Comparing Human and AI Emotion and Sentiment Analysis in Mixed Reality and Zoom Environments

University of Turku
Department of Computing
Master of Science (Tech) Thesis
TurkuNLP & FTLab
July 2024
Maryam Teimouri

UNIVERSITY OF TURKU
Department of Computing

MARYAM TEIMOURI: Comparing Human and AI Emotion and Sentiment Analysis in Mixed Reality and Zoom Environments

Master of Science (Tech) Thesis, 56 p., 4 app. p.
TurkuNLP & FTLab
July 2024

While Mixed Reality devices and platforms have entered the digital market, the question of their effectiveness in providing remote presence remains unanswered. In this thesis, we aim to challenge two trending and rapidly growing technologies: Mixed Reality and Large Language Models (LLMs). By designing a case study focused on education—a key target for Mixed Reality—we aim to measure immersion using sentiment analysis based on Plutchik's Wheel of Emotions. This analysis requires Finnish language skills since the study is conducted in Finland and in Finnish. Unlike our previous case study, which relied on translations by Finnish speakers, this research incorporates a pipeline that leverages LLMs to assist an English speaker in overcoming the language barrier. We then compare the results of the collaboration between an English speaker and AI against those of native Finnish speakers.
Eventually, the Mixed Reality experience is categorized as an immersive experience, achieving a sentiment analysis rate of 0.75 for immersion. Additionally, the English speaker's performance was found to be 19% less effective than that of the Finnish speakers and 24% less effective than using LLMs alone.

Keywords: Language Models, Mixed Reality, Sentiment analysis, Immersion

Contents

List of Figures

List of Tables

1 Introduction

By the evolution of digital communication, Mixed Reality (MR) platforms and video conferencing tools like Zoom have significantly altered the way individuals interact and engage with each other. This transformation is particularly notable among younger generation, where these technologies have become essential to not only entertainment but also education and socialization. The immersive nature of MR and the wide accessibility of video conferencing present convenient environments for communication, while these two have different features to offer. Understanding how these environments influence interaction patterns, language use, and social engagement can provide valuable insights into the development of effective digital communication tools and strategies.

This thesis, while focusing on education, seeks to explore the dynamics of children's communication behaviors by designing a case study that includes gameplay on a Mixed Reality (MR) platform. To better understand the impact of MR, the same design was also run on a non-reality, 2D communicative platform—-specifically, the Zoom online platform. There are many ways to analyze and study a design and experiment, but in this case, we decided to start with analyzing the dialogues from children's gameplay sessions in both MR and Zoom environments. Since the experiment design and conditions were not as controlled as a studio audio setup, we decided not to use signal processing techniques. Instead, we focused on the words and content of the dialogues. With the remarkable progress of Large Language Models, it is possible to convert the audio to transcription and text. Some data processing stages might be required for the text if the data quality is not convenient.

For exploring human language, researchers need to understand the language. In this

case, the experiment is conducted in Finland and in Finnish, while many researchers in the group do not speak the language. Previously in case study 1, Finnish speakers translated the entire study into English, which was a time-consuming process. Instead, we decided to use machine learning models that have knowledge of human languages to assist researchers, particularly English speakers, in conducting their research. To process the data from audio to usable English text, Large Language Models (LLMs) are used in several stages: transcription (converting audio to text), correction (refining the transcribed output to eliminate gibberish parts, which may occur due to differences between spoken and written Finnish), translation, speaker identification, and emotion detection. At each stage, the accuracy of the LLMs needs to be measured and analyzed to draw meaningful conclusions. The assessments are followed by the research questions. The thesis is guided by a set of research questions:

1. The Role of Transcription (audio to text): How does auto transcription affect the analytical insights derived from the data? To what extent does transcription accuracy influence the interpretive validity of linguistic patterns and communicative behaviors identified through NLP and GPT model analyses?

2. Impact of Translation on Analysis: How does the translation of dialogues from Finnish to English affect the analysis of communication patterns? Can these automatic tools effectively distinguish between native Finnish and English speakers, and what connections does this have for the study of bilingual or multilingual communication in digital settings?

3. Effectiveness of Automatic Tools: How effectively can current automatic tools, including NLP and GPT models, analyze and extract information from children's dialogues in digital environments? What are the limitations and strengths of these technologies in processing and understanding the use of child language and interaction?

4. Identification of Communication Contexts and Speakers: Can advanced NLP tools, particularly LLMs, accurately distinguish between dialogues occurring in MR and

Zoom environments? Is it possible to identify the speaker or the medium based solely on the analysis of their speech patterns and language use?

5. Differences Between Mixed Reality and Zoom Interactions: Are there noticeable differences in communication patterns, linguistic features, and social dynamics between groups of children interacting in MR versus those on Zoom? If so, to what extent can these differences be quantified, and what do they reveal about the influence of each platform on children's communication behaviors?

Through addressing these questions, this thesis aims to not only clarify the impacts of digital environments on children's communication but also to evaluate the potential of current NLP technologies to advance our understanding of language and interaction in the digital age. To answer these research questions, there are many tasks that can be done with prompting the GPT models, including:

- Translation and Transcription Refinement: Language models can assist with translating or cleaning up the experiment's transcripts that need to be clarified or are in a different language. This guarantees the data is prepared for examination.

- Sentiment Analysis: GPT models can perform a initial sentiment analysis by processing over the transcripts. This could entail interpreting the students' responses to determine if they were feeling happy, sad, or neutral, providing insight into how they were feeling while playing the activities [1].

- Theme Identification: Identifying recurrent topics or keywords by looking over the transcripts is also a task to be assisted by Language models. This may demonstrate how the MR and Zoom groups differ in terms of communication methods, degrees of participation, and emotional responses.

- Comparative Analysis: Highlighting significant variations or parallels between the two groups' interactions, linguistic styles, and degrees of engagement is an important piece of work in this research that GPT models can assist with.

- Summarizing: GPT models can produce summaries of the studies' long transcripts that preserve key information while distilling the spirit of the children's conversations and input.

- Coding for Qualitative Analysis: Language models can assist in creating a coding system for your study. As an example going through the transcripts and classifying important words or ideas in order to make theme analysis easier.

- Creating Interview Questions: Creating questions is also a task that LLMs are good at. That helps in going deeper into the experiences and perspectives of the children to do follow-up interviews.

- Literature Review: Including summarizing or explaining study methods, conclusions, and theories from previous psychological, educational, or technological studies that may be relevant to the project [2].

In this thesis, we are focusing on the first 4 tasks and assess how good LLMs and GPT models are at them. Flowchart in Figure 1.1 visualize how the data processing stages are organized and how the tasks are done.

Figure 1.1: Data flow

2 Background

In this chapter, we delve into the foundational knowledge necessary for understanding this thesis, providing a brief review. Additionally, references are provided for those interested in further exploring these topics. Since we are conducting sentiment analysis towards immersion using emotions, I begin this chapter by defining what emotions and sentiments are. Next, we review the AI and ML models, techniques, and structures used at each stage of our data flow (refer to 1.1). Finally, I discuss the technology and setup of our mixed reality platform, introducing each component in detail.

2.1 Sentiment and Emotion

Because they are both influenced by biology, cognition, and social context; emotions and sentiments are sometimes used as synonyms–yet, they are two different concepts. Short-lived, episodic reactions to certain situations or stimuli, emotions are marked by rapid changes in behavioral, autonomic, and brain functions. Sentiments, on the other hand, are long-lasting psychological predispositions or dispositions to respond to stable emotional, cognitive, and related responses to particular objects or circumstances. Compared to emotions, they are less dependent on particular occasions, more dispositional, and more consistent. While emotions can arise spontaneously and are not always directed toward an object, sentiments are developed and directed toward an object. All things considered, sentiments and emotions are essential components of the human experience, but they are not the same in terms of length, stability, or target orientation [3].

Words can be confusing when defining emotions and sentiments, but Figure 2.1 visualizes these concepts more clearly and easily.

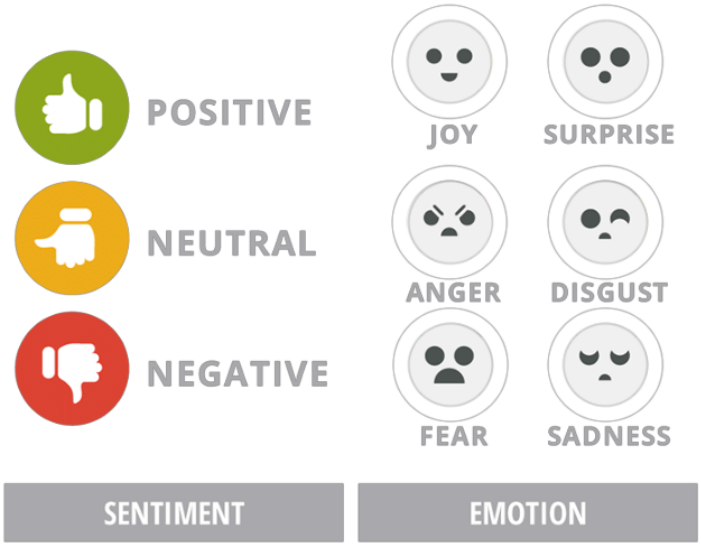


Figure 2.1: Comparison of sentiment and emotion[4]

2.2 Artificial intelligence and Machine learning

The science of creating computers and other devices with the ability to think, learn, and behave in ways that would typically need human intelligence or entail data sets too large for people to process is known as artificial intelligence. AI is a collection of technologies used for data analytics, forecasting, object classification, natural language processing, data retrieval, and other business applications. These technologies are mostly based on machine learning and deep learning. [5]

By supplying a system with enormous quantities of data, machine learning—a subset of artificial intelligence—allows it to learn and grow on its own utilizing deep learning and neural networks without needing to be explicitly programmed. Computer systems may continuously improve and adapt as they get additional "experiences" thanks to machine learning. Therefore, by giving these systems access to more and more diverse datasets to process, their performance can be enhanced [6]. Figure 2.2 visualizes the relationships and overlaps between AI, ML, and DL.

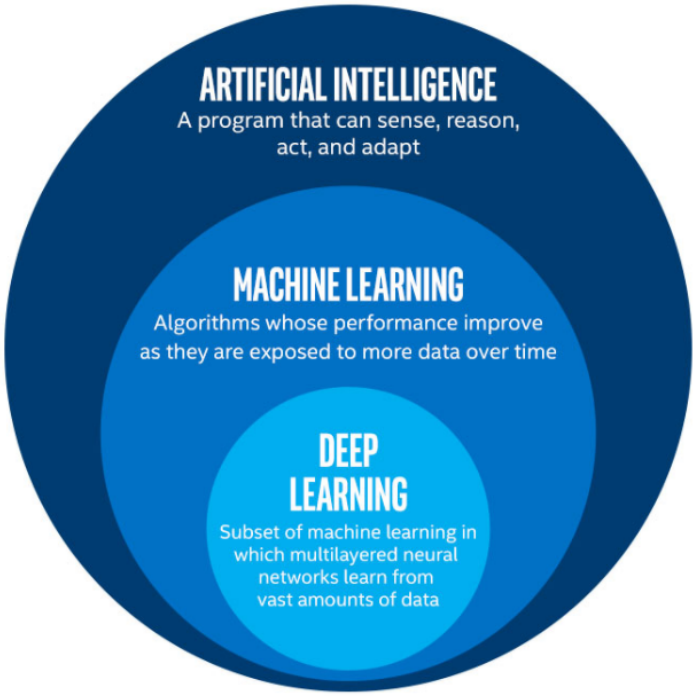


Figure 2.2: comparing artificial intelligence, machine learning, and deep learning[7]

2.3 Human Language Technology

Human Language Technology (HLT) involves using smart electronic devices and software to communicate with people by understanding human language and generating responses. One popular area within HLT is Natural Language Processing (NLP), which combines computer science and language study to develop methods for processing language automatically.

2.3.1 Natural Language Processing

A subfield of artificial intelligence (AI), called natural language processing (NLP), gives computers the ability to understand, produce, and modify human language. Natural language processing can be used to query the data using voice or text in natural language. NLP is applicable to all of the human languages and can be used with both written and spoken language and data. Here are some of the NLP applications that are used in this thesis: Sentiment Analysis, Text Classification, Machine Translation, and Speech Recognition.

2.4 Transfer Learning

In the context of machine learning models, transfer learning is using the structure or information from one learning problem to improve learning on a related problem. Using least squares regression, Jeremy West et al. created a formal explanation of inductive transfer for both linear and non-linear models. They demonstrated the sufficient and required conditions for inductive transfer to improve learning accuracy. Lastly, they used a variety of learning techniques to demonstrate transfer in both artificial and real-world scenarios, thereby providing empirical validation for the outcome [8].

2.5 Generative pretrained transformers

Sentiment analysis, document classification, question answering, and textual entailment are just a few of the activities that fall under the umbrella of natural language understanding. Large unlabeled text corpora are widely available, but labeled data for these tasks is scarce, which makes it difficult for task specific trained models to predict well. By generatively pre-training a language model on a wide corpus of unlabeled text first, and then discriminatively fine-tuning on each individual task, significant improvements on these tasks can be obtained. Effective transfer can be achieved with minimal changes to the model architecture by using input transformations for each task during fine-tuning [9].

2.5.1 Transformers

The Transformer model offers a structured memory that manages long-term dependencies in text more effectively than recurrent networks. It is well-known for its effectiveness in tasks like machine translation, speech to text, and text classification. Strong transfer performance is produced by this structural advantage in a variety of jobs. We use task-specific input adaptations for transfer learning that come from traversal-style methods, which handle structured text input as a single token sequence. Our tests show that these modifications allow for efficient fine-tuning with very small architectural changes to the

pretrained model [9].

2.5.2 Pretrained models

Reducing the need for supervised learning in natural language processing (NLP) requires effective learning from unprocessed text. Large amounts of manually labeled data are required for many deep learning algorithms, which limits their use in domains with insufficient annotated resources. In these situations, linguistic information extraction models from unlabeled data provide a useful substitute for expensive annotation efforts. Quality representations learned unsupervisedly can significantly improve performance even in the presence of extensive supervision. This is highlighted by the significant improvements in NLP task outcomes achieved using pretrained word embeddings [9].

