.nl
Natanael Djajadi	5228719	N.Djajadi@student.tudelft.nl
Matteo Fregonara	5022959	M.Fregonara@student.tudelft.nl
Siri Høystad	6284159	S.Hoeystad@student.tudelft.nl
Moniek Smink	6167667	M.C.Smink@student.tudelft.nl
Vassilis Varnas	6225934	V.Varnas@student.tudelft.nl
Sem Zeelenberg	6592570	E.S.Zeelenberg@student.tudelft.nl

Instructor: Jae Grant Maloney
Company: KLM Royal Dutch Airlines
Project Duration: September 2025 - November 2025

TU Delft KLM

Contents

Summary iii

Our Interdisciplinary Team iv

1 Introduction 1
- 1.1 Trainings at KLM . . . 1
  - 1.1.1 Technical and Human-Centered Trainings . . . 2
  - 1.1.2 VR Trainings . . . 3
  - 1.1.3 Limitations of Current VR Technology . . . 3
  - 1.1.4 Integrating Artificial Intelligence into VR . . . 3
  - 1.1.5 Deconstructing the AI Agent . . . 5
  - 1.1.6 KLM's Voice Agent Prototype: Caresse . . . 5
  - 1.1.7 Development Pathway Towards a Local AI Agent . . . 6
- 1.2 Problem Statement . . . 6
  - 1.2.1 Research Question . . . 6
  - 1.2.2 Evaluation . . . 7
  - 1.2.3 Broader Impact . . . 7

2 Development Process of our Prototype: OpenVoiceAgent 8
- 2.1 First Development Phase . . . 8
  - 2.1.1 Constraints . . . 8
  - 2.1.2 Technical Options & Model Choices . . . 8
  - 2.1.3 Pipeline Design . . . 12
- 2.2 Second Development Phase . . . 15
  - 2.2.1 Frontend . . . 15
  - 2.2.2 Latency Improvements . . . 16
  - 2.2.3 Removing Connection Dependency . . . 17
  - 2.2.4 Concluding the Second Phase . . . 17
- 2.3 Technical Future Work . . . 17
  - 2.3.1 Conversation Flow . . . 17
  - 2.3.2 Conversation Content . . . 18

3 Validating Our Prototype 21
- 3.1 Research Plan . . . 21
  - 3.1.1 Process . . . 21
  - 3.1.2 Research Questions . . . 21
- 3.2 PoC tests . . . 22
  - 3.2.1 Participants . . . 22
  - 3.2.2 Method . . . 22
- 3.3 Final Survey Tests . . . 26
  - 3.3.1 Survey Design . . . 26
- 3.4 Validation Results . . . 27
  - 3.4.1 PoC tests: Data Analysis . . . 27
  - 3.4.2 PoC Tests: Results . . . 29
  - 3.4.3 Survey: Data Analysis . . . 35
  - 3.4.4 Survey: Results . . . 35

3.4.5 Validation Summary . . . 37

4 Implementing our Prototype at KLM **39**
4.1 Implementation Analysis . . . 39
4.1.1 Precedent and Validation . . . 39
4.1.2 Our Speculations . . . 40
4.2 AI Adoption Interviews . . . 41
4.2.1 Limitations of Current Training . . . 41
4.2.2 Trust and Privacy . . . 41
4.2.3 Practical Applications . . . 41
4.2.4 Organizational Readiness and Culture . . . 42
4.2.5 Attitude and Outlook . . . 42
4.2.6 What These Interviews Meant For Us . . . 43
4.3 Potential Use-Cases of OpenVoiceAgent . . . 44
4.3.1 Within KLM . . . 44
4.3.2 Beyond KLM . . . 45

5 Value Created By Our Prototype **47**
5.1 Profit, People, Planet: An Overview . . . 47
5.1.1 Sustainable Development Goals Addressed . . . 47
5.1.2 People: Personal Merits to Prototype . . . 47
5.1.3 Profit: Costs of Prototype . . . 50
5.1.4 Planet: Potential Planetary Benefits to Prototype . . . 57
5.1.5 People, Planet, Profit: Closing Remarks . . . 60
5.2 Ethical Concerns & Risk Management . . . 61
5.2.1 Ethics . . . 61
5.2.2 Risk Management . . . 63
5.2.3 Conclusion . . . 65

6 Conclusion **66**
6.1 Project Outcomes . . . 66
6.1.1 For Our Team . . . 66
6.1.2 For KLM . . . 66
6.1.3 For Others . . . 66
6.2 Closing Remarks . . . 66

References **68**

A Extended Validation Results **73**
A.1 Extended Validation Figures . . . 73

Abstract

This report was written as part of TU Delft's Joint Interdisciplinary Project (JIP), a 10-week program that brings together students from diverse academic backgrounds to solve real-world challenges in collaboration with industry partners.

Our team, composed of six second-year master's students and one third-year bachelor's student from fields including Artificial Intelligence (AI), Computer Science, Data Science, Industrial Design, Civil Engineering, Management of Technology, and International Business Administration, worked in close partnership with KLM Royal Dutch Airlines, one of the oldest and most innovative airlines in the world. Together, we explored innovative AI-driven solutions aimed at improving training experiences within the aviation industry: specifically how AI can be used to create unpredictable, realistic voices that can hold conversations for training purposes.

A proof of concept (PoC) AI-driven voice agent called OpenVoiceAgent was created, capable of holding high-stakes conversations useful for training. Complex AI methods, such as speech recognition, text generation, and speech generation, as well as software engineering methods, such as multi-threading, were researched and integrated into a pipeline implemented on-premise at KLM. OpenVoiceAgent was fully handed over to KLM and has also been released publicly at https://github.com/matteo-fregonara/OpenVoiceAgent.git.

A detailed validation plan was carried out to test the workings of OpenVoiceAgent, assessing both technical performance and user experience. KLM employees from diverse backgrounds were involved in hands-on evaluation interviews to determine how realistic, adaptive, and emotionally varied the AI-driven conversations felt compared to natural human conversations. Furthermore, a more general validation survey was sent to participants around the world to gather a more public opinion. Results were quantitatively analyzed using advanced statistical techniques and qualitatively explored through a thematic analysis. The results demonstrated the usefulness of the system, providing an immersive and effective training environment.

An implementation analysis and value proposition according to the Triple Bottom Line framework, supported by cost-benefit and sustainability assessments, examined the potential of deploying the solution within KLM's existing training infrastructure. The findings indicate that the PoC could significantly reduce training preparation time and resource consumption while enhancing learning outcomes through training scalability. Beyond immediate operational gains, the project also explored long-term implications, including the scalability of OpenVoiceAgent across different training scenarios and its contribution to KLM's broader innovation and sustainability goals.

Overall, this project represents a forward-looking collaboration between academia and industry, combining cutting-edge AI technology with practical human-centered design to shape the future of aviation training while also teaching the developing students about interdisciplinary industry, research, and collaboration.

Our Interdisciplinary Team

We now introduce our interdisciplinary team. As our project was quite broad and we were a large team of seven, we organized our team into sub-teams based on specialty and interest:

- Development Team: responsible for researching, developing, and implementing OpenVoiceAgent.
  - Moniek Smink: a second-year Data Science & AI Technology master's student with more than six years of experience working within interdisciplinary AI, data analytics, and computer science across industry and academia.
  - Natanael Djajadi: a second-year Computer Science master's student specialized in cybersecurity and embedded systems, with around five years of applied experience in software engineering through academic, professional, and student organization projects.
  - Matteo Fregonara: a second-year Data Science & AI Technology master's student with a passion for AI/LLMs and over two years of industry experience as a software engineer.
- Quality Control Team: responsible for carrying out a user-centric research plan to validate the performance of OpenVoiceAgent.
  - Siri Høystad: a second-year Design for Interaction master's student with a background as a UX-designer and a passion for human centered design.
- Business Team: responsible for reporting on the financial and environmental feasibility of OpenVoiceAgent and VR trainings and outlining the implementation strategy.
  - Winnie Cheng: a second-year Construction Management & Engineering and first-year Business Administration master's student, with experience in an interdisciplinary student team that is developing the world's first liquid hydrogen-powered aircraft.
  - Vassilis Varnas: a second-year Management of Technology & Civil Engineering master's student with work experience as a construction engineer and a multimedia artist.
  - Sem Zeelenberg: a third-year International Business Administration bachelor's student with a strong interest in the aviation industry. Inspired by his father's career within the company, he is deeply passionate about aviation and the future of KLM.

Each of us is highly experienced in our own area of expertise, but throughout this project we learned to think, act, and build in an interdisciplinary way. First, our team was constantly sharing background knowledge of our fields, holding regular internal presentations where we shared our tricks-of-the-trade in topics such as AI, interviewing, statistics, and business in general. Next, our team maintained collaborative project management records and held regular check-ins so everyone always knew what others were working on. Finally, each of us was often involved in each others' processes: the business team developed the AI voice samples used by the development team, the development team aided the business team with AI cost estimation, and everyone was involved with the interviewing process led by our quality control team.

Our team's interdisciplinary collaboration was essential to the project's success. By combining our diverse expertise, we were able to approach complex problems from multiple perspectives and integrate technical, human, and strategic dimensions into one cohesive solution. The open exchange of ideas not only strengthened the technical quality of our proof of concept but also ensured that it remained relevant, usable, and valuable for KLM. Our ability to bridge disciplines, translate concepts across fields, and co-create under tight time constraints was what ultimately allowed us to transform an ambitious idea into a functioning, validated, useful prototype.

Figure 1: Our interdisciplinary team from left to right: Sem Zeelenberg, Natanael Djajadi, Siri Høystad, Winnie Cheng, Jae Maloney (our coach), Moniek Smink, Matteo Fregonara, and Vassilis Varnas.

1 Introduction

It's often difficult to find human connection in works of engineering, but it's hidden somewhere within the designs, the analysis, and the algorithms. Even the creators themselves sometimes are not aware of it, but making something wonderful and sharing it with people that you won't meet is a deeply selfless act. An understanding that ideas stretch beyond our lifespans and a quiet faith that people you won't shake hands with might expand on them and use them for good. When the Wright brothers lifted a fragile prototype machine off the sand dunes of North Carolina in 1903, they believed they had conquered gravity. Without realizing it, they had also set in motion a series of events involving human ingenuity, curiosity, and collaboration that would change humanity. From a single twelve-second flight, an industry worth over four trillion dollars emerged, involving airlines, manufacturing, logistics, tourism, and maintenance sectors.

1.1. Trainings at KLM

The aviation industry supports more than 90 million jobs worldwide, with professionals constantly trained to uphold the efficiency and safety of global flights. To this day, the people who work in this industry remain its backbone, and their training is of vital importance to all aviation companies, including KLM.

Across the aviation industry, employees undergo a wide range of trainings, technical, human-centered, or a combination of both. These trainings are conducted before they begin their duties or throughout their working careers to keep their skills sharp and up to date. Please see Table 1.1 for an overview of some of the trainings seen at KLM.

Category	Training Type	Description
Cabin Crew Training	Safety & Emergency Procedures	Operation of emergency exits, slides, fire extinguishers, and life vests.
	Crisis Management	Handling disruptive passengers and de-escalating emotional situations.
	Service Training	Customer service, communication, and hospitality for different passenger profiles.
Flight Crew (Pilots) Training	Simulator Training	Full-motion flight simulators for aircraft systems, takeoff/landing, and emergency maneuvers.
Ground & Maintenance Crew Training	Aircraft De-Icing Operations	Safety procedures for applying de-icing fluids on wings.

Category	Training Type	Description
	Maintenance & Engineering Training	Diagnostics, repairs, and component replacement procedures.
	Ramp Safety & Pushback Operations	Ground handling, towing, and safe ramp operations.
	Baggage & Cargo Handling	Load balancing, safety protocols, and hazardous materials awareness.
Customer Care & Service Training	Emotional Support & Communication Skills	Training for Care Team members to manage sensitive and emotional customer call
Emergency & Security Training	Firefighting & Smoke Evacuation	Simulations for onboard fires.
	Security Awareness & Threat Response	Recognizing and responding to potential security risks.
	First Aid & Medical Response	CPR, AED use, and managing in-flight medical emergencies.

Table 1.1: Overview of KLM's Main Training Categories and Their Objectives

Following this overview, it becomes clear that KLM's training programs span both technical and human-centered domains. The distinction between these categories is crucial when adopting immersive technologies.

1.1.1. Technical and Human-Centered Trainings

While many of these trainings focus on technical skills, such as operating aircraft systems or conducting de-icing procedures, others emphasize interpersonal and emotional competence, including communication, conflict management, and empathy under stress (Table 1.2). Technical trainings can often be replicated effectively in virtual environments, where tasks follow clear, procedural logic. However, interpersonal trainings are far more complex: when simulated in XR, they tend to feel scripted and predictable, limiting their ability to capture the spontaneity and emotional depth of real human interactions.

According to the conducted business exploration, technical trainings have been successfully replicated in XR, effectively replacing traditional methods that relied on mock airplane environments. Interpersonal trainings, however, continue to rely on conventional approaches, primarily through roleplay sessions with actors to simulate passenger interactions and emotionally charged situations.

Field of Training	Technical	Human-centered
Safety and Emergency Procedures	✓	□
Crisis Management	□	✓
Service Training	✓	✓
Simulator Training	✓	□
Aircraft De-Icing Operations	✓	□
Maintenance and Engineering	✓	□
Ramp Safety and Pushback	✓	□
Baggage and Cargo Handling	✓	□
Firefighting and Smoke Evacuation	✓	✓
Security Awareness and Threat Response	✓	✓
First Aid and Medical Response	✓	✓

Table 1.2: Field of training with a technical vs. human-centered checklist.

1.1.2. VR Trainings

In recent years, KLM has innovated and successfully moved beyond traditional training methods and embraced virtual reality (VR) to replicate technical scenarios. For instance, its subsidiary KLM Cityhopper rolled out a VR application for 2 types of aircrafts (Embraer 175 and 190). Pilots could familiarise themselves with cockpit procedures via a VR headset even from home. (KLM Newsroom, 2021) Other VR applications at KLM not limited to pilots include trainings for fire-safety, bridge operations and aircraft push-back handling. A clear pattern emerges within KLM's business operations, showing that VR-based trainings have proven highly effective and, since their initial rollout, have been increasingly adopted and integrated across different employee groups as depicted in Table 1.3 (KLM Tech Data, 2023)

1.1.3. Limitations of Current VR Technology

Technical trainings so far have been the primary focus of KLM's virtual reality simulations, from pilot simulations to cabin crew training and ground operations. These training modules have proven to be effective in learning technical skills, improving safety, and maintaining efficiency. The primary focus of KLM was technical trainings because these kinds of trainings can be simulated with the technology that exists today. On the other hand, trainings that involve communication with passengers and emotional depth cannot be simulated with static VR technology. So far, these trainings are conducted with actors who can produce this element of humanity and unpredictability. Therefore, the limitation of VR trainings to just technical simulations is a matter of feasibility, and there is a willingness, if the technology exists, to expand the trainings to human-centered ones as well.

1.1.4. Integrating Artificial Intelligence into VR

KLM recognizes the deep impact and improvement in operations that VR trainings bring, and they want to expand that innovation into all kinds of trainings, including human-centered ones. To achieve this, KLM plans to incorporate AI into VR trainings, allowing trainees to interact with virtual characters that can think, react, and respond in an unpredictable manner but within the context of the training.

Scenario	Roll-out (Date)	Notes
Pilot VR (Embraer 175/190)	5 Nov 2020	Pilots VR trainings simulating the Embraer 175 and 190 airplane (Business Review Europe, 2021).
Fire-Evacuation for Engineers	2017 (pilot)	VR simulation for safe evacuation from maintenance hangars, reported to have trained several hundred engineers (Computer Weekly, 2017).
Fire-Safety (Cabin/Crew)	Early 2019	Simulations of smoke alarms, passenger reactions, and emergency procedures (Capgemini Netherlands, 2019).
Passenger-Bridge Operations	Early 2019	VR module for operating the jet bridge (Capgemini Netherlands, 2019).
Aircraft Push-Back Handling	Early 2019	Simulated Schiphol apron scenario including arrival, taxi, unloading, and push-back sequences for ground-crew practice (Capgemini Netherlands, 2019).
De-Icing Simulation	2022	Demonstration of VR-based de-icing and ground-operations procedures (VROWL, 2022).

Table 1.3: Overview of KLM's VR Training Modules and Rollout

1.1.5. Deconstructing the AI Agent

The process of introducing AI to virtual environments involves deconstructing the very elements of an AI agent. The first step is to add unpredictability to what a participant hears through AI-generated speech, then how it adapts to the trainee's actions, and later on, what they see in a fully realistic virtual environment. The building blocks progress from voice to a fully integrated avatar.

Project Focus

Our project focuses on the auditory part of this vision. It explores how AI voices, also known as voice agents, can simulate emotionally responsive conversations in training situations that mirror real human conversations.

1.1.6. KLM's Voice Agent Prototype: Caresse

During KLM's exploration of AI-driven virtual trainings, they developed an experimental voice agent known as Caresse. This prototype was created using one of OpenAI's commercial models, which allowed natural conversations with quick responses. Their prototype proved that an AI-driven conversational agent was a promising solution.

Despite its positive responsiveness and usability, the implementation of Caresse is heavily based on OpenAI's commercial infrastructure, meaning that it is based entirely on external servers. This introduces several drawbacks regarding privacy, cost, and controllability. Without an enterprise agreement, KLM's project is not stable, and their data is not protected. Caresse proved that an AI-driven agent can be implemented and is useful, but also highlighted that there is a need for a locally hosted alternative that ensures full data sovereignty and long-term operational control.

1.1.7. Development Pathway Towards a Local AI Agent

The development of KLM's human-centered VR training follows a gradual process that allows testing, validation, and expansion over time. The phases unfold as follows:

- **Phase 1 — Proof of Concept (PoC):** The process begins with a simple AI-driven call scenario that demonstrates whether realistic and emotionally responsive speech can be produced in a training setting. The call scenario could be an incoming call from a distressed relative, in the aftermath of an accident. This phase proves that the concept is technically possible. The project which we will be working on focuses exclusively on this initial PoC phase.
- **Phase 2 — Minimum Viable Product (MVP):** Once the feasibility is confirmed, the AI voice agent is tested in an actual training session. This stage focuses on showing that the technology is useful for learning and that participants find it realistic and valuable. In this phase, the AI agent is in the form of an avatar.
- **Phase 3 — Full Integration:** After the system is proven to work in one context, it can be applied across to multiple KLM's human-centered trainings with the appropriate adjustments.

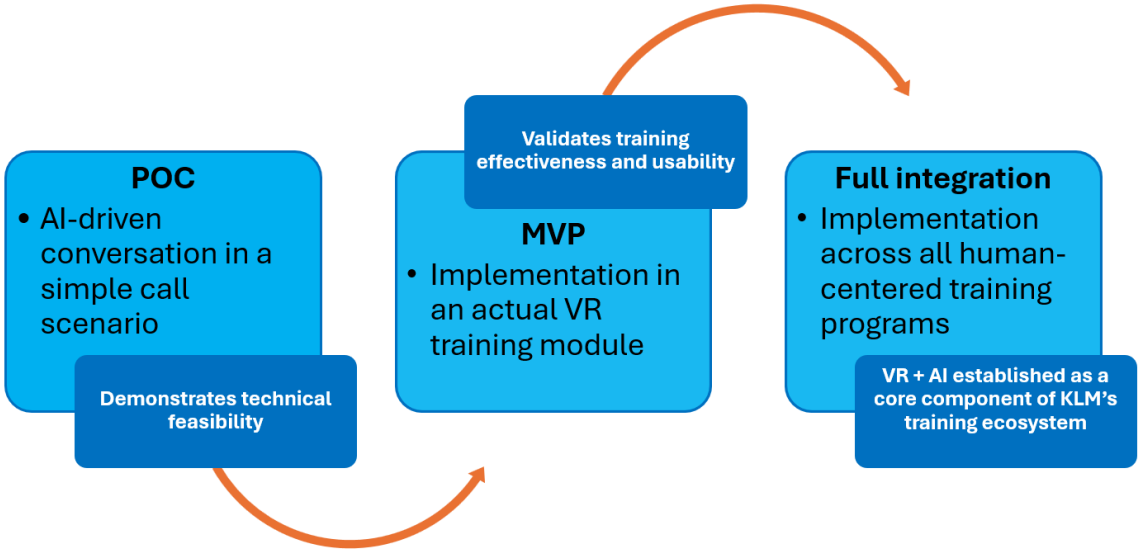


Figure 1.1: Development pathway (PoC → MVP → Full integration) for KLM's human-centered VR+AI training.

1.2. Problem Statement

The objective of this study is to explore the feasibility of an AI-driven conversational agent capable of recognizing and generating emotional responses in the form of speech.

The project we will be working focuses exclusively on the intial PoC phase as mentioned in 1.1.7. More specifically, the project aims to deliver a PoC AI voice agent operating entirely on a local system, ensuring full privacy, controllability, and scalability.

1.2.1. Research Question

The primary research question of this project is to examine if a PoC AI voice agent developed with limited resources and a narrow time-frame can be perceived as realistic and comparable with a human. A few consequent questions can be asked like:

- Is the developed PoC perceived as human-like and realistic in conversational interactions?

- How does the developed PoC's perceived humanness compare to KLM's existing AI system, *Caresse*?
- What parameters influence perceived humanness?

1.2.2. Evaluation

To evaluate the prototype, a series of tests were conducted comparing the developed voice agent to KLM's existing commercial model, Caresse. The purpose of these evaluations was to first assess the technical performance and conversational realism of the locally hosted prototype, and second, to gather user feedback to guide further refinement. By combining quantitative measures with qualitative impressions from KLM employees, the evaluation aimed to identify strengths, limitations, and opportunities for improvement in developing future emotionally responsive training agents.

1.2.3. Broader Impact

Beyond its technical goals, the project also seeks to broaden the flexibility of AI voice agents use within KLM's ecosystem, it sets the foundations for a broader use of AI at KLM and generally in the aviation industry.

2

Development Process of our Prototype: OpenVoiceAgent

Within ten weeks, our team moved through many phases of development while creating our conversational agent PoC. We started with the first development phase where we explored KLM's operational environment, researched potential solutions, and implemented our first working prototype, which was used in the validation tests discussed in Chapter 3. Next, we completed a second development phase where we addressed and researched solutions to limitations found in the first development phase, the final version of which was used in the survey discussed in Chapter 3. Finally, directions for future development are discussed.

2.1. First Development Phase

In the first development phase, over 50 AI models were researched and bench-marked. See below a zoomed out view of our model research in Figure 2.1.

By refining our problem statement, we quickly learned of several constraints imposed by KLM's operational environment, which we discuss below. Based on these constraints, we were able to narrow down our list of models according to feasibility and develop our prototype: OpenVoiceAgent. We will discuss our most important choices below.

2.1.1. Constraints

Our project was developed under several constraints set by KLM. First, the conversational agent was required to run entirely offline on a locally available NVIDIA GeForce RTX 5090 GPU. This restriction inherently limited our choices to open-source models that could be deployed without reliance on external APIs or cloud services. In addition, KLM did not provide any proprietary data for training or fine-tuning, which meant we had to rely entirely on pre-trained, out-of-the-box models. Lastly, all models and code used in the project were required to have a commercial license to allow KLM to use the pipeline.

2.1.2. Technical Options & Model Choices

We now introduce the technically feasible options for our conversational agent PoC that should be runnable on an NVIDIA GeForce RTX 5090 GPU on KLM premises. We introduce two general types of architecture: single model & pipeline of models, then go into more details on the specific model options for each.

Single Model: Speech-to-Speech (STS)

To build a conversational agent capable of emotional conversations, the simplest solution is to have a single model that can take in user speech waves and output agent speech waves that are coherent, logical, and contextual. This type of architecture requires one model to be responsible for all aspects of holding a conversation: hearing, thinking, and speaking. Such models tend to be larger in size and more brittle to new scenarios and voices. We examined

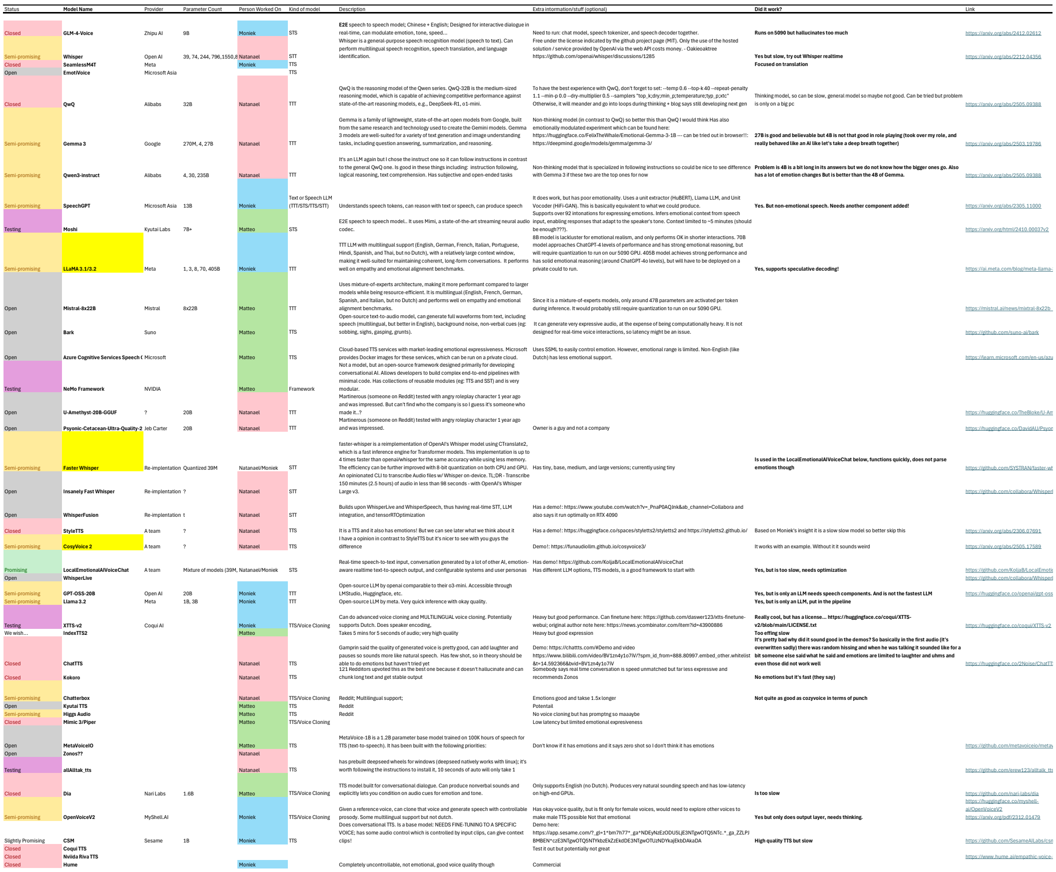


Figure 2.1: Research workbook of models researched for the first development phase of OpenVoiceAgent.

several such models but will now discuss two.

