

REFERENCES

- [1] I. Ahmed, G. Jeon, and F. Piccialli, "From artificial intelligence to explainable artificial intelligence in industry 4.0: A survey on what, how, and where," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 8, pp. 5031–5042, Aug. 2022.
- [2] W. Zhang, Y. Feng, F. Meng, D. You, and Q. Liu, "Bridging the gap between training and inference for neural machine translation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- [3] A. Gramegna and P. Giudici, "Why to buy insurance? an explainable artificial intelligence approach," *Risks*, vol. 8, no. 4, p. 137, 2020.
- [4] E. Darraj, C. Sample, and C. Justice, "Artificial intelligence cybersecurity framework: Preparing for the here and now with ai," in *Proceedings of the 18th European Conference on Cyber Warfare and Security*. Academic Conferences and Publishing Limited, 2019, pp. 132–141.
- [5] M.-C. Chang, J. Wei, Z.-A. Zhu, Y.-M. Chen, C.-S. Hu, M.-X. Jiang, and C.-K. Chiang, "Ai city challenge 2019-city-scale video analytics for smart transportation," in *Proceedings of the CVPR Workshops*, 2019, pp. 99–108.
- [6] OpenAI. (2024) Hello gpt-4o. Accessed on 24 Dec 2024, <https://openai.com/index/hello-gpt-4o/>.
- [7] S. Pichai. (2024) Google gemini ai. Accessed on 24 Dec 2024, <https://blog.google/technology/ai/google-gemini-ai/#sundar-note>.
- [8] Meta, "Llama 3.2: Revolutionizing edge ai and vision with open, customizable models," <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>, 2024, [Accessed: Jan. 8, 2025].
- [9] P. Yu, H. Xu, X. Hu, and C. Deng, "Leveraging generative ai and large language models: A comprehensive roadmap for healthcare integration," *Healthcare*, vol. 11, no. 20, p. 2776, Oct. 2023.
- [10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 2019, pp. 4171–4186.
- [11] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, Jul. 2020, pp. 7871–7880.

- [12] H. Y. Kwan, “User-focused telehealth powered by llms: Bridging the gap between technology and human-centric care delivery,” in *Proceedings of the 4th International Conference on Computing, Communication and Artificial Intelligence (CCAI)*, Xi’an, China, 2024, pp. 187–191.
- [13] A. García-Barragán, A. G. Calatayud, L. Prieto-Santamaría, V. Robles, E. Menasalvas, and A. Rodríguez, “Step-forward structuring disease phenotypic entities with llms for disease understanding,” in *Proceedings of the IEEE 37th International Symposium on Computer-Based Medical Systems (CBMS)*, Guadalajara, Mexico, 2024, pp. 213–218.
- [14] A. Ríos-Hoyo, N. L. Shan, A. Li, A. T. Pearson, L. Pusztai, and F. M. Howard, “Evaluation of large language models as a diagnostic aid for complex medical cases,” *Frontiers in Medicine*, vol. 11, 2024.
- [15] C. Shyr, Y. Hu, L. Bastarache, A. Cheng, R. Hamid, P. Harris, and H. Xu, “Identifying and extracting rare diseases and their phenotypes with large language models,” *Journal of Healthcare Informatics Research*, vol. 8, no. 2, pp. 438–461, 2024.
- [16] D. Brown and A. Green, “Prompt engineering: A cost-effective approach for ai adaptation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 7, pp. 1243–1255, 2022.
- [17] M. Yuan, P. Bao, J. Yuan, Y. Shen, Z. Chen, Y. Xie *et al.*, “Large language models illuminate a progressive pathway to artificial intelligent healthcare assistant,” *Med. Plus*, vol. 1, p. 100030, 2024.
- [18] G. Beri and V. Srivastava, “Advanced techniques in prompt engineering for large language models: A comprehensive study,” in *Proceedings of the 2024 IEEE 4th International Conference on ICT for Business, Industry and Government (ICTBIG)*, 2024.
- [19] G. McPeak, A. Sautmann, O. George, A. Hallal, E. A. Simal, A. L. Schwartz *et al.*, “An llm’s medical testing recommendations in a nigerian clinic: potential and limits of prompt engineering for clinical decision support,” in *Proceedings of the 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*, 2024, pp. 586–591.
- [20] S. Zhou, M. Lin, S. Ding, J. Wang, C. Chen, G. B. Melton *et al.*, “Explainable differential diagnosis with dual-inference large language models,” *npj Health Systems*, vol. 2, no. 12, 2025. [Online]. Available: <https://doi.org/10.1038/s44401-025-00015-6>
- [21] T. Savage, A. Nayak, R. Gallo, E. Rangan, and J. H. Chen, “Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine,” *npj Digital Medicine*, vol. 7, no. 1, p. 20, 2024.
- [22] V. Pawar, M. Gawande, A. Kollu, and A. S. Bile, “Exploring the potential of prompt engineering: A comprehensive analysis of interacting with large language models,” in *Proceedings of the 2024 8th International Conference on Computing, Communication, Control and Automation (ICCUBEA)*, Pune, India, August 2024, pp. 1–9.