2.5.3 Generative models in NLP

From Natural Language Processing (NLP) perspective, generative AI is the technology that makes it possible to produce text or voice that sounds like human language. Generative AI models can produce new material based on patterns they discover from large datasets, in contrast to standard AI models that examine and process already-existing data. These models make use of sophisticated neural network architectures and methods, frequently using Transformers or Recurrent Neural Networks (RNNs) to comprehend the complex linguistic structures. For a better understanding, refer to Figure 2.3.

By understanding context, grammar, and semantics, generative AI models are able to produce content that is coherent and relevant to the given context. They are vital resources for a wide range of applications, including code development, language translation, chatbots, and content production. [11]

2.5.4 GPT3 and GPT4

Both GPT-3 and GPT-4 are state-of-the-art models developed by OpenAI. In the following paragraphs, we will take a closer look into the differences between these two models.

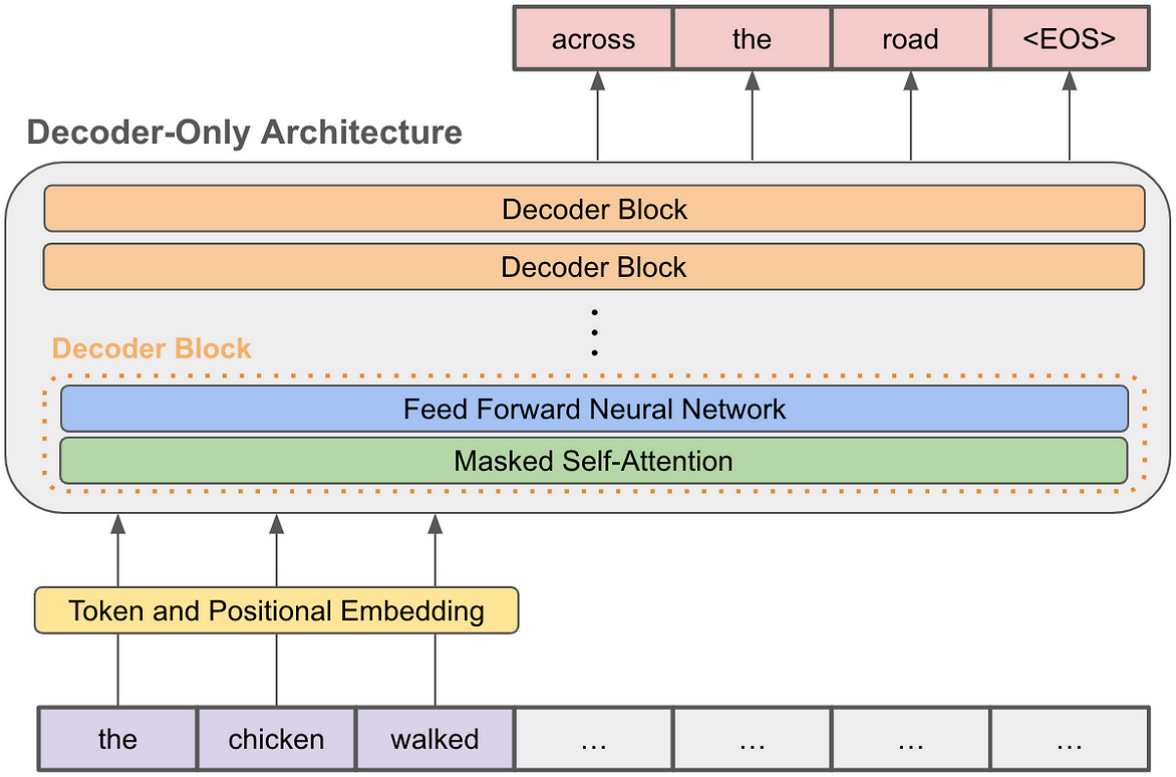


Figure 2.3: Generative pretrained transformers architecture [10]

- GPT3

  OpenAI created the cutting-edge GPT-3 language model. GPT-3 is based on the transformer architecture, similar to GPT-2, but larger in scale. With 175 billion parameters, GPT-3 has a much larger number of parameters than its predecessor, GPT-2. The model can execute a range of natural language processing tasks in few-shot, one-shot, and zero-shot scenarios while requiring little task-specific training data, including; language translation, summarizing, and generating code. Although GPT-3 is capable of producing findings that are believable yet sometimes outputs are inaccurate or biased.[1]

- GPT4

  The advanced language model GPT-4 replaces GPT-3.5. Improvements in factual accuracy, safety, and the capacity to manage challenging jobs are all included in GPT-4. Compared to GPT-3.5, GPT-4 has improved abilities, such as better results in professional and academic assessments. GPT-4 leverages a large-scale

web-scale corpus of licensed and publicly accessible data that has been refined by reinforcement learning from human feedback (RLHF). With the help of more than 50 specialists, significant improvements are made to lessen biases and harmful outputs, testing and refining the model through adversarial testing.[12]

Feature/Aspect	**GPT-3**	**GPT-4**
Release Date	June 2020	March 2023
Model Size	175 billion parameters	Not publicly disclosed, but significantly larger and more advanced than GPT-3
Training Data	Vast corpus of internet text	Expanded and more curated dataset, including diverse sources and licensed data
Learning Paradigm	Few-shot, one-shot, zero-shot learning	Enhanced few-shot, one-shot, zero-shot learning with better generalization
Factual Accuracy	Can generate plausible but incorrect info	Improved factual accuracy and reduced hallucinations
Safety Mitigations	Basic safety measures	Advanced safety measures with significant reduction in harmful outputs
Reasoning Abilities	Competent in various tasks	Superior reasoning and complex problem-solving abilities
Fine-Tuning	Limited fine-tuning with RLHF	Extensive fine-tuning with RLHF and additional safety reward signals
Benchmark Performance	Strong performance on NLP tasks	Outperforms GPT-3 on benchmarks such as TruthfulQA and professional exams
Applications	Language translation, summarization, code generation, etc.	Enhanced applications, including improved code generation and understanding of complex queries
Bias and Fairness	Contains biases from training data	Efforts to mitigate biases and improve fairness through expert testing and feedback

Table 2.1: Comparison between GPT-3 and GPT-4. Sources: [1], [12]

2.6 Annotation Agreement

In this study for comparing human and AI performance, a consistent set of annotations created by humans is required to be compared with AI and language models. It's important that all annotators agree on these annotations to build a thorough database. This means measuring agreement among annotators and improving it through the annotation process is crucial.

Inter-annotator agreement (IAA) and inter-annotator reliability (IAR) are two concepts that use various metrics to assess agreement among annotators. IAA metrics measure the extent to which annotations from different annotators are similar on the same data. Specifically, they indicate how well annotators adhere to guidelines, standards, and criteria. On the other hand, IAR delves deeper into annotations, also evaluating their correctness and validity.

Let's discuss some of the IAR metrics that will be used in this study. Cohen's kappa is a metric used to measure annotation agreement between two raters for classification and categorical tasks [13]. Fleiss' kappa assesses the reliability of annotation agreement for classification tasks over categorical data among multiple annotators [14].

- Cohen's kappa

  $\kappa = \frac{P_o - P_e}{1 - P_e}$

  - P_o is the observed agreement among raters.
  - P_e is the expected agreement (chance agreement).

- Fleiss' kappa [14]

  $\kappa = \frac{\bar{P} - \bar{P}_e}{1 - \bar{P}_e}$

  - $\bar{P}$ is the overall observed proportion of agreement among all raters.
  - $\bar{P}_e$ is the overall proportion of agreement that would be expected by chance.

In this study, Fleiss' Kappa is utilized. Let us examine the interpretation table outlining the range of values in this assessment and their corresponding meanings.

Kappa Value	Strength of Agreement
<0	Poor agreement
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

Table 2.2: Interpretation of Fleiss' Kappa Values: [15]

As the table 2.2 indicates, the values below 0 show no or poor agreement.

2.6.1 Tokenization

The process of dividing a text into smaller units is known as tokenization. The objective of this task is to prepare the text for subsequent text mining tasks through prepossessing. Given the variety of languages used in this study and the potential variation in sentence or dialogue lengths, it is essential to tokenize both the texts and their corresponding labels to ensure a valid assessment and apply the alignment.

Let's have an example. The sentence "I was supposed to be here last week, but I got a cold." will be tokenized into "I", "was", "supposed", "to", "be", "here", "last", "week", "but", "I", "got", "a", "cold".

2.7 Mixed Reality

Mixed Reality (MR) is a blend of Virtual Reality (VR) and Augmented Reality (AR) within the reality-virtuality continuum, a concept introduced by Paul Milgram and Fumio Kishino in 1994 [16]. VR immerses users entirely in virtual environments, while AR overlays digital content onto the physical world. Positioned between these two paradigms, MR facilitates interactive environments where physical and virtual objects coexist and interact, leveraging real-world physics. For a clearer understanding of the differences between MR, VR, and AR, refer to Figure 2.4.

To create a Mixed Reality platform, several components are essential:

1. **Sensors and Cameras:** Essential for capturing and tracking the user's environ-

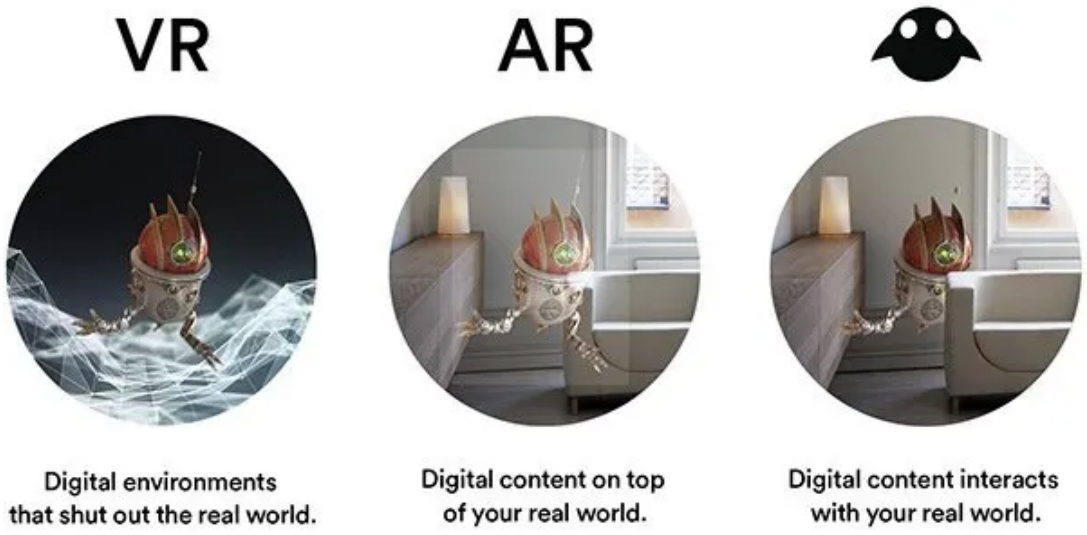


Figure 2.4: Comparing Mixed Reality, Augmented Reality, and Virtual Reality [17]

ment, facilitating communication within the platform. This study employs Lidar cameras for precise environmental mapping.

2. **Display Devices:** Devices such as the Microsoft HoloLens, utilized in this research, project digital images into physical space as three-dimensional holograms, enhancing immersive experiences.

3. **Software and Algorithms:** Crucial for real-time interaction and integration of data from the physical environment into virtual constructs [18], [19].

4. **Processing Power:** Requires robust computational capabilities to manage complex interactions between physical and digital elements, particularly in multi-user scenarios involving two computers.

Despite its potential, MR faces several challenges:

1. **Technical Limitations:** High processing power requirements and issues with latency and accuracy in digital overlays. In this experiment, we utilized powerful computers with dual graphic cards, yet encountered glitches and hangs. This raises concerns among researchers about the technology's readiness for rigorous studies.

2. **Cost:** Development of the platform and software, along with hardware costs, can

be prohibitive for widespread adoption [20]. This study is conducted in Finland; however, not every region globally may afford the necessary investments.

2.7.1 Unreal Engine

Unreal Engine, a comprehensive tool developed by Epic Games [1], facilitates the creation of interactive digital and virtual experiences. Originating in 1998, it has evolved into a competitive game engine widely utilized across diverse industries such as gaming and entertainment. Unreal Engine is distinguished by its high-quality graphics, advanced physics simulations, and real-time rendering capabilities, making it an invaluable resource for developing video games, virtual reality (VR) experiences, and Mixed Reality (MR). It supports a wide array of platforms, including desktops (Windows, Linux) and VR headsets such as Hololens. Its intuitive Blueprint visual scripting system enables rapid prototyping and development without extensive programming expertise, complemented by a robust C++ API that empowers experienced developers to create highly customized and optimized applications [21]. If you are interested in seeing how blueprints can be programmed, refer to Figure 2.5. The widespread adoption and continuous innovation of Unreal Engine underscore its pivotal role in modern digital content creation and interactive media development [22].