The first is OpenAI's `gpt-realtime` model (OpenAI, 2025a). KLM currently uses this model to create its Caresse PoC mentioned in Section 1.1.6. GPT-Realtime is a low-latency variant of OpenAI's GPT-4/5 series, optimized for streaming inference that enables real-time dialogue with natural turn-taking and immediate responsiveness. It is a commercial, uncontrollable model, meaning that KLM pays OpenAI to host and run the model, KLM's data is not protected without enterprise contracts, and if OpenAI decides to stop providing this service, KLM would be unable to continue with Caresse. GPT-Realtime is not a viable option for our PoC as it is not runnable on KLM premises but is still mentioned in this report to introduce KLM's Caresse PoC.

The second STS model we examined was `GLM-4-Voice`, developed by ZhipuAI (Zeng et al., 2024). This model is a bilingual end-to-end spoken chatbot, built on a 9B parameter `GLM-4` text backbone that enables audio streaming by alternating text and speech tokens. We tested this model, but found it could not be faithful to a prompt and had limited emotional variability. This means that if KLM were to use this model for training, the conversational agent would easily go off topic and sound monotone. Furthermore, `GLM-4-Voice` only has one voice and is not easily customizable to change between voices. As we lack the training data to finetune this model to accomplish our use-case's needs, this model was shelved.

Overall, we found single model approaches to be non-local, too unsteerable, or not equipped with enough customization.

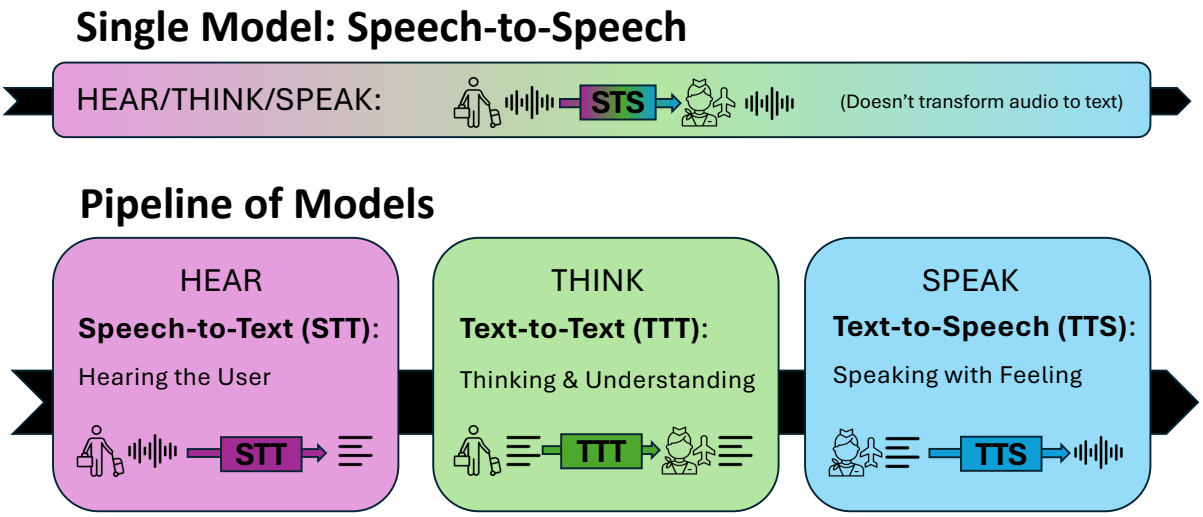


Figure 2.2: Comparison of 'Speech-to-Speech' and 'Pipeline of Models' approaches.

Pipeline of Models

The alternative to a single model approach is to have a pipeline of models that together achieve the same outcome. We split up the pipeline into different tasks including Speech-to-Text (STT), Text-to-Text (TTT), and Text-to-Speech (TTS). User speech is taken as input and agent speech is the output, but the tasks of hearing, thinking, and speaking are accomplished by separate models. For an overview of this approach and the STS approach discussed above, see Figure 2.2. Because of the customization and control allowed by such a modular pipeline, we chose to pursue this approach. We will now discuss our model options and choices for each task (STT, TTT, & TTS) within the pipeline.

Speech-to-Text (STT): Our first choice for the STT component of the pipeline was `Faster-Whisper` (Klein, 2023), a high-performance implementation of OpenAI's `Whisper` model (OpenAI, 2022) optimized for low-latency and resource-efficient inference. `Whisper` is a state-of-the-art, open-source automatic speech recognition (ASR) system trained on a large and diverse multilingual dataset. It offers strong robustness across a wide variety of accents, background noise conditions, and speaking styles, making it particularly suitable for real-world conversational settings. These properties are essential for a dialogue system like ours, where speech input can vary significantly depending on the user's tone, speed, or environment.

To ensure the system could operate efficiently in real time, we evaluated different `Faster-Whisper` model sizes, comparing their transcription accuracy and inference latency. After empirical testing, we selected the *small* variant, as it provided the best trade-off between performance and speed on our target hardware. The model's ability to deliver near real-time transcription while maintaining high word accuracy made it ideal for our application scenario. This model was integrated into the pipeline as the first stage: hearing, responsible for converting user speech input into text for the TTT model to generate responses.

Text-to-Text (TTT): The TTT task involves taking in user input and generating agent output that is grammatically coherent, follows logically from the user text, and takes into account the previous context of the conversation. The TTT model is essentially the 'brain' of our conversational agent PoC, responsible for generating an agent's textual and emotional response to a conversation. For example, if the user were to say `"I don't have any information about your missing brother."` an agent could respond with `"[frustrated] It's been hours, how could`

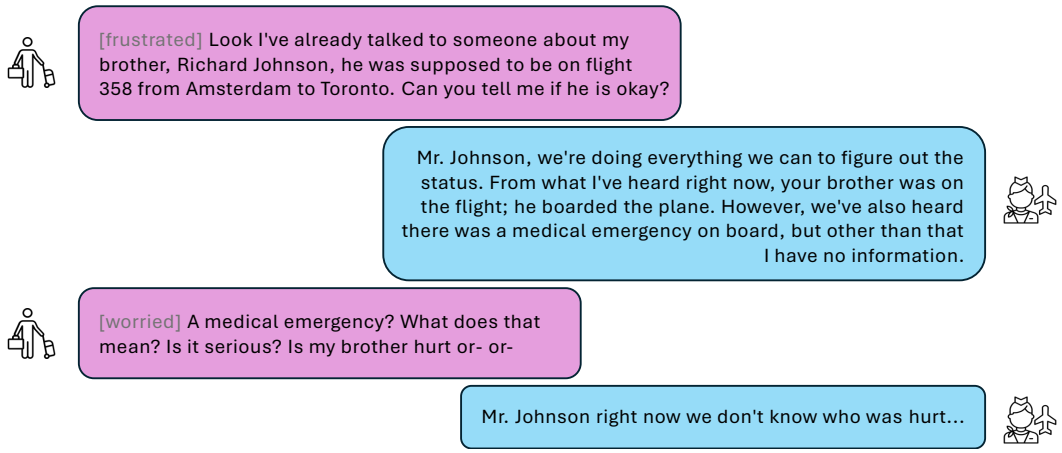


Figure 2.3: Example conversation user (right) had with OpenVoiceAgent as Mr. Johnson (left).

`you not have any information?"`. The first part of the response is named the emotion tag (i.e. `[frustrated]`) and is used in later steps to voice the provided message in a certain emotion. Therefore, not only is the TTT model responsible for the content of what is being said, but how it is said as well, making it a complex, critical task for our pipeline that needs to happen at low latencies. For an example conversation with emotion tags, please see Figure 2.3.

For this task, we considered various Large Language Models (LLMs), capable of taking in text and producing text while being steerable with system prompts. We considered several branches of LLMs including Llama (Grattafiori et al., 2024), Qwen (Qwen et al., 2025), and GPT-OSS (OpenAI et al., 2025). We found each of these LLM branches capable of generating adequate output, however, due to our latency limits, we chose to consider smaller model sizes and additional options for speed-ups. Two speed-ups we considered were quantization and speculative decoding. Quantization involves lowering the precision of model weights while speculative decoding is a text generation technique where a large LLM uses a small LLM to propose candidate tokens, speeding up the generation (Yan et al., 2025). We chose to use GGUF quantized `Llama-3.1-8B` as the main LLM with a GGUF quantized `Llama-3.2-1B` model as the smaller draft model for speculative decoding.

Finally, we realized that as models develop and become faster, our pipeline's TTT model is likely to be upgraded because it is the main 'brain' in our pipeline. Therefore, our optimized Llama models are deployed within LMStudio, a tool for local, fast, modular LLM deployment (Team, 2023). This tool allows a user to switch out the LLM used at any time, meaning that if a new, better model comes out or if a user wants to use a different LLM, future users of our conversational agent can switch the TTT model at any time, making our pipeline customizable.

Text-to-Speech (TTS): The TTS task involves taking the text and emotion tags generated by our TTT task, and outputting it as emotionally realistic audio. Low-latency options like `Kokoro` (Hexgrad, 2025b) were fast, but do not incorporate emotions. Thus, we looked at emotion-driven models like `ChatTTS` (Hexgrad, 2025a), `Higgs Audio` (Boson-AI, 2025), and `Cosyvoice2` (Du et al., 2024), where you can use tags such as `[sad]` to add emotion to generations, however, we found these not emotionally realistic enough and often indistinguishable.

Both `Higgs Audio` and `Cosyvoice2` additionally support *voice cloning*, a technique that lets a TTS model use a separate audio file to 'color' the generation, copying its voice and emotion. An illustrative example of voice cloning can be seen in Figure 2.4. `IndexTTS2` (Zhou & et al., 2025), for example, had very high-quality output, but took 5 minutes for 5 seconds of audio generation, making the latency way too extreme for real-time conversations. `XTTS-v2` (Casanova et al., 2024) can do advanced voice cloning and multilingual voice cloning with speaker encod-

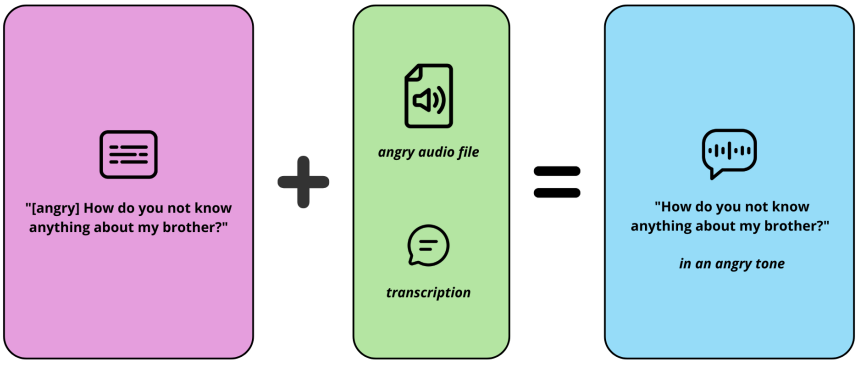


Figure 2.4: Voice cloning example

ing, but unfortunately, it has a non-commercial license. `Chatterbox` (Resemble AI, 2025) also has multilingual support, but we found its emotions too weak and inconsistent. `Higgs Audio` produced high-quality speech output with realistic emotional expression. However, just like with `IndexTTS2`, inference time proved to be too long for real-time conversations.

Finally, `Cosyvoice2` met our standards, as it best preserved the tone (intonation, pitch, rhythm, emphasis), emotion, and color of the voice (timbre) with acceptable latency. With streaming modeling technologies and quantization we were able to further improve the latency, although still not quite to real-time levels. `Cosyvoice2` is also multilingual, supporting English, Chinese, Japanese, Korean, and Chinese dialects.

Final Pipeline

For an overview of the models chosen in our final pipeline, please see Figure 2.5.

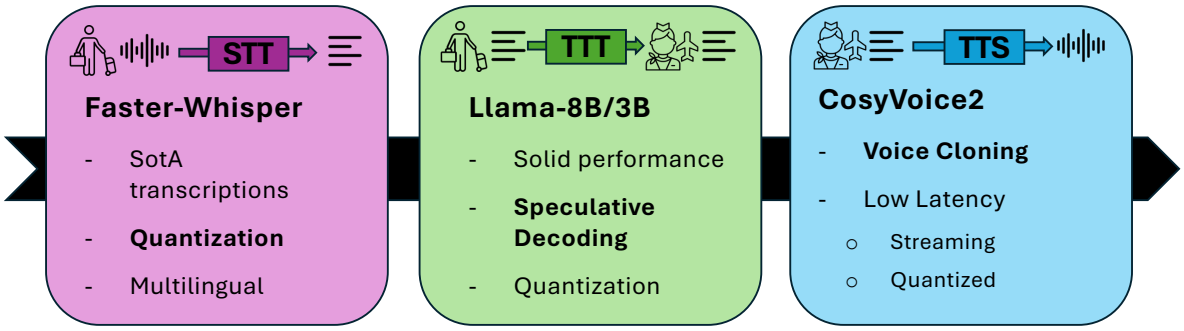


Figure 2.5: Final model choices and reasonings for each task within the pipeline of OpenVoiceAgent.

2.1.3. Pipeline Design

Now that we established which models are best suited for each part of the pipeline, we will now discuss how these models are orchestrated together into a coherent and responsive conversational agent. The design and architecture of the pipeline determines how information flows between the STT, TTT and TTS components, and how efficiently each component operates under real-time constraints.

Conversation Flow

Beyond which model to use, a conversational agent pipeline cannot feel truly realistic without some clever engineering of knowing *when* to use the models. For this, we use multi-threading and streaming to implement separate user listener, response generator, and response voicer threads. The user listener thread listens to and records any user speech, translating it into text using the STT model. The response generator thread processes any user input, generating agent responses using the TTT model. The response voicer thread processes any agent response output, generating the emotional speech in one sub-thread and playing the emotional speech in another. Furthermore, we allow the user listener thread to interrupt the other two which allows the user to interrupt the agent at any time, giving realistic conversational control to the user. Allowing the agent to interrupt the user is an ongoing question of whether we want to give the AI such control. For an overview of our multi-threading setup please see Figure 2.6.

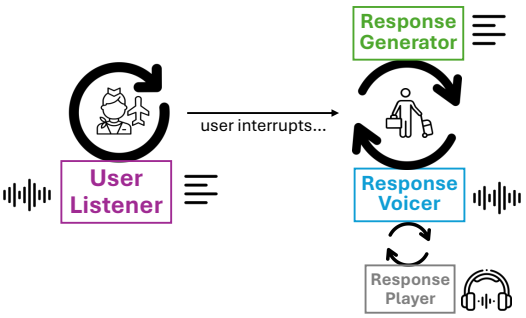


Figure 2.6: Multi-threading architecture of the pipeline of OpenVoiceAgent.

Prompt Engineering

Since we employed smaller LLMs (3B–8B), we designed structured and concise prompts to reduce confusion and maintain response consistency. Each prompt included three parts:

- A short *character description* describing the agent's identity and emotional disposition.
- A *scenario* describing the conversational context.
- A numbered list of *guidelines* specifying behavior and output format.

The guidelines instructed the model to begin each response with an emotion tag (e.g., `[sad]`, `[angry]`) and maintain logical and emotional consistency throughout the dialogue. Because smaller models often shifted emotions too abruptly, we refined the rules to promote smoother emotional transitions. To avoid confusion between user and agent roles, we made role definitions explicit and reinforced them within the guidelines. Finally, we added few-shot prompting with four example responses, which improved response consistency. For an example prompt given to our TTT model please see Figure 2.7.

Voice Samples

Since voice cloning requires reference audio, we used ElevenLabs (ElevenLabs, 2025), a state-of-the-art speech & AI voice generator, to generate reference audio clips in different emotions. We chose ElevenLabs because we found the voices to be very realistic and because the voice and emotion is controllable through prompting. Thus, we specified the voice gender and tone to create voice clips for 22 emotions from the Emotion Typology (Fokkinga & Desmet, 2022). The emotion typology is an initiative of the Delft Institute of Positive Design (Delft University of Technology), where many scientists from different universities (Delft University of Technology, Twente University, and University of Amsterdam), experts, and designers contributed over a time span of seven years (2015-2022) to create a comprehensive representation of distinct human emotions. The Emotion Typology can be seen in Figure 2.8.

Example Prompt for TTT model

You are Miss Johnson, a 31-year-old woman, worried and angry about her missing brother. Miss Johnson feels desperate, angry, hopeful, sad, etc., but emotions shift slowly, not suddenly. She speaks in short, blunt, casual sentences. Angry = aggressive. Sad = raw. Calm = terse. She holds grudges.

You are calling KLM on a hotline on the phone to ask whether they have located your brother, Richard Johnson, who possibly was on board KLM flight 358 from Amsterdam to Toronto. You have just heard that a serious incident has happened with this flight. You are not sure if your brother was on board, but it is very likely. You are now talking to a KLM employee. That person speaks to you as a KLM care team member. End the call once there is an agreement about a callback.

Rules:

1. Each answer starts with one emotion in [ ]. Use ONLY ONE of these emotions: `admire, amused, angry, annoyed, confused, crying, desperate, disappointed, distress, excited, frustrated, furious, gratitude, happy, hopeful, neutral, pride, relieved, resentment, sad, shock, shouting, thrilled, worried.`
2. Then write 1–2 short sentences (max 25 words), like natural phone speech.
3. Ask ONLY ONE question at a time.
4. Keep the same emotion for 2–3 responses. Then change only to a nearby emotion (e.g., shock to worried, frustrated to angry, sad to frustrated, etc.).
5. Escalate emotions step by step if no clear information.
6. Don't repeat exact words; always rephrase.
7. Always respond as the character; do not mention you are an AI. You are NOT in any way related to the KLM staff. You are NOT the KLM care team member.

Examples:

1. [frustrated] This is going in circles, why can't anyone at KLM just give me a straight answer about my brother?
2. [angry] Why haven't you found my brother yet? I demand answers NOW!
3. [distress] I still don't know what happened to my brother and I'm worried sick, can you please give me clear answers instead of only vague ones?
4. [desperate] I beg you, please, just tell me if Richard is safe or not!

Figure 2.7: An example of one of the prompts used for the TTT model (Llama). It includes a character description, a scenario, a numbered list of guidelines and sample responses.

To validate the voice clips to have the right emotion, we played an "emotion game" in the team, where teammates would guess the voice clip's emotion using the emotion typology without seeing the specified prompt. If the guesses matched, we kept the emotion that was specified, otherwise, we renamed it. The voice samples are accompanied by transcriptions. In the end, we created two sets of emotional voices: one female-American and one male-British voice, to demonstrate the flexibility of OpenVoiceAgent to accommodate most voices.

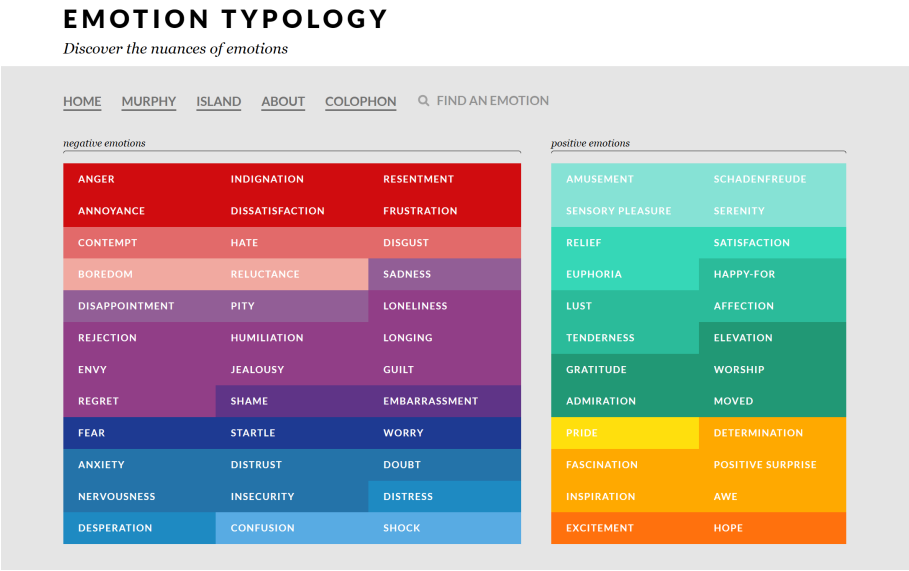


Figure 2.8: Emotion Typology

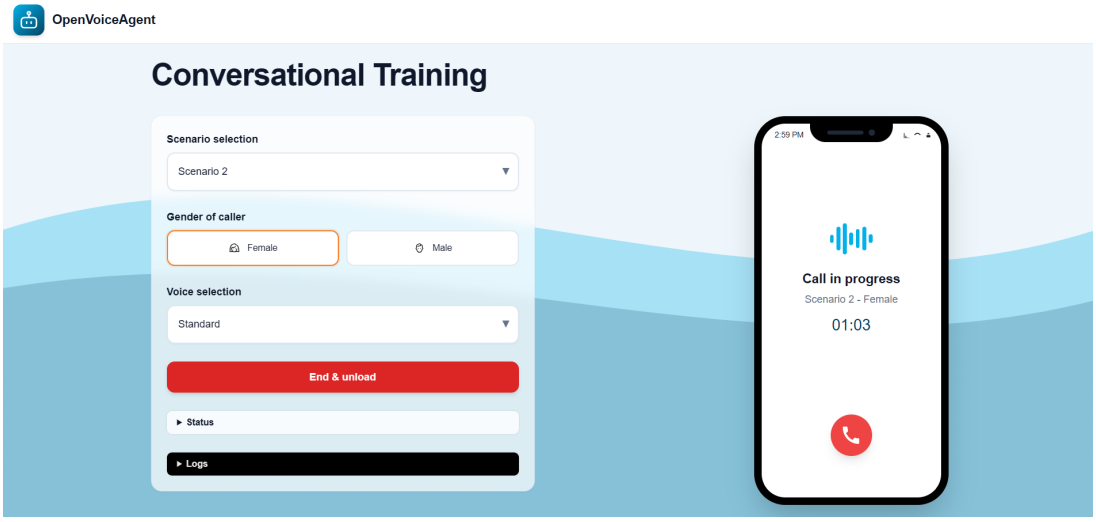


Figure 2.9: OpenVoiceAgent Frontend Interface

2.2. Second Development Phase

After delivering the first version of OpenVoiceAgent for the PoC tests, our development team returned to the drawing board to evaluate potential improvements and limitation mitigation strategies in a short two-week window. First, for a successful handover, a frontend was designed and implemented. Next, time-to-first AI generation response was identified as a major bottleneck and examined. Finally, an external dependency was ironed out to enable full on-premise deployment.

2.2.1. Frontend

A frontend interface was developed to simplify the testing of our pipeline in HTML, CSS, and JavaScript. This interface provides a more intuitive and user-friendly way to run experiments compared to manually entering commands in the terminal. Additionally, it allows us to effectively demonstrate the system's functionality during the final presentation. A screenshot of the interface during a call can be seen in Figure 2.9.

The key functionalities are that users can select a scenario, which is dynamically updated from

the `prompts/` directory, the caller gender, and the exact caller voice which is also dynamically updated from the `wavs/` directory. Every prompt scenario contains two sub-directories: `female_char/` and `male_char/`, containing the corresponding configuration files.

When the model is being loaded, the phone reflects this by displaying `Connecting` and `Please wait...`. When the model is ready, the user can start the call by clicking the green call button. The call can be ended either by clicking the red call button or selecting `End & Unload`.

For development and debugging purposes, we included collapsed sections labeled `Status` and `Logs`, where `Status` provides information such as the file path where the transcripts will be saved, and the `Logs` section displays output typically shown in the terminal.

The full implementation can be found on our GitHub repository, where the interface can be run and explored in action (Smink et al., 2025).

2.2.2. Latency Improvements

The overarching feedback from our own tests and our coach was that the response time of the AI agent was sub-human: participants were often waiting 2-5 seconds before an AI response. Our team evaluated three approaches for mitigation and made some simpler changes.

Flash Attention

First, flash attention was examined as a potential alternative GPU kernel for faster TTS performance. Flash attention is a hyper-optimized version of attention which keeps important computations in on-chip SRAM, significantly reducing the amount of GPU accesses, and smartly avoiding expanding unnecessary calculations (Dao et al., 2022). However, Python's built-in Flash SDPA doesn't support local or sliding windows which are used in `Cosyvoice2`'s `Qwen2` backbone. Furthermore, official flash attention implementations are not distributed for Windows. Within our short two week time frame, this option was infeasible.

