Conversational AI without the Cloud: A Lightweight, Local Dialogue Pipeline for Non-commercial Social Robots

Lewis Watson
Edinburgh Napier University
Edinburgh, United Kingdom
l.watson@napier.ac.uk

Emilia Sobolewska
Edinburgh Napier University
Edinburgh, United Kingdom
e.sobolewska@napier.ac.uk

Carl Strathearn
Edinburgh Napier University
Edinburgh, United Kingdom
c.strathearn@napier.ac.uk

Mayuko Morgan
Edinburgh Napier University
Edinburgh, United Kingdom
mayuko.morgan@napier.ac.uk

Yanchao Yu
Edinburgh Napier University
Edinburgh, United Kingdom
y.yu@napier.ac.uk

Abstract

A major limitation of current social robots is their dependence on cloud-based dialogue pipelines, which restricts use in settings with limited or unreliable connectivity. We present a lightweight, fully local spoken-dialogue system that runs on consumer-grade hardware and integrates open-source models for speech recognition, dialogue generation, and text-to-speech. The pipeline was deployed on Euclid, a non-commercial humanoid robot, across several public engagement events, enabling extended real-world interaction without internet access. We analyse over 5,000 dialogue turns recorded during these dialogues to characterise system behaviour, user interaction patterns, and challenges arising in noisy, multi-speaker environments. Our observations demonstrate the feasibility of privacy-preserving, on-device conversational robotics while highlighting limitations in turn-taking, response length, and environmental grounding. We outline planned improvements to support more robust and accessible social-robot interaction.

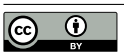
CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computer systems organization** → **Robotics**; • **Computing methodologies** → *Natural language generation.*

Keywords

local LLM, conversational AI, social robots, dialogue management, personalisation, privacy, edge computing

ACM Reference Format:
Lewis Watson, Emilia Sobolewska, Carl Strathearn, Mayuko Morgan, and Yanchao Yu. 2026. Conversational AI without the Cloud: A Lightweight, Local Dialogue Pipeline for Non-commercial Social Robots. In *Companion Proceedings of the 21st ACM/IEEE International Conference on Human-Robot Interaction (HRI Companion '26), March 16–19, 2026, Edinburgh, Scotland, UK*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3776734.3794421



HRI Companion '26, Edinburgh, Scotland, UK

ACM ISBN 979-8-4007-2321-6/2026/03
https://doi.org/10.1145/3776734.3794421

1 Introduction and Related Work

The widespread use of commercial robots in HRI has increasingly constrained research directions, limiting novelty in real-world conversational AI and modular social robotics [7]. This has prompted researchers to design projects around the capabilities of commercial robots, as opposed to building new robots to conduct more novel and impactful research [6]. Furthermore, researchers are often unable to develop, test and deploy advanced AI systems because of the restrictive nature of commercial robots, i.e. proprietary software, outdated libraries, inbuilt hardware and cloud software limitations [9]. This makes training and testing state-of-the-art large language models (LLMs) using commercial social robots problematic, and why AI simulation methods, such as the Wizard of Oz (WoZ), are still commonplace in HRI [16]. Thus, there is a need to move into a new research space where we build and test novel social robotics and AI systems for HRI using cost-effective materials and methods. In support, Google DeepMind AI released Gemini Robotics [3], a Vision Language Action (VLA) model that operates offline, emphasising the importance of developing locally run systems for robots. However, Gemini Robotics requires a high-performance computer to operate efficiently, which may make it less suitable for novel prototype robots, students or researchers with limited hardware resources.

This presents a significant gap in HRI, and the need for lightweight, low-resource and modular generative spoken dialogue systems for non-commercial social robots. The key contributions of this paper are a dialogue management pipeline for social robots that operates offline and on standard GPU laptops, a new dialogue dataset containing four field tests of this system, analysing 5,161 dialogue turns.