2.7.2 Lidar Cameras

Lidar (Light Detection and Ranging) cameras are integral components of Mixed Reality platforms, utilized to capture both the user and the surrounding environment for projection as holograms. These cameras employ laser technology to generate various types of images: RGB for capturing scene colors akin to standard 2D images, depth for creating spatial depth and facilitating the generation of 3D digital objects. By emitting laser pulses from a source and measuring their reflection time and speed, Lidar cameras accurately determine distances. For this experiment, Intel RealSense Lidar cameras [2]

[1] `https://www.unrealengine.com/en-US`

[2] `https://www.techinsights.com/blog/inside-intel-realsense-l515-lidar-camera`

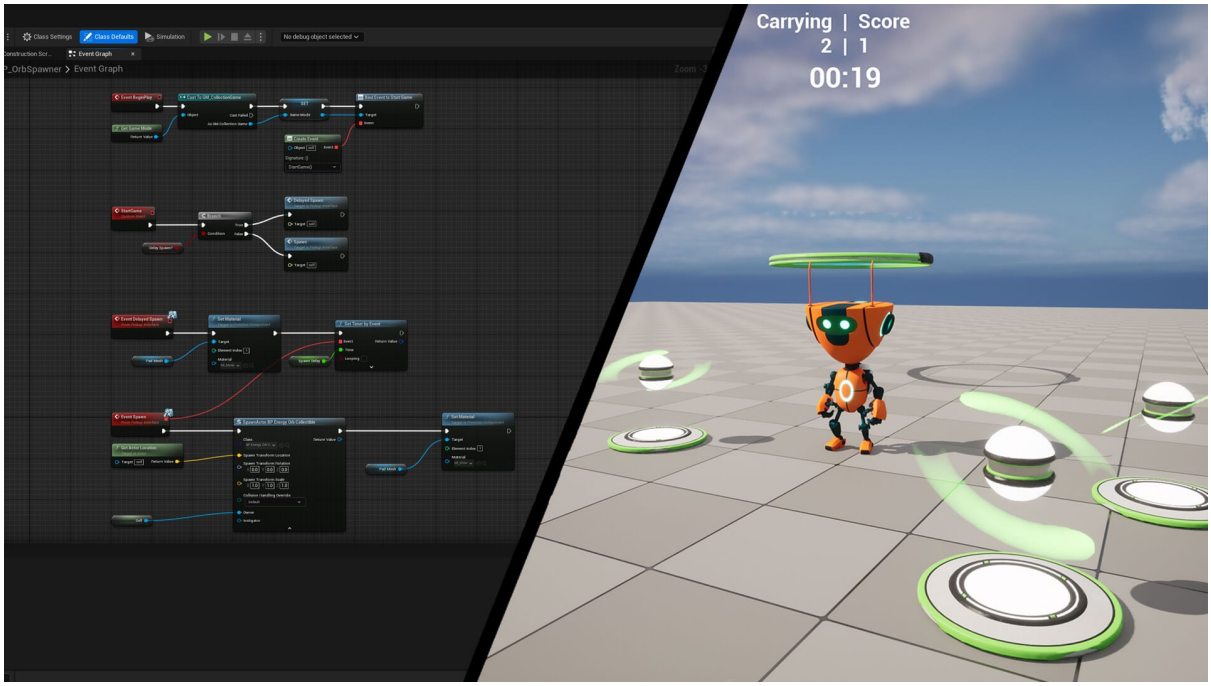


Figure 2.5: Blueprints and unreal engine [23]

were employed, known for their user-friendly text interface and straightforward configuration settings tailored for application development. You can see how the interface and configurations look in Figure 2.6.

2.7.3 Hololens headsets

Hololens, developed by Microsoft, serves the purpose of augmented reality (AR) and mixed reality (MR). Unlike virtual reality (VR) headsets, Hololens does not entirely transport users into a virtual environment; instead, it integrates virtual elements into the real world. Operating on its own Windows Mixed Reality OS, Hololens features a suite of hardware including microphones, laser sensors, 3D speakers, and high-quality cameras. These components enable the headset to capture both user vision and the surrounding environment, enhancing the mixed reality experience.

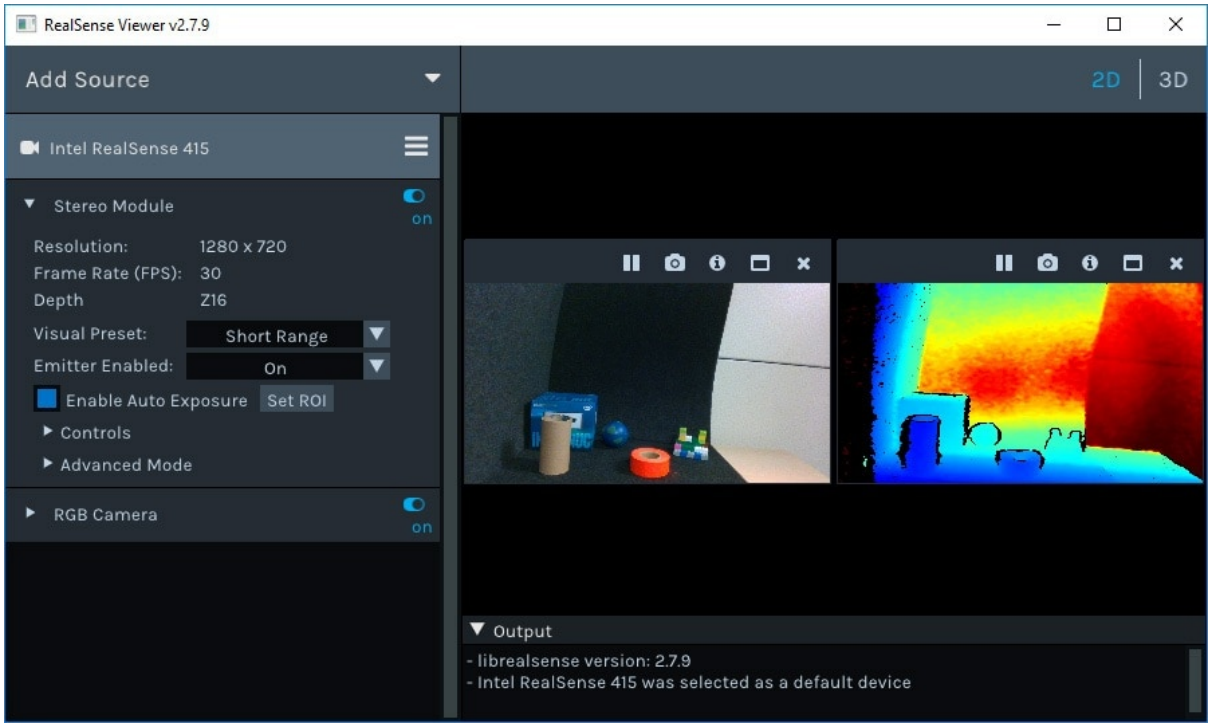


Figure 2.6: Realsense camera image viewer [24]

3 Related Work

In this section, I review the literature relevant to my thesis, focusing on both technological and scientific aspects. The discussed concepts and technologies include those employed in my research as well as previous state-of-the-art methods and related technologies utilized in earlier designs by the research group. In the 3.1 section, I review the Lidar Intel RealSense camera and Hololens headsets that are used in the Mixed Reality setup. Sections 3.2 and 3.3 include related papers for the tasks involved in this research; speaker identification and sentiment analysis towards immersion in education.

3.1 Technology

With an emphasis on their functional characteristics, operational methods, and performance expectations, the paper "Intel RealSense Stereoscopic Depth Cameras" provides a comprehensive review of Intel's stereoscopic RGBD imaging systems. It examines the optical properties and correlation algorithms utilized by these systems, illustrating their impact on applications such as gesture recognition and 3D reconstruction. Specifically, the study discusses the optical features, noise handling capabilities, and algorithmic performance of the Intel RealSense R200 and D400 series cameras using standard datasets like Middlebury. This detailed overview aids in understanding the practical limitations and potential of these depth cameras [25].

An in-depth review of Microsoft HoloLens's utilization in various industries from 2016 to 2020 is presented in the document "Review of Microsoft HoloLens Applications over the Past Five Years." It examines 44 research articles to detail the applications of HoloLens in fields such as industrial engineering, architecture, civil engineering, medical education,

and surgical and medical aids. The study emphasizes the enhancement of visualization methods and the utility of HoloLens across different sectors. This analysis provides significant insights into the effectiveness and potential of HoloLens technology, highlighting trends and the current state of research in a variety of industrial applications [26].

3.2 Immersion and Education

Speaking of sentiment analysis towards immersion, Ruíz Gándara África et al. conducted a study in a title of the immersive Van Gogh exhibition [27]. They applied text mining, emotion analysis, and sentiment analysis algorithms to identify the triggered cognitive feature while using the virtual reality experience.

Emotion analysis can be used for accelerating the learning process. Learning process can be challenging and instructors constantly look for ways to assess how students are doing. Students will learn easier if they experience positive emotion instead of negative. Applying sentiment analysis and monitoring emotions is one way to address this challenge [28].

3.3 Speaker Identification Task

To improve speaker identification, a novel approach titled "Enhancing Speaker Diarization with Large Language Models: A Contextual Beam Search Approach" integrates large language models (LLMs) with traditional acoustic-based speaker diarization systems. This method leverages the contextual expertise of LLMs to refine speaker probabilities in dialogue scenarios. By employing a beam search decoding process that combines both auditory signals and lexical information from an LLM, the proposed approach significantly enhances the accuracy of speaker diarization. The study demonstrates that incorporating LLMs can reduce diarization errors and improve the overall effectiveness of speaker identification systems across various conversational contexts [29].

4 Study setup and Data collection

The study was designed by the research team, with careful consideration given to room arrangement and game activities, taking into account the limitations of the technology. Participation in both the study and interviews was voluntary.

4.1 Study setup

We compared two groups in the study. In the first group, the subjects interacted with one another via TV displays as well as conventional cameras, microphones, and speakers. They viewed each other as life-size images on the screen. In the second group, participants saw each other as life-size holograms while wearing Microsoft Hololens 3D cameras and communicating with one another. Within their groups, the subjects took part in the study in pairs.

4.1.1 Setup preparation

With the use of our software, two rooms were linked. One of the students in room type A wore a Hololens and had a 2D camera. For this student, the other pupil was a holographic image. The students had access to a tablet with a shared whiteboard to draw. The tablet is supported by a stand in the room, allowing the students to set it down when they are acting out the role they have in the game.

A student in room type B was using a 3D camera to observe another student through a large television screen. We used one A-type room and one B-type room per pair in the study group (A–B). In the study's control group (B-B), both of the students in the pair

used a type B room.

Figure 4.1: A-B group and setup

4.1.2 Preparation of the study subjects

Before the study starts, subjects went through training sessions to get to know the technology, devices, and how to wear and use hololens. It is important that they know how to use the whiteboard to draw and interact with the camera.

Very little emphasis was placed on the equipment being used. Instead of instructing students to "test the system," the focus was on engaging in interaction with each other. The goal of the test is to find the depth of students' immersion in interaction with each other, whether facilitated by technology or not. Therefore, we aim to preserve this immersion by minimizing students' awareness of the technology.

In advance 4-8 animals or animated characters which students were already familiar with were prepared by printed photos of those.

Figure 4.2: Hololens headset view

4.1.3 Study action

After the equipment was tested and prepared, the student pairs were placed in their rooms. To be safe, in case of emergencies or unexpected surprises, such as a student becoming upset or just expressing their desire to interrupt the test, the experiment was monitored via one of the video feeds.

Both of the two pupils in each group were given four pictures initially. Each of them had a turn. One student role-played, imitating the gestures and speech patterns of the figures that were initially illustrated. After hearing and seeing explanations and enactments, their partner drew on the board and tried to figure out what was shown in the drawing. After that, they reversed roles, with one child acting out scenes and the other guessing and drawing. The examiners entered the room and ended the play after ten minutes of this procedure.

- Study group (A–B)

Students in room type B were initially given four photographs, and their task was to describe each one while mimicking the gestures and speech patterns of the characters. This student can watch what a student in room type A is doing in real time on a TV screen and there is a tablet screen that displays what the student is drawing on their whiteboard. Depending on what they perceive, this student then makes adjustments and cues a novelty in a timely manner.

Following the actions and descriptions of the students in room type B, students in room type A who were wearing Hololens also received four photos at the start of the activity. Based on their guesses, they sketched animated characters or animals on a digital whiteboard.

In the beginning, four photos were given to each student simultaneously. Students in this group advance to the next photo if they correctly guess the first one, and so on. Each couple had five minutes. Five minutes later, the two pupils' roles were reversed.

Students in room type A who were wearing Hololens placed the whiteboard down in their hands, took another set of photos (which were different from the first set), and repeated the previous students' activities from room type B.

Based on the TV screen's descriptions and actions of the students in room type A, students in room type B guesses the names of animated characters or animals to draw on a digital whiteboard.

- Control group(B1–B2)

  Students in room type B1, which was chosen at random, were given four photos and asked to describe each one while mirroring the gestures and facial expressions of the characters. This student had access to a shared whiteboard where they could view what another student in room type B2 was sketching in real-time, as well as the activities of the other student on a TV screen. Depending on what they perceive, this student then makes adjustments and cues a novelty in a timely manner.

  On a digital whiteboard in room type B2, students made guesses about animated

creatures or animals based on the descriptions and behaviors of other students in room type B1.

In the beginning, four photos were given to each student simultaneously. Five minutes later, the two students' roles were reversed.

4.2 Data collection

We recorded the video and audio feeds from all the devices. One challenge during setup was that both the Lidar cameras and Hololens headsets use lasers, which cannot operate in the same space without interfering with each other. To avoid this issue, we turned off the laser and deep image capturing features on the Lidar cameras when using them alongside the Hololens, resulting in the capture of simple 2D videos instead. Additionally, the Hololens headsets have cameras that recorded the experience.

The Hololens headsets have cameras that capture what the wearer sees, while the Lidar cameras record what the other person sees on the opposite side. Moreover, as a backup plan, we positioned a camera to capture the entire room and the overall experience.

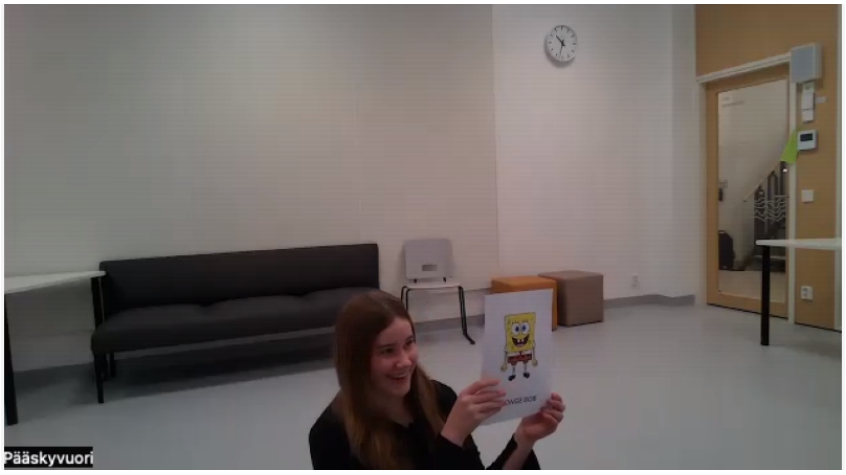
For the control group, the communication platform was Zoom, so Lidar cameras with 2D video capturing were used and the footage was transferred through the platform. To capture the experience, the video call was recorded on the university server [1].