Virtual Large Language Model (vLLM) Library

Next, vLLM was examined as a way to run our TTS model, `Cosyvoice2`, in an optimized way. VLLM is a high-performance inference engine library for LLMs, designed to maximize GPU throughput (Kwon et al., 2023). The creators of `Cosyvoice2` have integrated the model into the vLLM library. However, we were unable to verify the success of this integration because vLLM is only supported on Linux operating systems as well as having specific GPU, CUDA, and software requirements. Even while dual-booting a Linux operating system on our Windows RTX 5090 computer, we were unable to resolve the conflicts between `Cosyvoice2` and vLLM in our short two week time frame.

TensorRT

Finally, TensorRT was examined as another potential way to hyper-optimize our models. TensorRT is NVIDIA's deep learning inference runtime that compiles models into highly efficient GPU executables, often achieving large speedups through kernel fusion, quantization, and precision tuning (NVIDIA Corporation, 2024). However, TensorRT integration requires exporting and optimizing each model from PyTorch to ONNX to TensorRT formats, testing precisions, and matching CUDA/TensorRT versions. Our models and pipeline weren't profiled for TensorRT, and setting this up would require extensive rebuilding which was too intensive for our short timeline.

Miscellaneous

Besides specific speed up approaches, our development team looked at the existing pipeline and searched for simpler potential areas of latency improvement. Multi-threading sleep timers were optimized, a warm-up was sent to each model, and extraneous processing was removed.

2.2.3. Removing Connection Dependency

One of the constraints outlined in Section 2.1.1 required the agent to operate fully offline. During implementation, we discovered that our TTS model, `CosyVoice2`, attempted to download one of its dependencies, `wetext`, from the source repository each time the agent was initialized. This behavior caused the system to fail when executed in an offline environment. To address this, we applied a patch that redirected `CosyVoice2` to use a locally stored version of `wetext`, thereby ensuring the model could run successfully without an internet connection.

2.2.4. Concluding the Second Phase

In the second phase of development, we consolidated early feedback and implemented various improvements: adding a frontend, optimizing latency, and ensuring the agent can run fully offline. After completing these updates, we recorded new conversations with the improved agent and designed a survey to evaluate its performance to be discussed in Chapter 3. This marked the transition from development to the final evaluation phase.

2.3. Technical Future Work

Throughout these development phases, there have been many ideas and directions for technical improvements that we did not have the time to explore. We discuss ways to improve OpenVoiceAgent in general, not just for KLM specifically, below.

2.3.1. Conversation Flow

An important aspect of developing a realistic conversational agent is achieving a natural, human-like flow of interaction. Beyond generating coherent responses, this involves replicating aspects of human conversation such as timing, pacing, and interruptions. Improving conversational flow requires focusing on both technical performance, such as minimizing latency, and behavioral factors, such as enabling the agent to interrupt the user. The following subsections discuss future directions for improving these two aspects.

Latency

As discussed in Section 2.2.2, one of the main challenges identified during the PoC tests was the high delay between user input and the agent's spoken response. Fast response times are crucial to maintaining a natural conversational flow and making the agent feel more human-like. The current delay of 2–5 seconds was noticeably artificial for participants, so reducing this latency should be a primary goal for future iterations.

There are several ways to address this. Using vLLM or TensorRT could improve inference performance, as mentioned above. Running the system in a Linux-based environment would also simplify dependency management and enable better support for GPU-optimized libraries such as FlashAttention and vLLM.

In addition, future work could explore deeper quantization strategies to lower computational load. Currently, both the TTS and SST use 16-bit floating-point precision (FP16) to speed up inference with minimal accuracy loss. Further quantizing them to 8-bit or 4-bit precision using frameworks like BitsAndBytes (Dettmers, 2022) or TensorRT-LLM (NVIDIA Developer, 2024) could potentially further reduce inference time while maintaining performance similar to the original model. Model pruning (Han et al., 2015) or Low-Rank Adaptation (LoRA) (Hu et al., 2021) could also be applied to reduce model size and improve inference speed while preserving performance. In particular, combining LoRA with knowledge distillation, the process of training a smaller model to mimic a larger one, can yield smaller models that retain comparable performance which perform inference faster.

Another way to improve latency is by introducing a caching layer for the TTT model. Common user prompts, such as greetings or common questions, can be stored with their corresponding

model responses and returned instantly without inference. More advanced implementations can use *semantic caching*, where prompt embeddings are compared to previously stored ones to identify similar inputs.

While these optimizations are promising for improving and reducing latency, reaching true real-time performance with a fully local pipeline and consumer hardware remains challenging given the current generation of open-source models and available hardware acceleration. However, as newer models, quantization techniques, and hardware advancements continue to emerge, real-time local execution is likely to become increasingly feasible.

Agent Interruptions

An important feature that our coach and several PoC test participants often identified as interesting is to let the agent interrupt the human. In daily life, for example, when somebody is talking for too long without stopping, the other person will interrupt.

Because interruptions are a part of daily human interactions, researchers have done studies to explore why people interrupt, what silences really mean in a conversation, and its consequences (Cantrell, 2013; Lestary et al., 2018). However, due to the complexity of the task, we were unable to explore agent interruptions. This is not only because of technical implementation complexities, but also because the role of interruptions differs per person and per culture. Hilton (Shashkevich, 2018), a Stanford scholar who surveyed 5,000 American English speakers, identified two groups. High intensity speakers, who consider talking at the same time a sign of engagement and are uncomfortable with moments of silence, and low intensity speakers, who prefer people to speak one at a time in conversation and find simultaneous chatter rude. Furthermore, found in a paper from Mr. Murata in 1993 (Murata, 1994), there exist certain differences in the use of interruptions when comparing English and Japanese conversational styles, suggesting that interruptions differ between cultures as well.

There is also the complexity of when to interrupt. Usually, people interrupt at specific moments. In the same paper of Mr. Murata (Murata, 1994), interruptions were classified in two ways: co-operative and intrusive. Co-operative interruptions are when a conversational partner supplies or completes a word or phase when a speaker is searching. Intrusive interruptions appear more aggressive and threaten the ”territory” of the speaker. Intrusive interruptions can be further divided into topic-changing, floor-taking, and disagreement interruptions. Topic-changing interruptions (TCI) shift the discussion to a new subject introduced by the interrupter. Floor-taking interruptions (FTI) occur when a speaker takes the conversational floor to maintain a balanced turn-taking process, usually continuing or expanding on the current topic. Disagreement interruptions (DI) happen when the next speaker interrupts to express a contrasting or opposing viewpoint.

For future work, the complexity of agent interruptions can be adjusted as desired. A basic implementation could allow the agent to interrupt when the user speaks for an extended period. However, to create a more natural and human-like interaction, it would be valuable to study how and when people interrupt each other and replicate those patterns realistically in the system.

2.3.2. Conversation Content

Besides the flow of a training conversation, given more time, our team would also want to spend more time improving the content of the conversation. This means that the agent would say more meaningful, useful, and realistic things in response to a user. We discuss three of our ideas below.

Text-To-Text Fine-tuning

First, our team suggests fine-tuning the TTT model. The TTT model is OpenVoiceAgent’s ’brain’. It is responsible for responding appropriately to the user given the history of the conversation.

We noticed that our out-of-the-box TTT models often misused the emotion tags meant to control the voice cloning emotion. For example, they were often used to narrate actions, returning sentences such as '`[slams fist on table]`', or to return emotion tags that were nonexistent, such as '`[outraged]`'. By fine-tuning our TTT model to only use emotion tags correctly, the emotions within the voice agent could be more realistic and consistent. Furthermore, out-of-the-box TTT models would find it difficult to enact specific in-domain situations such as debates between lawyers or medical conversations. Fine-tuning in such specific situations may allow small, fast TTT models to act in such situations without needing to consult slower, larger TTT models.

Our team would suggest lightweight fine-tuning methods such as LoRA or model distillation to provide powerful training without high data requirements for easy, meaningful training. LoRA attaches lightweight adapter parameters to parts of a model which are then trained, allowing the rest of the model to stay frozen (Hu et al., 2021). Distillation involves tuning a small student model with a larger teacher model to provide the benefits of better quality generations at lower latencies (Zhang et al., 2024). Overall, fine-tuning a specialized TTT model for our task would create more realistic, controllable emotional situations, although the compatibility with speculative decoding and LMStudio would need to be re-assessed.

Emotional Realism

Next, our team would suggest smarter emotion handling within our pipeline.

Emotion Tag Sentiment Analysis: As mentioned above, we noticed that the TTT model would occasionally misuse emotion tags to those not included in the voice cloning sample emotion list. Currently, the emotion picked by the TTT model is compared to the list of available emotions via exact-keyword matching. This means that if the TTT model picked a nonexistent emotion, the default '`[neutral]`' emotion would be used as the voice cloning sample, often ruining the conversational immersion as such nonexistent emotion tags were often very descriptive such as '`[slams fist on table]`'. We propose training a lightweight sentiment analysis model to classify such emotion tags given by the TTT model back to the original list, to eliminate these edge cases. However, we must also consider that this would make our pipeline less portable as if a user would want to implement a new voice with different emotions, this sentiment analysis model would need to be modified.

Emotion Consistency: Furthermore, our team noticed that emotions were often quite variable within the conversations, sliding quickly between '`[angry]`' and '`[sad]`' and '`[worried]`'. In order to maintain more context while picking future emotions, we propose integrating previous emotion history and context into the TTT prompts, drawing more attention to the emotional history to allow the TTT model to make more informed decisions on future emotions. We believe this could regulate the emotional flow within a conversation and create a more realistic conversation. However, we must also consider that this could increase the latency of TTT generation and that this could potentially make such a conversation less emotionally realistic for some who want such fluctuations.

User Sentiment Analysis: Finally, our team would like to mention that the emotions of the users are currently not being parsed by our pipeline. The only input our pipeline receives is the text of what the user said. However, future developers could consider adding a sentiment analysis model to parse the input emotion of the user and giving that user emotion tag into the TTT model for more emotionally-intelligent agent responses. However, the EU AI act prohibits the usage of biometric data, such as vocal tone, to detect emotion in educational or workplace environments (European Parliament and Council of the European Union, 2024). As our pipeline may or may not be used in a workplace environment, we leave this enhancement to future deliberation.

Colorful Language

With out of the box TTT models, there are always some levels of censorship present. In our case, this means that it is quite difficult to get OpenVoiceAgent to use colorful language. Several PoC test participants identified this as unrealistic in emotional situations: humans often use swear words to express anger or frustration. Given more time, our team would be curious to explore uncensored TTT models, to provide a more realistic conversational experience. However, as the balance between realistic and unethical conversation must be carefully upheld, we leave this to future work. The ethics of OpenVoiceAgent will be discussed further in Section 5.2.1.

Multilingualism

Additionally, it must be mentioned that our current pipeline only processes and generates text in English. As KLM and other users could greatly benefit from using OpenVoiceAgent in other languages as well, given more time, we would aim to make our pipeline multilingual. This is easy in some ways and difficult in others. First, the STT model, `Faster-Whisper`, supports more than 90 languages, including Dutch, making it the most multilingual model in our pipeline. However, both the TTT (Llama) and TTS (`CosyVoice2`) models support limited languages. Llama supports German, French, Italian, Portuguese, Hindi, Spanish, and Thai (Grattafiori et al., 2024), while `CosyVoice2` supports Chinese, Japanese, Korean, and various Chinese dialects (Du et al., 2024). The TTT model can be easily exchanged for another within LMStudio, although there would be a latency hit without speculative decoding and quantization. `CosyVoice2` would be quite a bit harder with no easy alternative model. Future developers must evaluate whether it is possible to generate speech in Dutch with `CosyVoice2` given fine-tuning or if another TTS model can serve as an acceptable substitute. As new models continue to be developed, our team has hope that our pipeline will be expandable to be multilingual soon.

Multi-Speaker Conversations

Finally, our team would be curious to experiment with multi-speaker conversations where OpenVoiceAgent has to act as multiple different speakers within a dialogue. Our current pipeline is not equipped to act as multiple speakers because there is only one agent speaker prompt built in. However, future developers could expand OpenVoiceAgent to support multiple prompts with a unique voice for each, thus enabling OpenVoiceAgent to act as multiple roles and increasing the flexibility and usefulness of our PoC.

3
Validating Our Prototype

KLM's request to develop an unpredictable, locally hosted AI voice agent that can be used for trainings carries implicit expectations: the system must be convincing enough to substitute the human actors currently used in training. To know if the agent is perceived as good enough, it needs to be tested on humans.

3.1. Research Plan

3.1.1. Process

To ensure that the Human Research Ethics Committee (HREC) application could be approved within the project's limited timeframe, the scope of the validation study was quickly designed to involve interaction between participants and the AI, allowing for feedback on the agent's emotional maturity and perceived realism.

After extensive research, we found no established frameworks for evaluating perceived humanness in conversational agents, particularly those involving voice interaction. Therefore, we designed our own where we identify which dimensions of realism matter most for users and determine which technical factors influence that perception.

The Uncanny Valley Effect suggests that as artificial entities become more human-like, small imperfections in behavior or appearance can elicit discomfort rather than empathy (Mori et al., 2012). Recent studies, such as the work by Lomas (Lomas et al., 2025), argue that the boundaries of the uncanny valley are highly subjective and context-dependent. This motivated our decision to allow participants to use their own definition of human-like communication as a reference point for judgment, rather than imposing predefined measurement criteria.

While the evaluation could have been implemented as a straightforward perception test, for example asking participants to rate conversations on a simple humanness scale, this approach would not have generated meaningful insights for KLM's future development. Given that KLM frames this project as a technical exploration, the study was designed to provide more strategic value by identifying which specific aspects of the conversation most strongly shape the perception of humanness.

3.1.2. Research Questions

The primary validation question was:

1. Is OpenVoiceAgent perceived as human-like and realistic in conversational interaction?

   To ensure added value for KLM the following secondary questions were asked:
2. How does OpenVoiceAgent's perceived humanness compare to KLM's existing AI system, Caresse?
3. What parameters influence perceived humanness?

Answering the primary research question gives us a non-biased opinion if the agent is able to generate emotional responses and complete a conversation, as is defined in the problem statement in Section 1.2 The secondary research questions aim to give KLM insight into how they can work further. Comparing OpenVoiceAgent with Caresse contributes to KLMs ability to assess if they want to continue their exploration with a locally run AI, or explore other options. By exploring different parameters we start looking into what influences the perception of humanness, which is something KLM can continue researching later.

3.2. PoC tests

We first validated our prototype by performing prototype tests with KLM employees.

3.2.1. Participants

A total of nine participants were recruited through KLM's internal network. The sampling strategy was to represent both the intended users of the product and other relevant KLM staff who could provide their perspective on its potential use.

Four of the participants were members of the KLM Care Team, who are the primary target users of the Caresse training tool and the OpenVoiceAgent prototype. Their participation was essential, as they possessed first-hand experience in managing emotionally charged passenger interactions and could therefore evaluate the agent's realism and emotional adequacy from a professional standpoint.

An additional four participants were recruited from KLM's IT office. These participants did not have training in passenger-facing situations. However, they can still provide different perspectives with their backgrounds.

Finally, one participant was a KLM pilot, recruited through personal contacts of the research team. Including a pilot added a viewpoint from another professional group within KLM that frequently engages in procedural communication and crisis management scenarios.

Inclusion criteria

Participants had to be 18 years or older and fluent in English. Participants had little or no prior involvement in developing the Caresse or OpenVoiceAgent systems.

No formal screening questionnaire was used due to the limited timeframe for recruitment and testing. However, informal checks ensured that participants could be included.

All participants provided informed consent before participating in the study, and all recordings were anonymized prior to analysis to ensure confidentiality and compliance with ethical research standards.

Biases

Two of the participants had earlier experience with Caresse, since they were some of the initiators of the AI integration for Cabin Crew Training project. Two other participants stated they had seen other people use Caresse, but had never interacted with it themselves. In practicality, this means they potentially did not only use themselves as a human benchmark, but also Caresse.

3.2.2. Method

Research Design

The study was set up as a within-subject experimental design in which each participant interacted with different AI conversational agents. Nine versions in total were tested, eight of them OpenVoiceAgent and one of them Caresse. See Figure 3.1 for an overview.

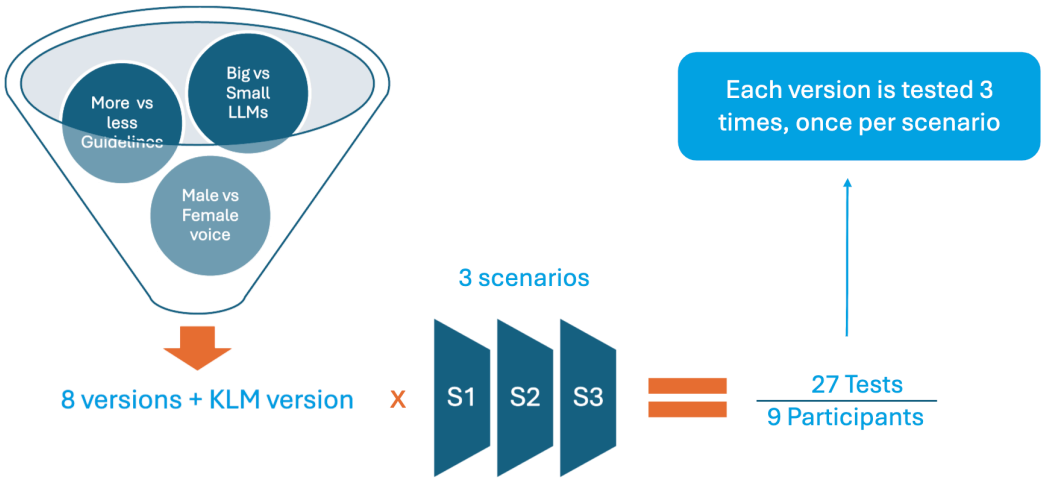


Figure 3.1: A visual overview of the validation research design displaying 8 OpenVoiceAgent versions and Caresse being tested across 3 scenarios with 9 participants.

The OpenVoiceAgent prototype included eight experimental versions, each defined by a unique combination of three binary parameters. Our independent variables were:

- **Prompt guidelines (more vs. less)**:
  Increasing the number of rules was expected to result in more predictable, consistent, and contextually safe responses, while fewer guidelines were expected to make the agent more spontaneous but occasionally incoherent.
- **Voice gender (female vs. male)**:
  Differences in the gender of the voice would automatically influence the tone, pitch and pacing. Gender-based expectations or biases may also shape perceived warmth or professionalism of the agent. We also wonder if the LLM used for the text-to-text could be biased by the variation in the callers name Mr. Johnson or Ms. Johnson. The different genders were created through voice cloning as explained in Section 2.1.2.
- **Model size (3B vs. 8B)**:
  The larger model was expected to produce more fluent and contextually appropriate responses but with slower latency, while the smaller model was expected to be faster but potentially less sophisticated in syntax and emotional phrasing and less faithful to the prompt given.

Each parameter was chosen to influence an aspect of perceived realism. See Table 3.1 for an overview of the eight OpenVoiceAgent versions.

Each participant completed three conversational scenarios with one agent per scenario, following a rotated order to minimize learning effects. See Table 3.2 for which participant tested what version in which scenario.

The scenarios depicted three scenarios where a relative of the caller was in a plane accident. It is the participants task to calm down the caller. Scenarios were always presented in the same progressive order to maintain a natural emotional build-up, preventing cognitive overload in early interactions:

Version	Prompt guidelines	Voice	LLM
V1	Less guidelines	Female	3B
V2	Less guidelines	Female	8B
V3	Less guidelines	Male	3B
V4	Less guidelines	Male	8B
V5	More guidelines	Female	3B
V6	More guidelines	Female	8B
V7	More guidelines	Male	3B
V8	More guidelines	Male	8B
KLM - Caresse	N/A	Male	N/A

Table 3.1: Parameter settings for each PoC test version.

Participant	Scenario 1	Scenario 2	Scenario 3
P1	KLM	V1	V2
P2	V1	V3	V4
P3	V2	V4	V5
P4	V3	V5	V6
P5	V4	V6	V7
P6	V5	V7	V8
P7	V6	V8	KLM
P8	V7	KLM	V1
P9	V8	V2	V3

Table 3.2: Test version distribution across nine participants in PoC tests.

- Scenario 1: No information available
- Scenario 2: Passenger confirmed on board
- Scenario 3: Passenger injured, transported to hospital

This set up of participant-agent rotation ensured that each OpenVoiceAgent version was tested three times across participants and scenarios. Meanwhile, Caresse appeared in three sessions to answer the research question “How does OpenVoiceAgent’s perceived humanness compare to KLM’s existing AI system, Caresse?”

Procedure

The testing was done at the KLM IT office. The setup included the facilitator, sitting in front of the participant, the note-taker, sitting diagonal of the participant, and the technical person, responsible for setting up the models and running them at the right time. See Figure 3.2 for a visual of our testing set up.

Each session began with a brief introduction and consent procedure, followed by a short pre-interaction questionnaire. The questionnaire gathered background data on the participant’s role at KLM, years of experience, English fluency (1-10 scale), and prior familiarity with AI voice assistants (e.g., Siri, Alexa, Google Home).

Participants were then guided through three scenarios. For each scenario, they received a printed scenario description and a short verbal briefing. The facilitator provided paper “call” and “hang-up” buttons, which participants used to start and stop the interaction. Participants had no other interface. See Figure 3.3a and 3.3b for visual examples.

Figure 3.2: A visual explanation of the testing setup of team members and participant placements.

(a) Session material

(b) Session in action

Figure 3.3: PoC testing session set up visual examples.

The participants could end the conversation at any moment if they desired, if not they were stopped after five minutes. After each interaction, participants completed a post-interaction questionnaire assessing perceived realism across five domains. Each domain were linked to

the earlier mentioned parameters and expected effect:

- Expected response (emotional and relevance)
- Voice realism (tone and pitch)
- Flow of conversation
- Word choice and phrasing
- Overall humanness

Each item was rated on a 1-7 Likert scale (as seen in Figure 3.3a, where 1 indicated not human at all and 7 indicated completely human. Participants were asked to elaborate verbally after each rating through semi-structured questions such as “What could have made this a 7?” or “Why did it sound robotic here?”.

After completing all three scenarios, a final comparative interview was conducted to elicit participants’ reflections across conditions. They were also asked to reflect about their own emotions using the Product Emotion Measurement Tool (PrEmo). PrEmo is a self-report instrument, adding a nonverbal emotional dimension to their responses. This tool can be used for mixed emotions and to make low intensity emotions explicit (Laurans & Desmet, 2017). Participants were asked to reflect about their emotion choice. At the end they were asked about their vision of a KLM care team training with such an AI voice agent.

Data Collection

The note-taker observed the participants, noting down e.g. body language and hesitation to supplement the responses from the participant. The note-taker also recorded their impression of the conversation (e.g. if there was friction or in which emotional tones the AI reacted) to contextualize the participants reflection after the session. Besides the conversational and observational data, the Likert scale ratings were recorded.

The conversations between the participant and the AI were transcribed automatically within the OpenVoiceAgent environment. Transcripts, survey data, voice recordings, and observational notes were anonymized and stored on TU Delft’s OneDrive in compliance with TU Delft’s data management guidelines.

3.3. Final Survey Tests

After the second development phase, we wanted to validate the final OpenVoiceAgent version the team handed over to KLM. In addition, the previous test had only 9 participants and, as later stated in Section 3.4.2, was not statistically significant. By testing the agents on a bigger user group, the quantitative results from the previous test could be compared to a bigger pool of opinions.

Caresse and three of the OpenVoiceAgent versions were tested in the following random order:

1. Version 4 (8B LLM, Male, Less guidelines)
2. Version 2 (8B LLM, Female, Less guidelines)
3. Caresse
4. Version 1 (3B LLM, Female, Less guidelines)

3.3.1. Survey Design

Respondents in the survey were asked to listen to a 1-1,5 minute long conversation between one of the team members and different AI agents, assuring that the conversation was long enough to include several passes back and forth. The videos were recorded with the same human calling, and each conversation started with the same opening line where the human asked ”Hello, who is speaking?”

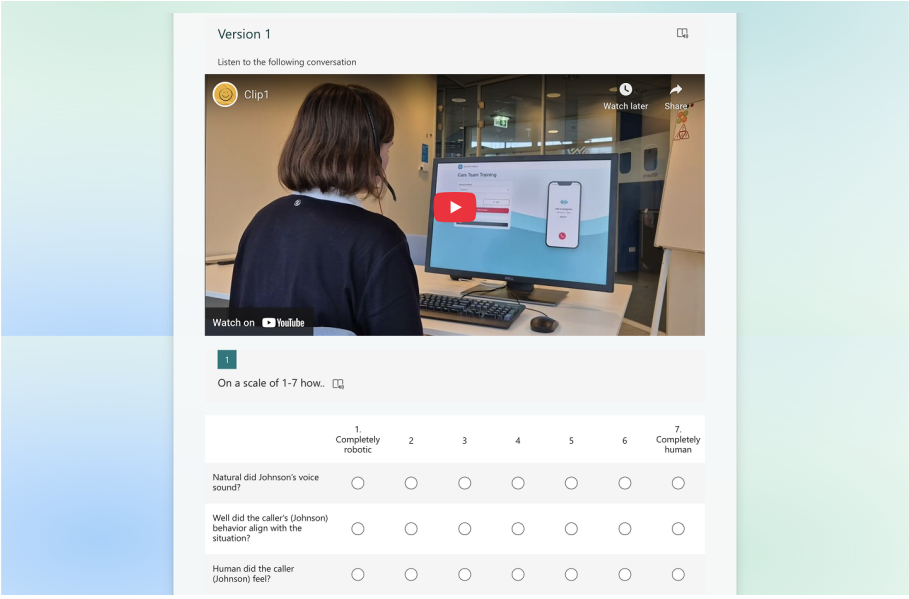


Figure 3.4: Screenshot of validation survey from Microsoft Forms completed by 78 participants.