2 Euclid the Humanoid Robot

Euclid (Figure 1) is a non-commercial social humanoid robot designed for natural human–robot interaction. It features an expressive face, microphone, and camera to support audiovisual tasks such as user tracking. Euclid runs a series of Arduino and Raspberry Pi microcontrollers that connect to a laptop or PC. These can be controlled over USB serial ports, allowing the AI system to be updated or adjusted in real time. Euclid has also been used as a test-bed for prior research on hybrid generative and rule-based dialogue systems trained on common-sense and object-centric datasets to

support domestic task guidance [19]. In our work, we use the same humanoid robot, but run our dialogue pipeline locally and fully offline.

Figure 1: Euclid the lifelike social humanoid robot

3 Conversational AI System

To address the limitations of cloud-dependent social robots, we developed a modular, privacy-centric dialogue pipeline designed to run entirely offline on consumer-grade hardware. This section details the architecture, component selection, and optimisation strategies used to deploy with Euclid.

3.1 Design Goals

The system was designed to allow the robot to operate reliably without external infrastructure, avoiding the limitations of cloud-based pipelines that depend on stable, high-bandwidth internet. This enables deployment across public events, classrooms, outdoor settings, and secure facilities with strict network policies. The key goals were:

Infrastructure Independence: All components run fully on-device, ensuring consistent operation without internet access or API keys.

Privacy by Design: Keeping all data local avoids transmitting speech to third-party servers and protects users' privacy.

Low-Latency Interaction: The system must prioritise fast response times to maintain a natural conversational flow.

Modularity and Adaptability: Components can be swapped or upgraded with minimal changes to the pipeline, and dynamic prompting enables adjustments to persona, context, and safety constraints at runtime.

3.2 Architecture Overview

The conversational pipeline follows a sequential *Speech-to-Speech* architecture executed entirely on-device: Audio Capture → Voice Activity Detection (VAD) → Automatic Speech Recognition (ASR) → Large Language Model (LLM) → Text-to-Speech (TTS) → Queued Audio Player.

Hardware Deployment Environment: To demonstrate the system's viability on accessible hardware, the pipeline was deployed on a standard laptop equipped with an NVIDIA RTX 2000 Ada Generation Mobile GPU (8 GB VRAM). This configuration represents a realistic baseline for modern consumer-grade laptops. Although this serves as our reference setup, the design is flexible: systems with more powerful GPUs or greater VRAM can run larger models without modification to the pipeline.

3.3 Audio Capture and ASR

For automatic speech recognition, we use OpenAI's Whisper V3 Large Turbo [14], a distilled Whisper model that offers a strong balance of accuracy and computational efficiency. Whisper models are known to hallucinate transcriptions on short or non-speech inputs, to mitigate this, the audio pipeline incorporates a two-stage gating process. We first estimate the root-mean-square (RMS) energy of incoming audio to filter quiet segments unlikely to contain speech. For segments above this threshold, we apply Silero VAD [18] to confirm the presence of human speech. Only audio that passes both checks is forwarded to the Whisper model for transcription.

We note that converting speech to text discards prosodic cues such as intonation and affect, which may influence conversational dynamics, incorporating paralinguistic information remains future work.

3.4 Modular Dialogue Management

The conversational logic is handled through a local OpenAI compatible API, hosted using LM Studio[1]. Hosting the LLM in a dedicated backend abstracts away tokenisation details and chat-templating differences (e.g., Harmony, ChatML, Alpaca, Llama-3 formats), allowing the Python pipeline to remain agnostic to model-specific formatting. This design enables models to be swapped without modifying the dialogue manager, and the standardised API ensures compatibility with alternative local engines such as Ollama or vLLM.

For deployment, we used a 4-bit quantised Qwen 2.5 7B model [21], selected for its instruction-following performance and efficiency; the weights require roughly 4.36 GB of VRAM (excluding chat context). Preliminary tests with the newer Qwen3 4B [20] indicate comparable or improved performance under the same hardware constraints.