Hololens headsets have microphones and speakers that work very well. However, for the participants using Lidar cameras, portable microphones were provided. Additionally, monitors had speakers for participants to hear the other party.

We also conducted a survey and interview with the subjects. The setup was designed to be simple and cozy so that interviewees felt welcomed and relaxed. A camera captured the whole conversation from a few meters away, and to ensure better voice quality for speech recognition tools, a microphone was set up at a central location within the group.

This material is saved on the University of Turku's Seafile storage. All the videos were edited and trimmed later for use in the study. For NLP processes, the video files

[1] `https://utu.zoom.us/`



(a) B1

(b) B2

Figure 4.3: B-B group and setup

were converted to audio files.

5 Data Processing and Tools

In this chapter, we will delve into the data flow and processing. Our objective is to generate English transcriptions from experimental videos originally in Finnish. We will thoroughly examine the decision-making process behind tool selection and usage.

5.1 Data policy

The original study materials are exclusively stored on the University of Turku Seafile service, with restricted access solely for the researchers involved in this study. The data will be removed after the study ends, and at the maximum, within 5 years. The materials will not be shared with anyone, including school students, teachers, and parents. Throughout the duration of the study, the materials will remain within Europe and will not be uploaded to any servers outside of this region.

5.2 Initial Data: Videos

The original material comprises multiple 10-minute videos. The study involved 5 pairs, with 3 pairs designated as the control group. Videos for the study pairs were recorded separately, whereas for the control groups, there is a single video for both. Additionally, following the experiment, four group interviews, each lasting 30 minutes, were conducted. In total, there were 17 files and a combined 250 minutes of data in Finnish. These materials are exclusively stored on the University of Turku Seafile service, with restricted access solely for the researchers involved in this study.

5.3 Speech to text

To utilize speech-to-text tools, we initially converted the experiment's videos to .wav files. This conversion was conducted using ffmpeg command lines via the Linux terminal, ensuring the preservation of sample rates and other acoustic features. Similarly, interview audio files, originally in .m4a format, were transformed to .wav using the same process.

To generate transcriptions from the audio files, we opted to utilize business tools. Among the various options available, three services showed promising potential: Google Cloud[1], Amazon Transcription[2], and Alto ASR[3]. While the former two have their own platforms, Alto ASR was accessed through CSC Puhti servers by importing the necessary modules. To assess the efficacy of these tools, I employed last year's study manual transcripts as a reference and calculated the Character Error Rate (CER) for each tool. However, Alto ASR's output was limited due to inadequate speech recognition, rendering it unsuitable for use. For Amazon Transcription, the CER value was 0.29, while for Google Cloud, it was 0.24. Providing further granularity, Google's CER for group discussions was approximately 0.31, whereas for conversations between two individuals, it was 0.17. Services' servers were chosen to be in Europe; Finland for Alto ASR, Ireland for amazon, and Sweden for Google.

Speech-to-Text Tool	**Character Error Rate (CER)**
Amazon Transcription	0.29
Google Cloud	0.24
Google Cloud (Group Discussions)	0.31
Google Cloud (Two-Person Conversations)	0.175
GPT-3.5	0.18
GPT-4	0.178

Table 5.1: Character Error Rates (CER) for Speech-to-Text Tools

Regarding the considerations in Table 5.1, transcription tools tend to perform better with two-person conversations compared to group discussions. Additionally, group discussions involve both children and adults, whereas the two-person conversations in-

[1] `https://cloud.google.com/speech-to-text?hl=en`

[2] `https://aws.amazon.com/transcribe/`

[3] `https://www.kielipankki.fi/tuki/aalto-asr-automaattinen-puheentunnistin/`

volve only adults. Consequently, we anticipate more challenging results from this year's experiment (CS2).

5.4 Text processing and translation

The output of the Google Cloud transcription service often included spoken language and occasional corruption (e.g., audio content was not detectable while children were moving intensely), potentially impacting the quality of the translation from Finnish to English. To address this issue, we employed both GPT-3.5[4] and GPT-4[5] models. Contextual descriptions were provided to both models to enhance their understanding of the transcription (See figure 5.1). Generally, the GPT-3.5 model primarily attempted to rectify syntactic issues, whereas the GPT-4 model's efforts resulted in more extensive edits based on the context provided. I evaluated the results using Character Error Rate (CER) values for both models' outputs. The CER for GPT-3.5 remained largely consistent, with a minor decrease in the third digit observed for two-person conversations, while for GPT-4, it increased to 0.18. However, we retained both sets of data as each model's distinct features—adherence to the original text versus contextual editing—may prove beneficial in translations.

For translations, the DeepL[6] platform and the GPT-4 API were used. DeepL has a user-friendly website for uploading files, while the GPT-4 API is accessible through Python. For the GPT-4 model, the context of the experiment and the file were provided.

5.5 Data annotation

As previously noted, multiple data versions were available for this study. We selected one English version and one Finnish version to commence our work. The English version represents the output of GPT-4, while the Finnish version is the initial output of Google

[4] `https://platform.openai.com/docs/models/gpt-3-5-turbo`

[5] `https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4`

[6] `https://www.deepl.com/translator`

GPT4

```
[ ] msg =  """I have transcribed text featuring a conversation between two kids
                speaking Finnish and playing guessing games. The transcription
                captures only one person's dialogue and includes ordinal numbers and timestamps.
                While the transcription is mostly accurate, there are sections with
                unclear or incorrect wording (gibberish). Could you help me identify
                and correct these errors, ensuring that each corrected segment is
                clearly associated with its ordinal number and original timestamp?
                Additionally, it's important to maintain the text's overall format,
                with each ordinal number on a separate line, followed by its timestamp
                and dialogue on subsequent lines. Not all sentences may end with a period.
                please apply corrections and print the whole text without any extra output.
                Here is the text:
                \n\n"""
```

Figure 5.1: Prompting GPT4 using open AI's API

Transcribe.

Annotation was carried out by three individuals, including two Finnish speakers and one English speaker.

5.5.1 Guidelines

In this study we had two types of annotations.

1. Speaker Marking: One of the annotation tasks was to mark which dialogue belongs to whom (S1 and S2)

2. The second task of annotation was to label each sentence by an emotion.

   Robert Plutchik categorized emotions into eight primary categories, evenly split between positive and negative. These emotions are viewed as opposites, which is evident in secondary emotions such as joy contrasting with sadness, surprise with anticipation, trust with disgust, and anger with fear. Plutchik provided detailed explanations for each emotion, further dividing them into subgroups and organizing them in a wheel-shaped model, considering them as secondary and tertiary emotions. He also observed that the intensity of an emotion is strongest at the center of the wheel and diminishes as one moves away from the center [30]. For a better understanding, please refer to Figure 5.2.

   Here are the definitions of the emotions for labeling:

Figure 5.2: Plutchik's wheel of emotions. Source: [31]

"Joy: A feeling of happiness, pleasure, or delight. It's characterized by a sense of contentment and positive emotion.

Sadness: A state of feeling sorrowful, unhappy, or despondent. It often involves a sense of loss or disappointment.

Anger: An intense emotional response often triggered by a perceived threat, injustice, or frustration. It involves feelings of hostility, irritation, or rage.

Fear: An emotional reaction to a perceived threat, danger, or harm. It triggers a sense of apprehension, nervousness, or anxiety.

Surprise: A sudden and unexpected emotional reaction to something unexpected or unfamiliar. It involves feelings of astonishment, amazement, or disbelief.

Disgust: A strong aversion or revulsion towards something unpleasant, offensive, or repulsive. It triggers feelings of nausea, repugnance, or contempt.

Trust: A positive emotional response characterized by confidence, reliance, or belief in someone or something. It involves feelings of security, faith, or dependence.

Anticipation: An emotion associated with looking forward to or expecting something in the future. It involves feelings of excitement, eagerness, or anticipation."[32]

5.5.2 Tools

We utilized the Doccano[7] platform for annotation. By enabling the sequence modeling mode with overlapping labeling, we were able to annotate both the speaker and emotions simultaneously, enhancing the depth of our analysis. For a clearer understanding of Doccano and it's features, refer to Figure 5.3.

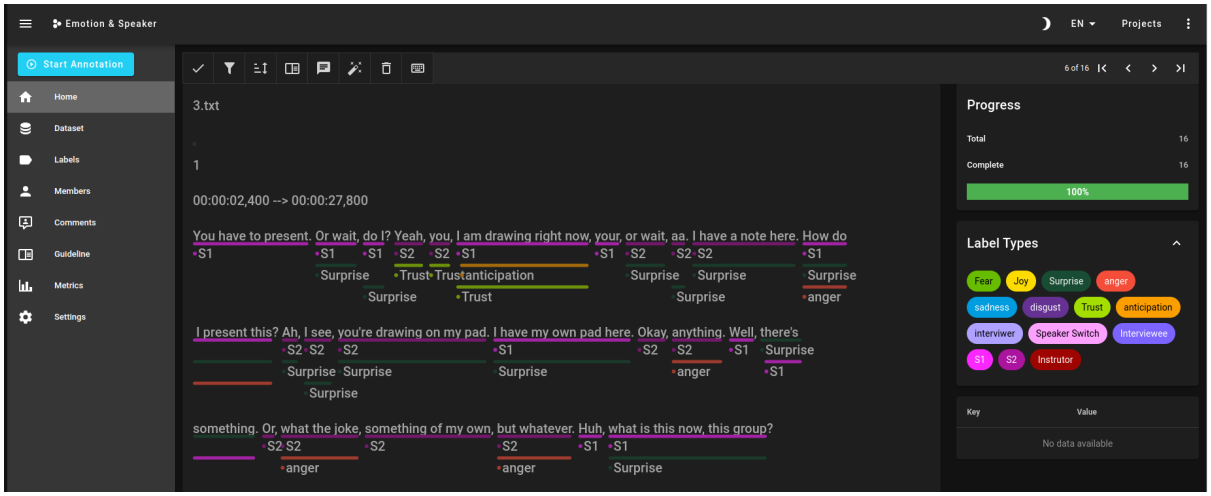


Figure 5.3: Doccano environment and features

[7] https://doccano.github.io/doccano/

6 Presentation of findings

In this chapter, we present the findings derived from data annotated by both human analysts and artificial intelligence systems. Furthermore, we aim to extract more complex insights from these foundational observations.

6.1 Annotation analysis

First, let's briefly examine the resources required for this annotation task.

Annotator	**Time spent**
English Speaker	16 hours
Finnish Speaker1	8 hours
Finnish Speaker2	8 hours
GPT4	8 minutes
GPT3	3 minutes

Table 6.1: Time spent by each annotator

Moreover, according to InfoFinland [1], the median minimum wage is €9.16 per hour. GPT-4-turbo costs $0.01 per 1000 characters of input (prompt) and $0.03 per 1000 characters of output. In a paid study, using humans annotators incurs higher costs and requires more time to produce results.

6.1.1 Visualization

The first step for every data analyst in a project is to plot and visualize the existing data to gain a meaningful understanding. These observations can provide insights into

[1] `https://www.infofinland.fi/en/work-and-enterprise/during-employment/wages-and-working-hours`

how annotations are conducted and what to expect, although they may not always be accurate due to using the number of labels used in the plotted graphs. In subsequent steps, token-level labels are employed, providing a more reliable basis for assessment.

One approach to gaining a deeper understanding of the data and assessing its quality is through visualization. By visualizing the data, we can identify correlations and relationships between various categories and data types. In this context, we will examine the distribution of labels across different tasks and evaluate the extent of overlap among different annotators.

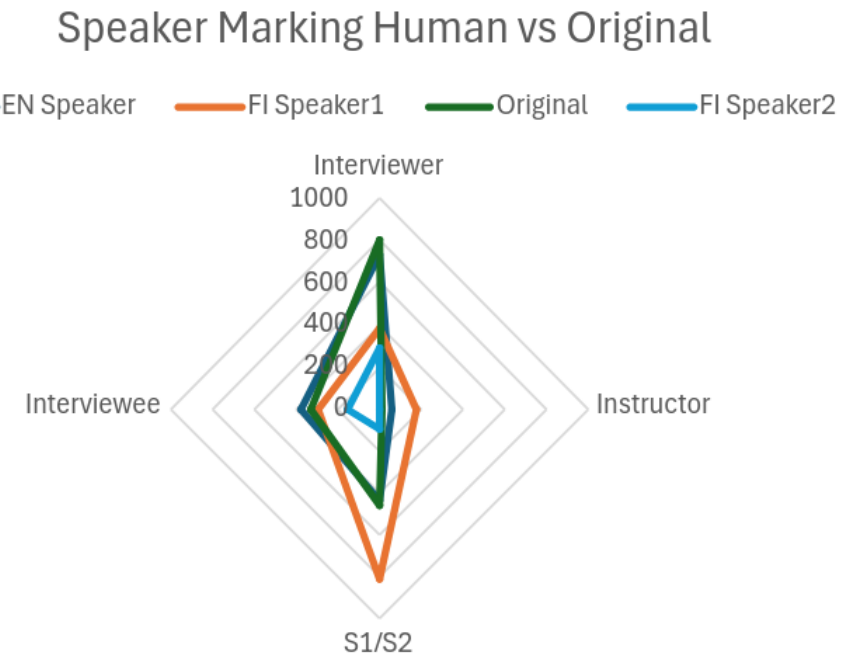


Figure 6.1: distribution of speaker marking in dialogues by human

Figure 6.1, we compare human annotation results with the actual annotations. The data used by the English-speaking individual had been previously edited and translated by GPT-4. As indicated by the graph, despite the original language of the experiment being Finnish, the English speaker's annotations has more overlap with the original annotations. Among the two Finnish speakers, the one involved in the experiment shows greater overlap, suggesting that the context of the experiment was challenging to grasp for those not directly involved.

