DAFTAR PUSTAKA

[1] G. Bansal, V. Chamola, A. Hussain, M. Guizani, and D. Niyato, "Transforming Conversations with AI—A Comprehensive Study of ChatGPT," *Cognit. Comput.*, vol. 16, no. 5, pp. 2487–2510, Sep. 2024, doi: 10.1007/s12559-023-10236-2.

[2] F. Fui-Hoon Nah, R. Zheng, J. Cai, K. Siau, and L. Chen, "Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration," *Journal of Information Technology Case and Application Research*, vol. 25, no. 3, pp. 277–304, Jul. 2023, doi: 10.1080/15228053.2023.2233814.

[3] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, S. Mishra, and A. Abraham, "AI-Based Conversational Agents: A Scoping Review From Technologies to Future Directions," *IEEE Access*, vol. 10, pp. 92337–92356, 2022, doi: 10.1109/ACCESS.2022.3201144.

[4] C.-C. Lin, A. Y. Q. Huang, and S. J. H. Yang, "A Review of AI-Driven Conversational Chatbots Implementation Methodologies and Challenges (1999–2022)," *Sustainability*, vol. 15, no. 5, p. 4012, Feb. 2023, doi: 10.3390/su15054012.

[5] R. Kandoi, D. Dixit, M. Tyagi, and R. S. Yadav, "Conversational AI," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 3, pp. 769–775, Mar. 2024, doi: 10.22214/ijraset.2024.58787.

[6] S. Meshram, N. Naik, M. VR, T. More, and S. Kharche, "Conversational AI: Chatbots," in *2021 International Conference on Intelligent Technologies (CONIT)*, IEEE, Jun. 2021, pp. 1–6. doi: 10.1109/CONIT51480.2021.9498508.

[7] J. Chatterjee and N. Dethlefs, "This new conversational AI model can be your friend, philosopher, and guide ... and even your worst enemy," *Patterns*, vol. 4, no. 1, p. 100676, Jan. 2023, doi: 10.1016/j.patter.2022.100676.

[8] W. Cui *et al.*, "Recent Advances in Speech Language Models: A Survey," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 13943–13970. doi: 10.18653/v1/2025.acl-long.682.

[9] Y. Song and W. Xiong, "Large Language Model-Driven 3D Hyper-Realistic Interactive Intelligent Digital Human System," *Sensors*, vol. 25, no. 6, p. 1855, Mar. 2025, doi: 10.3390/s25061855.

[10] H. Aldarmaki, A. Ullah, S. Ram, and N. Zaki, "Unsupervised Automatic Speech Recognition: A review," *Speech Commun.*, vol. 139, pp. 76–91, Apr. 2022, doi: 10.1016/j.specom.2022.02.005.

[11] K. Pondel-Sycz *et al.*, "A comparative study of deep *End-to-end* Automatic Speech Recognition models for doctor-patient conversations in Polish in a real-life acoustic environment," *International Journal of Electronics and Telecommunications*, pp. 1–8, Jul. 2025, doi: 10.24425/ijet.2025.153609.

[12] R. Pires Magalhães *et al.*, "Evaluation of Automatic Speech Recognition Approaches," *Journal of Information and Data Management*, vol. 13, no. 3, Sep. 2022, doi: 10.5753/jidm.2022.2514.

[13] M. Daouad, F. Ataa Allah, and E. W. Dadi, "Optimizing Whisper models for Amazigh ASR: a comparative analysis," *Int. J. Speech Technol.*, vol. 28, no. 1, pp. 27–37, Mar. 2025, doi: 10.1007/s10772-024-10158-w.

[14] S. Oyucu, "Comparing The Fine-Tuning and Performance of Whisper Pre-Trained Models for Turkish Speech Recognition *Task*," in *2023 7th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*, IEEE, Oct. 2023, pp. 1–4. doi: 10.1109/ISMSIT58785.2023.10304891.

[15] A. D. Nguyen, H. M. Phi, A. V. Ngo, L. H. Trieu, and T. P. Nguyen, "Investigating Recent Large Language Models for Vietnamese Machine Reading Comprehension," in *2024 16th International Conference on Knowledge and System Engineering (KSE)*, IEEE, Nov. 2024, pp. 198–204. doi: 10.1109/KSE63888.2024.11063515.

