ARCADA

QT Sermo

A conversation chatbot using synthetic data, open-source solutions and integrated on a QTrobot via a Jetson Orin Nano

Kristoffer Kuvaja Adolfsson

Master's Thesis

MEng in Big Data Analytics

2024

Master's Thesis

Kristoffer Kuvaja Adolfsson

QT Sermo. A conversation chatbot using synthetic data, open-source solutions and integrated on a QTrobot via a Jetson Orin Nano.

Arcada University of Applied Sciences: MEng in Big Data Analytics, 2024.

Commissioned by:

Arcada UAS

Abstract:

Previous research suggests that users desire conversations with humanoid robots. Advances in Large Language Models (LLMs) now enable more effective conversation capabilities. This thesis develops a conversation system on the QTrobot platform with a Jetson Orin Nano, utilizing free and open-source software (FOSS) to prioritize user privacy and transparency. Automatic speech recognition (ASR), a local LLM, and a text-to-speech solution enables real-time conversations. Experimentation with fine-tuning the LLM, using synthetic data generated from multiple open models, showed promise but also revealed limitations. This thesis proposes an accuracy metric for evaluating synthetic data generation of LLMs based on insights from the experiments, building a framework for creating larger datasets to enhance other models in the future. The conversation system's modular design allows for deployment on various robot platforms; however, further testing is required to validate its performance across different robotics systems.

Keywords: QTrobot, Artificial Intelligence, Large Language Models, Human-Computer Interaction, Embedded systems

Contents

Figures

Tables

Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
ASR	Automatic Speech Recognition
CBSE	Component-Based Software Engineering
CUDA	Compute Unified Device Architecture
CPU	Central processing unit
DL	Deep Learning
FOSS	Free and Open Source Software
GGML	Georgi Gerganov Machine Learning
GPT	Generative Pre-trained Transformer
GPU	Graphical Processing Unit
cRI	child-Robot interaction
HRI	Human-robot interaction
LLM	Large Language Model
ML	Machine Learning
NN	Neural Network
OOP	Object-oriented programming
RAG	Retrieval augmented generation
ROS	Robot Operating System
SOTA	State-of-the-art
STT	Speech-To-Text
TTS	Text-To-Speech
VAD	Voice Activity Detection

Foreword

As the field of science continues to evolve at an unprecedented pace, it has become increasingly clear that innovative research and development are essential for driving progress in our digital age. At Arcada University of Applied Sciences, we strive to foster a culture of innovation and excellence in modern society, equipping students with the knowledge and skills needed to tackle complex problems and push the boundaries of what is possible while keeping the human in the center.

This master's thesis represents a significant milestone in my own academic journey at Arcada, as I delve into the intricacies of AI and Robotics. Through this research, I aim to contribute to the pool of knowledge by exploring chatbots and QTrobot. The work presented in this thesis is a culmination of my academic and professional experience, as well as the encouragement, support, and guidance provided by my supervisor Leonardo Espoinosa-Leal. Additionally, I want to acknowledge the contributions of my peers and colleagues with their work and ideas and a special thank you to the people in Arcada's AI, Development and Robot labs for making this work come to fruition.

The following pages present my original research in QT Sermo, which I hope will contribute to the ongoing conversation on robots, AI and open-source. I am delighted to have you, the reader, share in my enthusiasm and to finally share this work.

Kristoffer Kuvaja Adolfsson

1. Introduction

Robotics is a broad study; finding a use case, a design or a purpose are all initial considerations at the beginning of a robot's life cycle. What will the robot do? How will it do the task? And what parts and components does the robot need? Simply scratching the surface of robotics opens too many questions to be answered. While constructing a robot is in essence the physical components a user can touch and see, it is software that binds components together and brings life to the interaction. Deploying the robot in the real-world while keeping the robot functional and iterating on feedback is the final phase of developing a robot (until the economy becomes more circular and recycling becomes more integral). in this thesis however, the focus lies on software. This thesis describes the full development for a conversation system called QT Sermo for the QTrobot platform by LuxAI S.A. (2024). This thesis will discuss robotics, the software development cycle of QT sermo and explore software solutions used for the conversation system: automatic speech recognition (ASR), large langauge model (LLM), text-to-speech (TTS), robot operating system (ROS) and LuxAI's own user interface QTrobot Studio.

At Arcada UAS human-robot interaction (HRI) has been researched for several years in a multitude of sectors integrating humanoid robots (robots designed with human-like characteristics, mimicking human functions) in health-, childcare, and consumer services (Hägglund et al. 2023, Tigerstedt et al. 2023a,b, Tigerstedt & Biström 2021), rehabilitation (Jeglinsky-Kankainen et al. 2024), indoor positioning and navigation (Heikel & Espinosa-Leal 2022, Majd et al. 2021) and trustworthy HRI (Pham & Espinosa-Leal 2024) . From our own observations and experiences, we see that stakeholders and users tend to try and converse with humanoid robots during initial interactions, even if humorously or out of curiosity, it is something not as prevalent in non-humanoid interactions. QT sermo conversation system is meant to bridge this gap between users and robots.

This thesis builds upon previous robot software development experiences and works, in Deploying Humanoid Robots in a Social Environment (Kuvaja Adolfsson et al. 2024) a three-phase method was described, drawing from service design (optimizing experiences via user inclusion) and action research (iterative cycles in development) to stabilize a general framework for robotics deployment at Arcada. Realizing the limitations, especially in

sensitive use cases like health- and childcare, with proprietary robots the team at Arcada has investigate data streams, sensors and the applicability of modifying an existing robot to run on open-source software (Biström et al. 2024). Also, in a forthcoming publication an application for improving the interaction between patient and a humanoid robot in a dental care scenario is laid out (Kristoffer et al. 2025).

Apart from a background in robotics and information technology, it is relevant to mention a previous vocational degree in *educational activities and care of children and young people* with over half a decade of experience in childcare with children in ages 0 - 7. QTrobot and child-robot interaction (cRI) as such naturally intertwine both past and current proficiencies and work in this field has previously been explored (Tigerstedt et al. 2024), with forthcoming future projects.

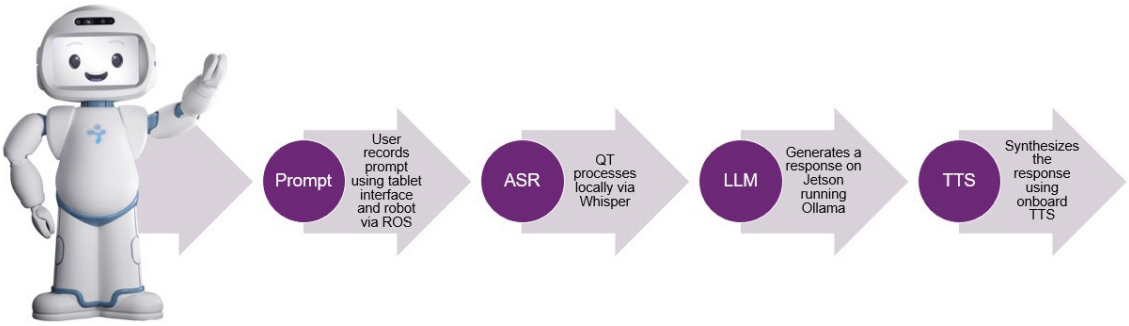


Figure 1. QTrobot (photo by LuxAI S.A.) and components for the conversation system.

This thesis explores the possibility of creating a functional conversation system (QT sermo) with a humanoid robot. The components for this conversational system, as seen in Figure 1, will be LuxAI QTrobot as the humanoid robot with incorporated tablets running the LuxAI user interface. Whisper by Radford et al. (2022) as the systems ASR. A Jetson Orin Nano hosting a custom fine-tuned model as the brain. And to give QTrobot a voice the built in Acapela will be used for TTS while built in gestures will bring motion and life to the robot.

The first order of operations will be to develop and deploy the conversation system on the robot platform in a functional state. The conversation system inposes limits on itself by primarily using open-source tools; or at a minimum to present viable alternatives in this thesis. The system is built initial for cRI use cases on the QTrobot platform but needs

to be flexible and serviceable for other robot platforms, research use cases and projects at Arcada such as service design, customer experience, healthcare or virtual realities. As such this thesis has 4 major goals:

- Can we develop a functional conversation system for the humanoid QTrobot using a Jetson Orin Nano?
- Can we leverage open-source software solutions to develop the system?
- Can we improve upon the system using state of the art techniques?
- Can we maintain modularity and flexibility to ensure the work is usable for future development or even other robots at Arcada?

The following section goes into related literature, bringing a short review on robotics while taking a special interest in humanoid robots followed by a more in-depth look into AI and it's recent developments. It also goes in on ASR and LLM technologies and how it all ties together via chatbots and robotics. In the methodology section the hardware and software are motivated and described in technical detail. While the experiments section brings up the two major development cycles, (1) developing and deploying the pipeline on QTrobot and (2) data generation and fine-tuning of the model. Results are then presented with an overview on the synthetic data, models, tables and graphs. Finally, a conclusive discussion that answers the thesis goals while also looking at future work and suggestions to be done in the field.

2. Related work

The conversation system developed in this work involves multiple elements intertwined to bring full functionality, as such related work will touch lightly upon robots and HRI. Due to the usage of QTrobot and its intended target group some literature will also relate to the cRI topic. Moreover AI is investigated, as it is the core of the conversation system. Researching the background and history of AI before moving on to more recent developments while finally moving over to more specifics related to LLMs and chatbots, core components of the conversation system.

2.1 Robots & Humanoids

In our current digital revolution, industrial robots are already in full swing: automating repetitive tasks in varying complexity - from logistics in warehouses and harbors, manufacturing in the automobile industry, or simply cleaning our floors (Gonzalez-Aguirre et al. 2021). The world we built however, is designed for humans and that is unlikely to change which poses a significant challenge for robotics development on the modern industrial robot (Hägele et al. 2016, Espinosa-Leal et al. 2020). Robots with bipedal locomotion and kinematic arms, two very complex problems, are argued to be more suitable for our environments and recent advances in AI have spurred a new boom for solving these problems, strongly benefiting a particular type of robot, the humanoid.

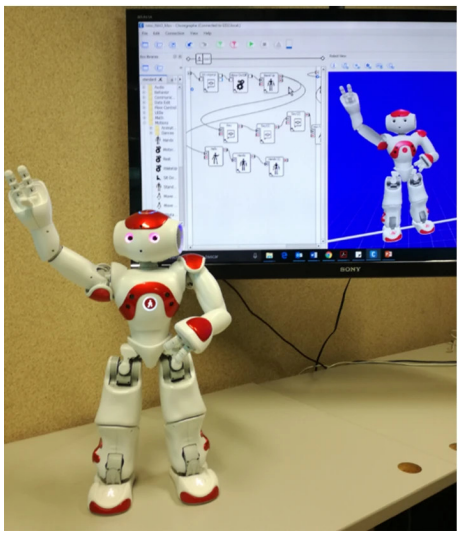


Figure 2. Nao robot and interface, image from Cruz-Ramírez et al. (2022).

The study between humans and robots, commonly referred as HRI, goes back to 1970 with theories like Mori's Uncanny Valley, however work done before the 2000s can be seen as seminal as stated by Vouloutsi et al. (2023). It wasn't until much later that interactions with robots really started to take shape, such as 2003 with Paro, a therapeutic seal-like robot or 2004 with the anthropomorphic robot Nao, see Figure 2. Vouloutsi also highlights the steady but slow progress of robotics development, as complex machines are progressing slower than perhaps anticipated. Kinetic arms need smaller and more powerful servos and bipedal walking needs more compute as machine learning (ML) algorithms use higher quantities of more complex data for precision and safety. Robots, however, aren't the only thing that has progressed slower than initially anticipated.

2.2 AI

In 1956, at the *Dartmouth Summer Research Project on Artificial Intelligence* (a six-week brainstorming session) a group of scientists optimistically suggested that machines could simulate every aspect of learning within a short time (a couple of months) if given enough resources; a lofty suggestion, considering the accomplishments we have yet to achieve today. Sheikh et al. (2023) marks this six-week session as the beginning of the modern phase of AI. In their text on the history of AI, they divide AI into three distinct phases starting from antiquity (800 B.C - 500 A.D) with mythical stories and fantasies on artificial life, providing examples like the automatons of the Greek god *Hephaistos* or *Daedalus and Talos*. The second era is represented with speculation and thinkers like *Galileo Galilei*, *Isaac Newton* and *René Descartes* whom created designs and inventions in their time, shaping artificial forms away from abstract myth and fantasies into more concrete mechanical forms. By the end of this second era *Karel Capek* wrote "R.U.R" (Rossum's Universal Robots) that coined the term robot from old Slavonic "rabota", meaning forced labor. Then our current era, from 1956 and forward, is perhaps better described in detail by Russell & Norvig (2010) whom clearly outline that AI became an industry really only in the 1980s and a few years later, Naural Networks (NN) made a return from the 1800s sparking one of many booms experienced in AI over the coming years.

