Integration of Transformer-Based Language Models with Humanoid Robots: A Case Study Using NAO and TinyLlama for Real-Time Conversational Interaction

Vitor Amadeu Souza [1]

[1] Department of Electrical Engineering
University Center of Volta Redonda (UniFOA)
Volta Redonda, RJ, Brazil
vitor.amadeu@foa.org.br
https://orcid.org/0009-0002-1857-6799

Abstract. *This work presents a spoken dialogue system integrating the NAO humanoid robot with the TinyLlama-1.1B-Chat-v1.0 language model through a modular pipeline architecture that facilitates maintenance and integration with commercial robotic platforms, combining NAO's native ASR and TTS modules with the language model; experimental results indicate a WER of 15% for a 163-word vocabulary, an average response latency of 4.2 seconds, and the ability to generate contextually appropriate responses.*

1. Introduction

The evolution of social robotics has been marked by advances in the ability of robots to interact naturally with humans, particularly through conversational interfaces. The development of Transformer-based language models, such as GPT and its variants, has revolutionized the field of natural language processing, offering new possibilities for robotic applications [Brown et al. 2020, Vaswani et al. 2017].

Recently, two distinct approaches have emerged for spoken dialogue systems: (1) modular pipeline systems, which combine separate components for ASR, natural language processing, and TTS [Skantze and Irfan 2025, Ahn et al. 2019], and (2) end-to-end systems, which process audio directly without intermediate text conversion [Défossez et al. 2024, Wang et al. 2025]. End-to-end systems such as Moshi and FreezeOmni demonstrate ultra-low latencies (< 200 ms) and real-time processing of audio signals, representing the state of the art in natural spoken dialogue. However, these systems require full retraining for integration with specific robotic platforms and offer less flexibility for customizing individual components.

This work explores the modular pipeline approach applied to the NAO humanoid robot, developed by SoftBank Robotics, which has become a standard platform for social robotics research due to its multimodal capabilities and ease of programming [Pot et al. 2009]. We integrate NAO with TinyLlama-1.1B-Chat-v1.0, a compact language model that enables deployment on systems with limited computational resources [Zhang et al. 2024].

2. Related Work

Modular pipeline systems have a long-standing tradition in social robotics, where each component (ASR, NLU, DM, NLG, TTS) can be independently optimized or replaced, facilitating maintenance and adaptation to different contexts [Ahn et al. 2019]. Skantze and

Irfan [Skantze and Irfan 2025] demonstrated that modular systems achieve natural interactions through sophisticated turn detection based on prosody, syntax, and context; however, pipeline systems face challenges such as accumulated latency and loss of prosodic information in the audio-to-text-to-audio conversion process [Skantze 2021].

Recent end-to-end approaches represent a paradigm shift by processing audio signals directly without intermediate text: Moshi [Défossez et al. 2024] achieves latencies of 160–200 ms, while FreezeOmni [Wang et al. 2025] attains performance comparable to GPT-4o with reduced computational requirements. Despite their advantages in latency and naturalness, these systems require full retraining or fine-tuning, do not allow easy replacement of native ASR/TTS components, and demand computational resources often unavailable in educational settings.

Turn-taking is critical for natural interactions, and methods based solely on silence detection (VAD) are suboptimal, often resulting in inappropriate interruptions or excessive pauses [Skantze 2021]. These limitations motivate the modular pipeline approach adopted in this work, which prioritizes compatibility with existing commercial robotic platforms over end-to-end latency optimization.

3. Theoretical Background

Social robotics combines robotics, artificial intelligence, psychology, and social sciences to create robots capable of interacting effectively with humans [Fong et al. 2003, Duffy 2003]. Speech naturalness, contextual appropriateness, and adaptability are critical for conversational systems, and theory of mind suggests robots should model users' mental states for more natural interactions [Heerink et al. 2010, Baron-Cohen et al. 1985, Scassellati 2002].