[16] L. Villa, D. Carneros-Prado, C. C. Dobrescu, A. Sánchez-Miguel, G. Cubero, and R. Hervás, "Comparative Analysis of Generic and Fine-Tuned Large Language Models for Conversational Agent Systems," *Robotics*, vol. 13, no. 5, p. 68, Apr. 2024, doi: 10.3390/robotics13050068.

[17] Y. Kumar, A. Koul, and C. Singh, "A *deep learning* approaches in text-to-speech system: a systematic review and recent research perspective," *Multimed. Tools Appl.*, vol. 82, no. 10, pp. 15171–15197, Apr. 2023, doi: 10.1007/s11042-022-13943-4.

[18] H. Sun, J. Song, and Y. Jiang, "Fast *Inference End-to-end* Speech Synthesis with Style Diffusion," *Electronics (Basel).*, vol. 14, no. 14, p. 2829, Jul. 2025, doi: 10.3390/electronics14142829.

[19] C. Hong, J. Hyuk Lee, and H. Kook Kim, "Leveraging Low-Rank Adaptation for *Parameter*-Efficient Fine-Tuning in Multi-Speaker Adaptive Text-to-Speech Synthesis," *IEEE Access*, vol. 12, pp. 190711–190727, 2024, doi: 10.1109/ACCESS.2024.3515206.

[20] Y. Shirahata, R. Yamamoto, E. Song, R. Terashima, J.-M. Kim, and K. Tachibana, "Period VITS: Variational *Inference* with Explicit Pitch Modeling for *End-to-end* Emotional Speech Synthesis," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP), IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096480.

[21] S. Dutta, S. Chandupatla, and J. Hansen, "Adapting Whisper for Light*weight* and Efficient Automatic Speech Recognition of Children for On-device *Edge* Applications," in *Workshop on Child Computer Interaction - WOCCI 2025*, ISCA: ISCA, Aug. 2025, pp. 36–40. doi: 10.21437/WOCCI.2025-8.

[22] Q. Zhao, G. Sun, C. Zhang, M. Xu, and T. F. Zheng, "Enhancing Quantised *End-to-end* ASR Models Via Personalisation," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, Apr. 2024, pp. 12426–12430. doi: 10.1109/ICASSP48485.2024.10448012.

[23] Md. M. H. Shuvo, S. K. Islam, J. Cheng, and B. I. Morshed, "Efficient Acceleration of *Deep learning Inference* on Resource-Constrained *Edge* Devices: A Review," *Proceedings of the IEEE*, vol. 111, no. 1, pp. 42–91, Jan. 2023, doi: 10.1109/JPROC.2022.3226481.

[24] P.-E. Novac, G. Boukli Hacene, A. Pegatoquet, B. Miramond, and V. Gripon, "*Quantization* and *Deployment* of Deep Neural Networks on Microcontrollers," *Sensors*, vol. 21, no. 9, p. 2984, Apr. 2021, doi: 10.3390/s21092984.

[25] R. Wang, Z. Gao, L. Zhang, S. Yue, and Z. Gao, "Empowering large language models to *edge* intelligence: A survey of *edge* efficient LLMs and techniques," *Comput. Sci. Rev.*, vol. 57, p. 100755, Aug. 2025, doi: 10.1016/j.cosrev.2025.100755.

[26] C. Surianarayanan, J. J. Lawrence, P. R. Chelliah, E. Prakash, and C. Hewage, "A Survey on Optimization Techniques for *Edge* Artificial Intelligence (AI)," *Sensors*, vol. 23, no. 3, p. 1279, Jan. 2023, doi: 10.3390/s23031279.

[27] S. Ding, P. Meadowlark, Y. He, L. Lew, S. Agrawal, and O. Rybakov, "4-bit Conformer with Native *Quantization* Aware *Training* for Speech Recognition," in *Interspeech 2022*, ISCA: ISCA, Sep. 2022, pp. 1711–1715. doi: 10.21437/Interspeech.2022-10809.

[28] K. Zhen, M. Radfar, H. Nguyen, G. P. Strimel, N. Susanj, and A. Mouchtaris, "Sub-8-Bit *Quantization* for On-Device Speech Recognition: A Regularization-Free Approach," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Jan. 2023, pp. 15–22. doi: 10.1109/SLT54892.2023.10022821.

[29] S. Kim *et al.*, "*Integer-Only* Zero-Shot *Quantization* for Efficient Speech Recognition," in *ICASSP 2022 - 2022 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), IEEE, May 2022, pp. 4288–4292. doi: 10.1109/ICASSP43922.2022.9747552.