After listening. participants were asked to answer the following three questions on a scale of 1-7, 1 being completely robotic and 7 being completely human.

1. How natural did Johnson's voice sound?
2. How well did the caller's (Johnson) behavior align with the situation?
3. How human did the caller (Johnson) feel?

The final question was "Do you have anything on your mind that you want to share about this AI agent, its applicability, or anything else? (optional)". See Figure 3.4 to see how the survey was set up, visually. Note that this test differs from the PoC test, not only by the amount of versions we were testing, but also the values we tested. Where the PoC test asked about expected response, voice realism, flow, word choice and overall humanness, in this test we are simply asking about the expected response, voice, and overall humanness. This is to ensure that the test remains short, since a high number of participants was desired.

Participants

The survey was sent out to participants from the previous test (with the request to share with KLM colleagues), students from the team members' respective faculties, friends, and family. No demographic data was collected, but given the network the survey was shared in, one can assume that a large portion of the respondents were students in their twenties. Over a six day period, a total of 78 responses were submitted.

3.4. Validation Results

We have introduced our two validation tests above, we will now review the results.

3.4.1. PoC tests: Data Analysis

The analysis combined a deductive approach with the quantitative analysis, and an inductive approach with the open questions. The intent was triangulation: to use qualitative findings to explain quantitative patterns and to use the descriptive statistics to prioritize which technical parameters merited follow-up testing to give solid advice to KLM.

Quantitative Analysis

The quantitative data described how participants rated each conversational agent version across the five dimensions, described in Section 3.2.2, on a seven-point Likert scale. These dimensions included whether the conversation content and emotions matched expectations, voice sounded realistic, conversation flowed naturally, agent's word choice was realistic, and conversation felt human overall. For conciseness, these dimensions are referred to as the expectedness score, voice score, flow score, word choice score, and humanness score, respectively.

To summarize participants' ratings across the five questionnaire items, we computed an overall perceived quality score by averaging the five scores. We justify this computation by testing the internal consistency between the five scores with a Cronbach's alpha reliability analysis. Cronbach's alpha evaluates how closely related a set of items are as a group; in other words, it tests whether the items measure the same underlying concept (Cronbach, 1951). A value of $\alpha = 0.86$ indicated strong internal consistency, as values above 0.7 are typically considered acceptable (Tavakol & Dennick, 2011), justifying our use of the averaged score as an overall measure of the system's perceived quality across the five dimensions.

To explore how the different PoC versions and values affected participants' ratings, mixed-effects linear models were fitted for each questionnaire outcome. Mixed-effects models are a flexible form of regression analysis that considers overall patterns (fixed affects) while accounting for individual differences between data points (random effects) simultaneously (Bates et al., 2015). This approach allowed us to compare scores across the different PoC versions and values while controlling for confounding biases between participants and scenarios.

For each question, the model estimated the average score for each PoC version or value while accounting for individual rating tendencies. Estimated marginal means and corresponding 95% confidence intervals were computed and plotted to visualize the uncertainty around each mean score. Pairwise differences between versions and values were tested using post-hoc comparisons of the estimated marginal means, with statistical significance established at a 95% confidence level ($p<0.05$). All quantitative analyses were completed in R (version 4.2.3) (R Core Team, 2024).

Qualitative Analysis

The goal of the qualitative analysis was to answer the *why* for the numerical data, while also uncovering topics and themes that the PoC test didn't explicitly test for.

To start, a quick cross-case analysis was done. The data that were originally sorted into notes per participant, were now sorted into insights per version. The highest and lowest ranked OpenVoiceAgent version, based on their overall perceived quality score, and Caresse were analyzed.

The main part of the analysis was conducted through a thematic analysis by affinity diagramming, in which themes are identified, coded, grouped, and clustered into higher-level patterns.

The difference between a cross-case analysis and a thematic analysis is minor but important. Cross-case examines the patterns per participant/version/question. In our case, we started with a within-case analysis per version, and then cross-case comparison to compare those analyses. Thematic analysis is used to find themes across the whole dataset. Therefore, the data is momentarily detached from the version and question.

The thematic analysis was approached in two ways:

1. **Facilitator-led manual thematic extraction**. Affinity diagramming is a method known for taking a lot of time to do manually. Because of time constraints and the rich contextual memory of the facilitator for a limited amount of participants, the facilitator skipped the detailed coding phase and moved directly to identifying recurring themes patterns. Once all data was summarized in patterns, they were sorted into themes. Prioritizing searching for

patterns that stood out while conducting the tests and when re-reading the transcripts afterwards ensured that contextual cues informed the themes created, while cutting down on manual work. The themes are discussed in Section 3.4.2.

2. **AI-assisted thematic extraction**. To mitigate individual bias and to expand analytic capacity, an additional team member also conducted a thematic analysis using an AI-assisted pipeline. The process followed the stepwise prompting approach described by (Naeem et al., 2025). The pipeline consisted of: cleaning transcripts, partitioning data into manageable chunks, generating code suggestions, clustering similar codes, and proposing thematic labels.

Limitations of AI-Assisted Thematic Extraction

Initially, the team considered using an AI tool to analyze the interview transcripts and notes in order to uncover subtle thematic patterns that might be difficult to detect manually. However, this approach proved to be less effective than expected. While the AI-generated thematic analysis confirmed some of the themes already identified through manual review, it also produced several inaccuracies and demonstrated significant limitations.

One of the main issues was related to the structure and quality of the data. The interview material was organized in Excel tables, and the recordings of the prototype interviews could not be transcribed in a format that the AI model could interpret reliably. Moreover, the methodological framework described by (Naeem et al., 2025), which inspired the prompts used in the analysis, was divided into six steps with suggested prompts that were often ambiguous and not directly applicable to the context of this project. Therefore, several adjustments were made to align the process with this study's objectives while remaining broadly consistent with the approach proposed in the referenced paper.

Suggestions to address these issues included collecting and formatting data with the specific goal of being compatible with AI prompting and designing custom prompts tailored to the unique context and objectives of the research.

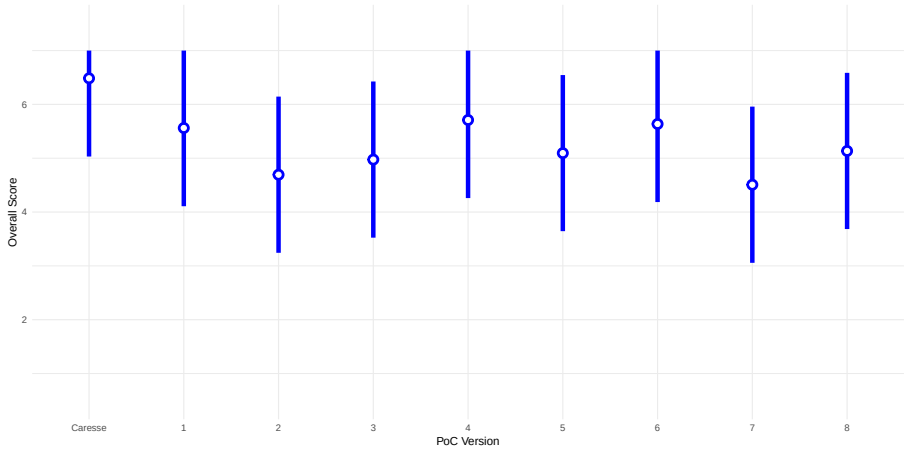
After both analyses were completed, the outcomes from each method were compared. Themes that appeared in both the manual and AI-assisted analyses were treated as high-confidence findings. Differences were discussed by revisiting the relevant transcript. This collaboration revealed a strength of the two approaches: the human analysis captured subtle contextual cues, while the AI analysis exposed less obvious linguistic regularities across sessions.

3.4.2. PoC Tests: Results

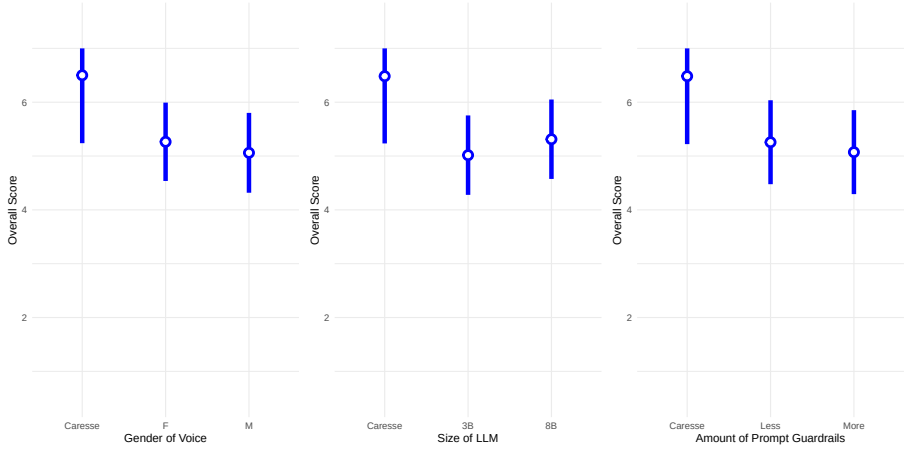
Quantitative

From our PoC tests with 9 participants, first, we examine our PoC versions. Figure 3.5a displays the estimated marginal means and 95% confidence intervals of the overall PoC quality metric, described in Section 3.4.1, for each of our 8 PoC versions and Caresse. Please see Figure A.1 in Appendix A.1 to see the means and confidence intervals for each of the five sub-scores used to compute the overall metric.

Our pairwise comparisons derived from our mixed-effects models showed no statistically significant differences between any of the versions. This indicates that the participants did not find our PoC significantly different from Caresse, quantitatively proving that OpenVoiceAgent can serve as a viable substitute to OpenAI's multi-billion dollar `gpt-realtime` model according to our nine KLM employees, an impressive result. Although, with more participants this could still change.



(a) Estimated least square means and 95% confidence intervals of the overall quality scores of the different PoC versions: 1 through 8 with Caresse as the control version.



(b) Estimated least square means and 95% confidence intervals of the overall quality scores of the different PoC values: gender of voice (female or male), size of the LLM (3B or 8B), and amount of prompt guardrails (less or more), with Caresse as the control.

Figure 3.5: PoC test results with nine participants: Estimated least square means and 95% confidence intervals of the overall quality scores, derived by averaging five quality sub-scores, from the nine PoC test participants. Statistically significant differences between means are labeled with '*' for $p<0.05$ and '**' for $p<0.01$. Confidence intervals are clipped to be between 1 and 7 because the five sub-scores were on 7-point Likert scales.

Next, we examine the impacts of our three values: gender of the voice, size of the LLM, and amount of prompt guardrails, on the overall quality score of the PoC (Figure 3.5b). Again, we see no statistically significant differences in our pairwise comparisons between our value levels. This indicates that the participants did not find our large LLM significantly different from our small LLM or Caresse, nor our female voice from our male voice or Caresse, nor our more guardrail prompt from our less guardrail prompt or Caresse.

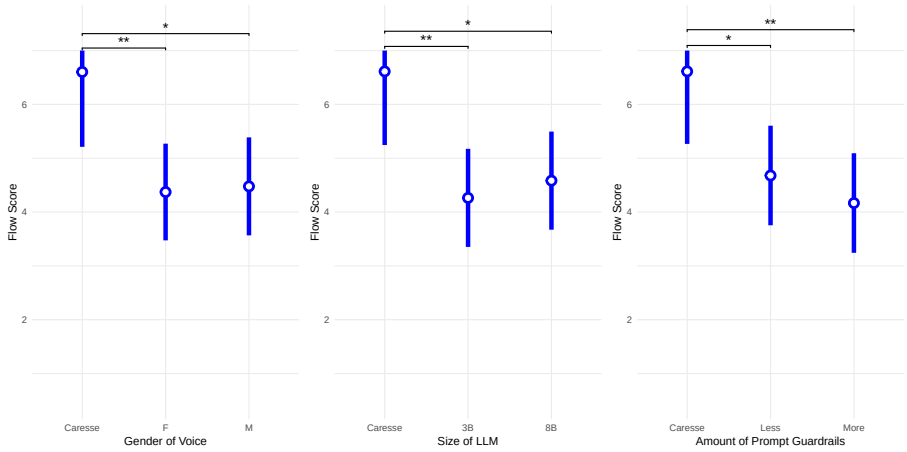


Figure 3.6: PoC test results with nine participants: Estimated least square means and 95% confidence intervals of the flow scores of the different PoC values: gender of voice (female or male), size of the LLM (3B or 8B), and amount of prompt guardrails (less or more), with Caresse as the control. Statistically significant differences between means are labeled with '*' for p<0.05 and '**' for p<0.01. Confidence intervals are clipped to be between 1 and 7 because the five sub-scores were on 7-point Likert scales.

Finally, we examine our three values for our five specific sub-metrics, including humanness, expectedness, voice, flow, and word choice, focusing only on significant associations. Figure 3.6 shows the estimated means and 95% confidence intervals of the flow score, described in Section 3.4.1, for each of our 8 PoC versions and Caresse.

Our pairwise comparisons derived from our mixed-effects models showed that there were statistically significant differences between the estimated flow score means between Caresse and the different value levels ($p_{Caresse,GenderF} = 0.009$, $p_{Caresse,GenderM} = 0.017$, $p_{Caresse,Size3B} = 0.006$, $p_{Caresse,Size8B} = 0.017$, $p_{Caresse,GuardrailsLess} = 0.023$, $p_{Caresse,GuardrailsMore} = 0.005$), likely because there were more data points for these values than the versions above. This indicates that the participants did find OpenVoiceAgent significantly different from Caresse in terms of how well the conversation flowed, motivating our second development phase latency work in Section 2.2.2. Beside the flow score, we found no other significant differences; see Figure A.2 in Appendix A.1 for our other sub-score figures.

Qualitative

To ensure that the numerical data aligned with what was said in the open questions we turned to the cross-case analysis as explained in Section 3.4.1.

When looking at what the participants said about the lowest ranked version 7 and the highest ranked version 4 there was little difference, which validates the numerical data. The main difference between OpenVoiceAgent and Caresse was the speed in which the agent responded. Caresse was also judged to be better at realistically displaying emotion. All the versions, including Caresse, were too polite and "perfect".

As explained in Section 3.4.1, in addition to the cross-case analysis a double Affinity Diagramming was conducted, once manually by the researcher and once with the aid of AI.

Human Thematic Analysis Results: The data was sorted into five final themes: Need for variation, Qualities that are amiss, Intelligently dumb, Feelings awakened, and Positives. See Figure 3.7 for a visual result of the second sorting. These themes were present in all versions, including Caresse, but to varying degrees. The latency and cadence were not an issue in Caresse. Besides these five main themes other insights from this test are discussed later in Section 5.1.3.

Figure 3.7: Themes created by resorting patterns in the thematic analysis through affinity diagramming

Need For Variation

> *"you would expect more fire, perhaps cursing" (P3)*

> *"It was not emotional enough because the content was very worried but the sound of the voice was not." (P9)*

Participants explained that the agent's emotional range was too limited. It stayed in a single mood, often angry, and didn't shift enough with the content of the conversation. This made the interaction feel flat. The few times it did change emotion, it was too abrupt, which was perceived as equally strange. In addition, the agent was perceived as too polite. In a real crisis situation participants explained that they would expect the human caller to be way more active instead of patiently waiting to hear what KLM has to say.

> *"I would expect him to be more anxious for information" (P9)*

Another pattern was that participants felt the caller was too narrow minded. It was searching for a specific type of information, responding as if focused on reaching the end goal rather than reacting as a person might, who would pursue multiple emotional threads at once. Once the agent reached its goal, it was described as backing down too soon, as if too satisfied.

Additionally, there were some requests for more variation. For example, letting the agent interrupt the user.

Qualities That Were Amiss

> *"It has well-structured sentences, which is unrealistic if you're emotional – too perfect." (P1)*

> *"The voice sounded like a real woman, but the cadence was really off, you can tell it's not a real person, the things she said were fine but the rhythm of the sentences was weird" (P1)*

This theme described the nuances that can feel uncanny when different from an actual human. The AI agent hesitated too little, it had few filler words, no stuttering or stammering. These are called linguistic disfluencies. As seen in one of the quotes, the rhythm of speech was off. An example that happened several times with different participants is the pronunciation of "khaahnyou" instead of "can you".

> *"The pauses in between the phrases and in between me and her talking made it robotic." (P7)*

The main factor in this theme was the speed in which the agent responded and the time it took to generate a response. It made people doubt if the agent had heard them, if the agent was finished talking, and if it was the participant's turn to talk now. Several participants stated that this was the main issue effecting the perceived humanness.

"It didn't bend with the emotion." (P7)

Where humans would normally feel acknowledged when you express their emotions and sympathize, some of the participants stated that the AI would do the opposite. They would get more upset if you tried to sympathize.

Finally, some participants noted logical breakdowns, when the agent suddenly shifted to self-narration or third-person speech which breake immersion.

Intelligently Dumb

"Like talking to a wall, I had a feeling she wasn't really listening." (P4)

"It doesn't feel human because there is no two-way conversation." (P6)

Even though the agent could phrase its sentences in a sophisticated manner, the participants noted that it was only surface level, because there was no real socially intelligent being behind the words. Participants stated that the agent didn't listen to what they were saying. It seemed like the agent only took one piece of the participants response and continued with that small piece of information. This is also related to the first theme, since the multiple emotional threads were missing.

Additionally, the agent's brain was perceived to remember everything in an unnatural way. In one example, the agent remembered details too perfectly, offering a phone number by request only after a long unrelated monologue. The participant stated this is something a real person would never do.

Feelings Awakened

"Bastard." (P8)

"She is feisty." (P4)

"Karen vibes." (P7)

These are all terms used to describe different agent versions. Despite the flaws mention earlier, the agent managed to evoke strong reactions. Participants expressed reluctance like "I'm not looking forward to talking to Mr. Johnson again" (P8). This shows that the interaction engaged users emotionally, even when those emotions were frustration. There were also cases where participants told stories about their own experiences as being a care team member, and how this made them think back to those moments.

The observers noted that participants became physically involved: using hand gestures, fidgeting with a pen or paper, and leaning forward or back as they thought. The conversation seemed to draw them in, suggesting a level of social presence which one would only expect in conversations perceived to be somewhat human.

"De-escalating is difficult" (P8)

Participant later told, through a prompted question, that some felt pride when managing to calm down the agent, and some a strong willingness to help. Some also felt little control and power over these agents, which by some participants was stated as possible, but was mainly seen as a non-human attribute, tying back to the theme Intelligently Dumb.

Positives

> *"The accent makes it realistic, it gives personality to someone." (P6)*
>
> *"I couldn't hear it was a computer." (P8)*

Participants still recognized potential in the system. Many appreciated the natural-sounding voices and even assigned personalities to them, greatly influences by the nationality of the voice. Some stated that it "sounded more like an actor and not a real person, a human actor, not a robot actor" (P2). Another pattern was that the wording and phrasing was generally seen as human.

Gender differences also influenced expectations for some participants: the male voice was expected to remain calm, while the female one was expected to be expressive and emotional. Participants valued having both options for training purposes, especially because it created a sense of safety for certain scenarios.

> *"She [the agent] delivered on something I said wrong. "Fuck!", I heard myself saying something wrong, and there she goes as expected" (P5)*

When the emotional timing did align, the experience felt authentic and educational. The system could effectively escalate or mirror emotion in realistic ways, hinting at its potential as a training tool once its listening and pacing issues are addressed.

Finally, the benefits of using AI in training were clear. Participants state that the prototype is useful for preparing the "What should I say?". Some participants stated that for training, it is important to have both genders to practice with. This is because personal experiences can make it easier or harder to deal with certain genders when facing conflict. The prototype gives a safe environment to train with both genders, which usually would not have been possible with an actor. Participants highlight several different use cases, which are described in Section 4.3

AI-Assisted Thematic Analysis

As previously stated to ensure that the analysis by the facilitator is not biased, a quality control was done by another team member using ChatGPT-4.5 according to a step by step process (Naeem et al., 2025).

Limitations: The outcomes of the AI-based thematic extraction reflected both the potential and the limitations of applying LLMs to qualitative data analysis. The results revealed technical issues that reduced the reliability of the generated themes. One major limitation was that the data were not sufficiently cleaned or codified before being processed. Notes from the interviews were not consistently categorized, which caused the AI to misinterpret comments referring to the participants as comments referring to the AI agent itself. Furthermore, during the step of code identification within the transcripts, the prompts used were long and complex, leading the AI to reduce its output and produce only a limited number of codes that satisfied the prompt's requirements rather than the project's research objectives. As a result, the AI occasionally arrived at false conclusions by confusing reflections on the experiment or participant behavior with evaluations of the system being tested.

AI analysis results: Despite these challenges, the AI thematic extraction produced four distinct themes that provided partial insight into the collected data:

1. Theme 1: Conversational Flow and Timing
2. Theme 2: Emotional Flatness and Over-Politeness
3. Theme 3: Repetition and Technical Breakdown
4. Theme 4: Moments of Human Connection

The first two themes overlapped closely with those identified through manual thematic analysis, providing some reassurance regarding the validity of the findings and showing that AI can help confirm existing insights. However, the last two themes were primarily the result of misinterpretation caused by unstructured data and poorly adapted prompting. Overall, while the AI-supported approach showed promise as a supplementary tool for thematic analysis, it should not be trusted as the primary tool of thematic extraction for our project.

Quantitative & Qualitative Combined

All the data from the PoC validation tests shows that there is no mention-worthy difference between the different OpenVoiceAgents, which is supported both by the numeric data and the cross-case analysis. In addition, the overall insights through the affinity diagramming method shows that the themes were present across all versions. Personal preference had a bigger effect on the overall score than the three parameters this test tried to pinpoint.

For there to be a clear difference between the different parameters with a small participant group, there would have needed to be an obvious perceivable difference. Even the team members, who have been part of each step in developing the agent, could not guess correctly if the LLMs were running with 3B or 8B, or if they had less or more prompt guidelines.

While the numeric data shows the slightest (not significant) preference for the female voice, the qualitative data was clear in a preference for the female voice. She was perceived as feistier and fuller of character, her emotions were also perceived as stronger. Since this gender preference didn't influence the final score, we can assume that the gender of the voice doesn't influence perceived humanness, but that this was rather an isolated preference.

While there are clear improvement areas for the agent before it is ready to be used, the overall feedback was positive and the potential for this tool was clear.

> *"This is beyond my expectation, I've never heard from an AI before. I look forward to seeing where this goes" (P8)*

One of the participants who had used Caresse before conducting the test even had a preference for OpenVoiceAgent:

> *"What you made is a lot better than what they [KLM] did." (P5)*

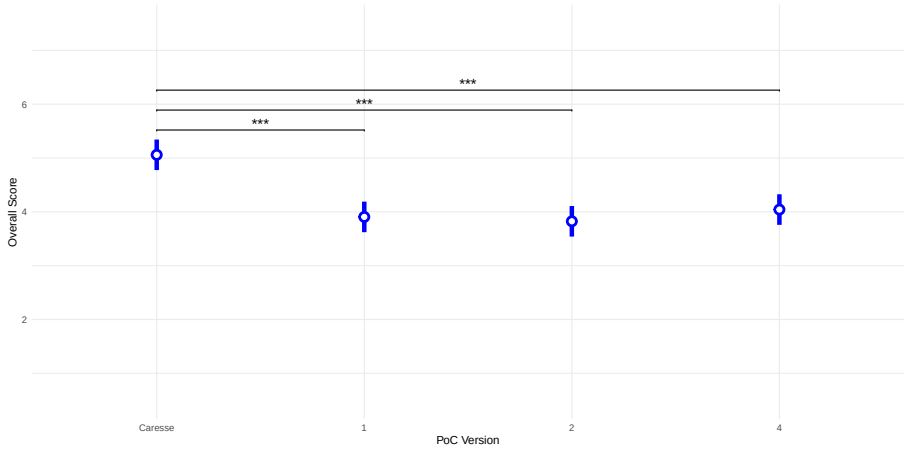
We now validate OpenVoiceAgent with a larger pool of respondents through a survey.

3.4.3. Survey: Data Analysis

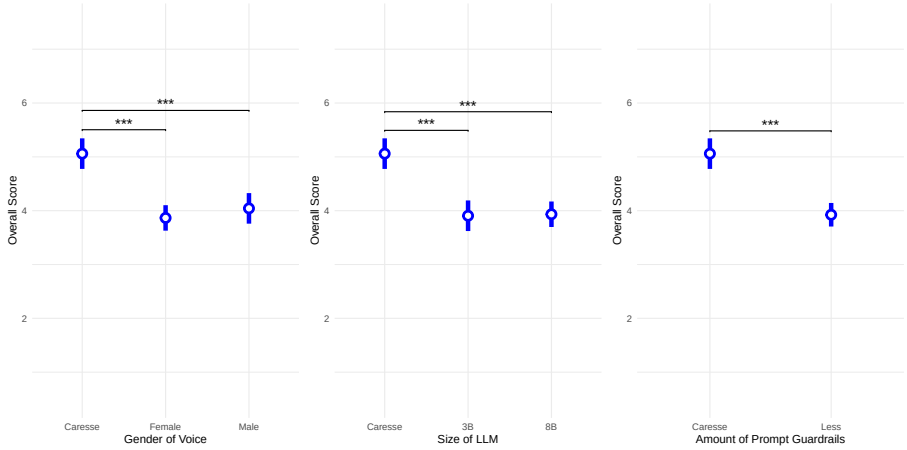
The survey was analyzed in the same manner as the PoC test with a mixed-effect model, see Section 3.4.1. Minor adjustments to the script had to be made since the survey tested three versions instead of nine, and three values instead of five. The survey also only depicted scenario two instead of all three.