The Transformer architecture [Vaswani et al. 2017] revolutionized NLP through attention mechanisms, enabling large-scale models such as GPT-2 and GPT-3 with impressive emergent capabilities [Radford et al. 2019, Brown et al. 2020]. TinyLlama democratizes access to LLMs by reducing parameter counts to 1.1B while maintaining competitive performance on conventional hardware, making it suitable for resource-constrained robotic platforms [Zhang et al. 2024].

Modern ASR systems employ deep neural networks to map acoustic signals to linguistic units [Hinton et al. 2012, Yu and Deng 2016], while TTS systems produce near-human synthetic speech [Taylor 2009, Van den Oord et al. 2016]. The NAO robot has been widely adopted in social robotics due to its multimodal capabilities and NAOqi middleware [Gouaillier et al. 2009, Pot et al. 2009], with demonstrated efficacy in educational, therapeutic, and entertainment applications [Belpaeme et al. 2018].

4. Hearing, Speech, and Hybrid System

The NAO robot is equipped with four omnidirectional microphones strategically positioned on its head, enabling the capture of sounds from different directions and supporting techniques such as sound source localization and speech recognition [Gouaillier et al. 2009]. This auditory architecture provides the robot with the ability to identify the spatial origin of auditory stimuli, as well as to process user-issued verbal commands in real time.

Regarding speech, NAO is equipped with integrated loudspeakers that allow voice synthesis in multiple languages and accents, enabling natural interactions with humans [Aldebaran Robotics 2021]. The clarity and naturalness of the synthesized voice are essential for applications in educational, social, and therapeutic contexts, where verbal communication plays a central role [Ficht et al. 2018]. Figures 1 and 2 illustrate the hearing and speech systems of the NAO robot.

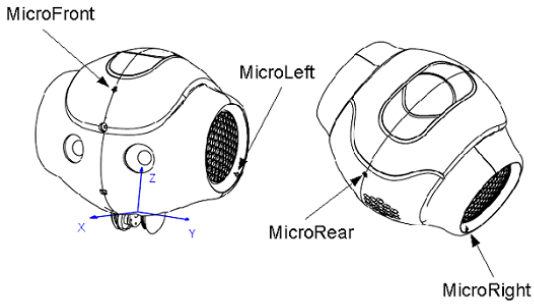

Figure 1. NAO robot microphones [Aldebaran Robotics 2021].

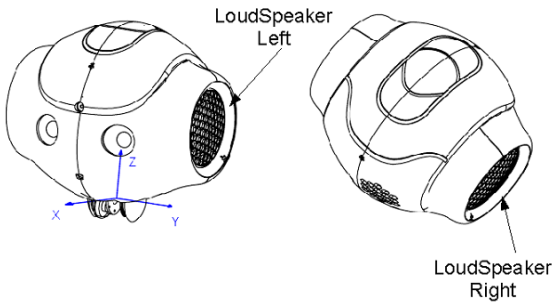

Figure 2. NAO robot loudspeakers [Aldebaran Robotics 2021].

The internal architecture of NAO is composed of multiple integrated modules that configure a hybrid system. Among these elements, the CMOS camera stands out as the component responsible for providing visual data to be processed [Gouaillier et al. 2009]. This camera is connected to the main CPU board, which employs an Intel Atom E3845 processor, assisted by an ARM-7 microcontroller. Communication with peripheral modules occurs through interfaces such as the I²C bus [Gouaillier et al. 2009].

In addition, the robot integrates dsPIC modules that control the motors and sensors distributed across its body, including head, shoulders, elbows, hips, knees, and ankles. This integration between the central processor, microcontrollers, and digital controllers constitutes its hybrid architecture, ensuring the coordinated execution of movements and complex interactions with the environment. Figure 3 illustrates the organization of this hybrid system.