A custom *Prompt Controller* dynamically assembles the system prompt at each turn. A stable template defines the robot's persona (e.g., an "old robot with rusty circuits'' who is "helpful, kind, chatty, and wise''), conversational style (e.g., speaking casually and avoiding disclaimers), and safety rules (e.g., maintaining PG-friendly content), while the controller injects limited real-time information (e.g., date and time) to guide the model's responses.

3.5 Text-to-Speech (TTS) Synthesis

To minimise GPU load, we use the Kokoro v1.0 TTS model (82M) [5], based on StyleTTS2 [10] and iSTFTNet [8], executed via ONNX Runtime [2] on the CPU. Running a compact TTS model on the CPU enables true parallelism with GPU-based LLM inference and avoids the overhead of large expressive TTS systems when their benefits are not required. Despite its small size, Kokoro provides more natural prosody than traditional rule-based synthetic voices while remaining far smaller than expressive models such as CSM-1 (1B) [17], Dia2 (2B) [13], or VibeVoice 7B [12], which can offer richer expressiveness but are generally impractical to run alongside an on-device LLM on consumer hardware. The pipeline is modular,

[1]LM Studio, https://lmstudio.ai

however, and could support such models on multi-GPU systems or devices with greater compute capacity.

Kokoro includes multiple language packs and voice presets, offering flexibility across deployment contexts. In practice, compact TTS models often struggle with non-Latin scripts or context-dependent pronunciations (e.g., Japanese particles), requiring additional G2P (grapheme-to-phoneme) preprocessing. Achieving acceptable pronunciation in another language typically also requires switching to a language-specific voice embedding, which can break vocal consistency and weaken the robot's vocal identity and weaken the intended character illusion. We additionally apply light text preprocessing to remove emojis and certain special characters, which can produce undesirable artefacts in small TTS models.

To minimise "Time-to-First-Audio" (TTFA), the TTS module uses a sentence-level streaming queue: once the LLM finishes generating, the text is segmented into sentences and synthesised asynchronously so the robot can begin speaking the first sentence while later sentences are still being produced. A planned optimisation is to stream tokens directly from the LLM backend and initiate synthesis as soon as the first sentence boundary appears, allowing speech to begin while the LLM is still generating the remainder of the response.

4 Public Engagement

Euclid was exhibited in a number of public-facing events, carrying out 5,161 conversations, some of which were open to the general public, including adults and children from a variety of nationalities, genders, cultures, languages, and educational levels (AI for Good Summit, United Nations, Geneva, July 2025; New Scientist Live, London, September 2025), while others were directed at academics in broadly understood STEM fields, such as robotics (The UK Robotics Expo, National Robotarium Edinburgh, September 2025) and engineering (Scottish Partnership in Energy and Engineering Research & Innovation, SPEERI; Edinburgh, November 2025). In one special event, Edinburgh Fringe "The Provocateurs" (Edinburgh, August 2025), Euclid was presented to the audience at a comedy show; however, he only interacted with the selected presenter.

4.1 General Public Interactions

In all but the last event, the environmental conditions were challenging, with large spaces distorting the sound, high levels of noise, multiple speakers, and general echoes of surrounding conversations.

A significant challenge was the system's 'always-on' microphone design in these noisy environments. Users often hesitated or 'froze' due to the pressure of the interaction, creating natural pauses that the VAD misinterpreted as turn-endings, leading to premature response generation, interrupting the user. Furthermore, the open microphone frequently triggered on background chatter or inquiries directed at the operators. To mitigate these false triggers and maintain conversational flow, operators had to manually mute the system whenever the robot was not being directly addressed.

Another interesting observation was the audience's reaction to the robot's perceived personality and playful exchanges, but also the flexibility of responses; some participants noted that they expected AI-like, ChatGPT responses (definitions, precise data, and hard facts), rather than contextualised and sometimes "cheeky" conversations. This added to a general sense of human-like dialogue and enhanced Euclid's characteristics.