3.4.4. Survey: Results

From our survey with 78 respondents, first, we examine our PoC versions. Figure 3.8a displays the estimated means and 95% confidence intervals of the overall PoC quality metric, described in Section 3.4.1, for each of our 3 survey PoC versions and Caresse. Please see Figure A.3 in Appendix A.1 to see the means and confidence intervals for each of the five sub-scores used to compute the overall metric. Our pairwise comparisons derived from our mixed-effects models showed that there were extreme statistically significant differences between Caresse and each of the PoC versions tested ($p_{Caresse,V1} < 0.001$, $p_{Caresse,V2} < 0.0017$, $p_{Caresse,V4} < 0.001$). This indicates that the participants did find our PoC significantly different from Caresse, quantitatively proving that OpenVoiceAgent is perceived as worse than OpenAI's multi-billion dollar `gpt-realtime` model, according to our 78 survey respondents.



(a) Estimated least square means and 95% confidence intervals of the overall quality scores of the different PoC versions: 1 through 8 with Caresse as the control version.



(b) Estimated least square means and 95% confidence intervals of the overall quality scores of the different PoC values: gender of voice (female or male), size of the LLM (3B or 8B), and amount of prompt guardrails (less or more), with Caresse as the control.

Figure 3.8: Survey results with 78 participants: Estimated least square means and 95% confidence intervals of the overall quality scores, derived by averaging five quality sub-scores, from the 78 survey participants. Statistically significant differences between means are labeled with '*' for $p<0.05$ and '**' for $p<0.01$. Confidence intervals are clipped to be between 1 and 7 because the three sub-scores were on 7-point Likert scales.

Next, we examine the impacts of our three values, gender of the voice, size of the LLM, and amount of prompt guardrails, on the overall quality score of the PoC (Figure 3.8b). Please see Figure A.3 in Appendix A.1 to see the means and confidence intervals for each of the five sub-scores used to compute the overall metric.

Again, we see extreme statistically significant differences in our pairwise comparisons between Caresse and our value levels ($p_{Caresse,GenderF} < 0.001, p_{Caresse,GenderM} < 0.001, p_{Caresse,Size3B} < 0.001, p_{Caresse,Size8B} < 0.001, p_{Caresse,GuardrailsLess} < 0.001$). This indicates that the participants did find significant differences between our PoC and Caresse. However, there were no statistically significant differences within the different versions of our PoC, indicating that our large LLM was not significantly different from our small LLM, nor our female voice significantly different from our male voice. We were not able to test the difference between more and less guardrails from the survey.

In addition to the scores, 23 comments were left on the final open question. The majority of the comments concern the speed of the responses, and the speed of which the agent talks. Other topics that were brought up all matched the same findings in the Affinity Diagram results in Section 3.4.1.

> *"One of the biggest factors in the voice feeling human is the waiting time before it answers. If a human receives news about a potential accident of a family member they would probably speak over the representative more, and be less polite in general." (Survey Respondent)*

Survey Insights

The survey delivered statistically significant results, indicating that OpenVoiceAgent was perceived less human than Caresse. Given established effectiveness of the existing alternative, and the limits in both resources and time in the development of OpenVoiceAgent, this outcome aligns with our expectations. The intention of the survey was to provide greater reliability and confidence in the findings, something the survey was able to do with 78 participants over the nine from the earlier test.

The survey showed that even with a bigger participant group OpenVoiceAgent is still perceived as somewhat human with a mean overall score of about 4. This score can be translated into "neutral" or "somewhat robotic and somewhat human". This means the agent was not perceived as completely robotic. Even though the scale is subjective, this indicates that KLM is on the right track in exploring their options with AI as a substitute or replacement of human actors, even when it is built locally.

When comparing the results of the survey and the PoC test in how the different versions are ranked in their overall score, the results are similar. V4 is scored the highest in the survey, V1 follows, and lastly V2. The same versions follow the same ranking in the PoC test. Even though the PoC test results should be interpreted with caution due to the limited number of participants, the two results point in the same direction, validating the PoC results more.

3.4.5. Validation Summary

With several different inputs, both quantitative and qualitative, from in-depth conversation to a quick survey, now we return to answer the three research questions.

1. Is OpenVoiceAgent perceived human-like and realistic in conversational interaction? Overall, OpenVoiceAgent is perceived as somewhere between robotic and human. On a scale of 1-7 it lands on the midpoint. The agent sparked responses both in body language and outspoken emotions. However, there were some aspects of the AI that broke the illusion of the agent being human, those were the following: The pauses were too long between turns

in the conversation, the agent did not seem to listen and would adjust their responses only to parts of what the participant said, the emotion was too one-toned, the way of speaking was “too perfect”, participants state they were missing linguistic disfluencies and colorful language, lastly the cadence. Other aspects were stated as believably human, those were: the quality of the voice (especially the female one). the word choice, and way of phrasing. The conversations were realistic enough that people could look past the realism disruptors to see the value this could bring for training and even other use cases. To conclude: OpenVoiceAgent produced social presence without genuine human believability, but with enough realism to spark a vision of future use.

2. How does OpenVoiceAgent’s perceived humanness compare to KLM’s existing AI system, Caresse?
   While the PoC tests are not statistically significant, where all versions can be seen as equal, Caresse does lean towards a higher result. Only on the parameter of “Flow” Caresse scored higher than any of the OpenVoiceAgent versions. This is also evident from the qualitative data, where the main issue of long pauses is not present for Caresse. The survey proved, with a statistically significant result, that Caresse is rated higher on perceived humanness than OpenVoiceAgent.
   This is not a response that surprises us, as both resources and time were limited for this project. We are, after all, competing with a multi-billion dollar company. We are however impressed by the little difference between Caresse and OpenVoiceAgent, proving the potential of the locally built agent over the OpenAI version.

3. What parameters influence perceived humanness?
   The study tested the following three parameters: gender of the voice, LLM size, and the guardrail prompts. None of these parameters were found to have a statistically significant influence on the values that were measured: expected response, voice, flow of conversation, word choice/phrasing, and overall humanness.
   While the results were not statistically significant, the tests did highlight some attributes that were missing or need to be removed from the agents. These are useful insights regardless of whether KLM continues with Caresse or OpenVoiceAgent.

Conclusion

OpenVoiceAgent shows realism to a certain extent, but remains limited by its polished, overly consistent delivery and mechanical timing. While it approaches human-like interaction, it does not yet embody it. Compared to Caresse, it is perceived as slower and emotionally monotone. To appear more human, future iterations must intentionally design for conversational imperfection, emotional variability, and social intelligence rather than rely on model size or gendered voice differences.

4

Implementing our Prototype at KLM

Building on the iterative development and validation of our conversational agent PoC, this chapter focuses on the practical implementation of the prototype within KLM's organizational context. To understand how the prototype could be effectively adopted, we engaged with key stakeholders and gathered insights into KLM's current approach to AI integration and digital training. Furthermore, we examined potential use cases that align with the company's training objectives, technological infrastructure, and organizational readiness. Together, these analyses provide a foundation for outlining a realistic and responsible implementation strategy for the prototype.

4.1. Implementation Analysis

Drawing on organizational precedented strategy and strategic considerations, we outline a structured approach to facilitate effective integration and adoption.

4.1.1. Precedent and Validation

As discussed in Chapter 1, KLM currently relies on conventional cabin crew training methods, including live instructors, physical mock-ups, and professional actors to simulate passenger interactions during routine and emergency scenarios. While effective, these methods face constraints related to logistical complexity, high operational costs, and limited scalability. Each session requires significant preparation and coordination, and the range of scenarios is often limited by the availability and performance of actors.

The KLM Cityhopper VR (KLC VR) tool, shown in Figure 4.1, provides a validated precedent for integrating new innovative technologies into aviation training. Initially deployed to support cockpit procedural training for the Embraer aircraft, KLC VR demonstrates that immersive VR can deliver training objectives efficiently while reducing dependency on physical simulators and scheduled instructor sessions (KLM, 2023a). KLM has also explored VR for cabin crew through initiatives like the Virtual Vitality Program, offering trainees immersive exposure to realistic cabin scenarios (KLM, 2023b). These experiences create a foundation of organizational knowledge and approval that enables the extension of VR into cabin interaction training with AI-driven unpredictability.

Research supports this approach, showing that XR and mixed reality technologies improve flight attendant proficiency, especially in procedural tasks and emergency management (Federal Aviation Administration (FAA), 2024; ScienceDirect, 2023). The addition of AI-driven scenario unpredictability further enhances learning outcomes by simulating human unpredictability, something that traditional methods cannot replicate efficiently (MDPI Electronics, 2023).

Figure 4.1: Air France-KLM immersive training solution | Meta for work

Before any transition, KLM conducts a feasibility study to evaluate the readiness and suitability of VR and Unpredictable Local AI for cabin crew training. This study examines operational requirements, technology readiness, cost implications, and potential learning outcomes compared to existing methods. Only when the feasibility study gives a positive recommendation does the company proceed with planning the switch (Maloney, personal communication, September 29, 2025).

4.1.2. Our Speculations

KLM's transition strategy aligns with practical and contractual considerations. Once feasibility is confirmed, the company waits for existing contracts with instructors, training facilities, and actors to expire. Upon contract completion, the training method switches directly from traditional approaches to VR+AI, minimizing overlap costs and operational disruption. Continuous assessment of learning outcomes, trainee confidence, and scenario realism guides iterative improvements in both VR realism and AI behavior. Once the VR+AI system reliably replicates realistic passenger interactions, live actors are retained only for tactile exercises and high-touch procedures.

This contractual transition strategy, provides multiple strategic advantages. Operational efficiency increases as reliance on external actors and cabin mock-ups declines. Training becomes scalable and flexible, enabling more frequent sessions across multiple locations. Exposure to unpredictable scenarios enhances skill retention, decision-making, improved communication under pressure, and enhances stress tolerance. Finally, by building on the validated KLC VR framework, the strategy ensures organizational confidence and continuity of training excellence.

This strategy positions KLM to transition from traditional actor-based cabin crew training to a hybrid VR+AI model, leveraging proven VR practices from the KLC tool while addressing previously unmet needs for dynamic, unpredictable passenger interaction training. By building this strategy on a precedented implementation by KLC, KLM can achieve a realistic and scalable training solution that aligns with industry best practices.

The contractual transition approach (grounded in feasibility evaluation and contract timing) ensures that innovation does not compromise training quality. Section 5 further examines the costs, benefits, and risks for implementing OpenVoiceAgent at KLM.

4.2. AI Adoption Interviews

Expanding on the technical and organizational feasibility reviewed above, we conducted a series of interviews on the adoption of artificial intelligence to understand how KLM employees perceive such innovations in practice. The interviews were conducted at KLM, see Figure 4.2 Successful implementation of our prototype requires technological readiness, as well as cultural and human acceptance within the organization.

To explore these dimensions, three semi-structured interviews were carried out with a cabin crew member, a cabin crew manager, and a management trainee with a technological background. The aim of these interviews was to capture diverse perspectives on how AI and XR are currently perceived as well as how they are envisioned within KLM's operations. These interviews focused on practical experiences with existing tools, as well as participants' views on privacy, trust, and realism. Overall, while the interviews did not contain any critical data for the PoC, they were vital in helping us sense the pulse on how KLM's personnel perceives these technologies. We will now discuss general themes seen throughout the interviews. For a thematic comparison of individual interviewees, please see Table 4.1.

4.2.1. Limitations of Current Training

Across the interviews, interviewees often pointed out weaknesses of current cabin crew training, especially the roleplay sessions with actors. Even though these exercises are valuable for practicing procedures, almost all respondents explicitly noted that the exercises feel predictable and that they could sometimes see through the acting of the role-players. On the other hand, virtual reality was praised for its immersion and how it helped learners "see the feel" the situation rather than imagine it.

4.2.2. Trust and Privacy

Privacy and safety emerged as recurring concerns. The manager and flight attendant both expressed a strong preference for locally hosted AI systems over cloud-based ones, linking them to feelings of safety and control. They felt that trainees and employees would express themselves more freely if they knew their data stayed within KLM. The management trainee confirmed this sentiment from a technical and organizational standpoint, stating that local AI aligns with KLM's strategic shift toward secure, internal innovation. However, a participant made a point that most people won't understand the difference between local and non-local AI systems. Collectively, this theme reflects that trust and transparency are important in the adoption of new technologies but so is the way they are introduced and explained.

4.2.3. Practical Applications

All interviewees reported using AI tools in their daily work or personal life. They showed a relatively advanced stage of digital familiarity. Typical uses included summarizing documents, drafting and rephrasing communication, translating text, and supporting project management tasks. Interestingly a participant also described using AI for reflection and emotional articulation, suggesting that AI is a way broader tool, not just a productivity enhancer. All in all, AI was viewed

as a supportive partner that enhances human productivity and expression.

4.2.4. Organizational Readiness and Culture

Despite overall optimism, interviewees acknowledged existing resistance within parts of the KLM organization. This hesitation was attributed more to knowledge gaps, particularly among older or less digitally oriented staff. One interviewee pointed out that misconceptions like equating automation with artificial intelligence, still create confusion about AI. Another interviewee emphasized the need for internal education and transparent communication about AI's role and limitations. Therefore, the interviewees believe that successful AI adoption at KLM will depend a lot on cultural transformation that is based on trust and inclusion.

4.2.5. Attitude and Outlook

When asked about the future of AI in aviation, interviewees consistently envisioned it as an enabler of safer and more efficient operations rather than a replacement for human expertise. Respondents saw strong potential in voice-based and offline systems integrated into existing tools such as crew iPads, emphasizing accessibility even without internet connectivity. At the same time, they underscored that empathy and human judgment remain irreplaceable in KLM's operations. Collectively, their reflections reveal a workforce that welcomes innovation but values technologies that complement and not overshadow the human factor of the workforce.

Figure 4.2: The adoption interviews were conducted at KLM inflight services building 107.

Theme	Cabin Crew Member	Cabin Crew Manager	Management Trainee (AI Background)
Authenticity in Training	Finds current actor-based training predictable. Believes AI and VR can make reactions more genuine	Says trainees can "see through" actors. Supports AI avatars and VR for realistic simulations.	Sees AI and VR as tools for personalization and engagement but emphasizes empathy must remain human-led.
Trust and Privacy	Prefers locally hosted AI for training and communication. Believes that privacy increases if data stay within KLM.	Strongly favors local AI for privacy and safety. They believe it creates a safer learning experience	Acknowledges privacy benefits of local AI but notes the scalability and stability of cloud tools, suggests hybrid models can be an option.
Practical Applications	Suggests AI for searching manuals, checking catering inventory on a flight, and personalizing passenger service. Also adds that an offline assistant would be ideal.	Proposes AI support for medical incidents and faster manual searches via onboard iPads.	Highlights automation, data-driven optimization, and voice bots as practical uses for efficiency and consistency.
Organizational Readiness and Culture	Views AI as normal in daily work but sees hesitation among peers.	Open to innovation but aware some colleagues resist digital tools; emphasizes awareness and upskilling.	Identifies resistance and knowledge gaps as main barriers. Advocates education and early involvement of key users.
Attitude and Outlook	Describes AI as positive, empowering, and time-saving; sees it as a helpful assistant.	Associates AI with curiosity and uses ChatGPT for reflection and clearer communication.	Expresses curiosity and optimism but warns against overreliance.

Table 4.1: Thematic Comparison of interview insights on AI adoption at KLM with three employees.

4.2.6. What These Interviews Meant For Us

In general, these interviews motivated our work: the interviewees found limitations in existing training methods, showed a clear preference for local AI, and would be open to innovation given inclusive and comprehensive education. We gained ideas and insights into the thoughts of KLM's employees for a more holistic view of the implementation of AI in KLM along with some forward-thinking ideas.

4.3. Potential Use-Cases of OpenVoiceAgent

Throughout this report, we have discussed and validated the technical capabilities of OpenVoiceAgent in the context of human-centered trainings (Section 1.1.1). However, a local conversational agent capable of holding emotional dialogue can be used in many more situations with widespread impact. Below we outline several potential use-cases, first focusing on applications within KLM, and later discussing broader applications. Insights from the PoC tests and survey results in Chapter 3 helped to inspire these use-cases. It must be mentioned that we have not rigorously tested OpenVoiceAgent for its performance in these scenarios, we are merely speculating about its future potential.

4.3.1. Within KLM

Crisis Management Trainings

Section 3.2.2 introduced an important use-case for OpenVoiceAgent: crisis management trainings, where KLM employees must learn to de-escalate emotional situations under pressure. In Section 3.2.2, this meant having to provide emotional information to family members worried for their brother's safety after an aviation accident. Because of OpenVoiceAgent's flexibility, many other crisis scenarios that require emotional and conversational intelligence as well as composure. These can be grouped into the following categories:

1. **Passenger-related emergencies**: training cabin and ground staff to de-escalate high-stress passenger interactions
    - Managing unruly or intoxicated passengers who resist instructions or behave aggressively.
    - Calming anxious or first-time flyers through reassurance and confident communication.
    - Handling in-flight medical emergencies, coordinating between passengers, crew, and medical professionals on the ground.
2. **Operational & flight disruptions**: helping staff handle emotionally charged conversations during service breakdowns.
    - Communicating with passengers stranded due to weather events, technical faults, or global crises such as wars or pandemics.
    - Delivering difficult messages during extended delays, cancellations, or rebookings.
    - Coordinating internally across ground operations, flight crew, and control centers when unexpected disruptions occur.
3. **Post-incident and emergency response**: preparing employees for sensitive communication after serious incidents.
    - Speaking with family members or next of kin in the aftermath of an aviation event.
    - Practicing calm, structured communication with first responders such as police, fire, or medical services.
    - Managing the media after a serious incident.

These examples demonstrate how OpenVoiceAgent can help KLM staff rehearse a large variety of emotionally complex interactions in realistic, low-risk simulations.

Other Trainings

Beyond crisis responses, OpenVoiceAgent can also aid in the practice of many other emotionally and communicatively complex scenarios central to KLM's daily operations. These can be grouped in the following categories:

1. **Customer service & communication quality**: training staff to deliver consistent, empathetic, and professional service across all customer touch-points.
    - Handling complaints and service recovery with calmness after delays, cancellations, or lost baggage.
    - Practicing phrasing when speaking to passengers of different backgrounds or emotional

states.
    - Responding to customer feedback or difficult reviews in a constructive and brand-consistent manner.
2. **Interdepartmental coordination & leadership communication**: developing teamwork, leadership, and communication skills in internal KLM contexts.
    - Simulating cross-departmental coordination between ground crew, flight crew, and operations control.
    - Training leaders in constructive feedback delivery, conflict mediation, and motivational communication.
    - Supporting team briefings and debriefings, helping supervisors communicate clearly with their teams.
3. **Diversity, equity, & inclusion scenarios**: building cultural awareness and inclusive communication across KLM's international workforce.
    - Engaging in simulated conversations involving cultural misunderstandings or differing communication styles.
    - Addressing language barriers or varying comfort levels with authority and hierarchy.
    - Practicing sensitivity in gender-related or accessibility-related interactions, ensuring respect in all contexts.
4. **Well-being & mental health support**: encouraging open, psychologically safe communication around stress, fatigue, and workload.
    - Rehearsing conversations about stress, burnout, or fatigue as a non-judgmental listener or speaker, acting as a form of exposure therapy to help employees gradually build or regain confidence.
    - Helping managers respond appropriately to signs of employee distress.
    - Integrating OpenVoiceAgent into peer-support programs, creating space for non-judgmental reflection and dialogue.

These additional scenarios demonstrate that OpenVoiceAgent is not limited to crisis management contexts. Its emotional adaptability, conversational realism, and scalability make it a versatile platform for improving communication and resilience across all levels of KLM's organization.

4.3.2. Beyond KLM

While OpenVoiceAgent was developed and validated in the aviation context, its conversational and emotional adaptability make it valuable in many other professional, private, and educational settings. The same capacity to simulation emotionally charged or high-stakes dialogue can be used for general communication and resilience training across a wide range of domains.

General Professional Communication

OpenVoiceAgent can serve as a personal or institutional tool for improving communication skills in emotionally complex or high pressure environments. Potential applications include:

1. Interview & negotiation practice: simulating job interviews, salary negotiations, or performance reviews where confidence and phrasing strongly influence outcomes.
2. Conflict management & mediation training: helping employees, leaders, or students practice resolving disagreements with empathy and composure.
3. Adversity & resilience training: allowing individuals to rehearse responses to emotionally stressful or confrontational scenarios in a safe, private, and repeatable setting.
4. Public-facing communication: preparing spokespeople or service representatives to handle complaints, criticism, or crisis-related questions with professionalism and composure.

Domain-Specific Applications

Beyond general workplace communication, OpenVoiceAgent can be tailored for domain-specific emotional dialogue training in professions where communication is critical:

1. Healthcare & veterinary medicine: practicing conversations around breaking bad news, discussing treatments, or managing anxious clients and patients.
2. Law & legal advocacy: rehearsing courtroom questioning, client consultations, or sensitive discussions involving trauma or confrontation.
3. Education & coaching: supporting teachers, trainers, or mentors in providing constructive feedback, handling difficult students, or navigating emotional classroom moments.
4. Emergency services & counseling: training responders or therapists to manage emotionally intense interactions while maintaining calmness and clarity.

Conclusion

In essence, OpenVoiceAgent extends beyond the aviation industry as a universal platform for emotional communication training, helping professionals across fields build empathy, confidence, and verbal resilience through realistic, emotionally grounded dialogue practice.

5

Value Created By Our Prototype

The previous chapters outlined the technical and organizational pathway for implementing OpenVoiceAgent. We now focus on the proposed OpenVoiceAgent-driven VR+AI crisis management training system. This chapter explains the ”why” behind the adoption of such a training system.

Using the Triple Bottom Line framework (Profit, People, Planet) (Harvard Business School, 2020), this chapter evaluates the broader value created by the system. Beyond its technical feasibility, the VR+AI prototype supports operational scalability (People), financial efficiency (Profit), and environmental sustainability (Planet). By analyzing economic, societal, and environmental benefits, this chapter provides a justification for implementation within KLM's training ecosystem.

5.1. Profit, People, Planet: An Overview

Before addressing each dimension in detail, the following section connects the prototype to the United Nations Sustainable Development Goals (SDGs).

5.1.1. Sustainable Development Goals Addressed

In addition to the Triple Bottom Line framework, the VR+AI training prototype contributes directly to the United Nations Sustainable Development Goals (SDGs), specifically SDG 6: Clean Water and Sanitation (United Nations, 2025b) and SDG 9: Industry, Innovation and Infrastructure (United Nations, 2025a). These goals align with both the environmental and technological innovation of the project.

Through the use of locally hosted AI (OpenVoiceAgent) integrated in the virtual training, the prototype supports SDG 9 by promoting innovation within aviation training and demonstrating how these new technologies can enhance operational efficiency by reducing reliance on human actors or a mock-up setup. At the same time, our prototype addresses SDG 6 by eliminating the use of commercial AI models (OpenAI), particularly their hidden water demands, by using a locally hosted AI that is a more sustainable alternative that minimizes data-center cooling water consumption.

The prototype not only advances innovation within KLM's training programs, but also contributes to the broader sustainability goals of the United Nations.

5.1.2. People: Personal Merits to Prototype

The People dimension of the Triple Bottom Line framework focuses on the social and human value generated by the VR+AI training system. In the context of KLM's training ecosystem, this aspect extends beyond technological or financial gains to examine how the system impacts those directly involved, particularly cabin crew and training staff. Evaluating the People component is essential, as the success of innovation depends on human acceptance, ethical integration, and its contribution to employee development and well-being. By addressing these factors, the analysis ensures that the implementation of the AI-driven training system supports

not only operational effectiveness but also a positive and responsible human experience.

Stakeholder Analysis

A stakeholder analysis was conducted to identify and evaluate the individuals, groups, and organizations that could influence, or be affected by, the development and testing of the AI-driven conversational training system for cabin crew. The analysis aimed to provide a structured understanding of stakeholder roles, resources, concerns, and potential influence on the PoC project (Mostert, n.d.).

This analysis served as an essential foundation for understanding the ecosystem surrounding AI adoption within the airline's training environment. By mapping stakeholders' roles, interests, and potential influence, it provided valuable insights into the human and organizational factors shaping the adoption process. Focusing on cabin crew as the primary context was particularly relevant, since the project's initial scope had always been centered on them rather than the care team. The aviation industry also poses unique emotional and communicative challenges, making it an ideal setting for testing AI-driven conversational systems.

Stakeholder Identification: The first step involved the identification of major stakeholders based on prior knowledge of airline operations, experience with similar training initiatives, and consultation with project team members.

- The primary stakeholders identified were Airline Management, Crew, Passengers, Trainers, and the Innovation Team (including developers and testers).
- Secondary stakeholders were determined to include Relatives, Investors, and Regulatory Organizations, such as the Federal Aviation Administration (FAA) and the European Union Aviation Safety Agency (EASA). These groups were selected due to their direct involvement in the project, potential to contribute resources, and the extent to which they could be affected by the outcomes and/or influence the outcomes of the training prototype.

Resources and contributions: Each stakeholder group was assessed in terms of the resources they could bring to the project:

- **Airline Management**: possesses strategic authority, budgetary control, and the capacity to allocate operational and organizational resources necessary for the development of the AI system. Their role was essential for the approval of project milestones and provision of institutional support.
- **The Crew**: as the primary end-users, contributed practical expertise, operational knowledge, and active participation in the training simulations. Their engagement was critical to testing the realism and effectiveness of the AI agent.
- **Trainers**: brought instructional experience, and insight into the behavioral competencies required for high-quality cabin crew performance.
- **The Innovation Team**: consisting of developers and testers, provided technical expertise, system design capabilities, and iterative development feedback, enabling the prototype to simulate realistic and emotional interactions and essentially building a local AI from the ground up.