- [23] N. Bui, G. Nguyen, N. Nguyen, B. Vo, L. Vo, T. Huynh *et al.*, “Fine-tuning large language models for improved health communication in low-resource languages,” *Computer Methods and Programs in Biomedicine*, vol. 263, p. 108655, 2025.
- [24] S. S. Z. Yang, G. M. Fye, W. C. Yap, and D. Z. Yu, “Enhancing medical summarization with parameter efficient fine tuning on local cpus,” in *Proceedings of the 5th International Conference on Electrical, Communication and Computer Engineering (ICECCE)*, Kuala Lumpur, Malaysia, October 2024.
- [25] H. K. Chaubey, G. Tripathi, S. K. Gopalaiyengar, and R. Ranjan, “Comparative analysis of rag, fine-tuning, and prompt engineering in chatbot development,” in *Proceedings of the 2024 International Conference on Future Technology and Smart Society (ICFTSS)*, 2024, pp. 169–172.
- [26] S. M. Shah, S. A. Gillani, M. S. A. Baig, M. A. Saleem, and M. H. Siddiqui, “Advancing depression detection on social media platforms through fine-tuned large language models,” *Online Social Networks and Media*, vol. 46, p. 100311, mar 2025.
- [27] W. Cui, K. Kawamoto, K. Morgan, and J. Finkelstein, “Reducing diagnostic uncertainty in emergency departments: The role of large language models in age-specific diagnostics,” in *2024 IEEE 12th International Conference on Healthcare Informatics (ICHI)*. IEEE, 2024, pp. 525–527.
- [28] J. Kainz, P. Seisl, M. Grob, L. Hauptfeld, J. Wahringer, A. Rappelsberger, and K.-P. Adlassnig, “Fine-tuning an existing large language model with knowledge from the medical expert system Hepaxpert,” in *Intelligent Health Systems – From Technology to Data and Knowledge*. IOS Press, 2025, pp. 143–147.
- [29] B. Chen, F. Yi, and D. Varró, “Prompting or fine-tuning? a comparative study of large language models for taxonomy construction,” in *Proceedings of the 2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, 2023, pp. 588–596.
- [30] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi *et al.*, “Retrieval-augmented generation for large language models: A survey,” *arXiv preprint*, vol. arXiv:2312.10997v5, Mar. 2024. [Online]. Available: <https://arxiv.org/abs/2312.10997>
- [31] F. G. Bingöl, D. Ağagündüz, and M. C. Bingol, “Accuracy of current large language models and the retrieval-augmented generation model in determining dietary principles in chronic kidney disease,” *Journal of Renal Nutrition*, 2025, article in Press.
- [32] J. Ge, S. Sun, J. Owens, V. Galvez, O. Gologorskaya, J. C. Lai *et al.*, “Development of a liver disease-specific large language model chat interface using retrieval augmented generation,” *medRxiv preprint*, Nov. 2023.
- [33] W. Shi, Y. Zhuang, Y. Zhu, H. J. Iwinski, J. M. Wattenbarger, and M. D. Wang, “Retrieval-augmented large language models for adolescent idiopathic scoliosis patients in shared decision-making,” in *Proc. 14th ACM Int. Conf. Bioinformatics*,

Computational Biology and Health Informatics (BCB '23), Houston, TX, USA, Sep. 2023, pp. 1–10.