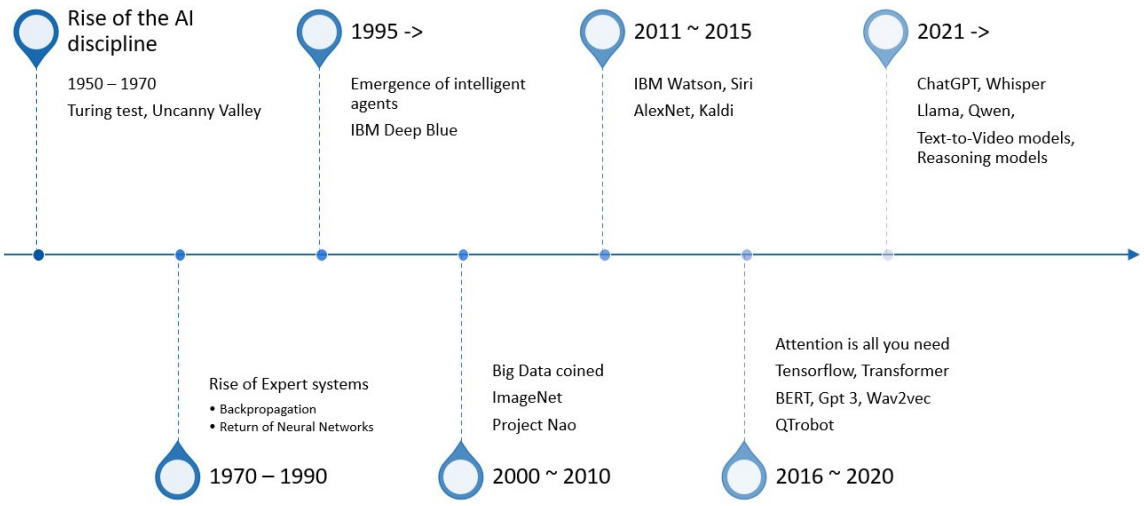


Figure 3. Snapshot of AI timeline and robots, adapted from Russell & Norvig (2010), Sheikh et al. (2023) and Vouloutsi et al. (2023).

It would then take until the late 1990s, see Figure 3, with the defeat of grandmaster Garry Kasparov by IBM's Deep Blue program before we really saw the hallmarks of modern AI. Followed by Big Data being coined in the early 2000s and intelligent agents not long before that. However, a real hallmark for today's AI came in 2016 when Google's AlphaGo used backpropagation, reinforcement learning and deep neural networks to defeat world champion Lee Sedol in the game "Go". This breakthrough is often credited too Yann LeCun and Andrew Ng for NNs and Geoffrey Hinton, David Rumelhart and Ronald Williams for publishing the backpropagation algorithm with a paper back in Nature in 1986. But recently it has also emerged that some credit for backpropagation should be contributed to a publication from Finland (Linnainmaa 1970), preceding Rumelhart and Williams.

Soon after, in 2017, the defining paper for LLMs would be published with the transformer by Vaswani et al. (2017) in *Attention is all you need*. The transformer allowed attention, typically used in recurrent NNs, to forgo the recurrent model and instead use attention of tokens in parallel. The significant factor here comes from the fact that parallelization is well utilized in GPUs with thousands of cores for matrix multiplications - compared to standard CPUs with only a few cores. It had already been established in 2012 with AlexNet (Krizhevsky et al. 2012) that NNs, specifically convolutional NNs in AlexNet, scale extraordinarily well with *simply* more compute and data. As such, the transformer

allowed several magnitudes of more compute power to train much larger and more complex NNs and not long after this the first Generative Pre-training Transformer (GPT) was released and in 2022 ChatGPT took the world by storm and ushered in the most recent AI boom.

2.2.1 Definition of AI

While these advancements are mostly confined in ML they're often contributed as AI but AI encompasses a much broader field as it is more the study of human intelligence and how machines might simulate it (TAIGA 2022). The relationship between ML and AI can be visualized as seen in Figure 4 and Prince (2023), describes it similarly as ML is "the part of AI that fits mathematical models to observed data". *Deeper* down in ML field we find deep learning (DL) that utilizes NNs to advocate for exactly that, advanced fits of mathematical models to observed data. Eventually advances within DL took us to what is today's encoder-decoder transformer architecture used in LLMs like ChatGPT.

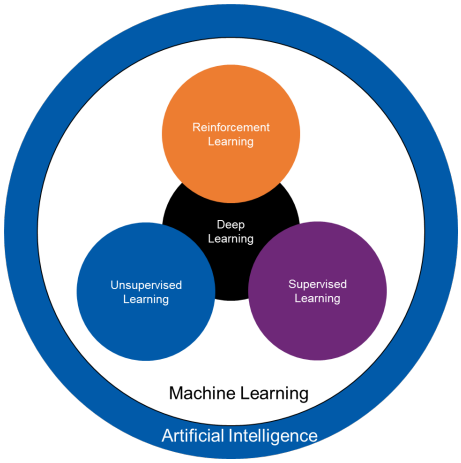


Figure 4. Relationship of ML within AI, adapted from Li (2018).

Generative AI and LLMs have seen a massive explosion in mass appeal, really hitting it off with with ChatGPT and GPT-4 (OpenAI et al. 2024). ChatGPT made headlines worldwide and ushered in a new AI boom that promise new heights for humanity in several areas like productivity, finance, liberty and creativity to name but a few. Some people even proclaim it will solve climate- and immigration crises or think it will take us to new planets, like Mars. However, a more cautious approach tells us that this is not the first time AI has experienced a boom and it probably won't be the last.

2.2.2 HRI and AI

Boom or not, recent AI advancements have introduced new paradigms in HRI by integrating LLMs into robot platforms and systems. The review paper by Zhang et al. (2023) provides an analysis, highlighting key developments in LLMs that enhance robot structures and their performance capabilities. Integrating LLMs into robotic systems allows for completion of more complex task, something explored in the review. While the review points to significant progress it also states that several challenges remain such as contextual understanding, data privacy, and ethics. Ethical challenges and data privacy are of special prevalence due to the target group of QTrobot being children.

Humanoids like QTrobot, often work as social robots (embodied robots that interact with humans) and have steadily been growing in popularity over the years. Today they can be seen in stores, restaurants and shopping malls and users find the overall interaction to be positive (Andtfolk 2022, Iwamoto et al. 2022, Tigerstedt & Fabricius 2023, Song & Kim 2022, Lu et al. 2020, Hägglund et al. 2023). Studies by Tung & Au (2018) and Fu et al. (2021) show that humanoid robots have more social acceptance compared to other types of robots. Sometimes users also feel safer around humanoids, in part due to less human contact (Wan et al. 2021, Rubagotti et al. 2022). Negative trends also exists, such as those found by Fusté-Forné (2021), that mentions a risk for dehumanization when using humanoids and that improper functions by robots lead to poorer experiences for users. Privacy concerns are also relevant risks as sensors can be intrusive and proprietary software is not always transparent on where data streams actually end up as found by Biström et al. (2022b).

2.2.3 Assistants and chatbots

While shape and form makes the humanoid - the use case must also be considered. Robot assistants have been used for game-based learning activities, such as done by Issa et al. (2023) that showed tailored experiences on robots had a clear benefit for users. Nasir et al. (2024) also studied how different social robots were perceived in learning, they used agents with varying degrees of domain knowledge in a subject and found users to engage more with the agent when the it had domain knowledge, but they could not conclude if this correlated with better learning. In Sweden Elgarf et al. (2021) made a custom game

for story telling in Unity with a Furhat robot to play stories but did not find that the Furhat robot added value based on user feedback.

Dünya & Durak (2023) found that chatbots can be useful in teaching, even if users (students) preferred real human interaction. In their work they cite design, content and algorithms as areas to improve chatbot interactions. Janson (2023), similarly investigated chatbots with users, trying to find more effective ways to implement chatbots and designs while Ko et al. (2024), researched chatbot anthropmorphism in children with autism closer and calls for a shift towards a more inclusive approach in chatbot interface design: one that considers the cognitive and social requisites of marginalized users.

Robot assistants have also been used in healthcare, like Ligthart et al. (2020) where a Nao robot was used to create stories with children being treated for cancer. They wanted children to be more engaged and have a higher sense of agency as it has shown to enrich the experience much like HRI researches previous mentioned. They found that when children could choose elements in the story, it would increase engagement and agency compared to traditional predetermined stories. To measure engagement Augello (2022) experimented with a novel approach, using a small light that would turn on during the interaction with a robot. The users were instructed to turn the light off as soon as they noticed and the responsive time was then used to measure engagement.

2.2.4 cRI

A short note on cRI is also warranted due to the target group for QTrobot. While cRI is a subfield of HRI and technicalities such as sensors, compute modules or servos themselves might be cross disciplinary the subfield itself naturally considers itself more sensitive, as an example data privacy laws in Finland is very strict often involving consent for data collection from legal guardians or parents (European Parliament & Council of the European Union 2016).

There is also the topic of evaluating interactions, it has been mentioned by Zaga et al. (2016) that questioners are of limited usage with children. Difficulties like gaining trust, both from children and guardians, is also a large factor in any child related research as found by Hervor Alma Arnadottir & Vis (2023). But there have been clear signs of ben-

efits in areas like promoting learning, especially compared to other digital tools, when robots work with children and children that have special needs (Costa et al. 2017, 2018, van den Berghe et al. 2019).

2.2.5 ASR

For chatbots to understand speech they need ASR, or speech-to-text (STT) as it is also sometimes called. There are several technological alternatives with Kaldi (Povey et al. 2011) having been the de facto standard in scientific research on speech recognition for years. However, in 2020, Meta released their Wav2vec 2.0 model (Baevski et al. 2020) introducing a new trend in speech recognition. Since then OpenAI has also released Whisper in 2022 (Radford et al. 2022) and unlike many previous hybrid systems that had to learn pronunciation, acoustic, and language with Hidden Markov Models and Gaussian Mixture Models techniques, Whisper uses a DL end-to-end system that learns all components jointly, making training easier.

2.3 LLM

The core component of the conversation system in our system is the LLM, built with the transformer (Vaswani et al. 2017), that takes prompts from the user via the ASR system. According to Dam et al. (2024) LLM-based chatbots, unlike traditional chatbots limited to basic conversational frameworks, have become a new way to generate knowledge, emerging as integral components across various domains and shifting how industries operate and interact with their users. Tokenizers play a crucial role in this process by breaking down user input into individual tokens, enabling the LLM to effectively capture context and nuances.

Tokenizers are software components that break down text (or speech) into smaller units called tokens, which can be words, subwords, characters, or other linguistic elements. Tokenizers save on compute during training but do have some side effects, like the recent famous issue of ChatGPT-4 trying to count occurrence of the letter "r" in Strawberry. Which tokenized with Tiktoken (the tokenizer used by ChatGPT) would look like "St | raw | berry", *which clearly has two "r"s*. Tokenization is essential for natural language processing (NLP) tasks and LLMs.

Continuing on how to best utilize an LLM when developing brings up the three most common techniques: (1) Prompt Engineering, a method of interacting with LLMs by providing instructions on what kind of information to generate. While it offers ease of use, cost-effectiveness, and flexibility, its effectiveness is limited by the model's pre-existing knowledge and may not provide the most up-to-date or highly specialized information. (2) Fine-tuneing updates an LLM with new information, allowing it to become an expert in a specific topic or domain. It requires more computational resources and time, but offers customization, improved accuracy, and adaptability. (3) RAG combines an LLM and a knowledge base to provide up-to-date and detailed answers by retrieving relevant information from the database. RAG offers dynamic information but requires integration of multiple systems which can be resource intensive.

Fine-tuning can further be divided into an additional four strategies as laid out by Parthasarathy et al. (2024), (1) task-specific fine-tuning, adapting the model for summarization, code generation, classification, and question answering on specific datasets. (2) Domain-specific fine-tuning, that aims to tailor the model to comprehend specific domains, such as medical, financial, or legal fields and generate domain specific content. (3) Parameter-efficient fine-tuning (PEFT) with techniques like low-rank adaptation (LoRA) by Hu et al. (2021), quantized-LoRA (QLoRA) by Dettmers et al. (2023), works by updating a small subset of model parameters and (4) Half Fine-Tuning, which aims to strike a balance between retaining pre-trained knowledge and learning new tasks by updating only half of the model's parameters during each fine-tuning round. As LLMs are manipulated a need to evaluate them arises, this will be discussed in the next section.