Secondary stakeholders were assessed for their indirect but influential roles.

- **Regulatory Organizations**: held the legal authority to approve training methodologies and ensure compliance with aviation safety standards. Their resources included: regulatory knowledge and the capacity to endorse or restrict experimental practices.
- **Investors**: could contribute financial support and strategic guidance, but had limited day-to-day involvement in the training operations themselves.

- **Passengers and relatives**: were included due to their indirect interest in the well-being and professional development of crew members, though their influence on project outcomes was minimal.

Concerns and perceptions: Stakeholder concerns and perceptions were also analyzed. Airline Management will be primarily concerned with return on investment and adherence to regulatory standards. Crew members were interested in the realism and usability of the system but expressed potential apprehension regarding performance evaluation and interaction with an emotionally responsive AI agent. Trainers were attentive to how the AI system could complement or replace traditional instructional methods, including the risk of reduced autonomy in delivering training. The Innovation Team focused on technical feasibility, financial feasibility, usability, and integration challenges, while Passengers were concerned with indirect benefits such as improved safety and quality of service. Regulatory bodies were attentive to compliance and the ethical application of AI in safety-critical training environments. Investors were primarily concerned with financial viability and scalability, whereas Relatives were concerned with the stress and workload implications for crew members during training exercises.

Power-interest Matrix: The analysis further evaluated each stakeholder's ability to contribute to, block, or be affected by the project. For the Power-Interest Matrix, see Figure 5.1.

1. **High-power, high-interest:** Stakeholders in this group, including **Airline Management, Crew, and Regulatory Organizations**, are central to project success, as their involvement is essential for providing resources, ensuring operational and regulatory compliance, and integrating the system into practical training routines.
2. **High-power, lower-interest:** This category comprises **Investors and the Innovation Team**, who can influence funding and technical development but are less engaged in day-to-day training operations. They primarily require periodic updates and assurance that the project aligns with strategic objectives.
3. **Low-power, high-interest:** Stakeholders such as **Passengers and Trainers** fall into this group; although they cannot directly influence project execution, they are highly invested in its outcomes and require consistent communication to ensure the system meets expectations and supports training goals.
4. **Low-power, low-interest:** Finally, **Relatives** belong to this category, requiring minimal engagement beyond general updates on training safety and well-being.

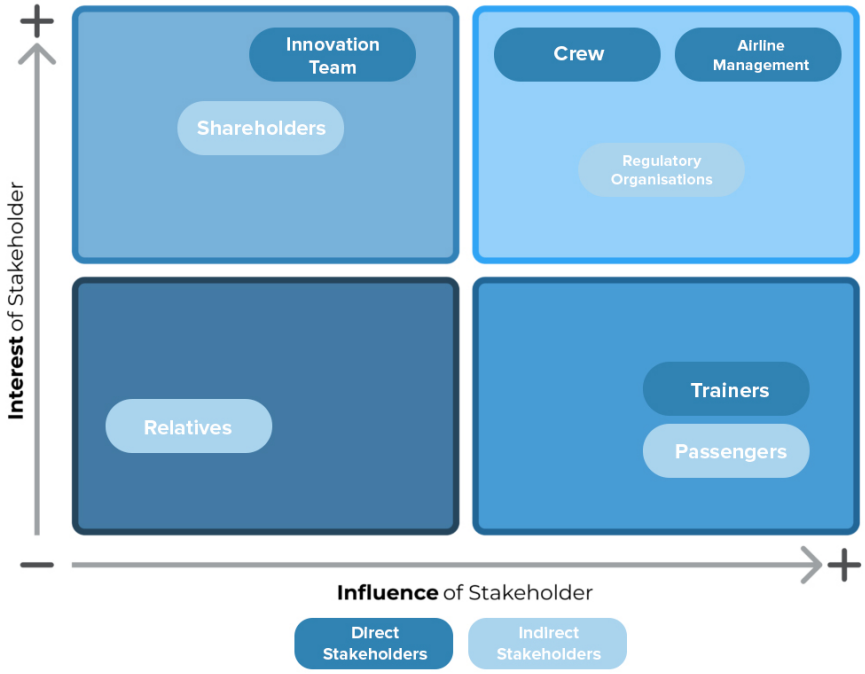


Figure 5.1: Power-interest matrix of stakeholders involved in our project.

Social Value: PoC Results

The interviews conducted in the PoC tests revealed some added values to the prototype. Participants reported a sense of pride when managing to calm down the agent and a strong feeling of willingness to help. Since the prototype creates a safe space to practice conversations, it is a space for cultivating positive emotions that strengthen well-being at work.

Another pattern was people's interest in internal growth. They asked questions such as:

> "*This makes me wonder how I can improve myself?*" (P7)

Participant 7 told about how the tool stimulated "internal reflection" and how one can "explore your own capabilities".

The participant who is a professional pilot stated that:

> "*I would be more prepared, and would like to be more prepared, by using your AI.*"

And further explained how pilots do not receive enough training to have these tough conversations or to make difficult announcements.

Besides the use cases that were explained earlier in Section 4.3, the People value is not only about expansion and efficiency, it all comes down to building the confidence of the user to feel prepared for their workday ahead.

5.1.3. Profit: Costs of Prototype

Following the People dimension, which explored the social value and experiential benefits of the VR+AI prototype, this section focuses on the Profit aspect of the Triple Bottom Line framework. Building on the implementation pathway described in Chapter 4, this analysis compares the current traditional, actor-based crisis management training with the proposed VR+AI approach. The results provide a foundation for understanding the economic viability of adopting the VR+AI model into KLM's infrastructure.

5.1.3.1 Financial Implementation Costs

The financial implementation costs of KLM's new VR+AI crisis management training were evaluated in comparison to the traditional, classroom-based method currently in use. The comparison includes both hardware configurations for VR headsets that could be used for VR training: the Meta Quest 3 and the Apple Vision Pro.

Each year, approximately 2.000 employees are required to complete the crisis management training. This estimated number of participants was used consistently for both training methods in the financial comparison.

The financial figures presented in this section are estimates. The actual costs may vary. These values are intended to provide an indicative comparison between the traditional and VR+AI training methods rather than exact financial projections. These estimates were given by a reputable KLM employee with 20+ years of experience in making innovation cost estimations.

As discussed in Section 4.1.1, KLM conducts a short feasibility study before adopting new VR-based training programs. These studies, such as the KLC VR cockpit, slide raft, and de-icing training programs, typically last a few months and are used to validate usability and learning effectiveness. Once a study confirms that the new method outperforms the existing one, KLM proceeds directly to implementation. In practice, this means that when existing contracts for actors expire (usually after one year), they are simply not renewed, and KLM transitions seamlessly to the VR+AI-based training system. This approach minimizes downtime and enables immediate integration into the new program.

Current Crisis Management Training

KLM's current crisis management training is conducted through physical classroom sessions and full-scale cabin mock-up simulations. Each session is done with 12 trainees. Each session requires instructors to supervise and assess trainees, hired actors to simulate passengers in distress, and the use of dedicated facilities.

On average, one complete training session costs around €1.500, broken down as follows:

- Instructor costs: €500
- Facility costs: €200
- Cabin mock-up simulator costs: €500
- Actor fees: €300

KLM conducts approximately 167 of these sessions annually, 10%-15% of the trainees need additional sessions, together with support costs/operations costs of €100.000, resulting in a total annual expenditure of roughly €480,000. These computations are visualized in Table 5.1.

Current Crisis Management	**Costs**	**Amount/Count**
Instructor Costs	€ 500,00	
Facilities	€ 200,00	
Cabin Mockup Simulator	€ 500,00	
Actors	€ 300,00	
Total one session	**€ 1.500,00**	
Total training sessions	€ 250.000,00	≈167 sessions/year
Additional Sessions	€ 130.000,00	≈200-250 trainees
Support/Operation costs	€ 100.000,00	
Total per year	**€ 480.000,00**	

Table 5.1: Cost breakdown for Current Crisis Management

VR+AI Crisis Management Training

As mentioned in Section 1.1.2, this new approach uses an immersive VR environment combined with an AI-driven system that enables emotionally realistic simulations of passenger interactions and crisis scenarios, without the need for live human actors.

Two VR headsets were evaluated for the VR+AI setup: the Meta Quest 3 and the Apple Vision Pro. These devices represent the latest generation of mixed-reality hardware, allowing the user to experience mixed reality, which includes elements of both virtual and augmented reality.

The initial investment of the VR+AI implementation will cost €150.000. This investment covers all essential components needed for successful deployment, including feasibility studies, hardware and software, and the development necessary to bring the product to a professional and operational level.

Meta Quest 3 Configuration: In the Meta Quest 3 implementation, the system consists of 16 VR headsets, instructors and local support infrastructure.

The first-year investment includes:

- Instructor costs: €62.500
- VR headsets (16 units at €550 each, 3-year depreciation): €2.933
- Initial MVP investment: €150.000
- Support and operation costs: €20.000

The total first-year cost is approximately €235.433, followed by only €88.139 (adjusted to inflation) in Year 2, once the initial investment is depreciated (Table 5.2). Compared to the traditional €480.000 annual cost, this results in annual savings of €244,567 in Year 1 and € 407.605 in Year 2, representing a reduction of more than 80% in yearly training expenditures.

VR+AI training Crisis Management: Meta Quest	**Year 1**	**Year 2**
Instructor Costs	€ 62.500,00	€ 64.550,00
VR headsets	€ 2.933,33	€ 2.933,33
Initial Investment MVP	€ 150.000.00	
Support/Operation costs	€ 20.000,00	€ 20.656,00
Total per year	**€ 235.433,33**	**€ 88.139,33**

Table 5.2: Annual Costs for VR+AI Training Crisis Management - Meta Quest 3

Apple Vision Pro Configuration: The Apple Vision Pro configuration follows the same training concept but with higher-end hardware; this might be required when the VR environment features more demanding visuals and requires greater processing power. Due to the increased headset price (€3.500 per unit), the initial costs are slightly higher:

- Instructor costs: €62.500
- VR headsets (16 units at €3.500 each, 3-year depreciation): €18.667
- Initial MVP investment: €150.000
- Support and operation costs: €20.000

The total first-year cost amounts to €251.166, decreasing to €103.872 (adjusted to inflation) in Year 2 (Table 5.3). Annual cost savings remain significant, estimated at €228.833 in Year 1 and €378.833 in Year 2 compared to the traditional program.

VR+AI training Crisis Management: Apple Vision	**Year 1**	**Year 2**
Instructor Costs	€ 62.500,00	€ 64.550,00
VR headsets	€ 18.666,67	€ 18.666,67
Initial Investment MVP	€ 150.000,00	
Support/Operation costs	€ 20.000,00	€ 20.656,00
Total per year	**€ 251.166,67**	**€ 103.872,67**

Table 5.3: Annual Costs for VR+AI Training in Crisis Management - Apple Vision

Financial Impact

To assess the long-term financial benefits of replacing the current roleplay-based crisis management training with the new VR+AI training, a comparative cost analysis was conducted. This evaluation includes both the short-term investment period and the long-term financial trajectory. Figure 5.2 illustrates the projected annual training costs for both methods in the first three years; this shows how inflation gradually increases the total expenditure of the traditional setup and how the VR+AI methods drop significantly after the initial investment. Based on the historical average inflation rate in the Netherlands (3.28% between 1970 and 2023), the annual cost of the current crisis management training is estimated to rise from approximately €480,000 to more than €750.000 over a 15-year period. In contrast, the VR+AI training model maintains costs well below €160.000 per year after the initial investment, demonstrating a substantial reduction also in long-term expenditure, as seen in Figure 5.3.

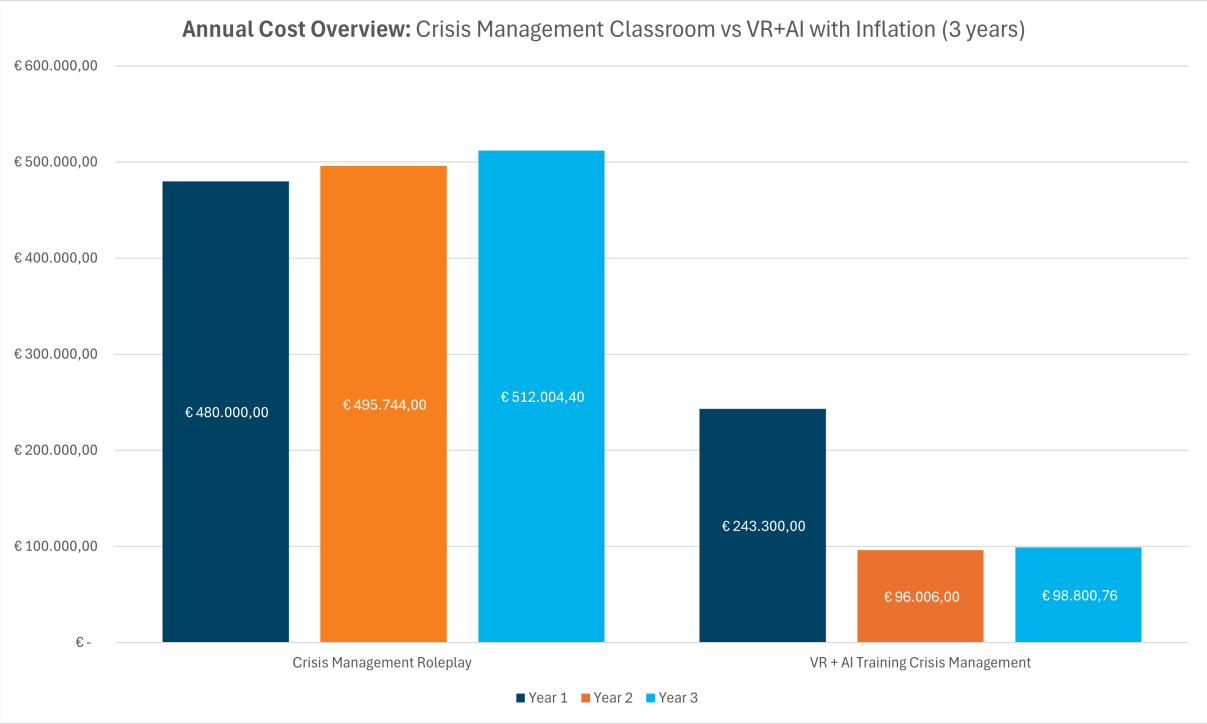


Figure 5.2: Annual Cost Overview Comparison (short-term)

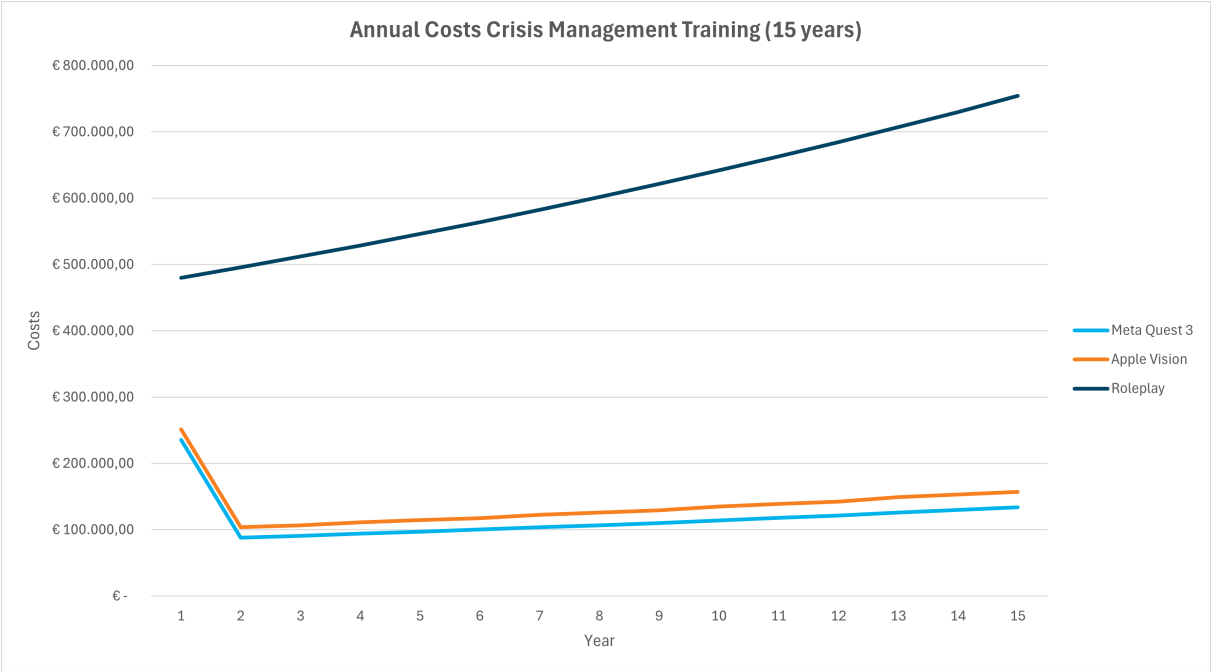


Figure 5.3: Annual Cost Overview Comparison (long-term)

After the initial investment, both VR+AI configurations, Meta Quest 3 and Apple Vision Pro, show a sharp decline in costs, stabilizing at approximately €90.000-€150.000 per year. This results in a consistent 79-82% annual cost reduction, as seen in Figure 5.4.

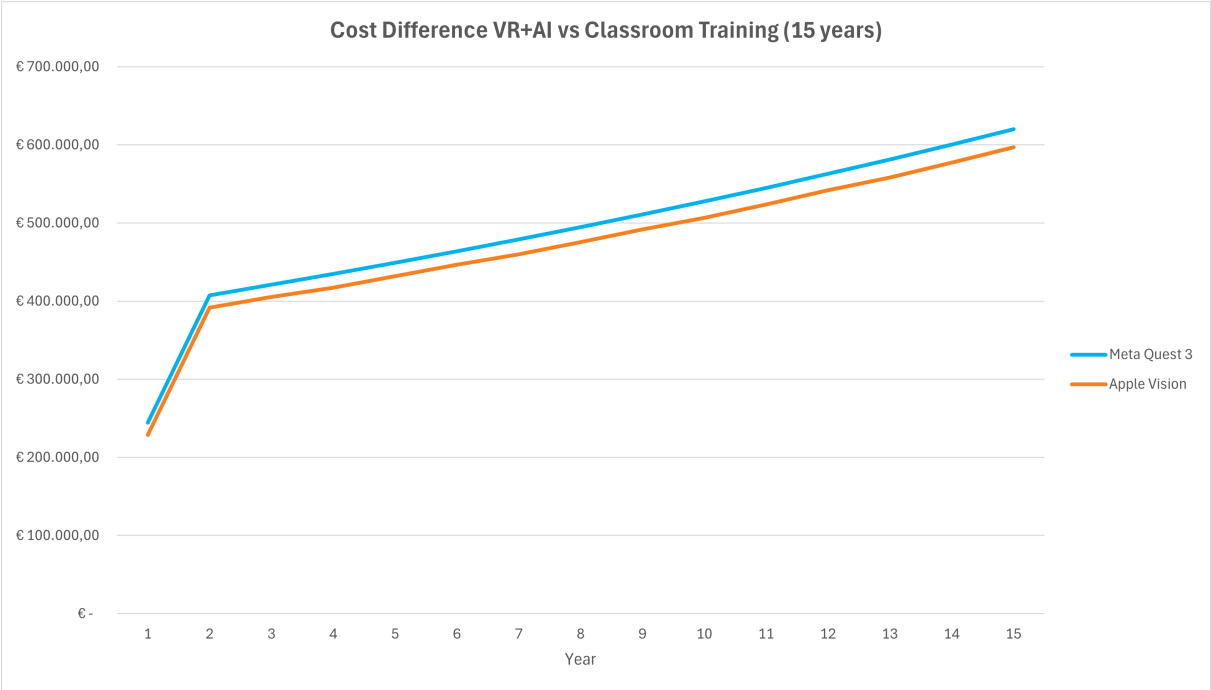


Figure 5.4: Annual Cost Difference (long-term)

Both the VR+AI models achieve a return on investment within eight months. In Figure 5.5, the costs are analyzed on a per-person basis, assuming 2.000 trainees annually, the difference is clearly noticeable: €248 per person in Year 2 in the traditional setup compared to just €44–€52 per person with the VR+AI training.

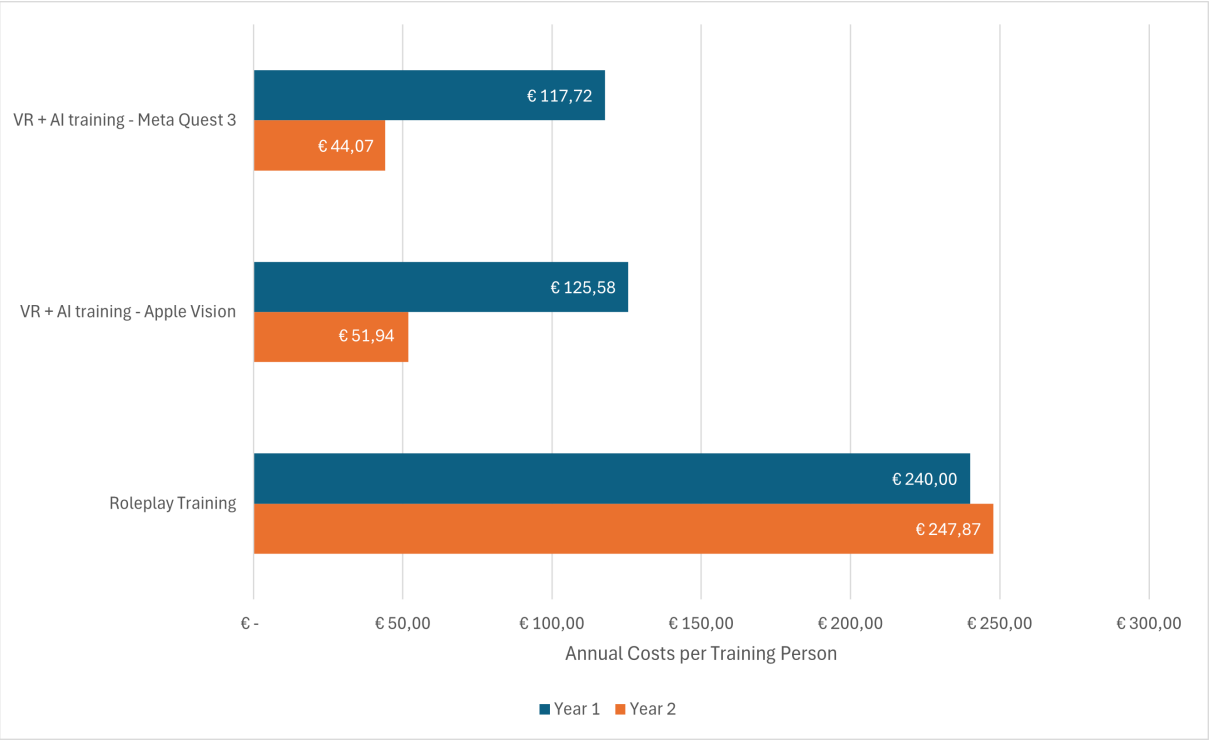


Figure 5.5: Annual Cost per person (short-term)

In conclusion, the financial analysis shows that adopting VR+AI crisis management training results in major long-term savings compared to the current method. Despite higher upfront investment, yearly costs decrease by over 80%, with a return on investment achieved within eight months.

5.1.3.2 AI Usage Costs

While the previous analysis addressed the financial investment of implementing the entire VR+AI training system, the use of the AI is also a recurring costs. The usage cost of AI is determined by using token-based billing, which reflects on the computational effort required to process and generate the information. In OpenAI's `gpt-realtime`, introduced in Section 2.1.2, every word, sound, and symbol processed by the system corresponds to a quantifiable number of tokens. Tokens are essentially the building blocks of text and audio that models, such as OpenAI's, process (OpenAI, 2025d). In the context of KLM's Caresse system (Section 1.1.6), which uses `gpt-realtime` for speech and text interaction, understanding the token usage is important for estimating financial usage costs.

We now examine how token-based pricing structures translate into actual operational costs and highlight how prompting size, conversation length, and the use of audio all influence the total usage costs. In this section, costs are expressed in U.S. dollars to reflect OpenAI's pricing model which is in U.S. dollars; this gives the most accurate representation of costs.

OpenAI Token Costs

A token represents a small chunk of data, which may consist of a word, sub-word, or even a single character, depending on the context. For example, the sentence "*The aircraft is boarding now.*" consists of approximately six tokens.

According to OpenAI, most English text contains approximately 0.75 words per token (OpenAI, 2025d). A typical conversational speed is 150 words per minute (WPM), which will generate about 200 tokens per minute in text form.

In audio models, it works differently: the token density increases significantly. Speech is generated by converting audio to text (to binary) which will be counted as tokens. The token count depends on the length of the audio rather than the size or type of the audio input (Google Collab, n.d.). The longer the audio file, the higher amount of tokens it will generate.

In Fu (2025)'s analysis of the `gpt-realtime` API, every time a longer or more descriptive prompt is used, such as instructing the model to act as an emotional passenger or simulate a crisis scenario, the system must process a significantly higher number of text input tokens before generating a response.

In practical terms, this means that a standard one-hour session without complex prompting costs roughly $12-$14, while scenario-based prompting can easily raise costs to $24 per hour or more.

Each AI session using speech involves multiple token streams:

- Audio input tokens: user speech converted to digital representations.
- Input text tokens: the written or transcribed user input (from the audio).
- Output text tokens: the model's generated response.
- Audio output tokens: the AI's spoken reply synthesized into sound.