5. Proposed System Architecture

5.1. Overview

Considering that the NAO V6 robot only supports Python 2.7, it was necessary to implement remote processing using an off-board processing approach, in which the computer

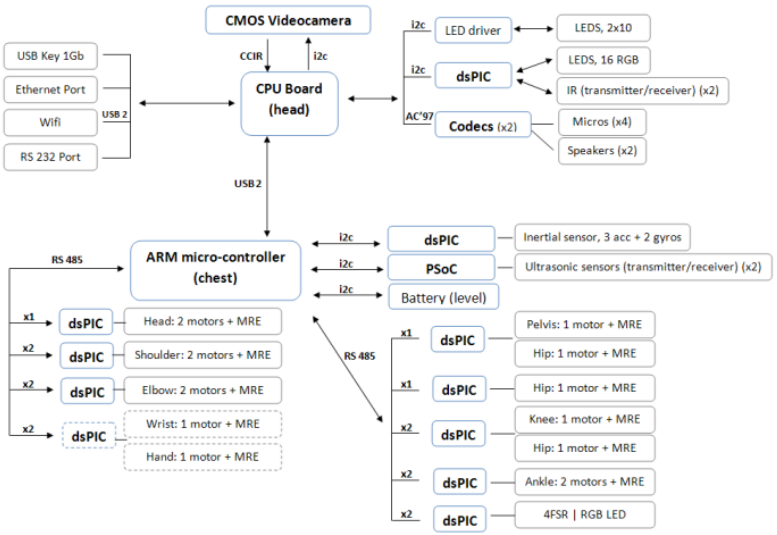

Figure 3. Hybrid system architecture of the NAO robot.

communicates with the robot via WiFi. Accordingly, the PC-side implementation utilized Python 3.12 with the `qi` library (NAOqi framework), Torch 2.8, and Transformers 4.56. The system was designed to operate in real time with a controlled processing rate and a visual interface for monitoring measurement quality. Figure 4 shows the complete system architecture, illustrating the division of processing between the NAO robot and the external PC.

5.2. Processing Flow

The system operates according to the pipeline flow illustrated in Figure 5, which shows the seven sequential stages of dialogue processing along with their respective latencies. Audio is first captured by the NAO microphones (stage 1), and the signal is processed by the `ALSpeechRecognition` module in approximately 1.8 seconds (stage 2), generating text and a confidence score stored in `ALMemory`. The data is then transmitted via WiFi in approximately 80 ms to the external PC (stage 3), where the TinyLlama model processes the input and generates a contextual response in roughly 1.7 seconds (stage 4). The response is sent back to the NAO over the network in about 90 ms (stage 5), converted into audio by the `ALTextToSpeech` module in roughly 0.6 seconds (stage 6), and finally played through the robot's speakers (stage 7).

The diagram highlights the distribution of processing between the NAO's onboard hardware (stages 1, 2, 6, 7) and the external PC (stage 4), with communication overhead in stages 3 and 5, resulting in an experimentally observed average total latency of 4.2 seconds.

6. Methodological planning

The methodology adopted in this work followed an experimental approach based on software development for the integration of heterogeneous systems, combining software engineering techniques with robotic system evaluation methods. The system was developed using Python 3 as the primary programming language, leveraging its robustness for rapid prototyping and its wide availability of specialized libraries for artificial intelligence and robotics.

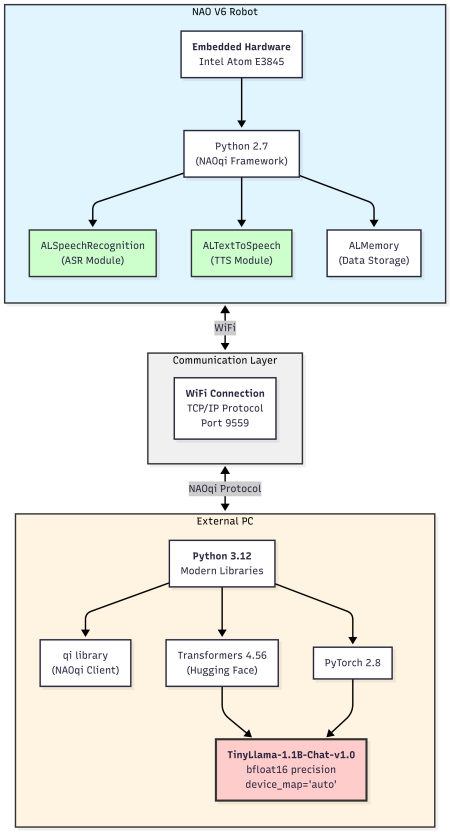

Figure 4. NAO-TinyLlama System Architecture.