4.2 Academic Engagement

The engagement involving members of academia was only different insofar that the environment was smaller and more self-contained. It was noticeable, however, that the audiences would be more likely to talk about the robot *with* the operators, without addressing it, rather than take part in an active conversation with Euclid.

At the UK Robotics Expo, where many other robots were exhibited, conversations revolved around comparing different AI systems (both in terms of conversational skills and appearance). Whilst at SPEERI, although the audience was more neutral and curious about the robot, without referring to other machines, the questions were largely addressed to the operators, rather than the robot itself.

4.3 Performance Scenarios

The final event at the "The Provocateurs" show at The Edinburgh Fringe tested Euclid's capacity for scripted performance. The robot participated in a stand-up comedy routine discussing the ethics of robotics and the utility of hands, a topic chosen to satirise Euclid's own lack of hands. The show's script was based on the current landscape of robotics, as well as research and anecdotal evidence regarding instances when robots have caused accidents and harmed people due to working arms. Subsequently, the audience was encouraged to ask questions via the presenter, who would relate them to Euclid.

This interaction took place on two occasions. In the first performance, the robot suffered a configuration issue where the system prompt retained context from the UN event; consequently, it hallucinated that it was still at the conference, requiring the operator to step in and readjust. Whereas on the second attempt, the dialogue was more focused and representative of the topic at hand.

4.4 Data Collection and Ethics

The study involved voluntary public-engagement interactions. Participants were informed on-site that the robot was operating autonomously and that anonymised, text-only dialogue logs would be analysed for research. No audio, video, or demographic data were stored. For interactions involving children, guardians were present and dialogue was limited to public, non-sensitive content. All logs were manually screened to remove identifying information, stored on encrypted drives, and transferred to university-managed cloud storage, with no attempt to link data to individuals.

5 Interaction Analysis

To explore performance and user interaction patterns across these diverse settings, we performed analysis on the recorded dialogues.

5.1 Data Overview & Pre-processing

The analysis is based on raw JSON dialogue logs generated by the system across five events. In these logs, assistant denotes Euclid's generated responses and user the human participant, following standard LLM chat conventions. After filtering empty files and

single-turn fragments (e.g., isolated "hello" inputs without an assistant reply), a total of 5,161 dialogue turns were retained, with the breakdown by venue shown in Table 1. The text data then underwent standard NLP pre-processing using SpaCy, including normalisation, stopword removal, and lemmatisation, to optimise the corpus for topic modelling.

Table 1: Distribution of Dialogue Turns by Venue and Role

Event Venue	Assistant	User	Total
Edinburgh Fringe (2025)	58	58	116
New Scientist Live (2025)	947	947	1,894
Robotarium Expo (2025)	214	214	428
United Nations (AI for Good)	1,360	1,363	2,723
Total	**2,579**	**2,582**	**5,161**

5.2 Topic Modelling

To characterise the dominant themes in the corpus, we applied BERTopic [4] separately to user and assistant utterances. Each utterance was encoded using the **all-MiniLM-L6-v2** sentence-transformer model [15], followed by UMAP [11] for dimensionality reduction and HDBSCAN [1] for density-based clustering. This combination allowed us to group semantically similar utterances into topics while treating unrelated items as outliers.

In total, users produced 3,942 utterances and Euclid produced 13,225. BERTopic assigned 72% of user utterances and 66% of Euclid (assistant) utterances to a topic cluster, with the remainder classified as outliers. For users, the most common words and topics centred on greetings and polite interaction (e.g., "hello", "thank", "Euclid", "introduce", "robot", "could"), indicating short, hesitant turns focused on meeting the robot and asking simple questions.

In contrast, the assistant's topics were more varied and content-rich, with frequent references to robotics and AI (e.g., "robot", "AI", "technology", "innovation"), event framing (e.g., "conference", "AI for good"), and engagement prompts (e.g., "chat", "feel free", "question"). Notably, one topic cluster contained persona-related terms such as "circuit", "rusty", and "buzz", indicating that the model frequently reflected elements of the "old robot" character described in the system prompt.