The total number of tokens determines the total usage cost for each session.

Commercial AI Usage Costs

According to OpenAI's official pricing, `gpt-realtime` is billed per million tokens (1M), with separate prices for each token type (OpenAI, 2025c). The following prices were used to calculate the usage costs for this project:

Token Type	Price (USD / 1M Tokens)	Description
Text input	$4.00	Newly provided prompt or message
Text cached input	$0.40	Reused or repeated context tokens
Text output	$16.00	Generated text reply
Audio Input	$32.00	Speech-to-Text conversion
Audio cached input	$0.40	Reused or repeated context tokens
Audio output	$64.00	Text-to-Speech generation

Table 5.4: Different token costs per type of `gpt-realtime`

A one-hour conversational training with KLM's Caresse PoC (that uses `gpt-realtime`) was used as the basis for calculating a reference scenario, where both the user and the AI speak for equal durations (30 minutes each) at 150 words per minute. This duration will be taken as a benchmark in comparison to our local AI: OpenVoiceAgent.

Based on the token costs, the following estimates were calculated:

- Total Tokens Processed (1 hour) without prompting: 127,786
- Total Tokens Processed (1 hour) with prompting: 263,239

Calculating the usage costs for one hour gives a total of **$12.39** without prompting and **$25.52** with prompting.

Local AI Usage Costs

For the local AI version, the token-based billing component is replaced by direct electricity costs, since the model runs entirely on KLM's own on-premise hardware. This means that instead of

paying per token, the only operational cost comes from the electricity required to power the GPU. Because the model is executed locally, no additional usage fees are incurred, making the AI operation free beyond electricity consumption.

Using an NVIDIA RTX 5090 GPU, as mentioned in Section 2.1.2, with a power draw of 575 W, and an average Dutch electricity price of €0,26/kWh (≈ $0.28/kWh) (ConsumentenBond, 2025), the total cost of operating the local model for one hour is approximately = **$0.17/hour**.

Usage Costs Comparison

Figure 5.6 presents a comparison between the commercial `gpt-realtime` (hosted) model and a local model on the RTX 5090 GPU, in terms of total usage costs over different duration periods. The results show a substantial cost gap between the hosted and local AI costs. As operational time increases, the difference grows linearly.

Prompting complexity substantially increases costs in commercial AI, since it raises the number of processed tokens significantly. The local AI scales linearly with time, tied only to energy use. Hosted models also scale linearly due to tokens accumulation, especially audio tokens over time. Since the usage costs per minute is more expensive when using commercial AI, local AI becomes more cost-efficient. For large-scale applications such as KLM's Crisis Management training program, adopting a local model provides a good financial alternative to commercial models.

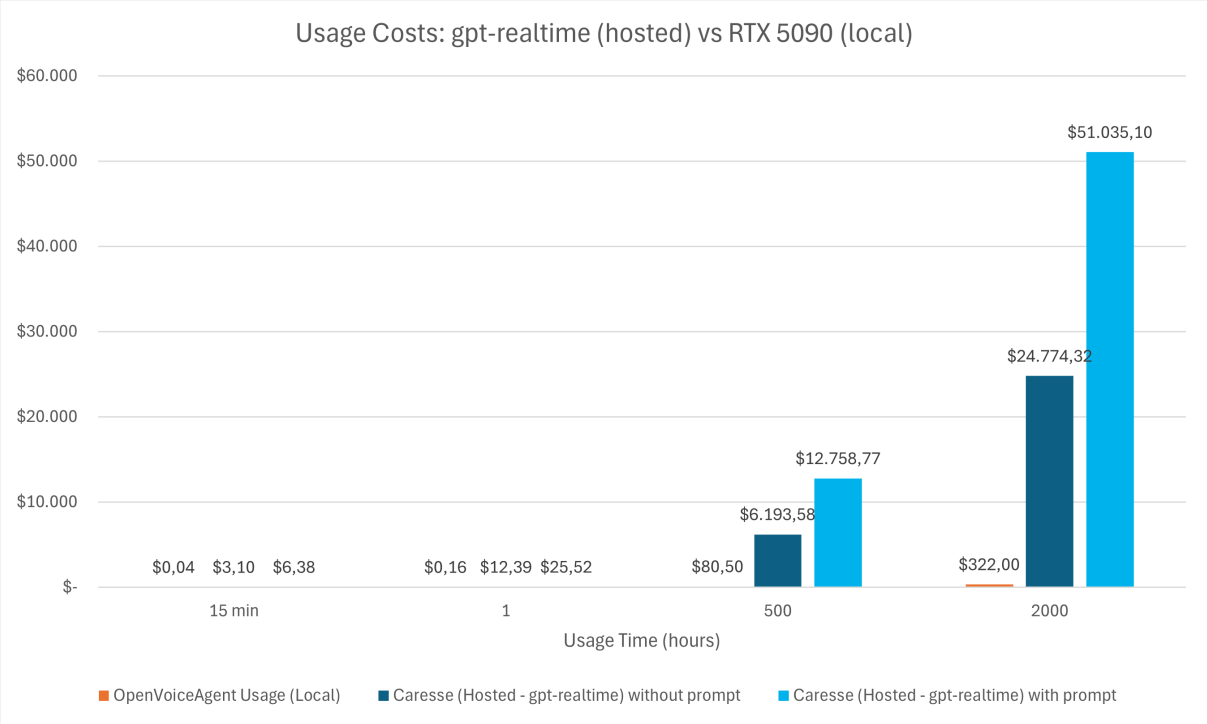


Figure 5.6: Usage Cost Comparison: Hosted AI (without and with prompting) and Local AI

5.1.4. Planet: Potential Planetary Benefits to Prototype

The previous subsection examined the Profit dimension by assessing the financial costs and savings associated with implementing the VR+AI training system. The Planet dimension extends this analysis by considering the environmental implications of using hosted and local AI models. This section evaluates the energy consumption and water usage associated with AI operation.

The environmental impact of using commercial AI models is not limited to token costs but also

includes significant electricity use and water consumption used in data center cooling. To quantify these effects, both the energy and cooling water requirements for one hour of AI operation were analyzed using official pricing data from OpenAI (OpenAI, 2025c), recent energy-efficiency research on large language models, and sustainability of data center resource use.

The electricity and water usage estimates are based on publicly available data and research. Actual energy consumption may vary depending on the type of model used, the prompting, and data center efficiency.

Electricity Consumption

Energy consumption varies depending on the specific AI model. In KLM's case, the Caresse system uses `gpt-realtime` for both audio and text token generation. Since no direct energy consumption data is currently available for `gpt-realtime`, we use the energy consumption for `gpt-4o mini`, as it shares a similar architecture and offers comparable performance for realtime tasks (OpenAI, 2025b).

According to a research paper of Jegham et al. (2025) GPT-4o mini consumes approximately 1,42 Wh for input processing and 0,33 Wh for output generation.

Based on these reference values, the estimated energy consumption for one hour of `gpt-realtime` operation, equivalent to the processing of 127.786 tokens (see Section 5.1.3), is approximately 0,11 kWh without additional prompting. When using extended prompting, for example, instructing the AI to act as a "worried brother of a passenger on a flight", total energy use increases to roughly 0,23 kWh per hour due to the additional workload associated with prompt interpretation (Fu, 2025).

Water Sustainability: SDG 6

Water plays a critical part of the cooling cycle in big scale data centers. Servers performing billions of operations generate significant heat, which must be cooled to avoid overheating in the data centers. The increasing water demand of cloud-based AI data centers poses growing challenges for Sustainable Development Goal 6 (Clean Water and Sanitation), which aims to ensure availability and sustainable management of water and sanitation for all.

In the EESI article, Yañez-Barnuevo (2025) highlight that large data centers can consume millions of gallons of water per day, comparable to water use for towns of 10.000-50.000 people. If demands increase for AI usage, the water demand will also go up. Around 80% of water withdrawn for the use of data centers is evaporated and never returning to local water sources.

These pressures directly challenge UN Sustainable Development Goal 6 (United Nations, 2025b).

- Target 6.1 (equitable access) is at risk when data centers monopolize municipal water.
- Target 6.3 (water quality and reuse) can be achieved if facilities adopt wastewater recycling or closed-loop liquid cooling.
- Target 6.4 (efficient and sustainable water use) is undermined when industrial cooling competes with human needs.
- Target 6.5 (integrated water management) requires that data centers be incorporated into regional water planning.

Water Consumption Commercial AI

Water use in data centers is typically measured via the metric Water Usage Effectiveness (WUE), which is defined as liters of water consumed (evaporated or unrecoverable) per kilowatt-hour of energy use. Many data centers aim for a WUE near 1,9 liters/kWh (Yañez-Barnuevo, 2025).

OpenAI `gpt-realtime` generates around 0,11 kWh for one hour of usage (see section 5.1.3), applying WUE = 1,9 L/kWh gives: **0,21 L**. And for prompting which uses 0,23 kWh it will use

0,44 L of water. That water demand may seem small, but this is just for a one hour of usage. If every trainee were to use the AI system for a 15-minute session, and assuming approximately 2.000 participants, the cumulative cooling water consumption would reach around 106 L without prompting, and approximately 219 L with prompting. If all participants would use it for one hour, this water usage will be even more, as seen in Figure 5.7.

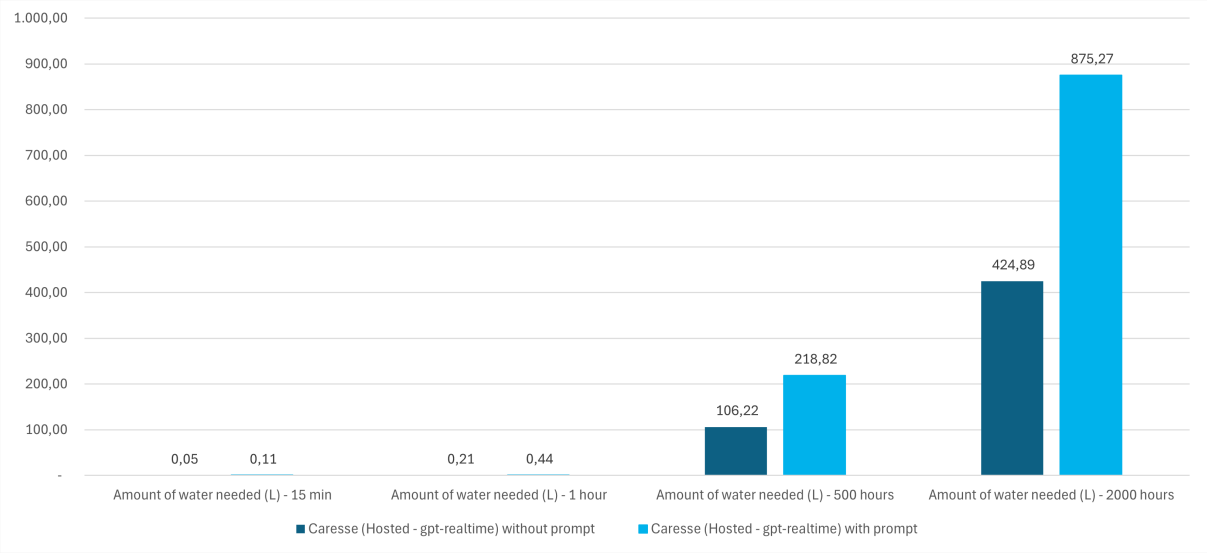


Figure 5.7: Water consumption of hosted AI with and without prompting.

Overview Costs and Usage

The comparison between KLM's Caresse PoC (hosted `gpt-realtime`) and the OpenVoiceAgent (local AI) highlights major differences in cost and sustainability.

On an hourly basis (Table 5.5), the hosted model costs approximately $12.39 without prompting and $25.52 with prompting, whereas the local AI requires only $0.16 per hour in electricity. If every trainee (in total 2.000 trainees) would do 15-min of AI training, this would result in over 500 hours of training in a year. At 500 hours, this results in a total of $12,758.77 for the hosted version versus $80.50 for the local system (Table 5.6).

While commercial AI appears more energy-efficient per hour, its hidden environmental costs, make it less sustainable overall. In contrast, local AI provides energy transparency, water independence, and the opportunity for circular resource integration by using renewable sources such as solar or wind energy.

1 Hour	**Caresse (Hosted - `gpt-realtime`) without prompt**	**Caresse (Hosted - `gpt-realtime`) with prompt**	**OpenVoiceAgent Usage (Local)**
Tokens Costs	$12.39	$25.52	-
Electricity Costs	$0.02	$0.04	$0.16
Costs of water	$0.00	$0.00	-
Amount of water needed (L)	0.21	0.44	-
Electricity Usage (kWh)	0.1116	0.22093	0.575

Table 5.5: Cost and usage data for 1 hour of use.

500 Hours	Caresse (Hosted - `gpt-realtime`) without prompt	Caresse (Hosted - `gpt-realtime`) with prompt	OpenVoiceAgent Usage (Local)
Tokens Costs	$6.193,58	$12.758,77	-
Electricity Costs	$9.77	$20.12	$80.50
Costs of water	$0.66	$1.36	-
Amount of water needed (L)	106.22	218.82	-
Electricity Usage (kWh)	55.91	115.17	287.50

Table 5.6: Cost and usage data for 500 hours of use.

Innovation: SDG 9

The VR+AI training prototype supports Sustainable Development Goal (SDG) 9: Industry, Innovation, and Infrastructure, by promoting innovation within aviation training and reducing reliance on the current training methods. It aligns with targets 9.1, 9.4, and 9.5, addressing infrastructure resilience, upgrading infrastructure, and digital access (United Nations, 2025a).

- **Target 9.1**: Develop quality, reliable, sustainable and resilient infrastructure. The implementation of a local, offline-capable AI enables KLM to train staff without dependence on internet connectivity or external data centers. This approach builds a more resilient digital infrastructure that remains operational even in the event of network disruptions. By transitioning from actor-based training to lightweight, portable VR setups, the prototype supports sustainable infrastructure that is easier to scale and maintain.
- **Target 9.4**: Upgrade infrastructure and retrofit industries to make them sustainable, with increased resource-use efficiency. By shifting to a VR-based environment, KLM significantly reduces its dependency on cabin mock-up simulators, physical classroom setups, and repeated actor hiring. The use of local AI promotes sustainability by eliminating the need for the use of datacenters which often requires high energy consumption and substantial water use for cooling.
- **Target 9.5**: Enhance scientific research and upgrade technological capabilities of industrial sectors. OpenVoiceAgent, developed in-house and optimized for realism and emotion, is a solution that can advance the technological capabilities of the aviation sector. OpenVoiceAgent will be open-source and this will encourage future academic and industrial research, supporting knowledge-sharing in the industry.

5.1.5. People, Planet, Profit: Closing Remarks

To conclude, the evaluation of the VR+AI training prototype through the Triple Bottom Line framework has shown that OpenVoiceAgent has incredible value. From a People perspective, OpenVoiceAgent enhances training quality by enabling emotionally intelligent, realistic simulations. By removing the need for live actors and mock-up simulators, it allows KLM employees to engage with crisis scenarios more frequently, safely, and on their own terms.

In terms of Profit, the prototype presents a highly cost-efficient alternative to KLM's current actor-based training setup. With an estimated return on investment within eight months and an estimated 80% reduction in yearly costs thereafter.

Under the Planet dimension, the prototype aligns with sustainability goals by reducing the environmental footprint of training. The local AI system consumes far less electricity and eliminates the need for water-intensive data center cooling. Compared to hosted AI models, the local setup shows more savings in long-term AI use. This aligns with the UN's SDGs.

5.2. Ethical Concerns & Risk Management

When innovating we must discuss the ethical concerns and risks associated with its development and deployment.

5.2.1. Ethics

As emerging technologies like AI and VR become increasingly integrated into training and operational contexts, addressing ethical considerations becomes essential. The implementation of an AI-driven conversational system, particularly one capable of emotionally charged interactions, raises important questions related to psychological safety, fairness, transparency, and accountability. This section discusses these ethical dimensions to ensure that the prototype's development and deployment align with both organizational values and broader regulatory frameworks. By critically examining these aspects, we aim to promote responsible innovation and safeguard user well-being.

Ethical Compliance and AI Alignment

The development and deployment of our crisis-response conversational AI system, designed to simulate emotionally intense interactions for staff training, adhere to the ethical principles and regulatory expectations outlined in the EU AI Act and related AI governance frameworks.

Although our AI operates in a training-only context and does not interact with real victims, we acknowledge that systems used in emergency preparedness fall within high-risk AI applications due to their impact on human decision-making in critical situations. Accordingly, our design philosophy is rooted in responsible innovation, embedding ethical safeguards across the system lifecycle and prioritizing AI alignment to ensure that the system's behavior consistently reflects human values, safety priorities, and regulatory standards. Though, we realize that certain ethical challenges require continuous supervision. For instance, the AI pipeline was built to be tailorable to different users whom may have the intention to adjust the pipeline in such a way that the AI can have uncensored outputs. However, these ambitions are outside our control, and we, ourselves have no intention to deploy uncensored technologies. The psychological realism of training scenarios, also require ongoing oversight and evaluation to maintain alignment with evolving ethical expectations.

In conclusion any potential unethical use would therefore depend on user behavior, and we rely on adherence to the EU AI Act to mitigate such risks. In addition, we strove to build the AI in accordance with applicable regulations and governance frameworks, reinforcing its alignment with both ethical and societal norms.

Risk Classification under the EU AI act

The EU AI Act and related standards classify AI systems as high-risk not based on superficial characteristics such as tone (e.g., rude versus polite), but on the nature of their application and potential impact on human decision-making (European Commission, 2023b). Specifically, high-risk classification depends on:

1. The domain of application: systems used in emergency response and/or safety training are considered high-risk;
2. Their influence on human decision-making. Thus, training staff for emotionally volatile scenarios means the AI is actively shaping critical professional behavior; and
3. The potential for psychological impact, even in controlled training environments, simulated emotions or emotional distress can induce stress or trauma (European Commission, 2023a).

Accordingly, our crisis-response conversational AI falls under the high-risk category as defined in Annex III of the EU AI Act.

Ethics Compliance Measures & Concerns

For an overview of our Ethical compliances and concerns, see the following table:

Ethical Principle (EU AI Act / OECD)	Identified Concern	Compliance Measures
Human Oversight & Control	-	The AI is used only as a training assistant, never an autonomous decision-maker. A human instructor is always present, with the ability to pause AI responses in real time.
Transparency & Disclosure	-	Trainees are explicitly informed that they are interacting with an AI simulation. Training sessions include pre-session briefings and post-session debriefs to contextualize responses.
Psychological Safety & Well-being	Response intensity levels are not configurable (e.g., calm, stressed, aggressive), therefore there is no controlled exposure. The AI is built with the intention to be unpredictable in emotions and the intensity it wishes to express those. Thus, constant human monitoring will be required in case of rising intensity levels making the trainee or user uncomfortable.	-
Fairness & Non-Discrimination	The AI is built to be unpredictable in behaviour and response. Although it has guidelines, these are merely prompts and therefore we cannot exclude the unethical responses the AI might give (racist remarks, sexist remarks, vulgar vocabulary).	-
Data Privacy & Security (GDPR Compliance)	-	No personal or sensitive trainee data is stored without consent. Interactions are anonymized, and all logs are securely stored for auditing purposes only. Since the AI is not running on a cloud-based service, all data will be stored within the company and therefore remain secured and private.
Accountability & Auditability	-	All AI-generated interactions are logged, with traceable metadata (scenario type, voice gender, timestamp) to support instructor review, error correction, and regulatory auditing.

5.2.2. Risk Management

Building upon the ethical considerations discussed in the preceding section, this section focuses on risk management for the AI deployment. It is evident that certain ethical concerns also manifest as operational and organizational risks. Specifically, issues related to psychological safety, well-being, fairness, and non-discrimination extend beyond moral responsibility and pose tangible threats to the project's integrity and successful implementation. In addition, a number of general risks, such as technological and regulatory risks, must be systematically managed to ensure the sustainable and responsible deployment of the local AI system within the airline context.

Psychological Safety and Well-being Risk

Risk Description: The AI system has been intentionally designed to display emotional unpredictability to simulate realistic, high-pressure training scenarios. However, the lack of configurable intensity levels may expose trainees to distressing or overwhelming experiences, compromising their psychological safety and well-being. Such exposure could diminish engagement and reduce the effectiveness of the training.

Mitigation Measures: Studies on AI-driven immersive training environments highlight the need for human oversight, adjustable intensity levels, and debrief sessions to protect users' psychological health (Zechner et al., 2023). Therefore, the following measures are recommended:

- Introduce configurable emotional intensity settings (e.g., stressed, aggressive) tailored to the trainee's experience level.
- Provide real-time human monitoring during high-intensity sessions to intervene when necessary.
- Conduct post-session debriefings to support emotional recovery and reinforce learning outcomes (Geraghty, 2023).

Fairness and Non-Discrimination Risk

Risk Description: Despite the presence of behavioural guidelines, the AI's semi-unpredictable nature introduces the possibility of producing biased or discriminatory responses. Such outputs could harm the organization's reputation, violate fairness principles, and compromise user trust.

Mitigation Measures: AI fairness research emphasizes bias detection, regular auditing, and human oversight as critical mitigation strategies (Bellamy et al., 2018). In line with these insights:

- Implement pre-deployment bias detection tools, such as the AI Fairness 360 Toolkit (Bellamy et al., 2018).
- Conduct continuous audits and retraining to identify and correct biased outputs.
- Maintain a transparent logging system to ensure traceability and accountability of AI interactions.
- Ensure human supervision to moderate and correct inappropriate responses.

Technology Failure Risk

Risk Description: As with most innovative technologies, there is a risk of bugs or system failures, especially during early development stages. Such failures may lead to project delays, cost overruns, or data loss.

Mitigation Measures:

- Apply structured IT project management methodologies for improved control.
- Execute a comprehensive testing plan, validated through multiple pilot rounds before live deployment.
- Maintain robust backup systems and employ strict version control.

Regulatory and Compliance Risk

Risk Description: Given the rapid evolution of EU AI and data protection regulations (e.g., EU AI Act, GDPR), the project may face new compliance obligations or restrictions. Sudden changes could delay deployment or increase administrative costs.

Mitigation Measures:

- Involve legal and compliance experts throughout the project lifecycle.
- Maintain direct communication with authorities (GDPR, EASA, ICAO) to anticipate policy changes.
- Regularly review internal protocols to ensure alignment with emerging EU legislation.

Safety and Reliability Risk

Risk Description: Inaccurate or misleading AI responses during medical or safety-related simulations may cause confusion or reduce confidence in the system. In high-stakes environments like aviation, reliability is paramount.

Mitigation Measures: According to the AI Risk Management Framework (Raimondo et al., 2023), effective mitigation includes:

- Retaining strict human oversight: the AI must serve only as a recommendation tool.
- Conducting iterative validation in controlled environments before operational use.
- Monitoring performance indicators to detect model drift or degradation over time.

Adoption and Trust Risk

Risk Description: Successful implementation depends on crew acceptance. Resistance may arise if AI is perceived as surveillance or a threat to expertise, while over-reliance could undermine decision-making confidence.

Mitigation Measures: User-centred AI adoption studies highlight the importance of transparency and voluntary engagement (Ehsan et al., 2021). To address this:

- Clearly communicate that AI serves as an assistive tool, not an authority.
- Begin deployment through voluntary participation and expand gradually.

Technical Maintenance and Obsolescence Risk

Risk Description: Without periodic retraining, AI systems risk becoming outdated or inaccurate, leading to decreased reliability and trust (Jain, 2025).

Mitigation Measures:

- Establish a governance framework including regular performance reviews and retraining cycles.
- Conduct periodic internal audits to ensure adherence to ethical and operational standards.

Legal and Liability Risk

Risk Description: When AI-generated recommendations influence critical training or medical outcomes, accountability can become ambiguous. This ambiguity may expose the organization to legal disputes or reputational harm.

Mitigation Measures: Legal studies on AI-assisted decision-making stress the importance of clear liability boundaries and disclaimers (Zech, 2021).

- Legally define the AI's role as a non-decisive, advisory system.
- Include embedded disclaimers clarifying that final decisions rest with human operators.

- Align insurance and legal frameworks to ensure comprehensive liability coverage before deployment.

5.2.3. Conclusion

Integrating ethical and operational risks within a unified risk management framework ensures that the AI system's deployment remains responsible, compliant, and sustainable. By prioritizing psychological safety, fairness, reliability, and accountability, the project aligns with both ethical principles and industry standards for AI governance. This holistic approach safeguards stakeholders while strengthening institutional trust in AI-assisted training and operational tools.

6
Conclusion

6.1. Project Outcomes

6.1.1. For Our Team

As a team, this project has brought us closer. We have a plethora of inside jokes, a successful escape room escape, and many days of laughter. Each of us has learned the others' way of working and has picked up some of the others' tricks-of-the-trade. For example, our 'non-coders' have learned how to work with Overleaf to generate presentation ready reports and about the intricacies of python packages; our 'excel newcomers' have learned the value of advanced excel formulas to generate meaningful figures from basic data entries; and our 'first-time interviewers' have learned about the mentally exhausting yet exciting nuances involved in connecting with interviewees. Each of us have learned more about our own fields as well. The development team has broadened their experience with speech technologies; the quality control team has learned about the frustration and helplessness of corporate interviewing; and the business team has learned about the difficulty of corporate data gathering. We have learned a lot from each other and this project and aim to stay well connected personally and professionally in the future.