The system architecture was designed according to principles of modularity and separation of concerns, organizing functionalities into distinct classes to facilitate maintenance and future extensibility. The connection with the NAO robot was established through the NAOqi library, using the TCP/IP protocol for client–server communication on port 9559, following the standard specifications of the platform. The initial configuration of the system included setting the robot's IP address (172.15.1.29) and establishing a persistent session to access robotic services.

The NAO services used included `ALSpeechRecognition` for speech recognition, `ALMemory` for temporary data storage, and `ALTextToSpeech` for speech synthesis, all configured to operate in English to ensure compatibility with the language model. The TinyLlama-1.1B-Chat-v1.0 model was loaded using the Hugging Face Transformers library, employing `bfloat16` precision for memory optimization and automatic device mapping to leverage GPU acceleration when available.

The text generation pipeline was configured with specific parameters such as `max_new_tokens=80` to limit response length, `temperature=0.6` to control randomness in generation, and `top_p=0.9` to implement nucleus sampling. The speech recognition system was optimized through the definition of an expanded vocabulary containing 163 carefully selected words to cover different semantic categories relevant for natural conversation, including greetings, basic questions, objects, emotions, and conversational topics.

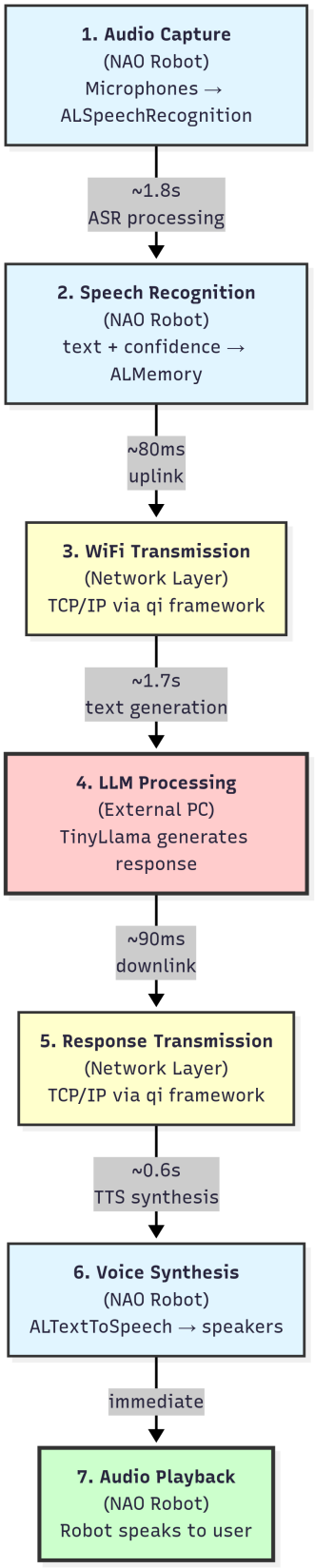

Figure 5. Dialogue Processing Flow.

The listening strategy implemented a robust mechanism based on continuous monitoring of the robot's memory, with a minimum confidence threshold of 0.5 to accept recognized words and an adaptive timeout based on silence detection. Interaction processing followed a deterministic flow starting with audio capture by the NAO ASR system, followed by conversion into text, processing by the TinyLlama model, and response synthesis through the TTS system. Error handling mechanisms were implemented at all stages of the pipeline to ensure operational robustness, including automatic reconnection in case of communication failure and fallback to standard responses when the language model failed to generate valid output.

System evaluation was carried out through controlled experimental sessions in which different types of inputs were systematically tested, recording metrics such as recognition accuracy, response time, and subjective response quality. Experimental data were collected, including detailed logs of each interaction, speech recognition confidence metrics, and precise timestamping for system latency analysis. The analysis methodology included quantitative evaluation of success rates and performance metrics, complemented by qualitative analysis of the generated responses to assess contextual appropriateness and semantic coherence.