2.3.1 Evaluation of LLMs

Hugging Face is a ML and data science platform and community with over a million users and one of, if not the, largest communities for AI and ML development. In their blog by Fourrier (2024), they mention 3 main ways of evaluating LLMs: (1) Humans as judges, which is manual work by human judges prompt and then score outputs, (2) canary-testing, were prompt evaluation is done by the developers themselves and (3) using other LLMs as judges and even prompters.

Chang et al. (2023) found that the appropriate evaluation method should be chosen according to the specific problem and data characteristics, for more reliable performance indicators. While models can be evaluated on many different metrics there is constantly new models released, but there is also a steady flow of benchmarks and evaluation pushing boundaries like Lingoly (Bean et al. 2024). Lingoly specifically evaluates reasoning models like o1. Other outlier evaluations, such as evaluating creativity, has been done by Zhao et al. (2024), that used a team of psychologist along GPT-3 to score creativity. As such there is not yet one defining way to evaluate LLMs.

2.4 Motivation

Building upon the existing literature on LLMs and robotic systems, this work will implement a conversation system, much like a chatbot, on the QTrobot that addresses some of the limitations mentioned. Specifically, our research will investigate the integration of LLMs into the QTrobot's dialogue to provide users with a more engaging and informative experience. Moreover, by utilizing open-source software and running on local hardware, our system will ensure a better chance for data privacy and security, reducing the risk of potential security breaches associated with cloud-based, proprietary or centralized systems as found by Biström et al. (2024) and Biström et al. (2022b) while also considering the cognitive and social requisites of marginalized users as mentioned by Ko et al. (2024) as designs can be adjusted and custom functionality implemented after user needs thanks to the systems open nature.

3. Research Methodology

3.1 Hardware

3.1.1 QTrobot

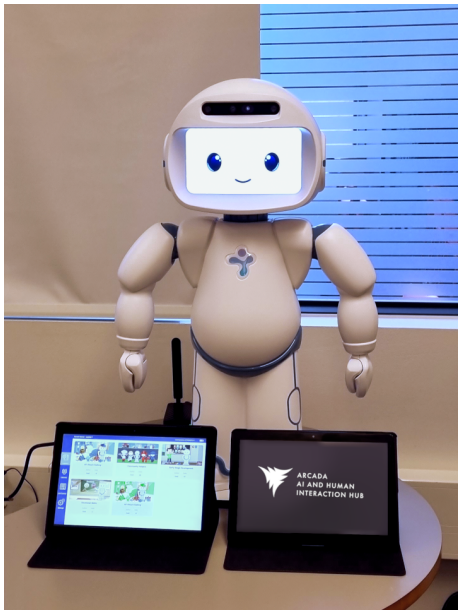


Figure 5. QT robot with educator (left) and learner (right) tablets (own photo).

QTrobot platform offers an extensive library of modules targeting cognitive, language, social, and emotional skills and is available in over 20 languages. The original curriculum offered is interfaced on the robot using 2 tablets, designated educator and learner. The robot is designed to facilitate various applications in education and has been used for research in areas of autism spectrum disorder, special needs and coaching (Jeglinsky-Kankainen et al. 2024, Costa et al. 2018, Spitale et al. 2023). In the context of education, QTrobot has been implemented as an assistive tool for teachers and educators seeking to enhance student engagement and progress monitoring (Zou et al. 2022, Issa et al. 2023).

García et al. (2020) notes that most commonly robotic software development is done following the principles of component-based software engineering (CBSE), similar in concept to object-oriented programming (OOP), for modularity and reusability. This is also the case for the QTrobot platform. As depicted in the accompanying diagram, see Figure 6, the QTrobot system consists of two interconnected computers: a Raspberry Pi-based

controller that is embedded within the robot's head (QTRP) and the primary compute component, an Intel NUC i7 PC with 32GB RAM within its body (QTPC). The two units are connected via an Ethernet cable with all hardware components connected to QTRP, save for the 3D camera integrated via the QTPC. The Wi-Fi module on QTRP broadcasts a hotspot with a unique SSID matching the robot's serial number (e.g., QTRD000101). Both computers run on Ubuntu/Debian Linux operating systems, leveraging ROS for a flexible architecture. From 2024 a newer version of the Robot, replacing the NUC with an Nvidia Jetson AGX, is also available with additional compute power.

When started, QTRP initializes the ROS environment and runs roscore, while also powering up QTPC via Wake-on-LAN. The QTrobot features two USB ports at its rear: one attached to QTRP and another connected to QTPC via a USB-C port. This USB-C port can be used to connect peripherals like keyboards, mice, and monitors using an extension hub, allowing for local development and troubleshooting. APIs are also provided for programming languages in C++, Python, and JavaScript for software development directly on the QTPC.

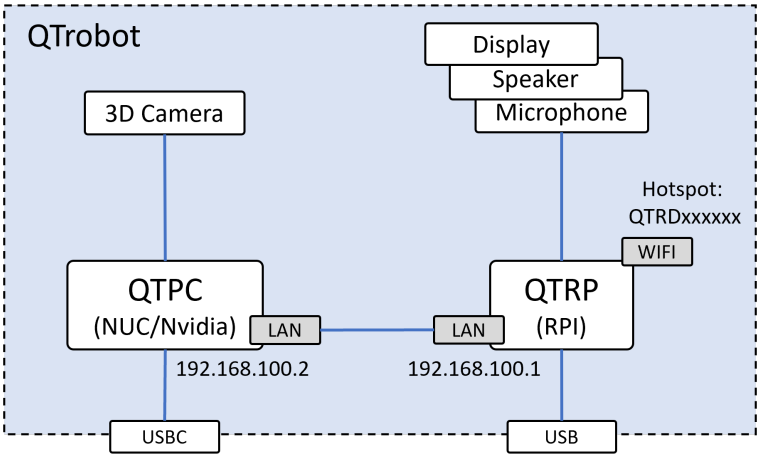


Figure 6. Overview of QT architecture, image from Luxai S.A. (2024).

The robot QT has been used in plenty of research such as Rodríguez-Lera (2018) where QT acted as a coach and emotional support to help long term young rehabilitation subjects. Together with the robot Misty, QT has also seen success in exploring robot coaching in mindfulness and well-being Spitale et al. (2023), Axelsson et al. (2023) adding further relevance to robots as useful research tools.

3.1.2 Jetson Orin Nano

The NVIDIA Jetson Orin Nano Developer Kit is made for entry-level AI development by offering performance in a compact package. With up to 40 TOPS of AI processing power and an array of connectors for robotics applications. The Jetson Nano Orin supports compute unified device architecture (CUDA) (NVIDIA et al. 2020) via it's shared 8GB of LPDDR5 DRAM and 68GB/s of bandwidth. It runs on a modified version of Ubuntu 20.04 and comes pre-installed with Jetson containers (a custom overlay for docker and containerization). Among the pre-approved containers is Ollama (Jeffrey Morgan Et al. 2023), which is a framework for running language models locally, ready with a local server and port. The Jetson Orin Nano used in this work has a PCIE 3.0 1TB SSD installed for enhanced performance and storage.

3.2 Software

3.2.1 ROS

QTrobot runs ROS Noetic Ninjemys builds on top of Ubuntu 20.04 (Focal) and uses C++ as the standard. It supports Python 3 exclusively. ROS Noetic is an updated distribution based on ROS Melodic, phasing out rosbuild in favor of catkin. ROS primarly works as a communications layer between components, laying much of the groundwork for CBSE in robotics.

3.2.2 Llama.cpp

Written by Gerganov & et al (2023), is a FOSS library written in C++ that enables inference on a wide range of LLMs, including Llama. Developed in tandem with Georgi Gerganov Machine Learning (GGML) to provide flexible and efficient model support for multiple back-ends, including x86, ARM, CUDA, Metal, Vulkan, and SYCL. A notable feature is its support for ahead-of-time model quantization, a technique that optimizes model performance by reducing precision requirements during inference time. In contrast to on-the-fly quantization, ahead-of-time quantization enables Llama.cpp to take advantage of specialized CPU extensions, specifically the library utilizes AVX, AVX2, and AVX-512 instructions on X86-64 platforms, as well as Neon on ARM architectures essential for optimizing runtime on many of the mentioned back-ends.

3.2.3 Ollama

Ollama, (Jeffrey Morgan Et al. 2023), is a FOSS framework that enables local execution of LLMs on user devices, ensuring data privacy and security while providing faster processing speeds. Ollama provides access to a range of pre-trained LLMs, including popular models like Llama 3, tailored to different tasks, domains, and hardware capabilities. Ollama integrates with various tools, frameworks, and programming languages like Python, LangChain, LlamaIndex and more, facilitating the incorporation of LLMs into workflows. Ollama is built on top of Llama.cpp and in this thesis work it runs on the Jetson Nano Orin in a container.

3.2.4 Llama 3.1 8B

Llama 3 (Grattafiori et al. 2024), is a family of large language models published by Meta AI. It was trained on a filtered dataset using heuristic filters, not safe for work (NSFW) filters, semantic deduplication approaches, and text classifiers powered by Llama 2. The data mix was carefully selected to achieve strong performance across various use cases and uses tokenizer, showing a 15% decrease in tokens allowing inference efficiency to stay on par with previous Llama 2 7B even with this model being 1B larger. On the 8B model Meta specifically found that training on larger datasets improved quality up to 15 trillion tokens.

Llama 3 uses a combination of supervised fine-tuning, rejection sampling, proximal policy optimization, and direct preference optimization for instruction fine-tuning. Instruction fine-tuning also plays a major role in ensuring safety, with comprehensive testing using human experts and automation methods to generate elicit problematic responses, iterating over releases to ensure safety on chemical, biological, cyber security, and other risk areas. While current work on safety is limited on Llama 3 much work has been done on the previous model, Llama 2. Khondaker et al. (2024), for example worked on detoxifying the model using ChatGPT to reformulate toxic content (harmful, offensive, or damaging content) while trying to retain agency in the content. The opposite is however also true, removing guardrails from models was already shown to be possible by Volkov (2024) via fine-tuning.

The choice of using Llama 3 for this project was heavily influenced by Meta's stance on having the model open as well as the popularity of previous versions. The Llama 3 models have also been further enhanced with 3.1 and as of writing 3.2 with vision support. While models such as Qwen 2.5 by Bai et al. (2023) from Alibaba Cloud show stronger scores on several leaderboards it falls some scrutiny due to it's strong ties with China. Moreover, models like Intellect-1, a open-source collaboratively trained model by Jaghouar et al. (2024) or OLMo from the team at the Allen Institute, Groeneveld et al. (2024), that passes the strict open-source definition from the Open Source Initiative Open Source Initiative (2024), could be argued to have stronger ethical claims do not have the same widespread support or popularity. Therefor Llama 3.1 is the definite choice at the time of writing.

3.2.5 MMLU-Pro

Hendrycks et al. (2021) released Measuring Massive Multitask Language Understanding (MMLU) with around 16 000 questions from different fields in biology, business, chemistry, computer science, economics, engineering, health, history, law, math, philosophy, physics, psychology and a finally category called other with questions that don't fit in previous categories for automating the process of evaluating LLMs. MMLU-Pro (Wang et al. 2024) was later released as a curated version, boasting more difficult questions to keep up with more capable models.

As Llama 3.1 is the model of choice and Meta AI used the MMLU-Pro in their original blog post for Llama3 (Meta Platforms, Inc. 2024) benchmarking on the same evaluation allows for more integrity. The repository Ollama MMLU-Pro ("Chigkim" 2024) also makes it simple to run the benchmark locally via Ollama. Other benchmarks like the EleutherAI Language Model Evaluation Harness (Sutawika et al. 2023) that is used on Huggingface's Open LLM Leaderboard was also considered for this thesis. However, given the fine-tuned models limited scope, benchmarking it on MMLU-Pro is a reasonable goal to demonstrate its adequacy.

3.2.6 Pandas

Chosen for familiarity; Pandas, made by pandas development team (2020), is a Python software library specialized for data exploration, manipulation and analysis using DataFrames.

It was used for exploring, cleaning and aggregating data.

3.2.7 Promptwright

Promptwright by StacklokLabs (2024), is a Python library used for generating large synthetic datasets using local LLMs and prompt enforcing (confining outputs for a specific format) and works various LLM service providers like OpenAI, Anthropic, and OpenRouter. Promptwright works with most LLM service providers and locally with Ollama, which is how it was used in this project.