[30] K. T. Chitty-Venkata, S. Mittal, M. Emani, V. Vishwanath, and A. K. Somani, "A survey of techniques for optimizing transformer *inference*," *Journal of Systems Architecture*, vol. 144, p. 102990, Nov. 2023, doi: 10.1016/j.sysarc.2023.102990.

[31] Sumaiya, R. Jafarpourmarzouni, Y. Luo, S. Lu, and Z. Dong, "Toward *Real-time* and Efficient Perception Workflows in Software-Defined Vehicles," *IEEE Internet Things J.*, vol. 12, no. 6, pp. 7240–7258, Mar. 2025, doi: 10.1109/JIOT.2024.3492801.

[32] M. S. Lai *et al.*, "*Real-time* Avatar-Base Speech-to-Speech Conversational AI Tutor on AI PC," in *2025 IEEE 15th Symposium on Computer Applications & Industrial Electronics (ISCAIE)*, IEEE, May 2025, pp. 108–113. doi: 10.1109/ISCAIE64985.2025.11080818.

[33] I. Froiz-Míguez, P. Fraga-Lamas, and T. M. Fernández-CaraméS, "Design, Implementation, and Practical Evaluation of a Voice Recognition Based IoT Home Automation System for Low-Resource Languages and Resource-Constrained *Edge* IoT Devices: A System for Galician and Mobile Opportunistic Scenarios," *IEEE Access*, vol. 11, pp. 63623–63649, 2023, doi: 10.1109/ACCESS.2023.3286391.

[34] Y. Song, R. C.-W. Wong, and X. Zhao, "Speech-to-SQL: toward speech-driven SQL query generation from natural language question," *The VLDB Journal*, vol. 33, no. 4, pp. 1179–1201, Jul. 2024, doi: 10.1007/s00778-024-00837-0.

[35] H. Shao, B. Liu, W. Wang, X. Gong, and Y. Qian, "DQ-Whisper: Joint Distillation and *Quantization* for Efficient Multilingual Speech Recognition," in *2024 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Dec. 2024, pp. 240–246. doi: 10.1109/SLT61566.2024.10832149.

[36] S. Rajput and T. Sharma, "Benchmarking Emerging *Deep learning Quantization* Methods for Energy Efficiency," in *2024 IEEE 21st International Conference on Software Architecture Companion (ICSA-C)*, IEEE, Jun. 2024, pp. 238–242. doi: 10.1109/ICSA-C63560.2024.00049.

[37] S. K. Kamaraju, J. Hegde, S. Panchal, T. Gandi, and S. Shrivastav, "Conversational Payments on UPI Apps: A *Pipeline* Approach Leveraging ASR and NLP Techniques," in *Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD)*, New York, NY, USA: ACM, Dec. 2024, pp. 410–416. doi: 10.1145/3703323.3704275.

[38] Z. Li and Q. Gu, "I-ViT: *Integer-only Quantization* for Efficient Vision Transformer *Inference*," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2023, pp. 17019–17029. doi: 10.1109/ICCV51070.2023.01565.

[39] M. Someki, Y. Higuchi, T. Hayashi, and S. Watanabe, "ESPnet-ONNX: Bridging a Gap Between Research and Production," in *2022 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, IEEE, Nov. 2022, pp. 420–427. doi: 10.23919/APSIPAASC55919.2022.9979824.

[40] V. Ethiraj, A. David, S. Menon, and D. Vijay, "Toward Low-*Latency End-to-end* Voice Agents for Telecommunications Using Streaming ASR, *Quantized* LLMs, and *Real-time* TTS," Aug. 2025.

[41] T. Dai, "Using *Quantized* Neural Network for Speaker Recognition on *Edge* Computing Devices," *J. Phys. Conf. Ser.*, vol. 1992, no. 2, p. 022177, Aug. 2021, doi: 10.1088/1742-6596/1992/2/022177.

[42] K. Coussement, M. Z. Abedin, M. Kraus, S. Maldonado, and K. Topuz, "Explainable AI for enhanced decision-making," *Decis. Support Syst.*, vol. 184, p. 114276, Sep. 2024, doi: 10.1016/j.dss.2024.114276.

[43] Muhammad Eid BALBAA and Marina Sagatovna ABDURASHIDOVA, "THE IMPACT OF ARTIFICIAL INTELLIGENCE IN DECISION MAKING: A COMPREHENSIVE REVIEW," *EPRA International Journal of Economics, Business and Management Studies*, pp. 27–38, Feb. 2024, doi: 10.36713/epra15747.