- [34] T. Y. C. Tam, S. Sivarajkumar, S. Kapoor, A. V. Stolyar, K. Polanska, K. R. McCarthy *et al.*, “A framework for human evaluation of large language models in healthcare derived from literature review,” *npj Digital Medicine*, vol. 7, no. 1, p. 258, Sep. 2024.
- [35] S. Zhou, Z. Xu, M. Zhang, C. Xu, Y. Guo, Z. Zhan *et al.*, “Large language models for disease diagnosis: A scoping review,” 2025. [Online]. Available: <https://arxiv.org/abs/2409.00097>
- [36] J. Risch, T. Möller, J. Gutsch, and M. Pietsch, “Semantic answer similarity for evaluating question answering models,” in *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*. Association for Computational Linguistics, Nov. 2021, pp. 149–157.
- [37] B. Mahesh, “Machine learning algorithms - a review,” *International Journal of Science and Research (IJSR)*, vol. 9, 2019.
- [38] N. Sharma, R. Sharma, and N. Jindal, “Machine learning, deep learning applications - a vision,” *Global Transitions Proceedings*, vol. 2, no. 1, pp. 24–28, 2021.
- [39] M. JayaSree and L. K. Rao, “A deep insight into deep learning architectures, algorithms and applications,” in *2022 International Conference on Electronics and Renewable Systems (ICEARS)*, Tuticorin, India, 2022, pp. 1134–1142.
- [40] K. Chowdhary and K. Chowdhary, “Natural language processing,” in *Fundamentals of Artificial Intelligence*, 2020, pp. 603–649.
- [41] D. Khurana, A. Koli, K. Khatter, and S. Singh, “Natural language processing: State of the art, current trends and challenges,” *Multimedia Tools and Applications*, vol. 82, no. 3, pp. 3713–3744, 2023.
- [42] G. Salton, A. Wong, and C.-S. Yang, “A vector space model for automatic indexing,” *Communications of the ACM*, vol. 18, no. 11, pp. 613–620, 1975.
- [43] K. Sparck Jones, “A statistical interpretation of term specificity and its application in retrieval,” *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.
- [44] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of the ICLR 2013 Workshop Track*, Scottsdale, Arizona, USA, 2013, presented at the International Conference on Learning Representations (ICLR 2013), Workshop Track.
- [45] J. Pennington, R. Socher, and C. D. Manning, “GloVe: Global vectors for word representation,” in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 1532–1543. [Online]. Available: <https://aclanthology.org/D14-1162>

- [46] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. [Online]. Available: <https://aclanthology.org/N18-1202>
- [47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [48] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.
- [49] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, "Progress in neural nlp: Modeling, learning, and reasoning," *Engineering*, vol. 6, no. 3, pp. 275–290, 2020.
- [50] B. Sindhu, R. P. Prathamesh, M. B. Sameera, and S. KumaraSwamy, "The evolution of large language model: Models, applications and challenges," in *2024 International Conference on Current Trends in Advanced Computing (ICCTAC)*, Bengaluru, India, 2024, pp. 1–8.
- [51] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix *et al.*, "Llama: Open and efficient foundation language models," *CoRR*, vol. abs/2302.13971, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2302.13971>
- [52] Meta, "Llama 3.2 | model cards and prompt formats," [https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/#-llama-3.2-lightweight-models-\(1b/3b\)/](https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_2/#-llama-3.2-lightweight-models-(1b/3b)/), 2024, [Accessed: Jan. 8, 2025].
- [53] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal *et al.*, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901.
- [54] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.
- [55] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 24 824–24 837.
- [56] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," *arXiv preprint arXiv:1801.06146*, 2018. [Online]. Available: <https://arxiv.org/abs/1801.06146>
- [57] S. Ruder, "Neural transfer learning for natural language processing," Ph.D. dissertation, National University of Ireland, Galway, 2019.