Among all annotators, Finnish Speaker 2, who was not involved in the experiment execution days, stated that the complexity of the concept and text without video and

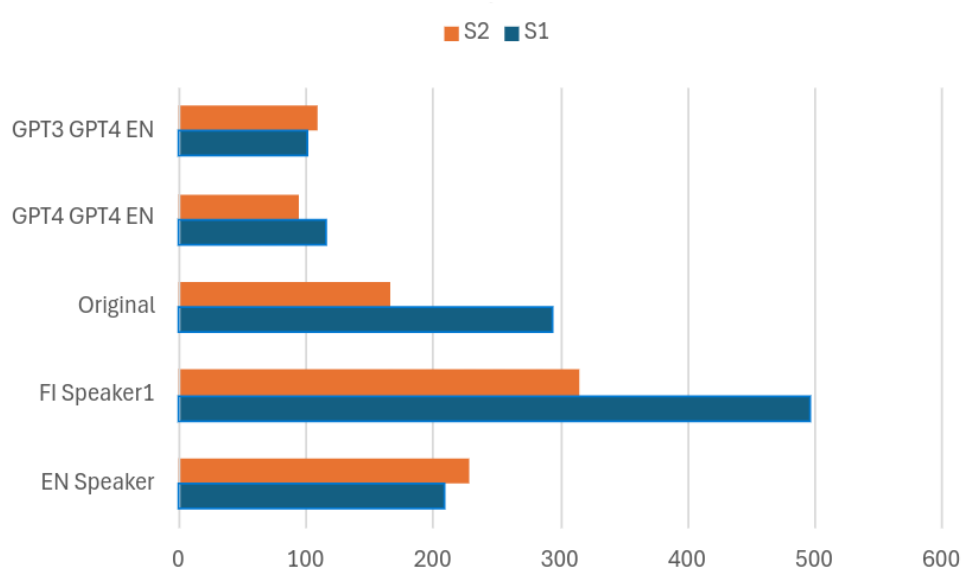


Figure 6.2: Speaker switch task

audio made it difficult to understand which speaker was talking. In Figure 6.2, it can be seen that none of the annotators closely matched the original annotations, indicating that the task was too complex.

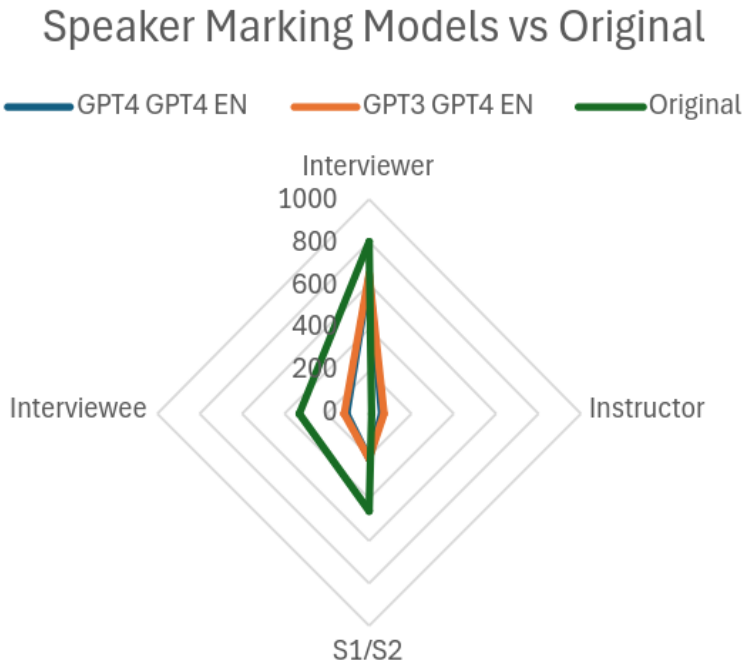


Figure 6.3: distribution of speaker marking in dialogues by models

In Figure 6.3, we compare the data edited by the GPT-3 model with that edited by the GPT-4 model, both of which have been annotated by GPT-4. The results indicate that both models yield similar outcomes.

Comparing these two figures, it can be concluded that the combination of GPT models and human input yields more competitive results than either human-only or AI-only annotations. Even when the annotator does not speak the original language, AI assistance can produce results that surpass those of native speakers. Additionally, despite GPT-4 being generally recognized as significantly more advanced than GPT-3, in this complex task, their performance statistics and outcomes are quite similar.

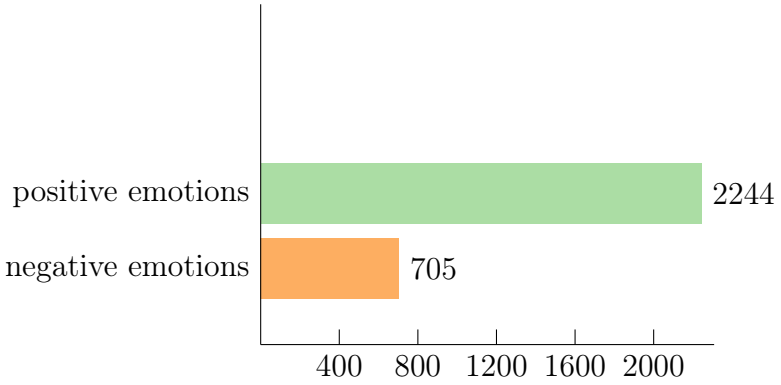


Figure 6.4: positive and negative emotions distribution

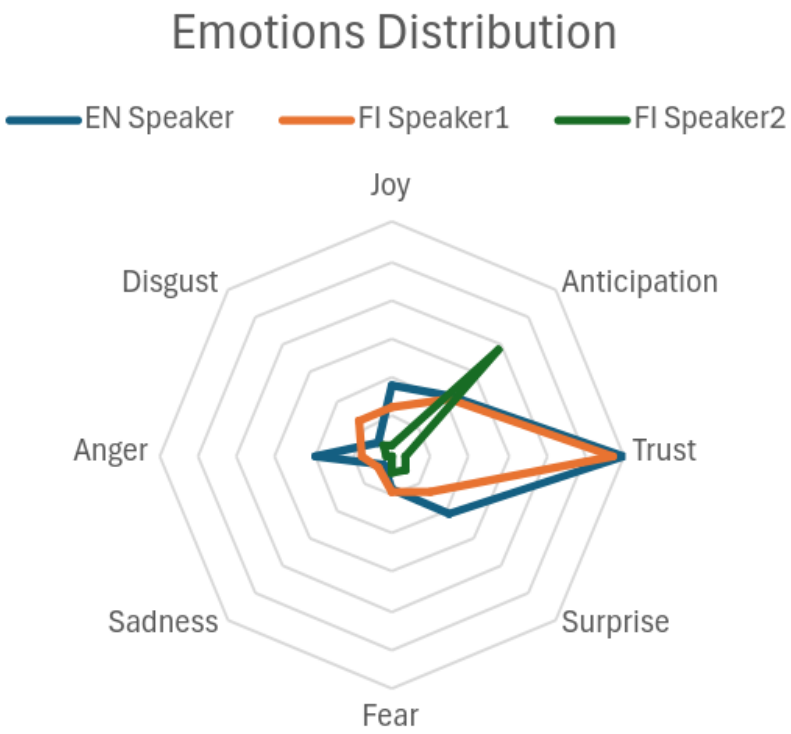


Figure 6.5: distribution of emotions among human annotations

To derive meaningful conclusions from the experiment, it is essential for annotators to achieve agreement in their annotations, which can then be compared with AI model

outputs. Given that the original language of the experiment is Finnish, the researchers expected that the highest quality annotations would be produced by native Finnish speakers. Contrary to these expectations, Figure 6.5 illustrates that the emotion distributions annotated by the English-speaking annotator and Finnish Speaker 1 have more overlap than those of Finnish Speaker 1 and Finnish Speaker 2. This difference is likely due to the involvement of the English-speaking annotator and Finnish Speaker 1 in the experiment's executive team, whereas Finnish Speaker 2 did not participate in CS2 execution days.

Despite these differences, all human annotators have mostly labeled the text with positive emotions rather than negative ones. Figure 6.4 also supports this idea.

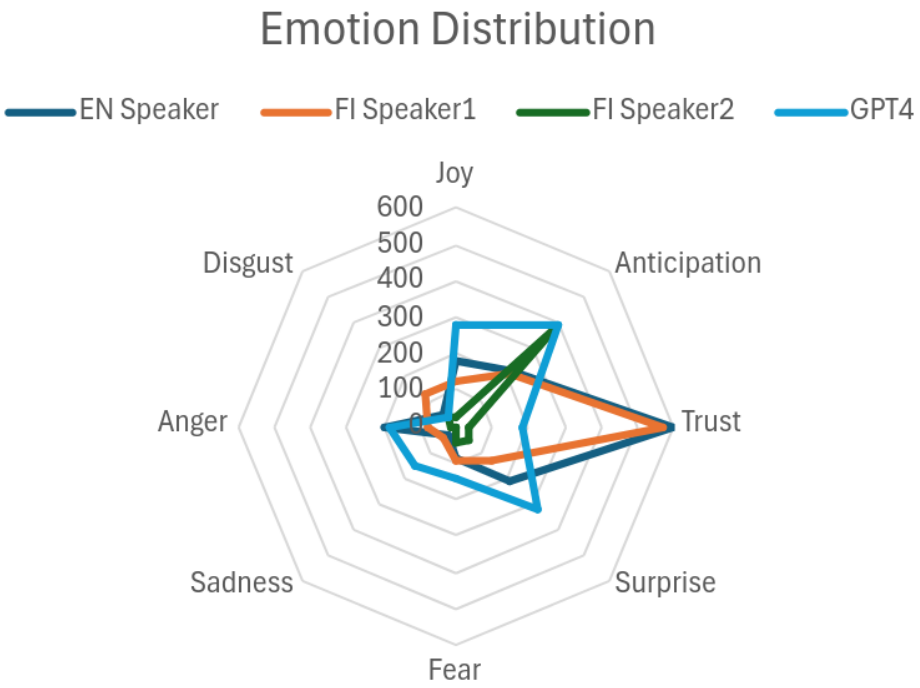


Figure 6.6: distribution of emotions among all annotations

In 6.6, by adding GPT4 model annotations to the distribution graph, more negative emotions can be seen. Also it seems like the model keeps overlaps with all annotators. It is expected to see have more overlapping emotions for this language model in the text, as the number of emotions and the area covered by its graph is larger.

Integrating GPT-4 model annotations into the distribution graph reveals a greater presence of negative emotions. Moreover, the model consistently overlaps with all annotators. Given the larger number of emotions and the expanded area covered by its graph,

it is anticipated that this language model would exhibit more overlapping emotions in the text.

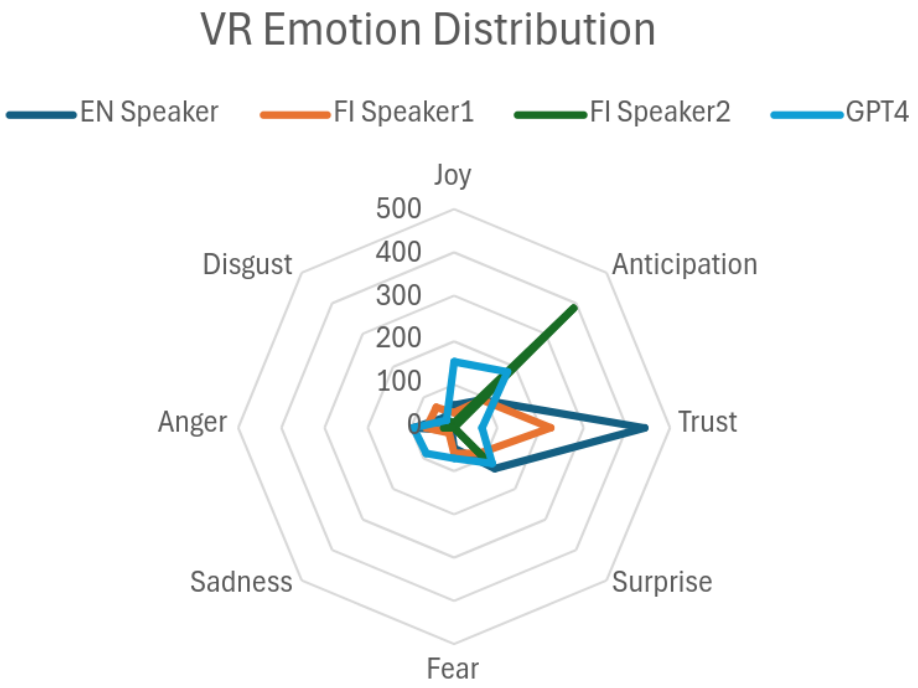


Figure 6.7: distribution of emotions in MR platform

The distribution of emotions is compared between two different platforms, Zoom and MR, as shown in Figures 6.7 and 6.8. Interestingly, the distribution for Zoom shows a higher prevalence of positive emotions, whereas MR exhibits a comparable amount of negative emotions.

6.2 Annotation agreement

After annotating a data set, the quality of the annotations need to be assessed. Besides visualization, one of the techniques for this assessment is using agreement metrics. As it was previously discussed 2.6, I am calculating fleiss kappa metric using python and statsmodels library among annotators for different tasks.

Let's take a look at them in table 6.2. In table 6.2, the speaker switch task has been evaluated based on interviews. This task specifically involves annotating each sentence with either 'interviewer' or 'interviewee' labels. The observed high level of agreement

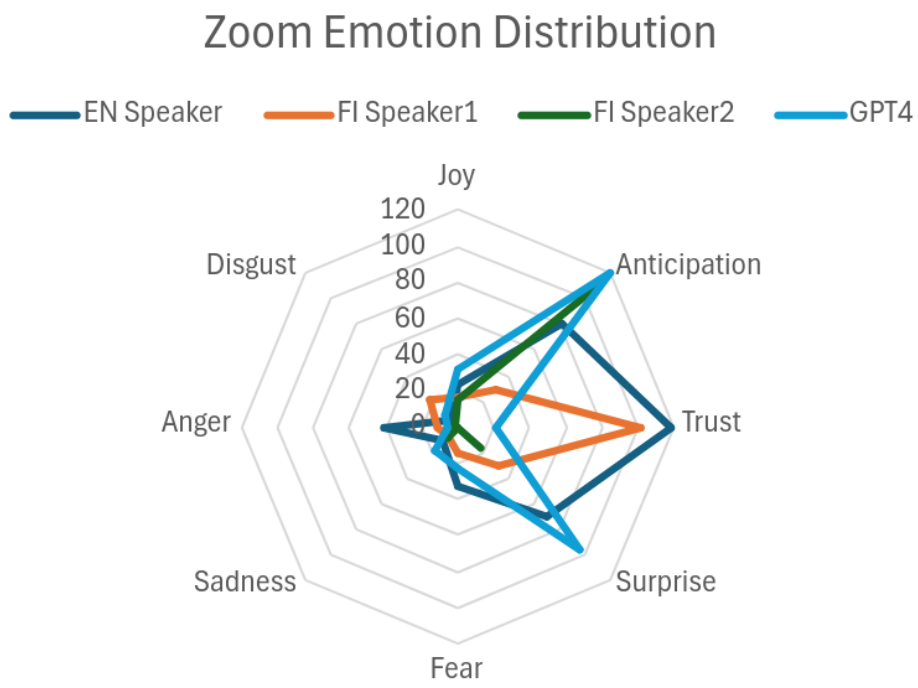


Figure 6.8: distribution of emotions in Zoom platform

Annotator 1	**Annotator 2**	**Fleiss kappa**
English Speaker	Finnish Speaker1	0.55
Finnish Speaker1	Finnish Speaker2	0.856
English Speaker	Actual switches	0.58
Finnish Speaker2	Actual switches	0.810
GPT-4	Actual switches	0.8130
GPT-3.5	Actual switches	0.8133

Table 6.2: Annotation agreement for speaker switch task in interviews

among Finnish speakers suggests that the annotations are of sufficient quality for comparative analysis. This reinforces the trustworthiness of human annotations in this study.