[44] F. Aslam, "The Impact of Artificial Intelligence on Chatbot Technology: A Study on the Current Advancements and Leading Innovations," *European Journal of Technology*, vol. 7, no. 3, pp. 62–72, Aug. 2023, doi: 10.47672/ejt.1561.

[45] A. B. Saka, L. O. Oyedele, L. A. Akanbi, S. A. Ganiyu, D. W. M. Chan, and S. A. Bello, "Conversational artificial intelligence in the AEC industry: A review of present status, challenges and opportunities," *Advanced Engineering Informatics*, vol. 55, p. 101869, Jan. 2023, doi: 10.1016/j.aei.2022.101869.

[46] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[47] P. Saxena, V. Saxena, A. Pandey, U. A. Prync Flato, and K. Shukla, *Multiple Aspects of Artificial Intelligence*. Book Saga Publications, 2023. doi: 10.60148/muasartificialintelligence.

[48] L. Alzubaidi *et al.*, "Review of *deep learning*: concepts, CNN architectures, challenges, applications, future directions," *J. Big Data*, vol. 8, no. 1, p. 53, Mar. 2021, doi: 10.1186/s40537-021-00444-8.

[49] S. Singhal, M. Deswal, and Er. S. Sharma, "AUTOMATIC SPEECH RECOGNITION USING DEEP NEURAL NETWORKS," *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, vol. 07, no. 11, pp. 1–11, Nov. 2023, doi: 10.55041/IJSREM27313.

[50] A. M. Pérez González de Martos, "Deep Neural Networks for Automatic Speech-To-Speech Translation of Open Educational Resources," Universitat Politècnica de València, Valencia (Spain), 2022. doi: 10.4995/Thesis/10251/184019.

[51] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of *deep learning* techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, Nov. 2023, doi: 10.1016/j.inffus.2023.101869.

[52] J. Datta, R. Rabbi, P. Saha, A. N. Zereen, M. Abdullah-Al-Wadud, and J. Uddin, "Deep representation learning using layer-wise VICReg losses," *Sci. Rep.*, vol. 15, no. 1, p. 27049, Jul. 2025, doi: 10.1038/s41598-025-08504-2.

[53] S. Dong, P. Wang, and K. Abbas, "A survey on *deep learning* and its applications," *Comput. Sci. Rev.*, vol. 40, p. 100379, May 2021, doi: 10.1016/j.cosrev.2021.100379.

[54] S. Li, F. Yu, S. Zhang, H. Yin, and H. Lin, "Optimization of Direct Convolution Algorithms on ARM Processors for *Deep learning Inference*," *Mathematics*, vol. 13, no. 5, p. 787, Feb. 2025, doi: 10.3390/math13050787.

[55] I. J. Ratul, Y. Zhou, and K. Yang, "Accelerating *Deep learning Inference*: A Comparative Analysis of Modern Acceleration *Frameworks*," *Electronics (Basel).*, vol. 14, no. 15, p. 2977, Jul. 2025, doi: 10.3390/electronics14152977.

[56] S. Thakur *et al.*, "Optimization of *deep learning* models for *inference* in low resource environments," *Comput. Biol. Med.*, vol. 196, p. 110615, Sep. 2025, doi: 10.1016/j.compbiomed.2025.110615.

[57] Z. Xu, H. Peng, and W. Wang, "AGO: Boosting Mobile AI *Inference* Performance by Removing Constraints on Graph Optimization," in *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*, IEEE, May 2023, pp. 1–10. doi: 10.1109/INFOCOM53939.2023.10228858.

[58] Z. Li, H. Li, and L. Meng, "Model Compression for Deep Neural Networks: A Survey," *Computers*, vol. 12, no. 3, p. 60, Mar. 2023, doi: 10.3390/computers12030060.

[59] A. Kuzmin, M. Nagel, M. van Baalen, A. Behboodi, and T. Blankevoort, "Pruning vs *Quantization*: Which is Better?," Feb. 2024.

[60] S. Makenali, B. Rokh, and A. Azarpeyvand, "Integrating Pruning with *Quantization* for Efficient Deep Neural Networks Compression," Sep. 2025.

[61] K. Kim, J. Jang, J. Choi, Y. Lee, K. Jin, and Y. Kim, "Plug-in and Fine-tuning: Bridging the Gap between Small Language Models and Large Language Models," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 5434–5452. doi: 10.18653/v1/2025.acl-long.271.