These patterns mirror the conversational asymmetry observed during deployment. While users tended to contribute brief, single-sentence turns, the assistant produced longer, multi-sentence responses to maintain the event context. Qualitatively, this occasionally manifested as excessive verbosity, where the model shifted into an overly instructional tone that, while informative, sometimes disrupted the natural pacing of social dialogue.

5.3 Sentiment Analysis

To estimate the emotional tone of the interactions, we applied a DistilBERT-based sentiment classifier (distilbert-base-uncased-finetuned-sst-2-english) to each utterance. The model assigns a positive or negative label with an associated confidence score.

Overall, the corpus was skewed towards positive sentiment: 79% of user utterances and 75% of assistant (Euclid) utterances were classified as positive. The analysis indicates that both roles are predominantly positive, though user utterances were slightly more strongly and consistently positive than Euclid's. This likely aligns with the public engagement context, where users typically greet the robot enthusiastically and the intentionally friendly, encouraging persona defined for the robot.

Table 2: Illustrative topics and keywords.

Source	Illustrative topic keywords
User	hello, thank, Euclid, tell, say, like, introduce, robot, yeah, could
Assistant	robot, AI, technology, innovation, conference, AI for good, chat, feel free, fantastic, pretty cool, circuit, rusty

6 Lessons, Limitations and Future Work

The deployments demonstrate that a fully local, offline spoken-dialogue pipeline can run reliably on a single consumer-grade GPU laptop and support hours of real-world public interaction without internet access, providing a viable path for privacy-first social robots.

Limitations remain clear: our analysis is based on text-only logs and operator observations, which may introduce observer effects, with no direct user feedback or visual context. Turn-taking frequently broke down in noisy, multi-party settings; responses were often too long or lecture-like; and the absence of scene awareness occasionally led to replies that did not align with the immediate interaction or environment. Furthermore, the analysis was based on all events rather than exploring the individual instances, which limited the depth of topic modelling, over contextualisation of the overall data.

Future work focuses on four concrete directions: (1) more robust turn-taking via interruption (barge-in) detection and optional push-to-talk; (2) distilling concise, characterful personalities using collected data and manually authored responses; (3) adding lightweight visual grounding (e.g., presence or gaze detection) to anchor responses in the environment; and (4) running formal user studies with standard HRI questionnaires, interviews, and controlled local-vs-cloud comparisons to quantify effects on trust, warmth, and perceived intelligence.

In parallel, we plan to explore optional on-device tool access (e.g., for simple real-time information queries) while maintaining the same privacy-preserving, fully local design.

7 Conclusion

We demonstrated a lightweight, fully local conversational pipeline running on a single consumer GPU laptop and deployed it on a life-like social robot across four public events. The approach reduces dependence on cloud services, protects privacy, and increases robustness in settings where connectivity is unreliable or restricted. Our analysis of 5,161 turns shows largely positive interactions but highlights the need for better turn-taking, concision, and light visual grounding. This work supports HRI Empowering Society by making capable, privacy-preserving, and environmentally mindful social robots more accessible to researchers, educators, and practitioners without proprietary dependencies.

References

[1] Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In *Advances in Knowledge Discovery and Data Mining*, Jian Pei, Vincent S. Tseng, Longbing Cao, Hiroshi Motoda, and Guandong Xu (Eds.). Springer, Berlin, Heidelberg, 160–172. doi:10.1007/978-3-642-37456-2_14

[2] ONNX Runtime developers. 2021. ONNX Runtime. https://onnxruntime.ai/.

[3] Gemini Robotics Team. 2025. Gemini Robotics: Bringing AI into the Physical World. arXiv:2503.20020 [cs.RO] https://arxiv.org/abs/2503.20020

[4] Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794* (2022).