On the other hand, the agreement between English speakers and actual switches is moderate. This comparison with Finnish speaker annotations suggests that some information and data may have been lost or altered during translation. Further discussion on this topic can be found in Section 6.4.

The substantial agreement between language models and actual language switches further supports their efficacy in annotating for this classification task. It is noteworthy that despite being provided only with the concept and the English version of the experiment, language models performed better than the English-speaking annotator and as

well as a Finnish speaker who had no prior involvement in the experiment.

Annotator 1	**Annotator 2**	**Fleiss kappa**
English Speaker	Finnish Speaker1	0.59
English Speaker	Actual switches	0.58
Finnish Speaker1	Actual switches	0.78
GPT-4	Actual switches	0.8106
GPT-3.5	Actual switches	0.8109

Table 6.3: Annotation agreement for speaker switch task in experiments

Let's take a look at a more complex task, with the same criteria. In Table 6.3, the same annotators evaluated the data from the experiment days, attempting to identify whether dialogues belonged to speaker 1, speaker 2, or the instructor. Among these annotators, Finnish speaker 2, who was not involved in the preparation or execution of the experiment, expressed difficulty in comprehending the task. Therefore, their annotations are not included in the table.

Table 6.3 also indicates that the information provided may not have been adequate or valid for the English speaker to annotate accurately. Specifically, the relatively low agreement level between the English speaker and actual language switches indicates that some information in the text and context may have been missed. In contrast, the Finnish speaker, who had access to the original data, shows a higher agreement with the actual language switches.

Language models excel in this task, surpassing humans despite accessing English data only, even though the task was more challenging.

Annotator 1	**Annotator 2**	**Fleiss kappa**
English Speaker	Finnish Speaker1	0.55
Finnish Speaker1	Finnish Speaker2	0.84
GPT-4	Finnish Speaker1	0.52
GPT-4	English Speaker	0.66
GPT-3.5	Finnish Speaker1	0.845
GPT-3.5	Finnish Speaker2	0.897

Table 6.4: Annotation agreement for Plutchik's Wheel of Emotions task

Annotation of human emotions is highly challenging due to the difficulty in achieving agreement among humans. This study extends this challenge by involving language

models in the annotation process. Table 6.4 shows strong agreement among Finnish speakers. It's important to note that Finnish speaker 2, who did not participate in the experiment, annotated only 10% of the text due to a lack of familiarity with the content.

On the other hand, agreement among other annotators is not very high. As previously discussed, translating the text from Finnish to English may result in information loss. It is noteworthy that while language models excelled in the speaker switch marking task, detecting human emotions and achieving agreement with humans themselves appears to be significantly more challenging.

Remarkably, the GPT-3.5 model demonstrates high agreement with Finnish speakers. From my observations, it appears that GPT-3 had fewer annotations compared to GPT-4, suggesting a reduction in false positives.

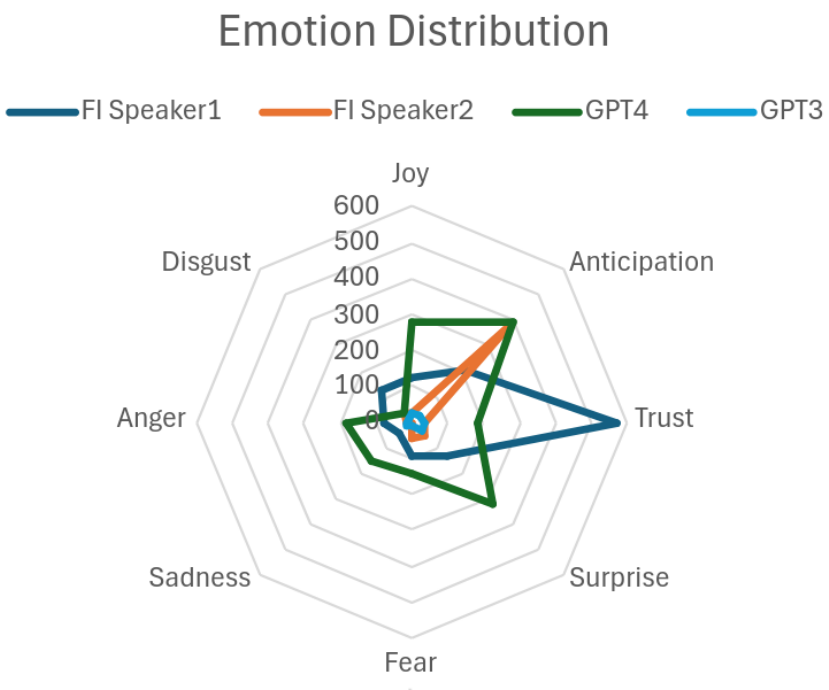


Figure 6.9: distribution of emotions between highest agreements

Following the notable results presented in Table 6.4, my curiosity was piqued regarding the distribution of emotions among annotators with high agreement levels (also see Tables 6.3, 6.2, and 2.2). Figure 6.9 illustrates that the number of annotations in GPT-3 is relatively low, precluding definitive conclusions or insights.

6.3 Sentiment analysis

$$sentimentTowardsImmersion = \frac{positiveEmotions}{positiveEmotions + negativeEmotions} \quad (6.1)$$

Let's take a look at Formula 6.1. This equation helps us determine the direction of emotions. I calculate this equation for each annotator, as well as a total one, to drive an explanation.

Annotator	**Sentiment towards immersion**
English Speaker	0.78
Finnish Speaker1	0.75
Finnish Speaker2	0.85
GPT-3.5	0.68
GPT-4	0.69
Total	0.74

Table 6.5: Sentiment analysis table for each annotator

Despite differences among annotators, sentiment towards immersion is consistently positive across each individual and collectively. Although Finnish speaker 2 exhibited more negative emotions compared to the other human annotators (see Figure 6.5) the highest rate in 6.5 belongs to Finnish speaker 2. The GPT-4 language model showed the lowest rate of negative emotions, which aligns with expectations given the distribution graph (Figure 6.6) indicating more negative emotions overall.

Annotator	**MR Sentiment towards immersion**
English Speaker	0.80
Finnish Speaker1	0.67
Finnish Speaker2	0.90
GPT-4	0.64
GPT-3.5	0.64
Total	0.75

Table 6.6: Sentiment analysis table for MR platform

In Table 6.6 and 6.7, all numbers indicate that both technologies are considered immersive by annotators. However, when comparing Table 6.6 and 6.7, unexpectedly, sentiment analysis towards immersion shows higher numbers for the Zoom platform compared to MR. The essence of MR lies in its ability to offer enhanced immersion for users, and losing to a 2D platform like Zoom is not the outcome we anticipated.

Annotator	Zoom Sentiment towards immersion
English Speaker	0.78
Finnish Speaker1	0.76
Finnish Speaker2	0.90
GPT-4	0.82
GPT-3.5	0.69
Total	0.80

Table 6.7: Sentiment analysis table for Zoom platform

6.4 Observations

In this section, I present notes that I collected while analyzing the data. By examining the transcripts, listening to the audio, and watching the videos, it becomes evident that adult voices are recognized more accurately than children's voices. This was anticipated, as discussed previously in Table 5.1.

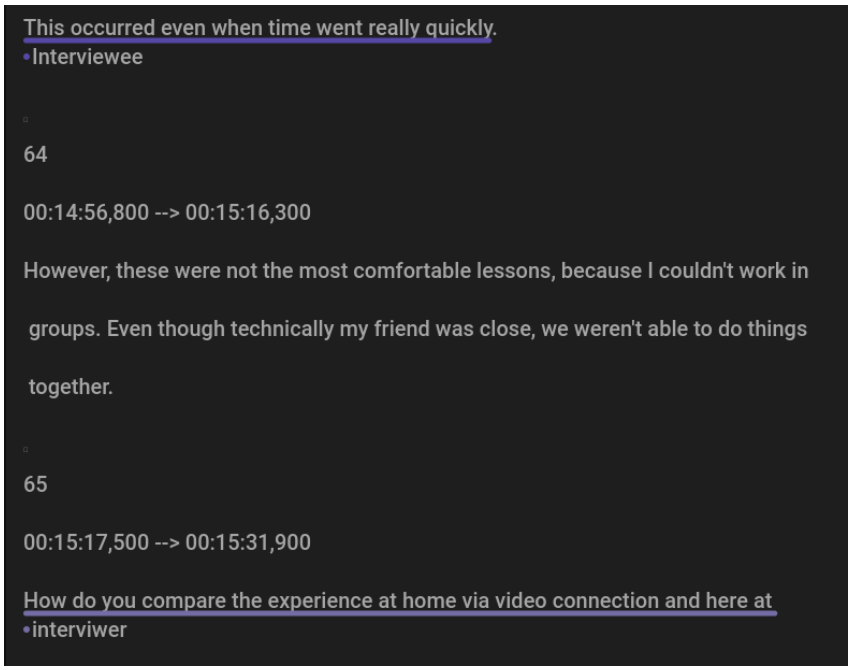
Although GPT models proved highly beneficial for editing corrupted text and translating from Finnish to English, they also presented certain challenges. Let us examine some examples. The model relies on preceding and subsequent dialogues for translating and editing. This approach occasionally results in the removal or alteration of certain words or phrases, thereby causing a loss of meaning.

In Figure 6.10, paragraph 68 should end with a question, but instead, the concept is combined with the next paragraph and its question. This makes it hard for the English speaker annotator to extract the exact original labels because there will be fewer of them.

Not only does this affect the number of annotations, but it also sometimes affects the number of each label. In Figure 6.11, the interviewee, Elina, says they do not use video calls much, just with their friends mostly, and the interviewer says "me neither." However, with GPT editing and translation, it looks like a single dialogue from the interviewer.

6.5 Challenges

Annotating human emotions and guessing who is talking only based on the text is super challenging for everyone. Something that were even more challenging during this study was to understand every annotators understanding of the data and trying to create a

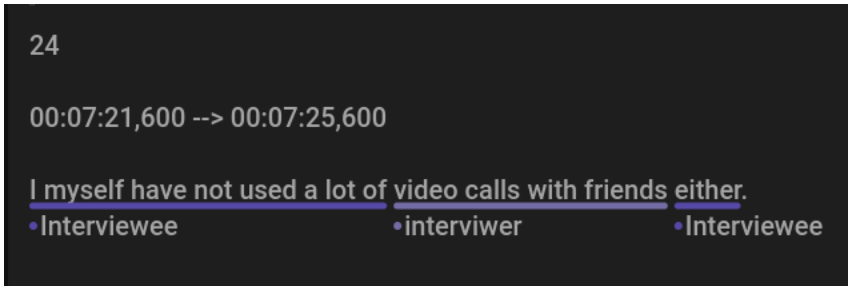


(a) Translation by GPT4

14:47.4 - 14:55.2	fi-fi	0.93	Mites ellino niinkun silloin se meni kyllä aika niinku nopeesti mut tota ne ei ollut niin niinku
14:56.8 - 15:16.3	fi-fi	0.95	Kivoin ja tunnit, kun ei niinku pystynyt tekee niinku esim. Mitään pari töitä tai ryhmätöitä piti aina tehdä yksin jotain tehtäviä Ja vaikka tässäkin niinku periaatteessa kaveri oli vaan niinku 10 metrin päässä, niin silti niin kun ei ollut kaverin kanssa silleen yhdessä. Oliko se kumminkin mitä niinkun?
15:17.5 - 15:31.9	fi-fi	0.91	Suhteessa jos sä mietit just niihin kotona on video yhteydessä koulussa ja sit suhteessa, että on täällä paikan päällä koulussa niin mihin kohtaan tää niinkun kokemus siinä. Onks tää mieluisampaa?

(b) Google transcription

Figure 6.10: Google transcription and GPT translation



(a) Translation by GPT4

07:16.3 - 07:20.3	fi-fi	0.8 7	Elina en paljoa kyllä.
07:21.6 - 07:25.6	fi-fi	0.7 8	Videopuheluita kavereiden kanssa en kauheasti itsekään.

(b) Google transcription

Figure 6.11: Effect of Google transcription and GPT translation on annotation

measurement base for charts and graphs.

Some examples:

1. Even though labels were established prior to starting the annotation project, one annotator did not use 'interviewee' at all; instead, they used 'Speaker 1'. Having to gather data from both the interviewee and experiment databases, I spent a considerable amount of time cleaning the database.

2. Another annotator found the task of marking speakers difficult and simply marked where a switch occurred without identifying whose turn it was.

3. Various forms of the same label were used, necessitating additional processing: 'Speaker 1', 'S1', 'speaker1', 'speaker 1'.

Despite these challenges, tackling a complex task led me to employ language models at various stages to enhance efficiency. Each stage involved processing data in chunks, with tailored and detailed prompts specific to each model, segmented by experiments and interviews.

Since I spent considerable time on preprocessing and prompting, although I cannot include them in this thesis, the codes related to annotation tasks, scripts, and other

preprocessing tools used for this study are available on my GitHub repository [2]. To have an understanding of the thesis's GitHub page, refer to Figure 6.12.