3.2.8 Unsloth

Developed by Han & Han (2023), Unsloth is a framework for fine-tuning of LLMs with techniques for accelerated training, reduced memory consumption, and enhanced accuracy. The architecture utilizes QLoRA (Dettmers et al. 2023), Flash attention (Dao et al. 2022), look-ahead mask (or casual masks) and cross entropy loss. As Llama 3.1, the model used in this project, is available and we have access to an Nvidia GPU for fine-tuning the benchmarks show a clear advantage in using Unsloth compared to alternatives like Axlotl (Ebrahimi et al. 2024).

3.2.9 Whisper and Whisper.cpp

Whisper.cpp by Georgi Gerganov Et al (2022) is based on the original paper from Radford et al. (2022) but optimized using C++. Whisper is an ASR system trained on 680,000 hours of multilingual data and enables transcription in multiple languages and translation from those languages into English. The Whisper architecture uses a simple end-to-end Transformer approach, with a decoder predicting text captions and performing tasks such as language identification and speech translation. From previous working experience with whisper we found it to perform adequate even on a minority language like Finland-Swedish as compared to more traditional ASR systems like Kaldi or Wav2vec (Espinosa-Leal et al. 2024). Whisper.cpp also has the advantage of not needing special hardware or separate pre-trained models and can be used directly on QTrobot via straightforward bash scripts and the project is open-source.

4. Experiments

4.1 Deploying components on the robot

Apart from Whisper.cpp and existing Python ROS API most logic and functionality had to be built from scratch. Starting with voice recordings, the existing API with the Respeaker used ros.spin, a functionality that occupies the current thread, which is a problem as the Respeaker is connected to the QTRP and not the QTPC, leaving it waiting and unable to handle other matters in the pipeline. To address this issue, each recording was run in its own thread with a handler function, that terminates once a stop signal is received from a handler function. The handler communicates with the tablet interface for user interactions. Using threads also allowed for extra redundancy, such as timeouts, in case users forget to initiate a stop command.

The recording is then locally saved and sent to whisper.cpp for inference. While Python wrappers exist for whisper.cpp, most were out of date since whisper.cpp was frequently updated. However, whisper.cpp can be run using bash commands, so a subprocess function in Python was written instead. A cleanup handler was implemented for any unavoidable bash syntax or whisper punctuation errors. An issue when using whisper.cpp is its tendency to hallucinate on silence; so, implementing voice activity detection (VAD) is necessary when using Whisper, fortunately the Respeaker already implements VAD so extra work was not necessary.

Once an interpretation was generated on the audio, it was forwarded to the Jetson Orin Nano running Ollama and hosting the fine-tuned model. Ollama has its own Python library for simple communication with standard request/respond protocols over HTTPS. The returning response from the LLM was then synthesized using QTrobot's built-in TTS while asynchronously performing a random movement from a list of predefined movements available directly through the ROS api. While TTS frameworks like TorTise (Betker 2023) and StyleTTS 2 (Li et al. 2023) where investigated for their high-quality speech synthesis and ability to be fine-tune for customized and tailored voices both models make heavy use of GPU acceleration for best performance, far beyond the capabilities of the Jetson Orin Nano. Lightweight TTS solutions do exist and have been around for

a long time and are accessible for local usage. However, Qtrobot's own Acapela voice specifically speaks on a wide range of voices and languages and brings everything needed for our system. While not open-source it does perform adequately for the use-case until more powerful embedded hardware can be acquired for utilizing the more demanding TTS solutions.

From previous work with social robots (Kuvaja Adolfsson 2022, Biström et al. 2022a) an ASR a technique utilizing keywords was implemented. The concept is to emphasis a specific keyword (or hot-word) to trigger an action or function. This allowed the whole process to be looped until the user exited, preferably by speaking one of two keywords 'bye' or 'goodbye'. On exit a pre-determined goodbye function was called, invoking the robot to say its farewells and wave at the user before returning to a neutral pose. The full flow can be seen in Figure 7.

Figure 7. Overview of component flow

The system used 8 classes to achieve the pipeline, including an asynchronous class directly from LuxAI for using gestures and TTS at the same time. The tablet interface was created using the standard LuxAI studio program accessible in a cloud management system and interfaces via a customly written sermo class. The sermo class enables modularity, enabling new features to be quickly added in the future and during development. All functions from the sermo class had to be uploaded as separate ROS service handlers to the QTRP, such as starting a conversation or stopping recording, to be usable on the QTPC via ROS on the tablets.

To track user interactions and assist with bug tracking, a tracker class was built that created a unique anonymous ID at the beginning of each conversation and timestamped all actions and responses. The data was then saved locally. A separate project was also built for viewing and analyzing the data using a PlotlyDash server.

To write and read files locally, a filename class was made, which kept the anonymity of all

files intact and allowed for reusability in line with CBSE principles by storing file paths in a separate location on the robot.

García et al. (2020) mentions integration testing is most preferable by robot developers for testing. In earlier works we have experienced the same (Kuvaja Adolfsson 2022), while separate tests or even unit tests can be usable on individual components it isn't until the whole system is deployed that you will find all the bugs and have an opportunity to straighten out the last small kinks. As robot development is overall very complex there are many things that can go wrong due to unforeseeable implementations on manufacturer choices or your own. During this thesis work and experiment the laboratory had a failing router that was only discovered when doing full integration testing as we troubleshooted connectivity issues between the tablets and the robot.

4.2 Synthesizing data and instruction-tuning

Initial experiments to collect data for fine-tuning mainly concerned searching for children's corpora or turning children's stories into instruction sets. Corpora specifically on children's language and speech was sparse, some projects such as PF STAR (Batliner et al. 2005), My Science Tutor (Pradhan et al. 2023) had difficult licenses or were not openly available or the rumored CMU Kids Corpus that was simply not found lead to several dead ends. Project Gutenberg. (n.d.) (1971) however, is a large online library with a large section of children's book, more precisely 749 children's books were found and downloaded in September 2024. After exploring and cleaning all books using Pandas the texts were tokenized using Tiktoken showing the average token length to be 54,000 tokens. While different methods exist, such as summarization by an LLM or simpler TF-IDF techniques to distill text, the data quality is certain to be questionable. As such pure synthetic data synthetically produced by LLMs was eventually chosen as a viable alternative as models fit for the Jetson Orin Nano's 8GB memory do not have the context length to support 54,000 tokens.

The Promptwright library allowed for quick and simple initiation of data generation and format enforcement. When using synthetic data, creativity was prioritized by setting a high temperature as suggested by Peeperkorn et al. (2024). This ensured as many new

data points as possible, aligning with the core issue at hand: a lack of data. For system prompt in Promptwright the following was used: *'You are the child friendly and helpful robot assistant QT. You provide clear and concise answers to user questions in a positive and child friendly way. You use simple language that is easy to understand.'* The assistant and user response samples given in the code were kept at default values, asking for a simple numeric question. This allows inherent biases but facilitates easy identification during data exploration and cleaning later.

AI models come in various forms, most commonly models are available in their normally trained form and as instruct-tuned. "Normally" trained models are typically trained to generate text based on a wide range of input data, focusing primarily on language fluency and coherence. In contrast, instruct models undergo additional fine-tuning to better understand and follow specific instructions provided by users. This fine-tuning allows them to perform tasks more accurately and contextually, making them particularly useful for applications that require precise and task-oriented responses. In this work, for all available cases, instruct models were used for synthesizing data. Models were chosen based on arbitrary popularity, sorted by Ollama's model interface. Some models were skipped due to lacking instruction models or being aimed at non-conversational tasks, such as coding. 20 models were tested, set to generate 500 rows of data each, with a total goal of 10,000 data points.

Fine-tuning has been done on as few as 2,000 data points in LIMA by Zhou et al. (2023), leaving room for failures in our experiment. Once generated, all outputs went through a custom filter: Using Pandas and regex to remove standard duplicates and any duplicates from the original numeric instructions to avoid bias. Structurally faulty outputs were also discarded, as well as outputs where the system instructed the user to ask for help constructing a "prompt" (or "prompts"). Statistics on models used and dropped rows can be seen in Table 1.

Future, using a simple mathematical algorithm to build an accuracy scale is suggested in this thesis for determining how creative a model is when being format enforced for synthetic data generation. Accuracy gives a score between 0 and 100 on how well the model

performs for synthetic data generation. Calculating accuracy is suggested to done by taking the Total number of generations (Tg), subtracting the number of accepted generations (ag), dividing it by Tg and multiply by 100, as seen in Equation 1.

$$accuracy = (Tg - ag)/Tg * 100 \tag{1}$$

The remaining dataset, just over 2100 rows with 'system', 'user' and 'assistant' prompts, was used to fine-tune Llama 3.1 8B model using Unsloth. Fine-tuning usually involves LoRA (Hu et al. 2021), which adds a new weight matrix that adjusts the original weights of the model, making it possible to adjust and even add knowledge.

A Quantized version of LoRA, developed by Dettmers et al. (2023), improves fine-tuning speed and memory utilization. For this project QLoRA was used, due to the availability of consumer hardware, an Nvidia RTX 4090, and the pre-Quantized Llama 3.1 8B model, which ensured sufficient VRAM resources were available. Fine-tuning on the dataset took around 8 minutes using AdamW optimizer, preferred for fine-tuning pre-trained models as Parthasarathy et al. (2024) laid out in their work. A batch size of 2x4 was used, the default recommendation, as smaller batches yield better results during fine-tuning. The resulting model was exported in GPT-Generated Unified Format (GGUF), a widely accepted format supported by Ollama.

Table 1. Table gives an overview of models used with short descriptions and statistics on filtered data, wrong output and an accuracy score.

Model	Description	Filtered	Wrong Structure	Accuracy
WizardLM2 7B	From Microsoft AI, aimed at complex chat, multilingual tasks, reasoning and AI agent use cases for Microsoft.	428	72	0.0
Solar-Ko-Recovery-11B	Aims to recover capabilities for Korean by rearranging embeddings and LM head, featuring an expanded vocabulary and bilingual corpus.	108	385	1.4
Qwen 2.5 7B	From Alibaba cloud with a focus on instruction tuning and enhanced safety measures.	489	1	2.0
OpenHermes 2.5 Mistral 7B	From Nous Research is trained on 1 million entries of primarily GPT-4 generated data and other high-quality open datasets, extensively filtered and converted to ShareGPT format using ChatML.	454	17	5.8
Gemma 9B	From Google, is a lightweight, SOTA open model suitable for text generation tasks like QA and summarization.	444	0	11.2
Orca2 13B	From Microsoft, built for research purposes only and provides single-turn responses and excels in tasks such as reasoning over user-provided data, reading comprehension, math problem solving, and text summarization.	379	45	15.2
Mistral Nemo 12B	SOTA LLMs from France, trained on extensively filtered public datasets and GPT-4 generated data.	285	123	18.4
TinyLlama 1B	Aims to bring Llama 2 performance in a smaller scale using the same architecture.	347	57	19.2
Llama 3 8B	Pretrained and fine-tuned generative text models from Meta utilizing QLoRA for improved instruction adherence.	74	325	20.2
Hermes3 8B	From Nous Research is a generalist language model with improved role playing, reasoning, long context coherence, and agentic capabilities. Focused on aligning LLMs to user needs with powerful steering capabilities.	378	13	21.8
GLM-4-9B	From Zhipu AI in China, an open-source pre-trained model with multi-language support up to 26 languages.	376	3	24.2
Mistral Small 22B	SOTA LLMs from France, trained on extensively filtered public datasets and GPT-4 generated data.	361	0	27.8
Meta Llama 2 7B	Pretrained and fine-tuned generative text models from Meta utilizing QLoRA for improved instruction adherence.	170	140	38.0
Mistral 7B	SOTA LLMs from France, trained on extensively filtered public datasets and GPT-4 generated data.	165	126	41.8
Zephyr 7B	A fine-tuned version of Mistral 7B v0.1 trained with DPO for enhanced helpfulness in assistant roles.	242	32	45.2
Llama 2 7B Uncensored	An uncensored Llama 2 7B fine-tuned on Wizard-Vicuna conversation dataset	184	45	54.2
Qwen 2 7B	From Alibaba cloud with a focus on instruction tuning and enhanced safety measures.	168	10	64.4
InternLM2 7B	Is a high-quality adaptable base model for deep domain adaptation and supports ultra-long contexts up to 200,000 characters	130	43	65.4
Qwen series 2.5 3B	From Alibaba cloud with a focus on instruction tuning and enhanced safety measures.	26	54	84.0

5. Results

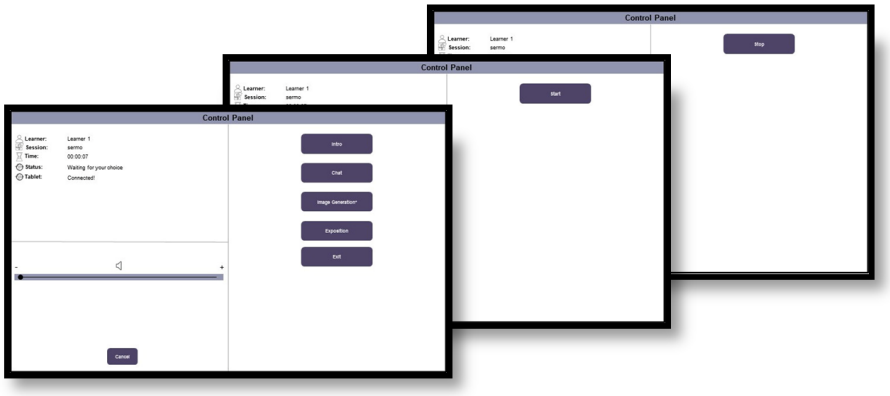


Figure 8. Graphic of Sermo user interface as presented on LuxAI educator tablet, showing main menu (top), start recording (middle), stop recording (bottom).