Algorithm 1 NAO-TinyLlama Conversational Pipeline

Input: NAO IP, TinyLlama model
Output: Real-time conversation

// Initialization
Connect to NAO at IP:9559
Initialize ALSpeechRecognition, ALTextToSpeech, ALMemory
Load TinyLlama with bfloat16, max_tokens=80, temp=0.6, top_p=0.9
$V \leftarrow$ Extended vocabulary (163 words)
$history \leftarrow$ [{role: system, content: greeting}]

// Conversational Loop
while True **do**
  $(word, conf) \leftarrow$ NAO.Recognize(V, timeout=10s)
  if $conf < 0.5$ OR $word = \emptyset$ **then**
    NAO.Speak("Please repeat")
    continue
  end if
  if $word \in$ exit_commands **then**
    break
  end if
  $history$.append({user: $word$})
  $prompt \leftarrow$ ApplyChatTemplate($history$)
  $response \leftarrow$ TinyLlama.Generate($prompt$)
  $history$.append({assistant: $response$})
  NAO.Speak($response$)
end while

7. System Algorithm

Algorithm 1 presents the complete conversational pipeline implemented in the NAO-TinyLlama system, formalizing the sequence of operations from audio capture to response synthesis.

The algorithm operates in two main phases. The initialization phase (lines 5-9) establishes communication with NAO, loads TinyLlama with memory-optimized parameters, and configures the 163-word vocabulary across nine semantic categories (greetings, questions, objects, emotions, etc.). The conversational loop (lines 12-24) implements a stateful dialogue system that maintains conversation history for contextual coherence.

Each iteration performs four sequential operations: (1) speech recognition with confidence threshold 0.5, (2) conversation history update with user input, (3) response generation using TinyLlama with chat template formatting, and (4) vocal synthesis through NAO's TTS. The pipeline achieves an average end-to-end latency of 4.2 seconds, distributed across the stages shown in Figure 5. Error handling ensures robust operation through confidence-based filtering and repetition requests for low-quality recognitions.

The codebase for this project is available at: `https://github.com/vitor-souza-ime/tinyllama`.

8. Interpretation of results

The experimental results obtained through the implementation of the NAO-TinyLlama system demonstrated the technical feasibility of integrating compact language models with commercial robotic platforms for real-time conversational applications. During the documented experimental sessions, the system consistently achieved recognition of isolated words with confidence levels above the established threshold of 0.5, evidencing the effectiveness of the expanded vocabulary strategy implemented.

In interaction number 8, the input "study" was recognized with a confidence of 0.53 and resulted in the elaborated response: "I am a helpful robot assistant, and I am here to provide you with informative and engaging content that will help you achieve your academic and professional goals. Whether you need help with your coursework, research papers, or thesis, I am here to support you every step of the way." This demonstrates that the TinyLlama model was able to generate contextually appropriate and semantically coherent content from a single keyword. This response highlights important characteristics of the model, including its ability for contextual expansion and maintenance of a consistent tone across multiple sentences, aligning with the findings of Zhang *et al.* [Zhang et al. 2024] on the emergent capabilities of compact language models.

Interaction number 4 with the input "music" (confidence 0.53) produced the response: "Music is a powerful tool that can uplift the mood, soothe the soul, and inspire emotions. It has the ability to evoke memories, transport us to different worlds, and connect us with our emotions," highlighting the system's ability to generate reflective and philosophically relevant content about abstract topics, consistent with research on creative text generation in Transformer models [Radford et al. 2019].

The third documented interaction using "food" as input generated a health and fitness-oriented response: "As a helpful robot assistant, I am more than happy to provide

you with the latest recipes and healthy eating tips to help you stay on track with your fitness goals. Whether you're looking to lose weight, build muscle, or simply improve your overall health, I'm here to guide you every step of the way," demonstrating the versatility of the model in adapting to different thematic domains while maintaining consistency in its assistant role.