- [58] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith, “Don’t stop pretraining: Adapt language models to domains and tasks,” *arXiv preprint arXiv:2004.10964*, 2020. [Online]. Available: <https://arxiv.org/abs/2004.10964>
- [59] C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?” *arXiv preprint arXiv:1905.05583*, 2019. [Online]. Available: <https://arxiv.org/abs/1905.05583>
- [60] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein *et al.*, “On the opportunities and risks of foundation models,” *arXiv preprint arXiv:2108.07258*, 2022. [Online]. Available: <https://arxiv.org/abs/2108.07258>
- [61] C. Jeong, “Fine-tuning and utilization methods of domain-specific llms,” *Journal of Intelligence and Information Systems*, vol. 30, no. 1, pp. 93–120, Mar. 2024, [Online]. Available: <http://dx.doi.org/10.13088/jiis.2024.30.1.093>.
- [62] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, “LoRA: Low-rank adaptation of large language models,” 2021, [Online]. Available: <https://arxiv.org/abs/2106.09685>.
- [63] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, “Parameter-efficient transfer learning for nlp,” 2019, [Online]. Available: <https://arxiv.org/abs/1902.00751>.
- [64] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “Qlora: Efficient fine-tuning of quantized llms,” 2023, arXiv preprint arXiv:2305.14314.
- [65] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.
- [66] D. Jurafsky and J. H. Martin, *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. USA: Prentice Hall PTR, 2009.
- [67] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, “Bertscore: Evaluating text generation with BERT,” in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: <https://openreview.net/forum?id=SkeHuCVFDr>
- [68] M. AI, “Llama 3.2 3b model card,” Hugging Face Model Hub, 2024, [Online]. Available: <https://huggingface.co/meta-llama/Llama-3.2-3B>.
- [69] J. Chen, Z. Cai, K. Ji, X. Wang, W. Liu, R. Wang, J. Hou, and B. Wang, “Huatuogpt-o1, towards medical complex reasoning with llms,” 2024. [Online]. Available: <https://arxiv.org/abs/2412.18925>
- [70] R. S. Porter and J. L. Kaplan, *The Merck Manual of Diagnosis and Therapy*, 19th ed. Whitehouse Station, NJ, USA: Merck & Co., 2011.

- [71] E. Braunwald, A. S. Fauci, D. L. Kasper, S. L. Hauser, D. L. Longo, and J. L. Jameson, *Harrison's Principles of Internal Medicine*, 15th ed. New York, NY, USA: McGraw-Hill Professional, 2001.
- [72] "Merck manual of diagnosis and therapy," https://en.wikipedia.org/wiki/Merck_Manual_of_Diagnosis_and_Therapy, 2025, wikipedia, Accessed: May 30, 2025.
- [73] "Awards | medical doctorate program - unc school of medicine," <https://www.med.unc.edu/md/events-awards/awards/>, 2025, university of North Carolina School of Medicine, Accessed: May 30, 2025.
- [74] "Accessmedicine and other access databases," <https://lib.arizona.edu/hsl/materials/access-databases>, 2025, university of Arizona Health Sciences Library, Accessed: May 30, 2025.
- [75] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation." USA: Association for Computational Linguistics, 2002. [Online]. Available: <https://doi.org/10.3115/1073083.1073135>
- [76] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*. Barcelona, Spain: Association for Computational Linguistics, 2004, pp. 74–81. [Online]. Available: <https://aclanthology.org/W04-1013>