[62] S. Makridakis, F. Petropoulos, and Y. Kang, "Large Language Models: Their Success and Impact," *Forecasting*, vol. 5, no. 3, pp. 536–549, Aug. 2023, doi: 10.3390/forecast5030030.

[63] Z. Zhao *et al.*, "Recommender Systems in the Era of Large Language Models (LLMs)," *IEEE Trans. Knowl. Data Eng.*, vol. 36, no. 11, pp. 6889–6907, Nov. 2024, doi: 10.1109/TKDE.2024.3392335.

[64] H. Naveed *et al.*, "A Comprehensive Overview of Large Language Models," *ACM Trans. Intell. Syst. Technol.*, vol. 16, no. 5, pp. 1–72, Oct. 2025, doi: 10.1145/3744746.

[65] R. Zhang, "A Comparative Analysis of LSTM and Transformer-based Automatic Speech Recognition Techniques," *Transactions on Computer Science and Intelligent Systems Research*, vol. 5, pp. 272–276, Aug. 2024, doi: 10.62051/zq6v0d49.

[66] A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T. K. Whangbo, "Improved Feature *Parameter* Extraction from Speech Signals Using Machine Learning Algorithm," *Sensors*, vol. 22, no. 21, p. 8122, Oct. 2022, doi: 10.3390/s22218122.

[67] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of *deep learning* techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, Nov. 2023, doi: 10.1016/j.inffus.2023.101869.

[68] X. Tan *et al.*, "*NaturalSpeech* : *End-to-end* Text-to-Speech Synthesis With Human-Level Quality," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 6, pp. 4234–4245, Jun. 2024, doi: 10.1109/TPAMI.2024.3356232.

[69] J. Yu *et al.*, "DIA-TTS: Deep-Inherited Attention-Based Text-to-Speech Synthesizer," *Entropy*, vol. 25, no. 1, p. 41, Dec. 2022, doi: 10.3390/e25010041.

[70] A. Radford, J. Kim, T. Xu, G. Brockman, C. McLeavy, and I. Sutskever, "Robust Speech Recognition via Large-Scale Weak Supervision,"

Proceedings of the 40th International Conference on Machine Learning, pp. 28492–28518, Jul. 2023.

[71] G. Team and G. Deepmidn, "Gemma 3 Technical Report," Mar. 2025.

[72] V. Pratap, A. Tjandra, B. Shi, and P. Tomasello, "Scaling Speech Technology to 1,000+ Languages," *The Journal of Machine Learning Research*, vol. 25, no. 1, pp. 4798–4849, Jan. 2024.

[73] K. Jaehyeon, J. Kong, and J. Son, "Conditional Variational Auto*encoder* with Adversarial Learning for *End-to-end* Text-to-Speech," Dec. 2021.

[74] R. Abdel-Salam, A. H. Abdel-Gawad, and A. G. Wassal, "VLCQ: *Post-training quantization* for deep neural networks using variable length coding," *Future Generation Computer Systems*, vol. 166, p. 107654, May 2025, doi: 10.1016/j.future.2024.107654.

[75] J. Lin *et al.*, "Adaptive Rounding Compensation for *Post-training Quantization*," 2023, pp. 74–85. doi: 10.1007/978-981-99-1642-9_7.

[76] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. W. Mahoney, and K. Keutzer, "A Survey of *Quantization* Methods for Efficient Neural Network *Inference*," in *Low-Power Computer Vision*, Boca Raton: Chapman and Hall/CRC, 2022, pp. 291–326. doi: 10.1201/9781003162810-13.

[77] J. Lin, J. Tang, H. Tang, S. Yang, G. Xiao, and S. Han, "AWQ: *Activation*-aware *Weight Quantization* for On-Device LLM Compression and Acceleration," *GetMobile: Mobile Computing and Communications*, vol. 28, no. 4, pp. 12–17, Jan. 2025, doi: 10.1145/3714983.3714987.

[78] G. Klein, D. Zhang, C. Chouteau, J. Crego, and J. Senellart, "Efficient and High-Quality Neural Machine Translation with OpenNMT," in *Proceedings of the Fourth Workshop on Neural Generation and Translation*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 211–217. doi: 10.18653/v1/2020.ngt-1.25.

[79] K. Egashira, R. Staab, M. Vero, J. He, and M. Vechev, "Mind the Gap: A Practical Attack on GGUF *Quantization*," 2025.