6.1.2. For KLM

The XR lab within KLM has gained a powerful, yet flexible prototype: OpenVoiceAgent, usable on their current infrastructure, fully installed, with detailed operating instructions (Smink et al., 2025). OpenVoiceAgent is open-source, private, and fully controllable; KLM can implement their own use-cases, voices, and models. OpenVoiceAgent has been verified to be useful by nine KLM employees and 78 external survey participants. Furthermore, KLM has received in-depth implementation and value analyses with social, monetary, and environmental benefits. A handover presentation was given at KLM presenting OpenVoiceAgent, including use and update instructions. Our coach and the other employees at KLM were beyond impressed with our work, professionalism, and presentational skills, calling us "the best JIP team they've ever had".

6.1.3. For Others

Like KLM, the world has also gained a powerful tool. OpenVoiceAgent is open-source and extends beyond the aviation industry as a universal platform for emotional communication training for conversations such as salary negotiations, treatment discussions, and conflict resolution. The applications of OpenVoiceAgent are endless and impactful.

6.2. Closing Remarks

Our team would like to sincerely thank Birgit de Bruin and the other coordinators at JIP, without whom we could not have enjoyed such a fulfilling project. More sincere thank you's to each of our interview participants at KLM and our many survey respondents for your insightful feedback and willingness to contribute to our project. Another thank you to Dr. Derek Lomas for his enthusiasm and sharing of technical resources. A big thank you to the employees at TU Delft's XR Zone for your sincere interest and help with hardware limitations. Finally, we especially

would like to thank our wonderful coach, Jae Maloney, for his unending enthusiasm, Welshness, and overall joy; we loved working with you and hope to stay in touch.

Our team is incredibly proud to have “replicated a multi-billion dollar commercial model in just 5 weeks”. Thank you to all who have supported us. With that, we thank you for flying with us and please reach out with any questions.

References

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., & Zhang, Y. (2018). Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv.org*. https://arxiv.org/abs/1810.01943

Boson-AI. (2025). Higgs audio v2: Redefining expressiveness in audio generation. https://github.com/boson-ai/higgs-audio

Business Review Europe. (2021). *Klm trains its crew with virtual reality and plans to develop vr programs to replace theoretical training* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://business-review.eu/future-of-work/klm-trains-its-crew-with-virtual-reality-and-plans-to-develop-vr-programs-to-replace-theoretical-training-215715

Cantrell, L. (2013). The power of rapport: An analysis of the effects of interruptions and overlaps in casual conversation [Accessed 2018-11-03]. *Innervate*, *6*, 74–85. https://www.nottingham.ac.uk/english/documents/innervate/13-14/06-lucy-cantrell-q33103-pp-74-85.pdf

Capgemini Netherlands. (2019). *Success story: Klm cityhopper virtual reality training* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://www.capgemini.com/nl-nl/wp-content/uploads/sites/7/2020/12/1A-083.20-SUcces-story-KLM-Cityhopper_web.pdf

Casanova, E., Davis, K., Gölge, E., Göknar, G., Gulea, I., Hart, L., Aljafari, A., Meyer, J., Morais, R., Olayemi, S., & Weber, J. (2024). Xtts: A massively multilingual zero-shot text-to-speech model. https://arxiv.org/abs/2406.04904

Computer Weekly. (2017). *Virtual reality simulation helps klm engineers escape in an emergency* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://www.computerweekly.com/news/450414139/Virtual-reality-simulation-helps-KLM-engineers-escape-in-an-emergency

ConsumentenBond. (2025, September). *Stroomprijs per kwh*. Retrieved October 24, 2025, from https://www.consumentenbond.nl/energie-vergelijken/kwh-prijs

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

Dao, T., Fu, D., Ermon, S., Rudra, A., & Ré, C. (2022). Flashattention: Fast and memory-efficient exact attention with io-awareness. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in neural information processing systems* (pp. 16344–16359, Vol. 35). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2022/file/67d57c32e20fd0a7a302cb81d36e40d5-Paper-Conference.pdf

Dettmers, T. (2022). *Bitsandbytes: 8-bit optimizers and quantization for transformers*. Retrieved October 21, 2025, from https://github.com/TimDettmers/bitsandbytes

Du, Z., Wang, Y., Chen, Q., Shi, X., Lv, X., Zhao, T., Gao, Z., Yang, Y., Gao, C., Wang, H., Yu, F., Liu, H., Sheng, Z., Gu, Y., Deng, C., Wang, W., Zhang, S., Yan, Z., & Zhou, J. (2024). Cosyvoice 2: Scalable streaming speech synthesis with large language models. https://arxiv.org/abs/2412.10117

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). Expanding explainability: Towards social transparency in ai systems. *Proceedings of the ACM on Human-Computer Interaction*, 1–19. https://doi.org/10.1145/3411764.3445188

ElevenLabs. (2025). Elevenlabs: The most realistic voice ai platform. https://elevenlabs.io/

European Commission. (2023a). *Annex iii: High-risk ai systems* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission. (2023b). *Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act)* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Parliament and Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending various Regulations and Directives (Artificial Intelligence Act) [[2024] OJ L 1689/1]. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689

Federal Aviation Administration (FAA). (2024). *Extended reality in flight attendant training* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://www.faa.gov

Fokkinga, S., & Desmet, P. (2022). Emotion typology. https://emotiontypology.com

Fu, F. (2025, February). *Openai realtime api pricing breakdown: Cost per minute analysis optimization guide*. Retrieved October 24, 2025, from https://frankfu.blog/openai/openai-realtime-api-pricing-breakdown-cost-per-minute-analysis-optimization-guide/

Geraghty, T. (2023, January). *Psychological safety: Artificial intelligence*. https://psychsafety.com/psychological-safety-90-artificial-intelligence/

Google Collab. (n.d.). *Counting audio tokens*. Retrieved 2025, from https://colab.research.google.com/gist/SivaMalasani/1ba936c13636acd9bacff63dd45ac8da/tocken-counting.ipynb

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., & et al., A. F. (2024). The llama 3 herd of models. https://arxiv.org/abs/2407.21783

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural networks [Available at https://arxiv.org/abs/1506.02626]. *Advances in Neural Information Processing Systems (NeurIPS)*, *28*.

Harvard Business School. (2020). *The triple bottom line: What it is & why it's important*. Retrieved October 28, 2025, from https://online.hbs.edu/blog/post/what-is-the-triple-bottom-line

Hexgrad. (2025a). Chattts: A generative speech model for daily dialogue. https://github.com/2noise/ChatTTS

Hexgrad. (2025b). Kokoro. https://github.com/hexgrad/kokoro

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2021). Lora: Low-rank adaptation of large language models. https://arxiv.org/abs/2106.09685

Jain, S. (2025, February). *Why data decay puts your ai strategy at risk*. https://bloomfire.com/blog/data-decay-impact-on-ai-strategy/

Jegham, N., Abdelatti, M., Elmoubarki, L., & Hendawi, A. (2025, May). *How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference.* arXiv (Cornell University). https://doi.org/10.48550/arxiv.2505.09598

Klein, G. (2023). Faster-whisper: Faster and simpler whisper inference [Optimized CTranslate2 implementation of OpenAI's Whisper model]. https://github.com/SYSTRAN/faster-whisper

KLM. (2023a). *Klm cityhopper vr tool* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://careers.klm.com

KLM. (2023b). *Virtual vitality program* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://careers.klm.com

KLM Newsroom. (2021). *Klm cityhopper introduces virtual reality training for pilots* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://news.klm.com/klm-cityhopper-introduces-virtual-reality-training-for-pilots

KLM Tech Data. (2023). *Vr training at klm — our work* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://techdata.klm.com/en/our-work/vr-training/

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., & Stoica, I. (2023). Efficient memory management for large language model serving with pagedattention. https://arxiv.org/abs/2309.06180

Laurans, G., & Desmet, P. (2017). Developing 14 animated characters for non-verbal self-report of categorical emotions]. *Journal of Design Research (online)*, *15*(3/4).

Lestary, A., Krismanti, N., & Hermaniar, Y. (2018). Interruptions and silences in conversations: A turn-taking analysis. *PAROLE: Journal of Linguistics and Education*, *7*, 64. https://doi.org/10.14710/parole.v7i2.64

Lomas, J. D., Maden, W., Bandyopadhyay, S., Lion, G., Patel, N., Jain, G., Litowsky, Y., Xue, H., & Desmet, P. (2025). Evaluating the alignment of ai with human emotions. *Advanced Design Research*. https://doi.org/10.1016/j.ijadr.2024.10.002

MDPI Electronics. (2023). Artificial intelligence in aviation [Accessed on 21 October 2025, from mdpi.com]. *MDPI Electronics*.

Mori, M., MacDorman, K. F., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics Automation Magazine*, *19*(2). https://doi.org/10.1109/MRA.2012.2192811

Mostert, E. (n.d.). *Stakeholder analysis* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://www.tudelft.nl/citg/over-faculteit/afdelingen/watermanagement/onderzoek/chairs/water-resources/water-resources-management/research/tools/stakeholder-analysis

Murata, K. (1994). Intrusive or co-operative? a cross-cultural study of interruption. *Journal of Pragmatics*, *21*(4), 385–400. https://doi.org/https://doi.org/10.1016/0378-2166(94)90011-6

Naeem, M., Smith, T., & Thomas, L. (2025). Thematic analysis and artificial intelligence: A step-by-step process for using chatgpt in thematic analysis [First published online April 21, 2025]. *International Journal of Qualitative Methods*. https://doi.org/10.1177/16094069251333886

NVIDIA Corporation. (2024). *Nvidia tensorrt: High-performance deep learning inference optimizer and runtime* [Version 10.0]. https://developer.nvidia.com/tensorrt

NVIDIA Developer. (2024). *Tensorrt-llm: High-performance inference for large language models* [Official NVIDIA documentation and source code at https://github.com/NVIDIA/TensorRT-LLM]. Retrieved October 21, 2025, from https://developer.nvidia.com/tensorrt-llm

OpenAI. (2022, September). Whisper: Openai's speech recognition model [Open-source automatic speech recognition model by OpenAI]. https://openai.com/research/whisper

OpenAI. (2025a, August). Introducing gpt-realtime and realtime api updates for production voice agents. https://openai.com/index/introducing-gpt-realtime/

OpenAI. (2025b). *Introducing gpt-realtime and realtime api updates for production voice agents*. Retrieved October 24, 2025, from https://openai.com/index/introducing-gpt-realtime/

OpenAI. (2025c). *Pricing*. Retrieved October 24, 2025, from https://platform.openai.com/docs/pricing

OpenAI. (2025d, August). *What are tokens and how to count them?* Retrieved October 24, 2025, from https://help.openai.com/en/articles/4936856-what-are-tokens-and-how-to-count-them

OpenAI, : Agarwal, S., Ahmad, L., Ai, J., Altman, S., Applebaum, A., Arbus, E., Arora, R. K., Bai, Y., Baker, B., Bao, H., Barak, B., Bennett, A., Bertao, T., Brett, N., Brevdo, E., & et al., G. B. (2025). Gpt-oss-120b & gpt-oss-20b model card. https://arxiv.org/abs/2508.10925

Qwen, : Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., … Qiu, Z. (2025). Qwen2.5 technical report. https://arxiv.org/abs/2412.15115

R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Raimondo, G. M., Locascio, L. E., U.S. Department of Commerce, N. I. o. S., & Technology. (2023). *Artificial intelligence risk management framework (ai rmf 1.0)* (tech. rep.). National Institute of Standards and Technology. https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

Resemble AI. (2025). Chatterbox-TTS [GitHub repository].

ScienceDirect. (2023). Machine learning and mixed reality in aviation training [Accessed on 21 October 2025, from sciencedirect.com]. *ScienceDirect*.

Shashkevich, A. (2018, May). *Exploring the dynamics of interruption in conversation*. Stanford University. Retrieved October 22, 2025, from https://news.stanford.edu/stories/2018/05/exploring-interruption-conversation

Smink, M., Fregonara, M., & Djajadi, N. (2025). Openvoiceagent: On-premise conversational ai for realistic and emotional dialogue. https://github.com/matteo-fregonara/OpenVoiceAgent

Tavakol, M., & Dennick, R. (2011). Making sense of cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Team, L. S. (2023). *Lm studio: Discover, download and run local llms* [Desktop app for running large language models locally (Windows, macOS, Linux).]. https://lmstudio.ai/

United Nations. (2025a). *Build resilient infrastructure, promote inclusive and sustainable industrialization and foster innovation*. Retrieved October 28, 2025, from https://sdgs.un.org/goals/goal9#targets_and_indicators

United Nations. (2025b). *Ensure availability and sustainable management of water and sanitation for all*. Retrieved October 24, 2025, from https://sdgs.un.org/goals/goal6#targets_and_indicators

VROWL. (2022). *The possibilities: 5 examples of vr training in aviation* [Accessed on 21 October 2025]. Retrieved October 21, 2025, from https://vrowl.io/blog/the-possibilities-5-examples-of-vr-training-in-aviation

Yan, M., Agarwal, S., & Venkataraman, S. (2025). Decoding speculative decoding. https://arxiv.org/abs/2402.01528

Yañez-Barnuevo, M. (2025, October). *Data centers and water consumption*. Retrieved 2025, from https://www.eesi.org/articles/view/data-centers-and-water-consumption#:~:text=While%20%E2%80%9C0%E2%80%9D%20is%20the%20ideal,a%20great%20goal%20to%20beat.

Zech, H. (2021). Liability for ai: Public policy considerations. *ERA Forum*, *22*(1), 147–158. https://doi.org/10.1007/s12027-020-00648-0

Zechner, O., Pretolesi, D., Jaespert, E., Guirao, D. G., & Tscheligi, M. (2023). Ethical considerations for ai-driven adaptive virtual environments in xr training for first responders: An industry perspective. *2022 IEEE International Conference On Metrology For Extended Reality, Artificial Intelligence And Neural Engineering (MetroXRAINE)*, 775–780. https://doi.org/10.1109/metroxraine58569.2023.10405605

Zeng, A., Du, Z., Liu, M., Wang, K., Jiang, S., Zhao, L., Dong, Y., & Tang, J. (2024). Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. https://arxiv.org/abs/2412.02612

Zhang, R., Shen, J., Liu, T., Wang, H., Qin, Z., Han, F., Liu, J., Baumgartner, S., Bendersky, M., & Zhang, C. (2024). Plad: Preference-based large language model distillation with pseudo-preference pairs. https://arxiv.org/abs/2406.02886

Zhou, S., & et al., Y. Z. (2025). Indextts2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *arXiv preprint arXiv:2506.21619*.

A

Extended Validation Results

A.1. Extended Validation Figures

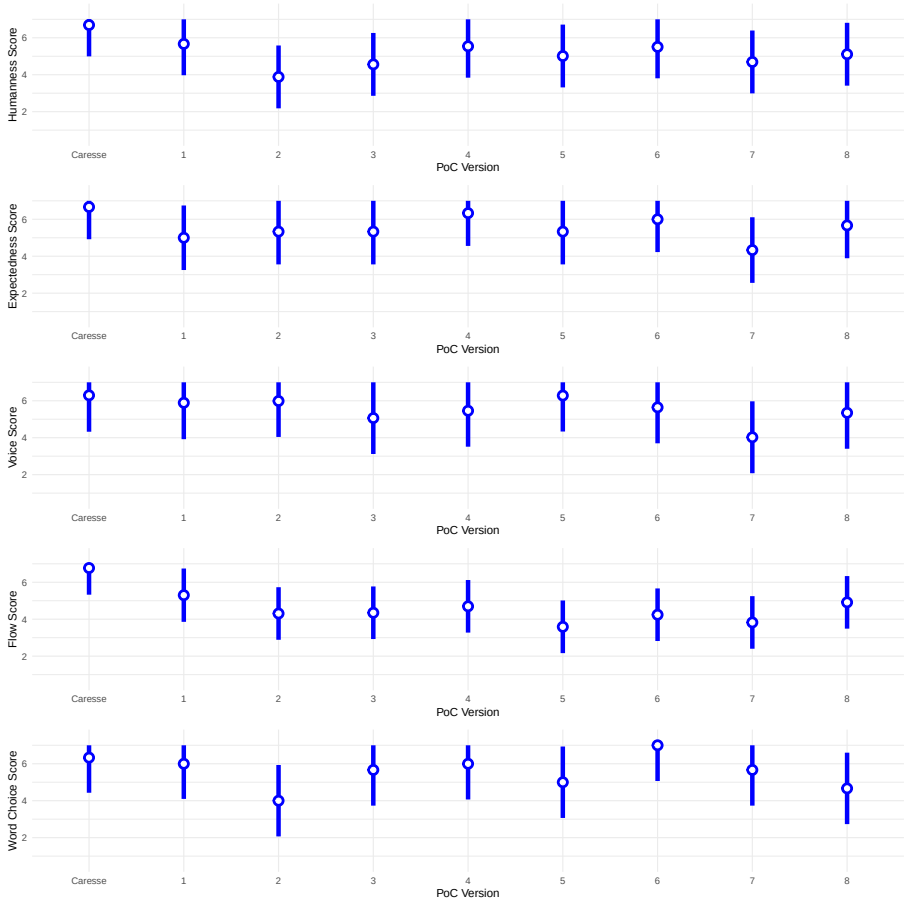


Figure A.1: PoC test results with nine participants: Estimated least square means and 95% confidence intervals of the humanness, expectedness, voice, flow, and word choice scores of the different PoC versions: 1 through 8 with Caresse as the control version. Statistically significant differences between means are labeled with '*' for p<0.05 and '**' for p<0.01. Confidence intervals are clipped to be between 1 and 7 because the five sub-scores were on 7-point Likert scales.

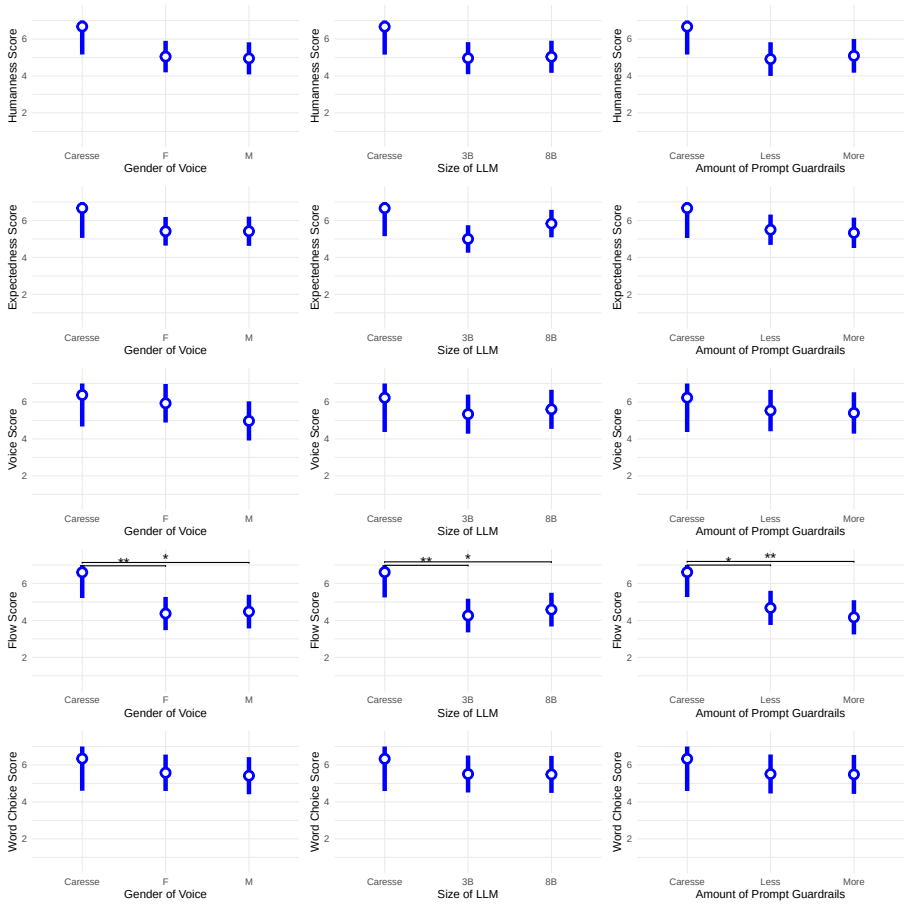


Figure A.2: PoC test results with nine participants: Estimated least square means and 95% confidence intervals of the humanness, expectedness, voice, flow, and word choice scores of the different PoC values: gender of the voice (female or male), size of the LLM (3B or 8B), and amount of guardrails in the prompt (less or more), with Caresse as the control. Statistically significant differences between means are labeled with '*' for p<0.05 and '**' for p<0.01. Confidence intervals are clipped to be between 1 and 7 because the five sub-scores were on 7-point Likert scales.

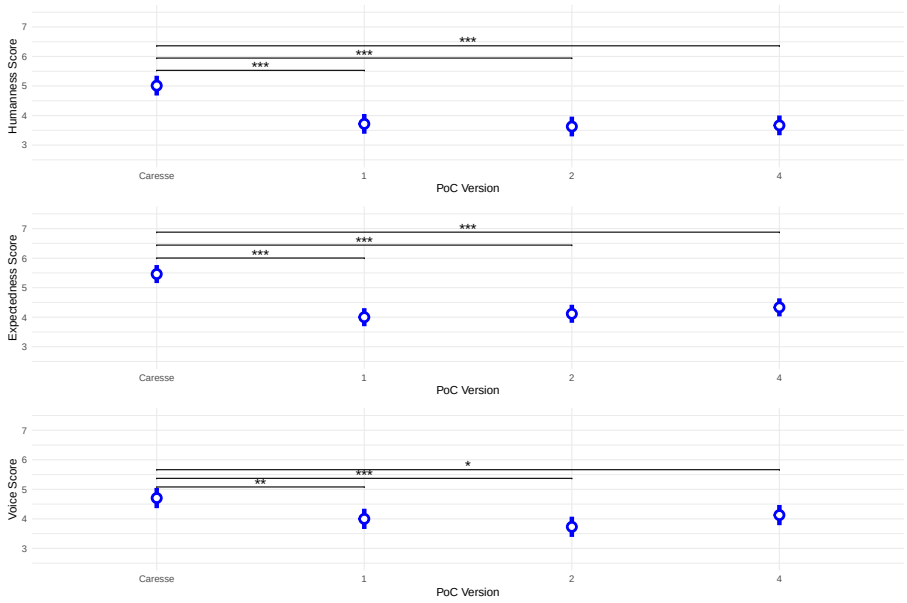


Figure A.3: Survey results with 78 participants: Estimated least square means and 95% confidence intervals of the humanness, expectedness, and voice scores of the different PoC versions: 1 through 8 with Caresse as the control version. Statistically significant differences between means are labeled with '*' for p<0.05, '**' for p<0.01, and '***" for p<0.001. Confidence intervals are clipped to be between 3 and 7 because the five sub-scores were on 7-point Likert scales and no intervals went below 3.

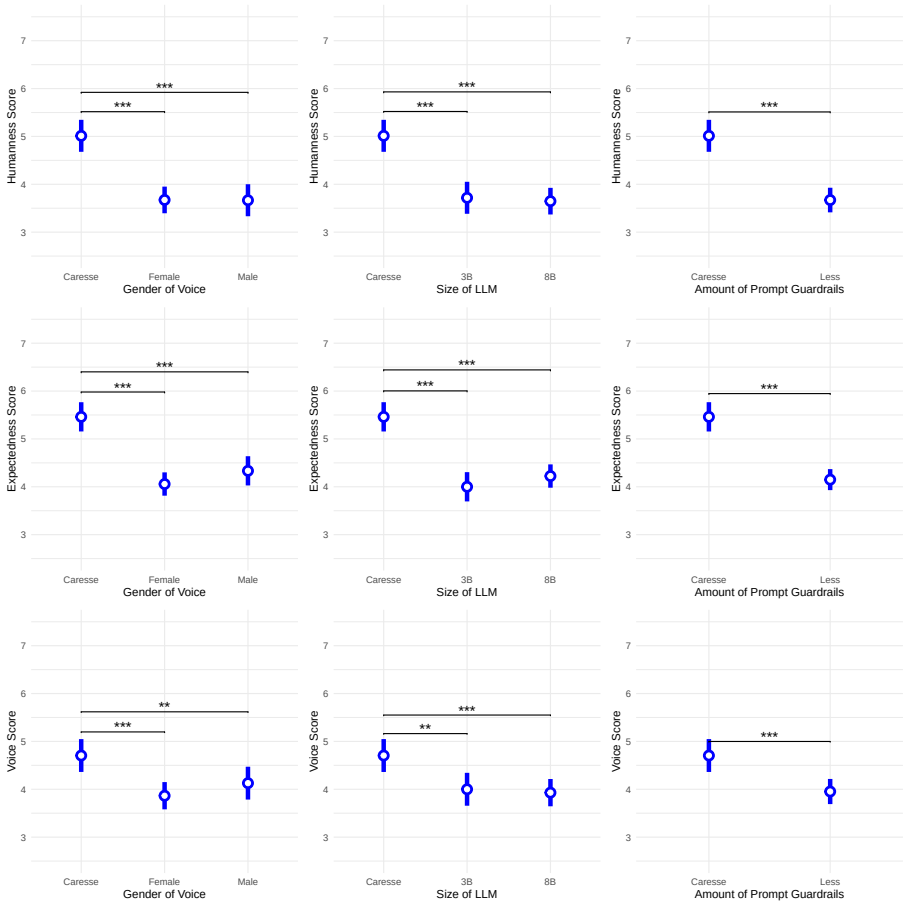


Figure A.4: Survey results with 78 participants: Estimated least square means and 95% confidence intervals of the humanness, expectedness, voice, flow, and word choice scores of the different PoC values: gender of the voice (female or male), size of the LLM (3B or 8B), and amount of guardrails in the prompt (less or more), with Caresse as the control. Statistically significant differences between means are labeled with '*' for p<0.05, '**' for p<0.01, and '***" for p<0.001. Confidence intervals are clipped to be between 3 and 7 because the five sub-scores were on 7-point Likert scales and no intervals went below 3.