The conversation system's full pipeline operates through ROS, starting from the LuxAI educator tablet with a custom program, developed using LuxAI Studio, see Figure 8. The program allows the users to run the conversation system by choosing "Chat" via the interface. The start button beings recording and is replaced with a stop button once pressed. A separate Python thread handles communication between the tablet interface and the robot during speech and recording. Additionally, a timeout was implemented to cater to end-users who might be unfamiliar with the two-step process of starting and stopping recording.

Once the audio recording is done it gets stored and transcribed on the QTPC. Processed locally using Whisper.cpp, called via a Python *subprocess* and bash commands we ensure that data handled locally. The results from Whisper is then passed through a simple regex cleaning function to check for specific keywords. Currently implemented keywords include 'bye' and 'goodbye'. If these keywords are detected, a default end message is spoken with a 'wave' gesture from the robot. Otherwise, interaction continues as follows: a random gesture from the preconfigured QTrobot library is picked, such as wave, point or dance, via a custom function and played while a query is sent to the Ollama server. Once the Ollama server gives a response its sent back to QTPC and spoken via Acapela TTS. Simultaneously, a new random gesture and QTrobots default facial animation for speech is played. Users can repeat the conversation as needed by recording again or ending the

interaction by speaking one of the two keywords. Conversations are kept in the models memory until a new "Chat" is initiated and an overview of the full flow, starting from the user interaction, can be seen in Figure 9.

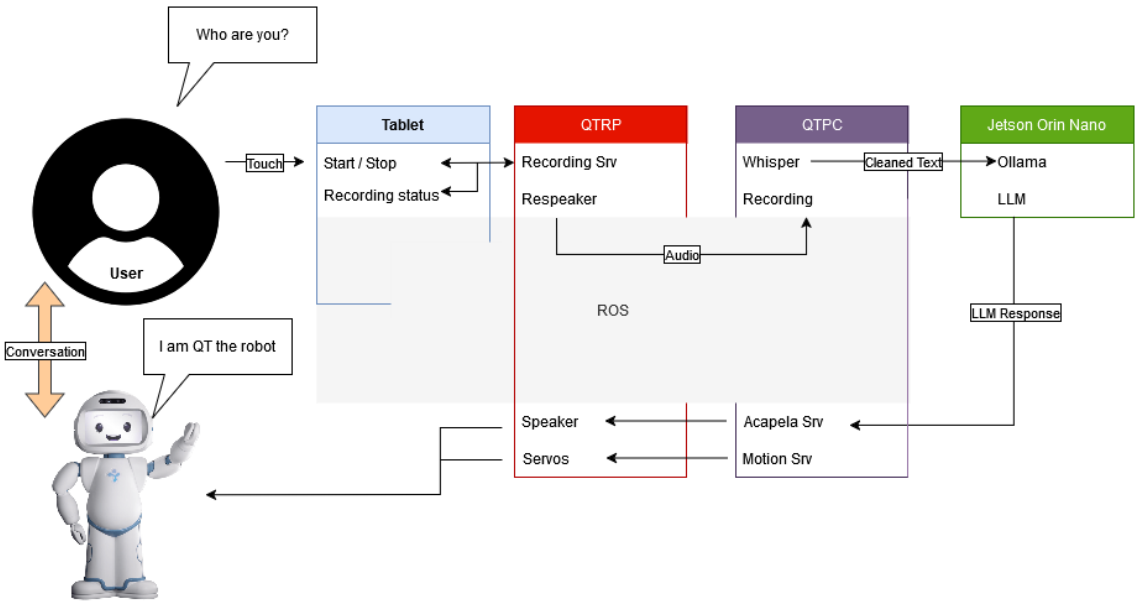


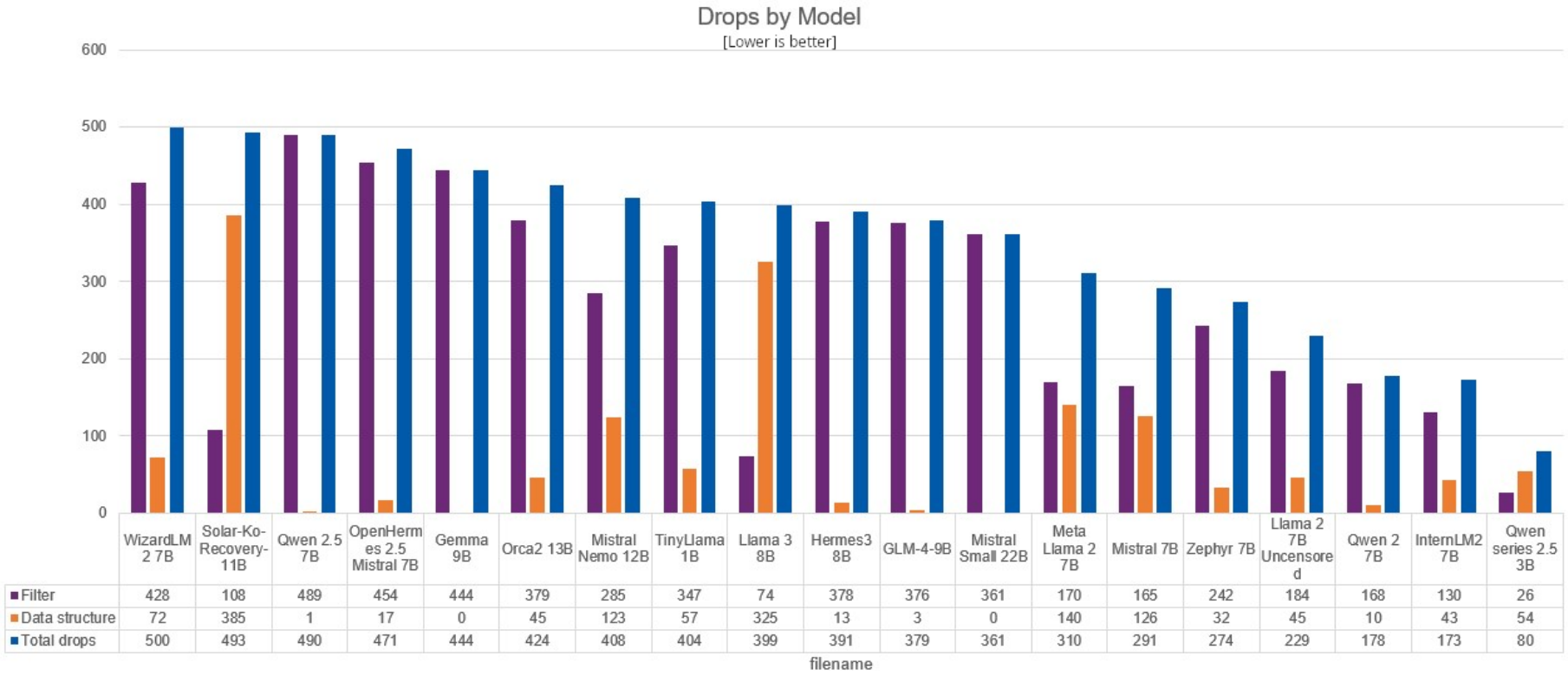
Figure 9. Overview of full conversation system and flow, starting from the user interaction.

5.1 Data synthesis evaluation

Constructing a bar chart from our previous Table 1 gives us Figure 10. We can see clear differences in the models' ability to create suitable synthetic data. Especially Solar-Ko and Llama 3.1 8B have a strong drop in correct data structure. This means that the generated synthetic data was not being enforced. Upon closer inspection, it was common for Llama to use conversation chains of several "assistant" interactions (almost 300 extra) rather than the requested single generation, while skipping over 100 "user" generations. In contrast, Solar-Ko had a tendency to generate extra system prompts, almost doubling the number of system messages with 890 system prompts compared to the expected 500. Both cases could be worked around by simply dropping or reformatting the outputs if needed. Other frameworks than Promptwright might also be more successful.

The custom filter with regex removed all mentions of the original question: "2 + 2" and additionally "prompt" as it was the highest occurring word after auxiliary words. "2 and 2" was however not filtered, with 40 occurrences in the final dataset, an oversight as it's almost 2% of the dataset. Using Natural Language Toolkit (Bird & Klein 2009) to count

words, we also find that "story" occurs 491 times, followed by "apples" at 459 occurrences. When giving mathematical examples it is quite common to suggest counting with "apples", which might explain the high occurrence considering the default mathematical prompt used, whereas "stories" are more child-inspired, it aligning more with the default system prompt. Further evaluation would be time consuming and challenging considering the scope of the project, but the conclusion that models like Qwen, Llama2, and InternLM2 are well-suited for generating synthetic data with a high success rate should be noted as seen in Figure 10.



	WizardLM 2 7B	Solar-Ko-Recovery-11B	Qwen 2.5 7B	OpenHermes 2.5 Mistral 7B	Gemma 9B	Orca2 13B	Mistral Nemo 12B	TinyLlama 1B	Llama 3 8B	Hermes3 8B	GLM-4-9B	Mistral Small 22B	Meta Llama 2 7B	Mistral 7B	Zephyr 7B	Llama 2 7B Uncensored	Qwen 2 7B	InternLM2 7B	Qwen series 2.5 3B
Filter	428	108	489	454	444	379	285	347	74	378	376	361	170	165	242	184	168	130	26
Data structure	72	385	1	17	0	45	123	57	325	13	3	0	140	126	32	45	10	43	54
Total drops	500	493	490	471	444	424	408	404	399	391	379	361	310	291	274	229	178	173	80

Figure 10. Bar plot on model dropout from synthetic data generation.

5.2 LLM evaluation

Wang et al. (2024) author of MMLU-pro, which is built from MMLU (Hendrycks et al. 2021), was used to benchmark the fine-tuned model and the original quantized model via local inference at a temperature of 0 to avoid variance. Table 2 shows the MMLU-Pro scores and using standard deviation (σ) we can see a small divergence in the scores however nothing impeccable. The largest differences being in health and psychology in the original models favor. Considering the rather small data amount and modest curation of the dataset used for fine-tuning we can draw the conclusion that the fine-tuning had a very small, if any, effect. While prompting the models individually, in an attempt to be judge, the chats could be said to have minor nuance, but this might as well be due to the technology of LLMs varied response rather than the fine-tuning. Drawing any conclusions would simply be speculation without very extensive testing which considering the results was not deemed of interest. Of final note is that the model does run within similar performance as the original model on the Jetson Nano Orin, generating around the same 15 tokens per second.

Table 2. MMLU-Pro scores for all categories for fine-tuned and original models and standard deviation (σ) between scores.

Name	overall	biology	business	chemistry	computer science	economics	engineering	health	history	law	math	philosophy	physics	psychology	other
Fine-tuned 8B Q4	41.46	64.30	42.59	32.51	43.41	52.96	32.09	48.66	44.09	27.52	38.42	40.48	33.26	56.89	44.59
Llama 3.1 8B Q4	42.88	61.65	44.49	34.45	43.17	52.25	31.37	53.42	43.04	29.88	39.45	44.29	35.64	59.90	46.43
σ	1.00	1.87	1.34	1.37	0.17	0.50	0.51	3.37	0.74	1.67	0.73	2.69	1.68	2.13	1.30

6. Conclusions

The field of technology is developing extraordinarily fast, AI is making new developments at a rapid rate and deploying solutions can take longer than technologies stay cutting edge such as Llama 3.1 already having been iterated upon in Llama 3.2 and Llama 3.3. As such be mindful that technologies presented in this work might have been surpassed. However, FOSS solutions with large communities have been utilized and these commuunities have so far been quick on integrating new cutting edge discoveries.