[80] K. A. K. Borgersen and M. Goodwin, "English K_*Quantization* of LLMs Does Not Disproportionately Diminish Multilingual Performance," 2026, pp. 106–115. doi: 10.1007/978-3-032-11402-0_8.

[81] O. H. Anidjar, R. Marbel, and R. Yozevitch, "Whisper Turns Stronger: Augmenting Wav2Vec 2.0 for Superior ASR in Low-Resource Languages," Dec. 2024.

[82] T. Rosenbaum, E. Winebrand, O. Cohen, and I. Cohen, "Deep-Learning *Framework* for Efficient *Real-time* Speech Enhancement and

Dereverberation," *Sensors*, vol. 25, no. 3, p. 630, Jan. 2025, doi: 10.3390/s25030630.

[83] S. Sagi, "Optimizing LLM *Inference*: Metrics that Matter for Real Time Applications," *Journal of Artificial Intelligence & Cloud Computing*, pp. 1–4, Feb. 2025, doi: 10.47363/JAICC/2025(4)446.

[84] J. Conde, M. González, P. Reviriego, Z. Gao, S. Liu, and F. Lombardi, "Speed and Conversational Large Language Models: Not All Is About Tokens per Second," *Computer (Long Beach. Calif).*, vol. 57, no. 8, pp. 74–80, Aug. 2024, doi: 10.1109/MC.2024.3399384.

[85] S. Raschka, J. Patterson, and C. Nolet, "Machine Learning in Python: Main Developments and Technology Trends in Data Science, Machine Learning, and Artificial Intelligence," *Information*, vol. 11, no. 4, p. 193, Apr. 2020, doi: 10.3390/info11040193.

[86] A. Kapoor, "Analysis of Code Editors: Features, Evolution, and Impact on Software Development," 2024. doi: 10.2139/ssrn.5021567.

[87] P. Arisaputra, A. Zahra, and A. Handoyo, "Indonesian Automatic Speech Recognition with XLSR-53".

[88] J. M. Springer, V. Adlakha, S. Reddy, A. Raghunathan, and M. Mosbach, "Understanding the Influence of Synthetic Data for Text Embedders," in *Findings of the Association for Computational Linguistics: ACL 2025*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2025, pp. 22551–22567. doi: 10.18653/v1/2025.findings-acl.1160.

[89] S. Gondi and V. Pratap, "Performance and Efficiency Evaluation of ASR *Inference* on the *Edge*," *Sustainability*, vol. 13, no. 22, p. 12392, Nov. 2021, doi: 10.3390/su132212392.

[90] M. Mensah, I. Wiafe, and A. Ekpezu, "Benchmarking Akan ASR Models Across *Domain*-Specific Datasets: A Comparative Evaluation of Performance, Scalability, and Adaptability".

[91] C. Arriaga, A. Pozo, J. Conde, and A. Alonso, "Assessing *Latency* in ASR Systems: A Methodological Perspective for *Real-time* Use," 2024.

[92] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A Survey on Model Compression for Large Language Models," *Trans. Assoc. Comput. Linguist.*, vol. 12, pp. 1556–1577, Nov. 2024, doi: 10.1162/tacl_a_00704.

[93] Y. Sheng, L. Zheng, and B. Yuan, "FlexGen: high-*throughput* generative *inference* of large language models with a single GPU," *ICML'23: Proceedings of the 40th International Conference on Machine Learning*, 2023.

[94] N. Brown, “Enhancing Trust in LLMs: Algorithms for Comparing and Interpreting LLMs,” 2024.

[95] Y. Cui, P. He, and J. Zeng, “Stepwise Perplexity-Guided Refinement for Efficient Chain-of-Thought Reasoning in Large Language Models,” 2025.

[96] V. Ethiraj, A. David, S. Menon, and D. Vijay, “Toward Low-*Latency End-to-end* Voice Agents for Telecommunications Using Streaming ASR, *Quantized* LLMs, and *Real-time* TTS,” 2025.

[97] S. Braun and H. Gamper, “Effect of Noise Suppression Losses on Speech Distortion and ASR Performance,” in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, May 2022, pp. 996–1000. doi: 10.1109/ICASSP43922.2022.9746489.

[98] A. Söhler, J. Irigoyen, and A. S. Kirkedal, “Quantizing Whisper-small: How design choices affect ASR performance,” Nov. 2025.