[5] hexgrad. 2025. Kokoro-82M (v1.0). https://huggingface.co/hexgrad/Kokoro-82M Released 2025-01-27.

[6] Katherine Isbister, Peter Cottrell, Alessia Cecchet, Ella Dagan, Nikki Theofanopoulou, Ferran Altarriba Bertran, Aaron J. Horowitz, Nick Mead, Joel B. Schwartz, and Petr Slovák. 2022. Design (Not) Lost in Translation: A Case Study of an Intimate-Space Socially Assistive "Robot" for Emotion Regulation. *ACM Trans. Comput.-Hum. Interact.* 29, 4, Article 32 (mar 2022), 35 pages. doi:10.1145/3491083

[7] S. Joshi, W. Kamino, and S. Šabanović. 2025. Social robot accessories for tailoring and appropriation of social robots. *International Journal of Social Robotics* 17 (2025), 917–936. doi:10.1007/s12369-023-01077-y

[8] Takuhiro Kaneko, Kou Tanaka, Hirokazu Kameoka, and Shogo Seki. 2022. iSTFTNet: Fast and Lightweight Mel-Spectrogram Vocoder Incorporating Inverse Short-Time Fourier Transform. In *2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 6207–6211. doi:10.1109/ICASSP43922.2022.9746713

[9] Manuel Lauchenauer, Ali Mert, Simon Schlegel, Davide Miceli, Dario Hächler, and Mike Krey. 2025. *Social Robots: A Challenge for Technology Design*. 524–537. doi:10.1007/978-3-031-92611-2_34

[10] Yinghao Aaron Li, Cong Han, Vinay S. Raghavan, Gavin Mischler, and Nima Mesgarani. 2023. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models. In *Advances in Neural Information Processing Systems*, Vol. 36. Curran Associates, Inc., 19594–19621. https://proceedings.neurips.cc/paper_files/paper/2023/file/3eaad2a0b62b5ed7a2e66c2188bb1449-Paper-Conference.pdf

[11] Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software* 3, 29 (2018), 861. doi:10.21105/joss.00861

[12] Microsoft Research VibeVoice Team. 2025. VibeVoice-7B. https://huggingface.co/vibevoice/VibeVoice-7B. Long-form multi-speaker text-to-speech model, 7B parameters.

[13] Nari Labs. 2025. Dia2-2B. https://github.com/nari-labs/dia2. Streaming dialogue text-to-speech model, 2B parameters.

[14] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. doi:10.48550/ARXIV.2212.04356

[15] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, Hong Kong, China, 3982–3992. doi:10.18653/v1/D19-1410

[16] Finn Rietz, Alexander Sutherland, Suna Bensch, Stefan Wermter, and Thomas Hellström. 2021. WoZ4U: An Open-Source Wizard-of-Oz Interface for Easy, Efficient and Robust HRI Experiments. *Frontiers in Robotics and AI* Volume 8 - 2021 (2021). doi:10.3389/frobt.2021.668057

[17] Johan Schalkwyk, Ankit Kumar, Dan Lyth, Sefik Emre Eskimez, Zack Hodari, Cinjon Resnick, Ramon Sanabria, Raven Jiang, and the Sesame team. 2025. CSM-1B. https://huggingface.co/sesame/csm-1b. Conversational Speech Model, 1B parameters.

[18] Silero Team. 2024. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad.

[19] Carl Strathearn, Yanchao Yu, and Dimitra Gkatzia. 2023. TaskMaster: A Novel Cross-platform Task-based Spoken Dialogue System for Human-Robot Interaction. In *Proceedings of the Workshop on Human-Robot Conversational Interaction (HRCI) at the ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. Stockholm, Sweden. https://cui.acm.org/workshops/HRI2023/pdfs/HRCI23_paper_5.pdf

[20] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] https://arxiv.org/abs/2505.09388

[21] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] https://arxiv.org/abs/2412.15115

Received 2025-12-08; accepted 2026-01-12