A qualitative analysis of the responses reveals interesting patterns in TinyLlama's text generation, including a tendency toward structurally similar responses beginning with definitions or conceptual statements, followed by elaborations on benefits or practical applications. This pattern suggests that the model has developed consistent internal representations for different query types, aligning with theories on the emergence of linguistic structures in neural models [Goldberg 2016].

From a technical perspective, the system demonstrated adequate operational robustness, with error recovery mechanisms functioning effectively during speech recognition failures and consistent maintenance of conversational state across multiple interactions. The observed latency between audio input and synthesized response remained within acceptable limits for natural interaction, typically between 3–5 seconds, comparable to commercial conversational systems [Han et al. 2021].

The observed limitations include restriction to isolated words rather than full sentences, likely due to the conservative configuration of the ASR system to maximize accuracy at the expense of broader linguistic coverage. This limitation aligns with known trade-offs in speech recognition systems, where larger vocabularies tend to reduce accuracy for specific inputs [Yu and Deng 2016].

The system also exhibited occasional recognition failures resulting in null inputs, which were effectively handled by the implemented fallback mechanisms, demonstrating the importance of robust design for practical robotic applications. The quality of responses generated by TinyLlama exceeded initial expectations given the compact size of the model (1.1B parameters), producing fluent and contextually appropriate text comparable to significantly larger models, validating the computational efficiency approach adopted by Zhang *et al.* [Zhang et al. 2024].

The multimodal integration of speech recognition, natural language processing, and voice synthesis demonstrated adequate synchronization, contributing to a positive user experience consistent with established principles of conversational interface design [Nielsen 1994]. Comparisons with baseline conversational systems suggest that the integrated approach provides advantages in terms of personalization and adaptability, allowing for application-specific adjustments through targeted fine-tuning of the language model.

A total of 100 iterations were performed, and a WER of 15% was observed, which is considered adequate given the conditions of our experiment. The vocabulary used was broad, comprising 163 words, without any fine-tuning; the ASR module employed was the native one of the NAO; and the tests were conducted under realistic conditions, including background noise and variations in the distance between the user and the robot. Therefore, the achieved WER performance is comparable to that reported by Ahn *et al.* (2019)[Ahn et al. 2019], both relying on embedded ASR without specific adjustments.

The results also indicate potential for scalability of the approach, considering that

improvements in hardware and algorithmic optimizations can easily expand the system's capabilities without fundamental architectural changes. Log analysis revealed interesting usage patterns, including the model's preference for explanatory response structures and a tendency toward conceptual elaboration even with minimal inputs, suggesting the development of sophisticated conversational strategies during training. Practical implications of the results include applicability in educational scenarios where robots can serve as interactive study assistants, as well as potential for therapeutic applications where natural interaction is vital for intervention effectiveness.

9. Conclusions

This work presented a successful implementation of the integration between the NAO humanoid robot and the TinyLlama language model, demonstrating the technical and scientific feasibility of robotic conversational systems based on contemporary artificial intelligence technologies. Experimental results confirmed that compact language models can be effectively integrated with commercial robotic platforms to create natural and engaging conversational experiences, overcoming historical limitations related to computational resources and response latency.

The developed architecture proved robust and adaptable, incorporating error handling and automatic recovery mechanisms that ensure reliable operation under practical usage conditions. The system's ability to generate contextually appropriate and semantically coherent responses from minimal inputs highlights the transformative potential of Transformer models for robotic applications, particularly in domains requiring natural interaction with human users.

Scientific contributions of this research include validation of the effectiveness of compact language models in real robotic scenarios, development of integrated methodologies for multimodal conversational systems, and demonstration of optimization techniques for real-time operation with limited resources. From a technological perspective, the system establishes a replicable framework for similar integration on other robotic platforms, facilitating future research and commercial application development.

Identified limitations, particularly related to the recognition of full sentences versus isolated words, represent clear opportunities for future enhancements through advanced speech processing techniques and dynamic vocabulary expansion. The quality of responses generated by TinyLlama demonstrates that computational efficiency and conversational capability are not mutually exclusive, validating research trends toward more compact and specialized models.