6.1 Thesis goals

While overcoming preference for human interaction over chatbots as Dünya & Durak (2023) set out to do might be too tall a task for this thesis and far from anything this system achieved, this conversation system will serve as a base to build future projects upon now that it is deployed and working. QT sermo did manage to bring the interaction sought, the ability to converse with a robot. The conversation system as it stands serves as a foundation for future development due to its modularity. Support for integrating new AI services, such as image generation or retrieval-augmented generation is possible as Ollama evolves and new capable models, like Llama 3.2 with vision is released. Ollama tool handling, a way for models to directly use actions via code, remains an unexplored feature in the scope of this project, but the goal to implement such features by continuing this project with more powerful hardware, such as a Jetson AGX Orin with 64 GB of RAM is next on the agenda.

The fine-tuned model has also shown adequate performance but leaves room for significant improvements in subsequent iterations. Based upon the experimental results there is a possibility of generating larger and more qualitative datasets using the same methods and using the most accurate models from Table 1. Something that is much-needed considering the lack of available datasets aimed at children. Utilizing the collected and explored data from Project Gutenberg. (n.d.) (1971) the early experiments should not be dismissed either as it might prove useful for generating more qualitative data at a later date.

Evaluating what this thesis set out to do, it can be concluded that the first two questions (1) to develop a functional conversation system on QTrobot with a Jetson Orin Nano and

(2) leverage open-source software solution to develop the system had a positive outcomes. The system is working and it does so with mostly FOSS solutions, only the TTS, limited by hardware, was glossed over.

As for (3) improving the LLM using SOTA proved unaccomplished with fine-tuning considering the lackluster experiments and benchmarking results. The synthetic data generation experiments do prove that the means to generate larger data sets exists, which remains the largest challenge at this time. When evaluation a model, it should be considered that benchmarking via MMLU-Pro might not be the most appropriate when it comes to cRI or even HRI use cases, rather using humans as judges and prompters should be preferable as the interaction differs for each use case.

The last goal of the thesis was to keep the conversation system modular so it can be deployed on other robot platforms, and while the conversation system certainly utilizes modularity it would only be honest to conclude that future projects' success can really only be answered once proven. As such it would be necessary to deploy the conversation system on another robot platform to have legitimate proof, even if principles of CBSE and OOP was followed. In conclusion this was goal remains to be seen.

6.2 Discussion

6.2.1 ASR

A perhaps more convenient way to implement the ASR system would be to use live transcription and keeping the robot actively listening. While this was investigated there were two factors standing in the way. First and most importantly any live transcription done with Whisper has a recommended VRAM beyond the Jetson Orin Nano's capability. Secondly having a microphone constantly open with an interpretation in sensitive uses cases, such as environments with children, might not be feasible. As such the solution presented in this work was a good compromise.

Running the ASR live would most likely strike a more user-friendly experience, while Apple's Siri or Amazon's Alexa never achieved full market share, they are still familiar for many users and having built a similar "activation word" to initiate the conversation system might have been preferable. This is something that will be explored but not with

urgency much due to privacy concerns listed in the previous section.

Since the project began more SOTA models have also arrived making better progress however, it is yet to be seen if performance is up to par with the Whisper.cpp project as Georgi Gerganov Et al (2022) has done remarkable optimizations and much of their hard work stands for the success of QT sermo.

6.2.2 LLM

While a large limiting factor comes down to the hardware of Jetson Orin Nano the performance of the fine-tuned model can't be blamed on the hardware. Here, the goal of creating a more QTrobot aligned interaction was simply not ambitious enough. RAG would have for example offered a more substantial result if given a clear ambition, perhaps reading specific fairy tales from the Project Gutenberg. (n.d.) (1971) collection could have been a lofty goal. As such only investing time in engineering a performative prompt would have been better. Perhaps benchmarking different models inference time on the Jetson Orin Nano might have benefited the research community more in the long run as finding relevant information for embedded systems and LLMs was difficult.

The MMLU-Pro benchmarks showed little deviation, while scoring higher on benchmarks wasn't the goal of the fine-tuned model, it deviated so little from the original source it gives at least clear proof that it does perform and is fully usable. The model interestingly performs better in some areas, most notably biology. This is probably more due to the system instructions than that of the dataset used for fine-tuning as experiments such as the word counting the dataset showed high relevance of stories and apples, something that is unlikely to have an effect on biological questions considering the nature of the MMLU-Pro benchmark.

The data synthesizing, while not pushing the edge in research provided much valuable insights into creating synthetic datasets and perhaps some novelty in local cRI data generation. Purposing the accuracy metric for data generation also opens the door for others to experiment and find relevant models. The metric can also be expanded and used in the future for synthesizing more data for different use cases where data is lacking outside cRI.

Any projects in the near future most likely involves larger models as the Jetson Orin Nano is replaced with a Jetson Agx Orin. This means fine-tuning needs to be carefully considered as more performative models are rapidly released and generating data for fine-tuning is very time consuming. The field of LLMs also keep pushing forward, just recently Qwen by Bai et al. (2023), LG by Research et al. (2024) and Meta have been releasing new models that push the frontier forward more rapidly, and who knows what might come tomorrow?

6.2.3 TTS

While unfortunately the TTS implementation never changed from the default Acapela system used on the QTrobot due to hardware and scope limitation, investigations were made and possible solutions found. Tortise especially seems like a very flexible alternative once capable hardware is aquired. While speech synthesis is nothing new, just like ASR new transformer-decoder DL models have brought life to the subject. Currently on the table is a project for creating a natural sounding native model on a minority language in the region. The goal is to deploy this on the QTrobot and bring a new personality and if successful perhaps the other robots available at Arcada UAS will follow.

6.3 Contribution

In conclusion, the proposed conversation system has successfully integrated SOTA and FOSS technologies to provide a transparent chatbot for users on QTrobot available at an open repository[1]. Leveraging ROS, Jetson Orin Nano, Ollama and Whisper, this thesis has developed an interface for users to engage with their preferred LLM in real-time. While the LLM failed to impress the system's pipeline ensures accurate transcription based on whisper and playback using the robot's original voice. Additional mechanisms have been implemented to accommodate end-users such as support for a timeout feature and a modular interface with room to expand for future projects. The modularity of the project means it can be moved to other robot platforms in the future. Finally, a log system that timestamps all interactions with unique identifiers that allows for valuable statistics, insights and bugfixes through the use of PlotlyDash, is also available as a side project in a separate repository[2].

[1] https://github.com/qt-reachy/qt_sermo

[2] https://github.com/qt-reachy/qt_sermo_analytics

Ultimately, this conversation system serves as a small testament to the potential of diverse perspectives in AI and robotics development while staying sustainable, transparent and trustworthy by using open-source solutions. It lays the foundation for continuous development for deploying flexible conversation systems, or chatbots, on modern robotics.

Thank you for reading

References

Andtfolk, Malin. 2022, *The Possibilities for Using Humanoid Robots as a Care Resource (1st ed)*. Available: https://urn.fi/URN:ISBN:978-952-12-4201-4.

Augello, Agnese. 2022, Unveiling the reasoning processes of robots through introspective dialogues in a storytelling system: A study on the elicited empathy, *Cognitive Systems Research*, vol. 73, , pp. 12–20. Available: https://doi.org/10.1016/j.cogsys.2021.11.006.

Axelsson, Minja; Spitale, Micol & Gunes, Hatice. 2023, Robotic Coaches Delivering Group Mindfulness Practice at a Public Cafe, In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, New York, NY, USA: Association for Computing Machinery, p. 86–90. Available: https://doi.org/10.1145/3568294.3580048.

Baevski, Alexei; Zhou, Henry; Mohamed, Abdelrahman & Auli, Michael. 2020, *wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations*.

Bai, Jinze; Bai, Shuai; Chu, Yunfei; Cui, Zeyu; Dang, Kai; Deng, Xiaodong; Fan, Yang; Ge, Wenbin; Han, Yu; Huang, Fei; Hui, Binyuan; Ji, Luo; Li, Mei; Lin, Junyang; Lin, Runji; Liu, Dayiheng; Liu, Gao; Lu, Chengqiang; Lu, Keming; Ma, Jianxin; Men, Rui; Ren, Xingzhang; Ren, Xuancheng; Tan, Chuanqi; Tan, Sinan; Tu, Jianhong; Wang, Peng; Wang, Shijie; Wang, Wei; Wu, Shengguang; Xu, Benfeng; Xu, Jin; Yang, An; Yang, Hao; Yang, Jian; Yang, Shusheng; Yao, Yang; Yu, Bowen; Yuan, Hongyi; Yuan, Zheng; Zhang, Jianwei; Zhang, Xingxuan; Zhang, Yichang; Zhang, Zhenru; Zhou, Chang; Zhou, Jingren; Zhou, Xiaohuan & Zhu, Tianhang. 2023, *Qwen Technical Report*. Available: https://arxiv.org/abs/2309.16609.

Batliner, Anton; Blomberg, Mats; D'Arcy, Shona; Elenius, Daniel; Giuliani, Diego; Gerosa, Matteo; Hacker, Christian; Russell, Martin; Steidl, Stefan & Wong, Michael. 2005, The PF_STAR children's speech corpus, pp. 2761–2764.

Bean, Andrew M.; Hellsten, Simi; Mayne, Harry; Magomere, Jabez; Chi, Ethan A.; Chi, Ryan; Hale, Scott A. & Kirk, Hannah Rose. 2024, *LINGOLY: A Benchmark of Olympiad-Level Linguistic Reasoning Puzzles in Low-Resource and Extinct Languages*. Available: https://arxiv.org/abs/2406.06196.

van den Berghe, Rianne; Verhagen, Josje; Oudgenoeg-Paz, Ora; van der Ven, Sanne & Leseman, Paul. 2019, Social Robots for Language Learning: A Review, *Review of Educational Research*, vol. 89, no. 2, pp. 259–295. Available: https://doi.org/10.3102/0034654318821286.

Betker, James. 2023, *Better speech synthesis through scaling*. Available at https://github.com/neonbjb/tortoise-tts.

Bird, Edward Loper, Steven & Klein, Ewan. 2009, *Natural Language Processing with Python*.

Biström, Dennis; Kuvaja Adolfsson, Kristoffer; Tigerstedt, Christa & Espinosa-Leal, Leonardo. 2024, Open-Sourcing a Humanoid Robot, In: Michael E. Auer; Reinhard Langmann; Dominik May & Kim Roos, eds., *Smart Technologies for a Sustainable Future*, Cham: Springer Nature Switzerland, pp. 381–389.

Biström, Dennis; Edgren, Johannes; Penttinen, Johan; Kankkonen, Markus & Kuvaja Adolfsson, Kristoffer. 2022a, *Arcada Social Robots Repository*. Available at https://github.com/socbots.

Biström, Dennis; Westerlund, Magnus; Duncan, Bob & Jaatun, Martin Gilje. 2022b, Privacy and security challenges for autonomous agents : A study of two social humanoid service robots, In: *2022 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, pp. 230–237.

Chang, Yupeng; Wang, Xu; Wang, Jindong; Wu, Yuan; Yang, Linyi; Zhu, Kaijie; Chen, Hao; Yi, Xiaoyuan; Wang, Cunxiang; Wang, Yidong; Ye, Wei; Zhang, Yue; Chang, Yi; Yu, Philip S.; Yang, Qiang & Xie, Xing. 2023, *A Survey on Evaluation of Large Language Models*. Available: https://arxiv.org/abs/2307.03109.

"Chigkim". 2024, *GitHub - chigkim/Ollama-MMLU-Pro*. Available: https://github.com/chigkim/Ollama-MMLU-Pro.

Costa, Andreia; Steffgen, Georges; Rodríguez Lera, Francisco; Nazarikhorram, Aida & Ziafati, Pouyan. 2017, *Socially assistive robots for teaching emotional abilities to children with autism spectrum disorder*.

Costa, Andreia P.; Charpiot, Louise; Lera, Francisco Rodríguez; Ziafati, Pouyan; Nazarikhorram, Aida; Van Der Torre, Leendert & Steffgen, Georges. 2018, More Attention and Less Repetitive and Stereotyped Behaviors using a Robot with Children with Autism, In: *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, pp. 534–539.