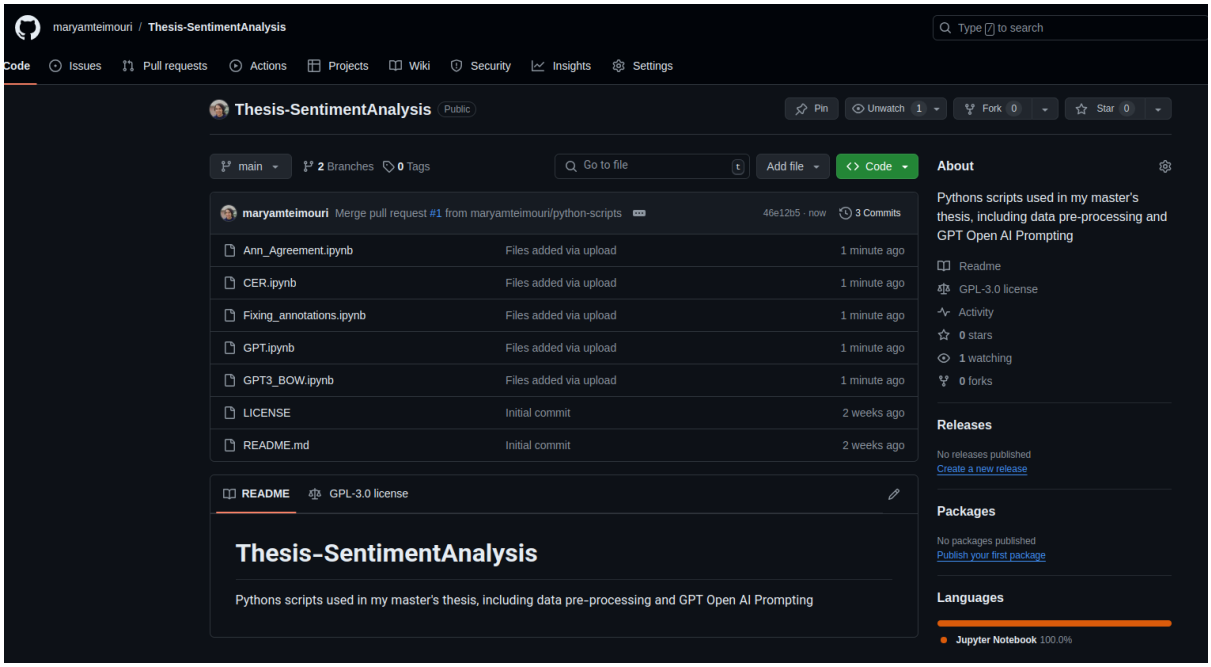


Figure 6.12: An overview of the thesis's GitHub page

[2] https://github.com/maryamteimouri/Thesis-SentimentAnalysis

7 Discussion

This research was guided by research questions introduced in Chapter 1. After exploring this topic across six chapters, the time has come to address these research questions.

1. The Role of Transcription: Competitive business products in the market excel in transcription (see Table 5.1). However, their effectiveness depends on specific conditions such as conversations involving two speakers and superior voice detection capabilities for adults. Annotations mostly feature interviewer labels with limited contributions from interviewee responses.

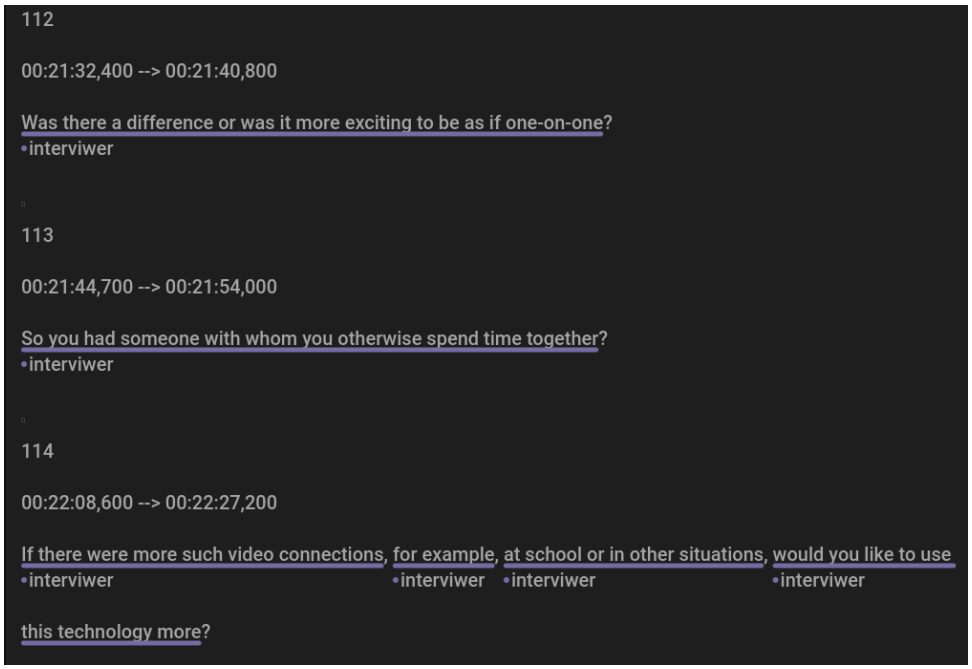


Figure 7.1: Lack of children answers in transcriptions

With these challenges in mind, the agreement between language models and Finnish speakers regarding speaker detection annotations was high, indicating that the retained information from the text is usable and the concept is understandable.

Therefore, transcription plays a crucial role in reducing data loss. Moreover, leveraging transcription tools can potentially overcome barriers, such as having a non-Finnish speaker working with Finnish data by transcribing and then translating.

2. Impact of Translation on Analysis: The English-speaking participant who analyzed translated data achieved the lowest scores in speaker identification tasks. However, language models (GPT-4 and GPT-3) fed on translated text showed high accuracy in identifying actual speaker switches. This suggests that while translation may result in some data loss, such loss can be compensated for by leveraging prior knowledge of the original language and literature.

   For studying bilingual or multilingual communication in digital settings, conducting reliable studies is feasible with the assistance of sophisticated tools such as large language models.

3. Effectiveness of Automatic Tools: According to Tables 5.1, 6.3, and 6.2, all automatic tools demonstrate effectiveness, but their efficacy can be further enhanced by optimizing the environment. For instance, Character Error Rate (CER) is lower in sessions with only two speakers and performs better with adult voices compared to children's voices. Similarly, speaker identification tasks are more straightforward for models in interviews due to similar reasons compared to experiments.

4. Identification of Communication Contexts and Speakers: In this study, the testing of this ability was pushed to the limit of human performance, and large language models performed as well as a Finnish speaker and better than an English speaker. For instance, in interviews, the task was simplified to identifying only the interviewer and interviewee. However, for more complex tasks such as distinguishing between multiple interviewees based solely on audio, additional contextual information like video or audio cues becomes necessary. Identifying the exact speaker remains challenging for humans and Finnish speakers relying solely on transcription.

5. Differences Between Mixed Reality and Zoom Interactions: Based on the numbers, both Zoom and MR technology provide immersive experiences, but Zoom shows a

5% higher effectiveness. This suggests that MR technology may not yet be fully developed, a concern raised by researchers in our group. However, detecting human emotions is a complex task even for humans, and apart from Finnish speakers, agreement levels were not high. This indicates that communication patterns and linguistic features are sensitive to their original language, and even with a large repository of information like GPT, the task remains highly complex and challenging.

Although considerable effort was made to analyze and draw conclusions, it is evident that additional varied experiments are necessary to strengthen the findings. This study represents a single and designed experiment.

8 Conclusion

Every beginning has an end, and to conclude this research and thesis, the information within this document will be utilized. First, let us examine the bag of words of this document as presented on Figure 8.1.

Figure 8.1: Bag of words for the document

Every day, new technologies emerge thanks to the internet and advanced gadgets and vehicles. However, it is more important to learn how, when, and where to use these technologies, and often, to figure out which technology is best for a given situation. To achieve this, we must first assess the available tools and technologies. Then, we should test these tools in various situations to determine their suitability.

With the advancement of the hardware industry, we now have faster internet and more powerful computers, all while devices become smaller in size. Smart gadgets are highly capable in terms of computation and user-friendly. Additionally, the rapid release

of various AI models is facilitated by robust and efficient infrastructure. Leading this progress are large language models and mixed reality equipment, which are becoming trends in entertainment, education, and healthcare.

In this chapter, I will present the results and conclusions from assessing various AI models and their suitability for different tasks in analyzing mixed reality immersion in education. Specifically, I will evaluate whether these models can make the impossible possible or simply enhance efficiency and ease of use. Additionally, if the models prove to be beneficial, I will explore the sentiment analysis regarding immersion on the mixed reality platform.

8.1 AI tools assessment

Unlike the expectations, GPT-4 did not make a significant difference in the tasks (the translation capability of the model was not assessed). GPT-3 and GPT-4 had similar Character Error Rates (CER) in editing transcription text (see Table 5.1). In the speaker identification task, the agreement rates were very close (see Tables 6.2 and 6.3). In the emotion annotation task, although GPT-4 had more annotations than GPT-3 (see Figure 6.9) , the number of false positives in agreement was high.

Both GPT-3 and GPT-4 performed remarkably well in speaker identification, especially considering the availability of the English version of the text, so translation information loss is not a concern. However, in the emotion detection task, both models were inefficient. Since translation was not assessed in this study, it is unclear whether this inefficiency is due to translation issues or the models' lack of understanding of human emotions.

This research, originally conducted in Finnish, involved an English speaker working on the project. The English speaker did not achieve high annotation agreement in the speaker identification and emotion annotation tasks. However, the agreement numbers indicate a level of consistency between the English speaker and the other annotators. Considering that without LLMs this task would have been impossible for the English

speaker, and given the observed level of agreement, it suggests that AI has made what was previously impossible, possible.

8.2 Sentiment analysis towards immersion

Based on the data presented in Table 6.5, the overall experiment and design were both engaging and immersive. However, the metrics in Table 6.6 indicate that while the Mixed Reality (MR) experience was immersive, it did not meet expectations. Furthermore, as shown in Table 6.7, the total sentiment scores and the scores from all annotators, except for the English Speaker annotator, are slightly higher for the Zoom platform.

Overall, these findings suggest that, for this particular experiment and design, the MR platform did not significantly enhance immersion. In contrast, the Zoom platform was not only more immersive but also more cost-effective and easier to use.

9 Future Work

Uncertainty about the future is inevitable, yet preparation is crucial for effective adaptation. Preparation necessitates planning. Although the scope of this topic provides adequate potential and data for evaluation, the constraints of a master's thesis timeline limit exhaustive exploration. Nonetheless, ongoing technological advancements and research expansion continuously redefine possibilities. Throughout this study, numerous new technologies emerged, potentially altering established frameworks and generating diverse scenarios.

This research suggests multiple avenues for future exploration, each warranting individual consideration and detailed elaboration.

- Comprehensive data exploration: At the outset of this research, various types of data were collected, but not all were utilized in this study (see Figure 9.1). For future work, it would be wiser and more sustainable to thoroughly analyze the existing data and derive additional conclusions before altering the experimental design.

- Data collection planning: The study employed specific technologies and addressed specific research questions. All technologies used were novel, making it challenging to gather additional data to ensure the validity of measurements and numbers. Furthermore, this experiment and research constituted Case Study 2 of the broader investigation into immersion in mixed reality. However, Case Study 1 had a different design, rendering its data unusable. Therefore, for projects aiming to address research questions involving large language models, careful planning for data collection becomes essential.

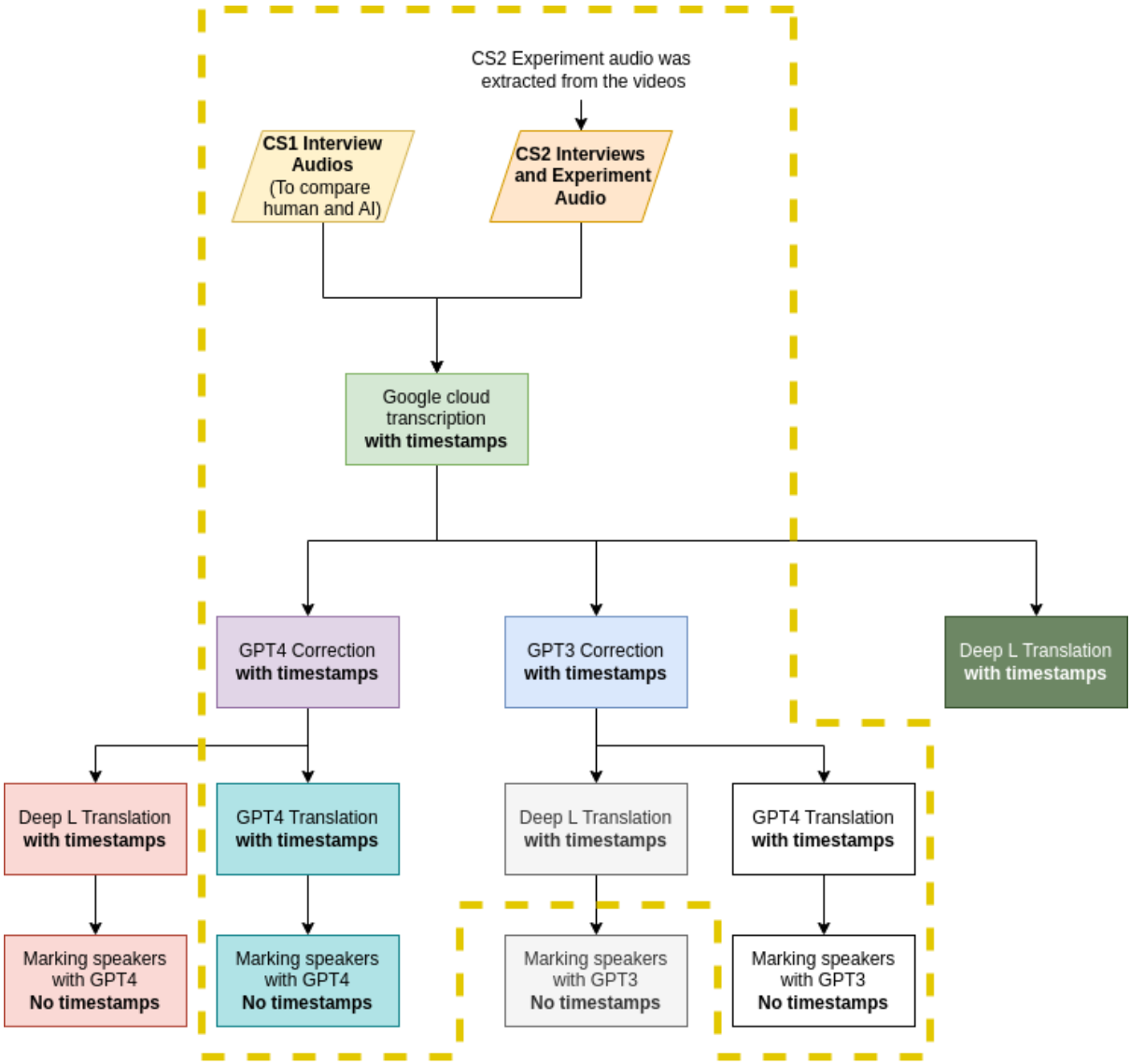


Figure 9.1: Used data in data flow

- Technology: This experiment utilized Unreal Engine versions 4.27, 4.27.1, 4.27.2, and 4.28, each incorporating incremental improvements aimed at enhancing reliability and user experience. Additionally, Unreal Engine 5.1, 5.2, and 5.3 introduced advanced features such as Nanite virtualized geometry and Lumen global illumination, significantly elevating graphical fidelity and engine performance [1].