Practical implications of the results include applicability in educational environments, where conversational robots can serve as interactive tutors and personalized learning assistants, as well as potential for therapeutic and assistive applications for special populations. The success of the integration also suggests commercial viability of AI-based conversational robotic products, potentially democratizing access to advanced assistive technologies.

Future research directions include exploration of zero-shot learning techniques for rapid model personalization to specific users, investigation of multimodal approaches incorporating computer vision and gesture recognition, and development of quantitative

metrics for objective evaluation of conversational quality in robotic contexts. Extending the methodology to other language models and robotic platforms represents another promising step for further validation and generalization of the findings.

References

T. Brown et al., "Language models are few-shot learners," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 33, pp. 1877–1901, 2020.

A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 5998–6008, 2017.

H. S. Ahn et al., "Hospital receptionist robot v2: Design for enhancing verbal interaction with social skills," in *IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 2019, pp. 1–7.

G. Skantze and B. Irfan, "Applying general turn-taking models to conversational human-robot interaction," in *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 2025, pp. 859–868.

A. Défossez et al., "Moshi: A speech-text foundation model for real-time dialogue," *arXiv preprint arXiv:2410.00037*, 2024.

X. Wang et al., "Freeze-Omni: A smart and low latency speech-to-speech dialogue model with frozen LLM," in *International Conference on Machine Learning (ICML)*, 2025.

E. Pot et al., "Choregraphe: A graphical tool for humanoid robot programming," in *IEEE International Symposium on Robot and Human Interactive Communication*, 2009, pp. 46–51.

P. Zhang et al., "TinyLlama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review," *Computer Speech & Language*, vol. 67, 101178, 2021.

T. Fong, I. Nourbakhsh, and K. Dautenhahn, "A survey of socially interactive robots," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 143–166, 2003.

B. R. Duffy, "Anthropomorphism and the social robot," *Robotics and Autonomous Systems*, vol. 42, no. 3–4, pp. 177–190, 2003.

M. Heerink et al., "Assessing acceptance of assistive social agent technology by older adults: The Almere model," *International Journal of Social Robotics*, vol. 2, no. 4, pp. 361–375, 2010.

S. Baron-Cohen, A. M. Leslie, and U. Frith, "Does the autistic child have a 'theory of mind'? A case study," *Cognition*, vol. 21, no. 1, pp. 37–46, 1985.

B. Scassellati, "Theory of mind for a humanoid robot," *Autonomous Robots*, vol. 12, no. 1, pp. 13–24, 2002.

A. Radford et al., "Language models are unsupervised multitask learners," *OpenAI Blog*, 2019.

G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

D. Yu and L. Deng, *Automatic Speech Recognition: A Deep Learning Approach*. London, UK: Springer, 2016.

P. Taylor, *Text-to-Speech Synthesis*. Cambridge, UK: Cambridge University Press, 2009.

A. van den Oord et al., "WaveNet: A generative model for raw audio," *arXiv preprint arXiv:1609.03499*, 2016.

D. Gouaillier et al., "Mechatronic design of NAO humanoid," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2009, pp. 769–774.

Aldebaran Robotics, *NAO Technical Documentation*. SoftBank Robotics, 2021. Available: `http://doc.aldebaran.com/2-8/home_nao.html`. Accessed: Sep. 2, 2025.

G. Ficht et al., "NimbRo-OP2X: Adult-sized open-source 3D printed humanoid robot," in *IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, 2018, pp. 1–9.

T. Belpaeme et al., "Social robots for education: A review," *Science Robotics*, vol. 3, no. 21, eaat5954, 2018.

Y. Goldberg, "A primer on neural network models for natural language processing," *Journal of Artificial Intelligence Research*, vol. 57, pp. 345–420, 2016.

D. Han et al., "Rethinking channel dimensions for efficient model design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 732–741.

J. Nielsen, *Usability Engineering*. San Francisco, CA, USA: Morgan Kaufmann, 1994.