Cruz-Ramírez, Sergio Rolando; García-Martínez, Moisés & Olais-Govea, José Manuel. 2022, NAO robots as context to teach numerical methods, *International Journal on Interactive Design and Manufacturing (IJIDeM)*, vol. 16, no. 4, pp. 1337–1356. Available: https://doi.org/10.1007/s12008-022-01065-y.

Dam, Sumit Kumar; Hong, Choong Seon; Qiao, Yu & Zhang, Chaoning. 2024, *A Complete Survey on LLM-based AI Chatbots*. Available: https://arxiv.org/abs/2406.16937.

Dao, Tri; Fu, Daniel Y.; Ermon, Stefano; Rudra, Atri & Ré, Christopher. 2022, *FlashAttention: Fast and Memory-Efficient Exact Attention with IO-Awareness*. Available: https://arxiv.org/abs/2205.14135.

Dettmers, Tim; Pagnoni, Artidoro; Holtzman, Ari & Zettlemoyer, Luke. 2023, *QLoRA: Efficient Finetuning of Quantized LLMs*.

Dünya, Beyza Aksu & Durak, Hatice Yıldız. 2023, Hi! Tell me how to do it: examination of undergraduate students' chatbot-integrated course experiences, *Quality & Quantity*, vol. 58, no. 4, pp. 3155–3170. Available: https://doi.org/10.1007/s11135-023-01800-x.

Ebrahimi, Sana; Chen, Kaiwen; Asudeh, Abolfazl; Das, Gautam & Koudas, Nick. 2024, *AXOLOTL: Fairness through Assisted Self-Debiasing of Large Language Model Outputs*.

Elgarf, Maha; Skantze, Gabriel & Peters, Christopher. 2021, Once Upon a Story: Can a Creative Storyteller Robot Stimulate Creativity in Children?, In: *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, IVA '21, New York, NY, USA: Association for Computing Machinery, p. 60–67. Available: https://doi.org/10.1145/3472306.3478359.

Espinosa-Leal, Leonardo; Adolfsson, Kristoffer & Shcherbakov, Andrey. 2024, *Automatic Speech Recognition of Finnish-Swedish Dialects: A Comparison of Three Cutting-Edge Technologies*, Springer International Publishing, pp. 309–315.

Espinosa-Leal, Leonardo; Chapman, Anthony & Westerlund, Magnus. 2020, Autonomous industrial management via reinforcement learning, *Journal of intelligent & Fuzzy systems*, vol. 39, no. 6, pp. 8427–8439.

European Parliament & Council of the European Union. 2016, *Regulation (EU) 2016/679 of the European Parliament and of the Council*. Available: https://data.europa.eu/eli/reg/2016/679/oj.

Fourrier, Clémentine. 2024, *Let's talk about LLM evaluation*. Available: https://huggingface.co/blog/clefourrier/llm-evaluation.

Fu, Changzeng; Liu, Chaoran; Ishi, Carlos Toshinori; Yoshikawa, Yuichiro; Iio, Takamasa & Ishiguro, Hiroshi. 2021, Using an Android Robot to Improve Social Connectedness by Sharing Recent Experiences of Group Members in Human-Robot Conversations, *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6670–6677.

Fusté-Forné, Francesc. 2021, Robot chefs in gastronomy tourism: What's on the menu?, *Tourism Management Perspectives*, vol. 37, , p. 100774. Available: https://www.sciencedirect.com/science/article/pii/S2211973620301410.

García, Sergio; Strüber, Daniel; Brugali, Davide; Berger, Thorsten & Pelliccione, Patrizio. 2020, Robotics software engineering: a perspective from the service robotics domain, In: *Proceedings of the 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2020, New York, NY, USA: Association for Computing Machinery, p. 593–604. Available: https://doi.org/10.1145/3368089.3409743.

Georgi Gerganov Et al. 2022, *whisper.cpp*. Available at https://github.com/ggerganov/whisper.cpp.

Gerganov, Georgi & et al. *2023*. Available at https://github.com/ggerganov/llama.cpp.

Gonzalez-Aguirre, Juan Angel; Osorio-Oliveros, Ricardo; Rodriguez-Hernandez, Karen L; Lizárraga-Iturralde, Javier; Morales Menendez, Ruben; Ramirez-Mendoza, Ricardo A; Ramirez-Moreno, Mauricio Adolfo & Lozoya-Santos, Jorge de Jesus. 2021, Service robots: Trends and technology, *Applied Sciences*, vol. 11, no. 22, p. 10702.

Grattafiori, Aaron; Dubey, Abhimanyu et al.. 2024, *The Llama 3 Herd of Models*. Available: https://arxiv.org/abs/2407.21783.

Groeneveld, Dirk; Beltagy, Iz; Walsh, Pete; Bhagia, Akshita; Kinney, Rodney; Tafjord, Oyvind; Jha, Ananya Harsh; Ivison, Hamish; Magnusson, Ian; Wang, Yizhong; Arora, Shane; Atkinson, David; Authur, Russell; Chandu, Khyathi Raghavi; Cohan, Arman; Dumas, Jennifer; Elazar, Yanai; Gu, Yuling; Hessel, Jack; Khot, Tushar; Merrill, William; Morrison, Jacob; Muennighoff, Niklas; Naik, Aakanksha; Nam, Crystal; Peters, Matthew E.; Pyatkin, Valentina; Ravichander, Abhilasha; Schwenk, Dustin; Shah, Saurabh; Smith, Will; Strubell, Emma; Subramani, Nishant; Wortsman, Mitchell; Dasigi, Pradeep; Lambert, Nathan; Richardson, Kyle; Zettlemoyer, Luke; Dodge, Jesse; Lo, Kyle; Soldaini, Luca; Smith, Noah A. & Hajishirzi, Hannaneh. 2024, *OLMo: Accelerating the Science of Language Models*. Available: https://arxiv.org/abs/2402.00838.

Hägele, Martin; Nilsson, Klas; Pires, J Norberto & Bischoff, Rainer. 2016, Industrial robotics, *Springer handbook of robotics*, pp. 1385–1422.

Han, Daniel & Han, Michael. *2023*. Available at https://github.com/unslothai/unsloth.

Heikel, Edvard & Espinosa-Leal, Leonardo. 2022, Indoor scene recognition via object detection and TF-IDF, *Journal of Imaging*, vol. 8, no. 8, p. 209.

Hendrycks, Dan; Burns, Collin; Basart, Steven; Zou, Andy; Mazeika, Mantas; Song, Dawn & Steinhardt, Jacob. 2021, *Measuring Massive Multitask Language Understanding*. Available: https://arxiv.org/abs/2009.03300.

Hervor Alma Arnadottir, Sissel Seim, Guðrún Kristinsdottir & Vis, Svein. 2023, Challenges for researchers when getting access to children and young people and their consent in research. A scoping review, *Nordic Social Work Research*, vol. 0, no. 0, pp. 1–15. Available: https://doi.org/10.1080/2156857X.2023.2290129.

Hu, Edward J.; Shen, Yelong; Wallis, Phillip; Allen-Zhu, Zeyuan; Li, Yuanzhi; Wang, Shean; Wang, Lu & Chen, Weizhu. 2021, *LoRA: Low-Rank Adaptation of Large Language Models*.

Hägglund, Susanne; Tigerstedt, Christa; Biström, Dennis; Wingren, Mattias; Andersson, Sören; Kuvaja Adolfsson, Kristoffer; Penttinen, Johan & Espinosa-Leal, Leonardo. 2023, Stakeholders Experiences of and Expectations for Robot Accents in a Dental Care Simulation: A Finland-Swedish Case Study, In: Raul Hakli; Pekka Mäkelä & Johanna Seibt, eds., *Social Robots in Social Institutions*, *Frontiers of Artificial Intelligence and Applications*, vol. 366, IOS Press, pp. 125–134, robophilosophy 2022 ; Conference date: 16-08-2022 Through 19-08-2022. Available: https://cas.au.dk/en/robophilosophy/conferences/rpc2022.

Issa, Ilyas; Nurgazy, Symbat; Madeniyetov, Maksat & Sandygulova, Anara. 2023, Robot-Assisted Word-to-Picture Matching Game for Language Learning, In: *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, New York, NY, USA: Association for Computing Machinery, p. 711–715. Available: https://doi.org/10.1145/3568294.3580179.

Iwamoto, Takuya; Baba, Jun; Nakanishi, Junya; Hyodo, Katsuya; Yoshikawa, Yuichiro & Ishiguro, Hiroshi. 2022, Playful Recommendation: Sales Promotion That Robots Stimulate Pleasant Feelings Instead of Product Explanation, *IEEE Robotics and Automation Letters*, vol. 7, , pp. 1–8.

Jaghouar, Sami; Ong, Jack Min; Basra, Manveer; Obeid, Fares; Straube, Jannik; Keiblinger, Michael; Bakouch, Elie; Atkins, Lucas; Panahi, Maziyar; Goddard, Charles; Ryabinin, Max & Hagemann, Johannes. 2024, *INTELLECT-1 Technical Report*. Available: https://arxiv.org/abs/2412.01152.

Janson, Andreas. 2023, How to leverage anthropomorphism for chatbot service interfaces: The interplay of communication style and personification, *Computers in Human Behavior*, vol. 149, , p. 107954. Available: https://www.sciencedirect.com/science/article/pii/S0747563223003059.

Jeffrey Morgan Et al. 2023, *Ollama*. Available at https://github.com/ollama.

Jeglinsky-Kankainen, Ira; Hellstén, Thomas; Karlsson, Jonny & Espinosa-Leal, Leonardo. 2024, Rehabilitation with Humanoid Robots: A Feasibility Study of Rehabilitation of Children with Cerebral Palsy (CP) Using a QTRobot, In: *International Conference on Science and Technology Education*, Springer, pp. 390–400.

Khondaker, Md Tawkat Islam; Abdul-Mageed, Muhammad & Lakshmanan, Laks V. S. 2024, *DetoxLLM: A Framework for Detoxification with Explanations*. Available: https://arxiv.org/abs/2402.15951.

Ko, Kuan-Chou; Lin, Chia-Wei & Yeh, Zhi-Jun. 2024, Chatbot anthropomorphism might not be the design for all: examining responses to anthropomorphized chatbots by autistic individuals, *Marketing Letters*. Available: https://doi.org/10.1007/s11002-024-09754-2.

Kristoffer, Kuvaja Adolfsson; Dennis, Biström & Leonardo, Espinosa-Leal. 2025, FLOSSA: An Application for Improving the Interaction between Patient and a Humanoid Robot in a Dental Care Scenario, In: *Smart Technologies for an All-Electric Society*, "forthcoming".

Krizhevsky, Alex; Sutskever, Ilya & Hinton, Geoffrey E. 2012, ImageNet Classification with Deep Convolutional Neural Networks, In: F. Pereira; C.J. Burges; L. Bottou & K.Q. Weinberger, eds., *Advances in Neural Information Processing Systems*, vol. 25, Curran Associates, Inc. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

Kuvaja Adolfsson, Kristoffer. 2022, *Apputveckling för Roboten Alf : Ett arbete inom MäRI projektet*. Available at https://urn.fi/URN:NBN:fi:amk-2022053013052.

Kuvaja Adolfsson, Kristoffer; Tigerstedt, Christa; Biström, Dennis & Espinosa-Leal, Leonardo. 2024, Deploying Humanoid Robots in a Social Environment, In: Michael E. Auer; Reinhard Langmann; Dominik May & Kim Roos, eds., *Smart Technologies for a Sustainable Future*, Cham: Springer Nature Switzerland, pp. 373–380.

Li, Yinghao Aaron; Han, Cong & Mesgarani, Nima. 2023, *StyleTTS: A Style-Based Generative Model for Natural and Diverse Text-to-Speech Synthesis*.

Li, Yuxi. 2018, Deep reinforcement learning, *arXiv (Cornell University)*. Available: https://arxiv.org/abs/1810.06339.

Ligthart, Mike E.U.; Neerincx, Mark A. & Hindriks, Koen V. 2020, Design Patterns for an Interactive Storytelling Robot to Support Children's Engagement and Agency, In: *Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '20, New York, NY, USA: Association for Computing Machinery, p. 409–418. Available: https://doi.org/10.1145/3319502.3374826.

Linnainmaa, Seppo. 1970, The representation of the cumulative rounding error of an algorithm as a Taylor expansion of the local rounding errors, *Master's Thesis (in Finnish), University of Helsinki*, pp. 6–7.

Lu, Vinh Nhat; Wirtz, Jochen; Kunz, Werner H.; Paluch, Stefanie; Gruber, Thorsten; Martins, Antje & Patterson, Paul G. 2020, Service robots, customers and service employees: what can we learn from the academic literature and where are the gaps?, *Journal of Service Theory and Practice*, vol. 30, no. 3, pp. 361–391. Available: https://doi.org/10.1108/jstp-04-2019-0088.