  This research utilized GPT-3 Turbo and GPT-4 for annotation, editing, and translations. Additionally, there are other models available for text processing within this setup, such as GPT-3, GPT-4O, Gemini, and LLAMA3.

[1] `https://www.unrealengine.com/en-US`

- Machine learning models: While numerous large language models exist today, not all data from interviews and experiments can be effectively extracted from text only data. The use of other models, such as Convolutional Neural Networks (CNNs) for processing images or videos, presents exciting opportunities. Moreover, various models for motion extraction and skeleton detection enable the analysis of posture and related emotions. With current computational capabilities, it is feasible to implement multimodal structures and utilize multiple models simultaneously.

  With the release of GPT-4O, there is now the capability to integrate transcription, translation, and annotation tasks into a unified model and interface. However, from the findings of this experiment, it appears that Large Language Models (LLMs) may perform better when tasks are divided into smaller steps. Therefore, this observation also warrants further investigation.

- Tasks performed by Large Language Models (LLMs): This study assessed the capabilities of LLMs in emotion annotation, text editing, and speaker identification (text only). However, certain tasks, such as translation, were not considered in this study.

- Setup: This experiment was primarily designed to analyze visual features rather than being specifically tailored for text and Large Language Models (LLMs). Improving the setup could involve using higher quality microphones or adjusting their placement to ensure effective recording during children's activities, minimizing the need for them to move far from the microphone. Additionally, transcription tools encountered difficulties with children's voices due to their datasets being predominantly composed of adult voices. Therefore, evaluating these tools with a target group of a higher age range might facilitate more accurate assessments.

  Additionally, more interactive features could have been utilized in this experiment. Currently, children were only connected through microphones and cameras to draw on shared screens. In the mixed reality platform, there exists a movable 3D Canvas that could have been integrated. Moreover, a web interface was designed to send

emoji 2D textures to each pair, aiding in emotion analysis; however, these features were not utilized in the study. For a clearer understanding of this feature, refer to 9.2.

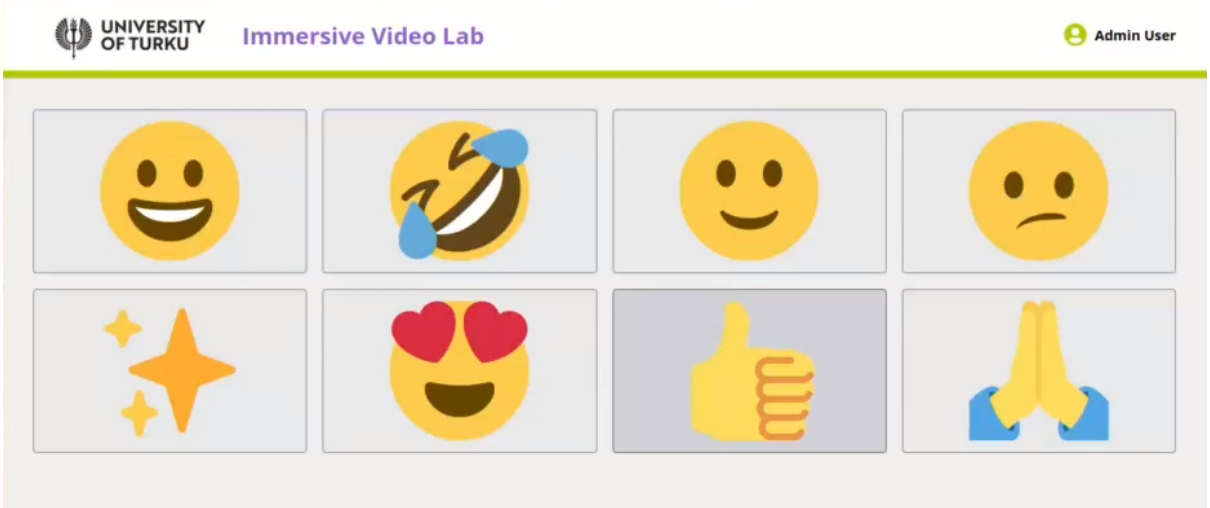


Figure 9.2: An overview of the interaction page

References

[1] T. B. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners", *arXiv preprint arXiv:2005.14165*, 2020.

[2] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need", in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[3] M. Munezero, C. S. Montero, E. Sutinen, and J. Pajunen, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text", *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014. DOI: 10.1109/TAFFC.2014.2317187.

[4] E. Adam, "Improve the user experience of chatbots with tone analysis", Ph.D. dissertation, Feb. 2019. DOI: 10.13140/RG.2.2.20569.57444.

[5] [Online]. Available: https://cloud.google.com/learn/what-is-artificial-intelligence.

[6] [Online]. Available: https://cloud.google.com/learn/what-is-machine-learning.

[7] S. Srivastav, *Artificial intelligence, machine learning, and deep learning. what's the real difference?*, 2020. [Online]. Available: https://medium.com/swlh/artificial-intelligence-machine-learning-and-deep-learning-whats-the-real-difference-94fe7e528097.

[8] J. West, D. Ventura, and S. Warnick, "Spring research presentation: A theoretical foundation for inductive transfer", *Brigham Young University, College of Physical and Mathematical Sciences*, vol. 1, no. 08, 2007.

[9] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *et al.*, "Improving language understanding by generative pre-training", 2018.

[10] T. D. Science, *Understanding the open pre-trained transformers (opt) library*, Accessed: 2024-07-18. [Online]. Available: `https://miro.medium.com/v2/resize:fit:1200/1*-4bqsnXpB1XZZ49ifw-n7A.png`.

[11] J. Pathak, *Generative AI in Natural Language Processing — packtpub.com*, `https://www.packtpub.com/article-hub/generative-ai-in-natural-language-processing`, [Accessed 13-05-2024].

[12] OpenAI, "Gpt-4 technical report", *arXiv preprint arXiv:2303.08774*, 2023.

[13] J. Cohen, "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46, 1960. DOI: `10.1177/001316446002000104`.

[14] J. L. Fleiss, "Measuring nominal scale agreement among many raters", *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971. DOI: `10.1037/h0031619`.

[15] J. R. Landis and G. G. Koch, "An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers", *Biometrics*, pp. 363–374, 1977.

[16] P. Milgram and F. Kishino, "A taxonomy of mixed reality visual displays", *IEICE Transactions on Information and Systems*, vol. 77, no. 12, pp. 1321–1329, 1994.

[17] Medium, *Difference between virtual reality and augmented reality*, Accessed: 2024-07-18. [Online]. Available: `https://miro.medium.com/v2/resize:fit:600/1*G_AcUtku9GLZx92eyUzpOA.jpeg`.

[18] L. Johnson and S. Adams Becker, *Horizon Report: 2017 Higher Education Edition*. EDUCAUSE, 2017.

[19] R. Raskar, G. Welch, and H. Fuchs, "Spatially augmented reality", in *Proceedings of the First IEEE Workshop on Augmented Reality*, IEEE, 1999, pp. 11–20.

[20] J. Chen and C. Liu, "Mixed reality technology and its applications: A review", *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 1, pp. 1–14, 2020.

[21] Epic Games, *Unreal engine*, 2024. [Online]. Available: https://www.unrealengine.com/.

[22] P. A. Laplante, *Technical Writing: A Practical Guide for Engineers and Scientists*, 2nd ed. CRC Press, 2017, pp. 165–167, ISBN: 978-1-4987-3566-5.

[23] E. Games, *Unreal engine gameplay mechanics*, Accessed: 2024-07-18. [Online]. Available: https://cdn2.unrealengine.com/unreal-engine-gameplay-mechanics-1920x1080-5796c04aeee3.jpg?resize=1&w=1920.

[24] Calvert, *Intel realsense depth camera*, Accessed: 2024-07-18. [Online]. Available: http://www.calvert.ch/wp-content/uploads/2018/06/office.png.

[25] L. Keselman, J. I. Woodfill, A. Grunnet-Jepsen, and A. Bhowmik, *Intel realsense stereoscopic depth cameras*, 2017. arXiv: 1705.05548 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1705.05548.

[26] S. Park, S. Bokijonov, and Y. Choi, "Review of microsoft hololens applications over the past five years", *Applied Sciences*, vol. 11, no. 16, p. 7259, 2021. DOI: 10.3390/app11167259. [Online]. Available: https://www.mdpi.com/2076-3417/11/16/7259.

[27] M. Munezero, C. S. Montero, M. Mozgovoy, and E. Sutinen, "Exploiting sentiment analysis to track emotions in students' learning diaries", in *Proceedings of the 13th Koli Calling International Conference on Computing Education Research*, 2013, pp. 145–152.

[28] R. G. África, M. R. González-Rodríguez, and M. C. Díaz-Fernández, "Salient features and emotions elicited from a virtual reality experience: The immersive van gogh exhibition", *Quality & Quantity*, pp. 1–20, 2023.

[29] F. Weninger, S. P. Kumar, T. Fritsch, M. Schröder, and B. Schuller, *Enhancing speaker diarization with large language models: A contextual beam search approach*, 2023. arXiv: 2309.05248 [cs.CL]. [Online]. Available: https://arxiv.org/abs/2309.05248.

[30] M. A. Mohsin and A. Beltiukov, "Summarizing emotions from text using plutchik's wheel of emotions", in *Proceedings of the 7th Scientific Conference on Information Technologies for Intelligent Decision Making Support (ITIDS 2019)*, Atlantis Press, 2019, pp. 291–294.

[31] E. Meyer, J. Green, E. Arnold, and M. H. Wickstrom, "Understanding how and when graduate student instructors break through challenges with active learning", *International Journal of Research in Undergraduate Mathematics Education*, pp. 1–33, May 2024. DOI: 10.1007/s40753-024-00240-8.

[32] R. Plutchik, "A general psychoevolutionary theory of emotion", in *Theories of emotion*, Elsevier, 1980, pp. 3–33.

Appendix A Command lines

```
ffmpeg -i file.mp4 ../path/file.wav
ffmpeg -ss 02:15 -to 02:25 -i file.wav new_file.wav
```

Listing A.1: ffmpeg - creating audio and trimming

```
find *.wav | sed 's:\:\\\:g'| sed 's/^/file/' > fl.txt; ffmpeg -f
   concat -i fl.txt -c copy output.wav; rm fl.txt
```

Listing A.2: ffmpeg - concatenating existing files for online services

Appendix B Python codes

```
from jiwer import cer
import numpy as np
import docx2txt

reference = docx2txt.process("Case study 1 interviews.docx")
f = open("Amazon_transcrip_CS1.txt", 'r')
amazon_hypothesis = f.read()

error = cer(reference, amazon_hypothesis)
print("Amazon Error: " , error)
```

Listing B.1: CER assessment

```
from openai import OpenAI
client = OpenAI(api_key='####')

gpt_prompt = """
Hi, I have a text which is the auto-transcription of an interview in
    Finnish.
The interview happens between an adult and a few kids.
They are talking about an experiment they did last week,
which included playing games on zoom or on a VR platform.
The text looks fine, but it needs some editing.
Please correct those parts in the text for me and print
the entire text with the corrections made.
Ensure that any corrections blend seamlessly with the
original content, preserving the overall structure and timestamps.
Do not introduce any extra outputs, it is a must.
```

```
Here is the text:\n\n"""

def process_file_in_chunks(file_path, chunk_size):
    """Process the file in chunks with a specified size."""
    output = ""
    try:
        with open(file_path, 'r', encoding='utf-8') as file:
            file_content = file.readlines()

        # Process in chunks
        for i in range(0, len(file_content), chunk_size):
            chunk = "".join(file_content[i:i+chunk_size])
            # Define process_chunk to use the API
            processed_chunk = process_chunk(chunk)
            output += processed_chunk

    except Exception as e:
        print(f"An error occurred: {e}")

    return output

def process_chunk(chunk_text):
    """Process a single chunk of text. Implement API calling and
   response handling here."""
    # This function is a placeholder
    #for where you would make the API call.
    processed_text = ""
    stream = client.chat.completions.create(
          model="gpt-4",
          messages=[{"role": "user", "content":f"{gpt_prompt}{
   chunk_text}"}],
          stream=True,
    )
    for chunk in stream:
          if chunk.choices[0].delta.content is not None:
```

```
                processed_text += chunk.choices[0].delta.content
    return processed_text

file_paths = ['2_interview.srt', '3_interview.srt', '1_interview.srt',
    '4_interview.srt']
chunk_size = 100
for path in file_paths:
    # Adjust chunk_size as needed
    output = process_file_in_chunks(path, chunk_size)
    with open(f"corrected_{path}", 'w', encoding='utf-8') as file:
        file.write(output)
```

Listing B.2: GPT chat API