Luxai S.A. 2024, *Quick start with coding on QTrobot - QTrobot Documentation*. Available: https://docs.luxai.com/docs/intro{_}code.

LuxAI S.A. 2024, *Using Autism Robot to Engage Students with Autism and Other Special Educational Needs*. Available: https://luxai.com/.

Majd, Amin; Biström, Dennis; Tigerstedt, Christa & Espinosa-Leal, Leonardo. 2021, Social and Service Robots Deployed for Social Distancing-Optimization and Placement, In: *2021 International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, IEEE, pp. 1–9.

Meta Platforms, Inc. 2024, *META (2024) Introducing Meta Llama 3: The most capable openly available LLM to date, AI at Meta.* Available: https://ai.meta.com/blog/meta-llama-3/.

Nasir, Jauwairia; Bruno, Barbara & Dillenbourg, Pierre. 2024, Social robots as skilled ignorant peers for supporting learning, *Frontiers in Robotics and AI*, vol. 11, . Available: https://doi.org/10.3389/frobt.2024.1385780.

NVIDIA; Vingelmann, Péter & Fitzek, Frank H.P. 2020, *CUDA, release: 10.2.89*. Available at https://developer.nvidia.com/cuda-toolkit.

Open Source Initiative. 2024, *The Open Source AI Definition – 1.0.* Accessed: 2024-12-9. Available: https://opensource.org/ai/open-source-ai-definition.

OpenAI et al. 2024, *GPT-4 Technical Report*.

Parthasarathy, Venkatesh Balavadhani; Zafar, Ahtsham; Khan, Aafaq & Shahid, Arsalan. 2024, *The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities*. Available: https://arxiv.org/abs/2408.13296.

Peeperkorn, Max; Kouwenhoven, Tom; Brown, Dan & Jordanous, Anna. 2024, *Is Temperature the Creativity Parameter of Large Language Models?* Available: https://arxiv.org/abs/2405.00492.

Pham, Truong An & Espinosa-Leal, Leonardo. 2024, How Indirect and Direct Interaction Affect the Trustworthiness in Normal and Explainable Human-Robot Interaction, In: *International Conference on Science and Technology Education*, Springer, pp. 411–422.

Povey, Daniel; Ghoshal, Arnab; Boulianne, Gilles; Burget, Lukas; Glembek, Ondrej; Goel, Nagendra; Hannemann, Mirko; Motlicek, Petr; Qian, Yanmin; Schwarz, Petr; Silovsky, Jan; Stemmer, Georg & Vesely, Karel. 2011, The Kaldi Speech Recognition Toolkit, In: *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, IEEE Signal Processing Society, iEEE Catalog No.: CFP11SRW-USB.

Pradhan, Sameer S.; Cole, Ronald A. & Ward, Wayne H. 2023, *My Science Tutor (MyST) – A Large Corpus of Children's Conversational Speech*. Available: https://arxiv.org/abs/2309.13347.

Prince, Simon J.D. 2023, *Understanding Deep Learning*, The MIT Press, 1-2 p.. Available: http://udlbook.com.

Project Gutenberg. (n.d.). *1971*. Available: https://www.gutenberg.org/.

Radford, Alec; Kim, Jong Wook; Xu, Tao; Brockman, Greg; McLeavey, Christine & Sutskever, Ilya. 2022, *Robust Speech Recognition via Large-Scale Weak Supervision*.

Research, LG AI; :; An, Soyoung; Bae, Kyunghoon; Choi, Eunbi; Choi, Stanley Jungkyu; Choi, Yemuk; Hong, Seokhee; Hong, Yeonjung; Hwang, Junwon; Jeon, Hyojin; Jo, Gerrard Jeongwon; Jo, Hyunjik; Jung, Jiyeon; Jung, Yountae; Kim, Euisoon; Kim, Hyosang; Kim, Joonkee; Kim, Seonghwan; Kim, Soyeon; Kim, Sunkyoung; Kim, Yireun; Kim, Youchul; Lee, Edward Hwayoung; Lee, Haeju; Lee, Honglak; Lee, Jinsik; Lee, Kyungmin; Lee, Moontae; Lee, Seungjun; Lim, Woohyung; Park, Sangha; Park, Sooyoun; Park, Yongmin; Seo, Boseong; Yang, Sihoon; Yeen, Heuiyeen; Yoo, Kyungjae & Yun, Hyeongu. 2024, *EXAONE 3.0 7.8B Instruction Tuned Language Model*. Available: https://arxiv.org/abs/2408.03541.

Rodríguez-Lera, Francisco Javier. 2018, Emotional Robots for Coaching: Motivating Physical Rehabilitation using Emotional Robots. Available: https://api.semanticscholar.org/CorpusID:209898527.

Rubagotti, Matteo; Tusseyeva, Inara; Baltabayeva, Sara; Summers, Danna & Sandygulova, Anara. 2022, Perceived safety in physical human–robot interaction—A survey, *Robotics and Autonomous Systems*, vol. 151, , p. 104047. Available: https://www.sciencedirect.com/science/article/pii/S0921889022000173.

Russell, Stuart & Norvig, Peter. 2010, *Artificial Intelligence: A Modern Approach*, 3 edn., Prentice Hall, 16-28 p..

Sheikh, Haroon; Prins, Corien & Schrijvers, Erik. 2023, *Artificial Intelligence: Definition and Background*, Cham: Springer International Publishing, pp. 15–41. Available: https://doi.org/10.1007/978-3-031-21448-6_2.

Song, Christina Soyoung & Kim, Youn-Kyung. 2022, The role of the human-robot interaction in consumers' acceptance of humanoid retail service robots, *Journal of Business Research*, vol. 146, , pp. 489–503. Available: https://www.sciencedirect.com/science/article/pii/S014829632200323X.

Spitale, Micol; Axelsson, Minja & Gunes, Hatice. 2023, Robotic Mental Well-being Coaches for the Workplace: An In-the-Wild Study on Form, In: *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '23, New York, NY, USA: Association for Computing Machinery, p. 301–310. Available: https://doi.org/10.1145/3568162.3577003.

StacklokLabs. 2024, *GitHub - StacklokLabs/promptwright: Generate large synthetic data using an LLM*. Available: https://github.com/StacklokLabs/promptwright.

Sutawika, Lintang; Gao, Leo; Schoelkopf, Hailey; Biderman, Stella; Tow, Jonathan; Abbasi, Baber; ben fattori; Lovering, Charles; farzanehnakhaee70; Phang, Jason; Thite, Anish; Fazz; Aflah; Muennighoff, Niklas; Wang, Thomas; sdtblck; nopperl; gakada; tttyuntian; researcher2; Chris; Etxaniz, Julen; Kasner, Zdeněk; Khalid; Hsu, Jeffrey; AndyZwei; Ammanamanchi, Pawan Sasanka; Groeneveld, Dirk; Smith, Ethan & Tang, Eric. 2023, *EleutherAI/lm-evaluation-harness: Major refactor*. Available: https://doi.org/10.5281/zenodo.10256836.

TAIGA. 2022, *AI for good?* Quote from inauguration of TAIGA - AI for good?, Umeå University's center for transdisciplinary AI. Available: https://www.umu.se/en/centre-for-transdisciplinary-ai/.

pandas development team, The. 2020, *pandas-dev/pandas: Pandas*. Available: https://doi.org/10.5281/zenodo.3509134.

Tigerstedt, C.; Edgren, E.; Kuvaja-Adolfsson, K.; Biström, D. & Espinosa-Leal, L. 2024, The robot with the heart-shaped eyes - A preliminary insight into a study about child-robot interaction on early childhood education, In: *ICERI2024 Proceedings*, 17th annual International Conference of Education, Research and Innovation, IATED, pp. 1919–1927. Available: https://doi.org/10.21125/iceri.2024.0549.

Tigerstedt, C. & Fabricius, S. 2023, Humanoid robots in service contexts - insights from sessions with end users, In: *ICERI2023 Proceedings*, 16th annual International Conference of Education, Research and Innovation, IATED, pp. 1413–1420. Available: https://doi.org/10.21125/iceri.2023.0456.

Tigerstedt, Christa & Biström, Dennis. 2021, *Teaching and learning with humanoid social and service robots in higher education - Learning from service design and IT framework use case development modules*.

Tigerstedt, Christa; Dennis Biström, Dennis & Kuvaja Adolfsson, Kristoffer. 2023a, *Underhållning och trygghet - inblick i en robots dag i äldre omsorgen*. Available at https://www.arcada.fi/sv/permalink/1353.

Tigerstedt, Christa; Dennis Biström, Dennis & Kuvaja Adolfsson, Kristoffer. 2023b, *Waiter please! The capable service robots Amy and Alex*. Available at https://www.arcada.fi/en/permalink/1308.

Tung, Vincent & Au, Norman. 2018, Exploring customer experiences with robotics in hospitality, *International Journal of Contemporary Hospitality Management*, vol. 30, .

Vaswani, Ashish; Shazeer, Noam; Parmar, Niki; Uszkoreit, Jakob; Jones, Llion; Gomez, Aidan N.; Kaiser, Lukasz & Polosukhin, Illia. 2017, Attention Is All You Need, *CoRR*, vol. abs/1706.03762, . Available: http://arxiv.org/abs/1706.03762.

Volkov, Dmitrii. 2024, *Badllama 3: removing safety finetuning from Llama 3 in minutes*. Available: https://arxiv.org/abs/2407.01376.

Vouloutsi, Vasiliki; Cominelli, Lorenzo; Dogar, Mehmet; Lepora, Nathan; Zito, Claudio & Martinez-Hernandez, Uriel. 2023, Towards Living Machines: current and future trends of tactile sensing, grasping, and social robotics, *Bioinspiration & Biomimetics*, vol. 18, no. 2, p. 025002. Available: https://doi.org/10.1088/1748-3190/acb7b9.

Wan, Lisa C.; Chan, Elisa K. & Luo, Xiaoyan. 2021, ROBOTS COME to RESCUE: How to reduce perceived risk of infectious disease in Covid19-stricken consumers?, *Annals of Tourism Research*, vol. 88, , p. 103069. Available: https://www.sciencedirect.com/science/article/pii/S0160738320302139.

Wang, Yubo; Ma, Xueguang; Zhang, Ge; Ni, Yuansheng; Chandra, Abhranil; Guo, Shiguang; Ren, Weiming; Arulraj, Aaran; He, Xuan; Jiang, Ziyan; Li, Tianle; Ku, Max; Wang, Kai; Zhuang, Alex; Fan, Rongqi; Yue, Xiang & Chen, Wenhu. 2024, *MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark*. Available: https://arxiv.org/abs/2406.01574.

Zaga, Cristina; Lohse, Manja; Charisi, Vicky; Evers, Vanessa; Neerincx, Marc; Kanda, Takayuki & Leite, Iolanda. 2016, 2nd Workshop on Evaluating Child Robot Interaction, In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, HRI '16, IEEE Press, p. 587–588.

Zhang, Ceng; Chen, Junxin; Li, Jiatong; Peng, Yanhong & Mao, Zebing. 2023, Large language models for human–robot interaction: A review, *Biomimetic Intelligence and Robotics*, vol. 3, no. 4, p. 100131. Available: https://www.sciencedirect.com/science/article/pii/S2667379723000451.

Zhao, Yunpu; Zhang, Rui; Li, Wenyi; Huang, Di; Guo, Jiaming; Peng, Shaohui; Hao, Yifan; Wen, Yuanbo; Hu, Xing; Du, Zidong; Guo, Qi; Li, Ling & Chen, Yunji. 2024, *Assessing and Understanding Creativity in Large Language Models*. Available: https://arxiv.org/abs/2401.12491.

Zhou, Chunting; Liu, Pengfei; Xu, Puxin; Iyer, Srini; Sun, Jiao; Mao, Yuning; Ma, Xuezhe; Efrat, Avia; Yu, Ping; Yu, Lili; Zhang, Susan; Ghosh, Gargi; Lewis, Mike; Zettlemoyer, Luke & Levy, Omer. 2023, *LIMA: Less Is More for Alignment*. Available: https://arxiv.org/abs/2305.11206.

Zou, Jianling; Gauthier, Soizic; Anzalone, Salvatore M.; Cohen, David & Archambault, Dominique. 2022, *A Wizard of Oz Interface with Qtrobot for Facilitating the Handwriting Learning in Children with Dysgraphia and Its Usability Evaluation*, Cham: Springer International Publishing, pp. 219–225.