Endnotes

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I., 2017, Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
2. Wikipedia, 2024, Word n-gram language model. Available at: https://en.wikipedia.org/wiki/Word_n-gram_language_model.
3. Sutskever, I., Vinyals, O., & Le, Q. V., 2014, Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems,* 27.
4. Gu, A., Goel, K., & Ré, C., 2021, Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*.
5. Jalammar, J. (n.d.). The illustrated transformer. Available at: https://jalammar.github.io/illustrated-transformer/.
6. Ba, J. L., Kiros, J. R., & Hinton, G. E., 2016, Layer normalization. *arXiv preprint arXiv:1607.06450*.
7. He, K., Zhang, X., Ren, S., & Sun, J., 2016, Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
8. HuggingFace., 2024, Byte Pair Encoding. Available at: https://huggingface.co/learn/nlp-course/chapter6/5?fw=pt.
9. Kudo, T., & Richardson, J., 2018, Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.
10. HuggingFace, 2024, Unigram tokenization. Available at: https://huggingface.co/learn/nlp-course/chapter6/7?fw=pt.
11. Goodfellow et. al., 2016, Deep Learning. MIT Press. Available at: http://www.deeplearningbook.org.
12. Radford, Alec et al., 2019, Language models are unsupervised multitask learners.
13. Brown, Tom, et al., 2020, Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.
14. Devlin, Jacob, et al., 2018, BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

15. Radford, A., & Narasimhan, K., 2018, Improving language understanding by generative pre-training.

16. Dai, A., & Le, Q., 2015, Semi-supervised sequence learning. *Advances in Neural Information Processing Systems*.

17. Ouyang, Long, et al., 2022, Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems,* 35, 27730-27744.-27744.

18. OpenAI., 2023, GPT-3.5. Available at: https://platform.openai.com/docs/models/gpt-3-5.

19. OpenAI., 2023, GPT-4 Technical Report. Available at: https://arxiv.org/abs/2303.08774.

20. Thoppilan, Romal, et al., 2022, Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*.

21. Llama 3.2: Revolutionizing edge AI and vision with open, customizable models. Available at: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/.

22. Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., ... & Irving, G., 2021, Scaling language models: Methods, analysis & insights from training Gopher. Available at: https://arxiv.org/pdf/2112.11446.pdf.

23. Du, N., He, H., Dai, Z., Mccarthy, J., Patwary, M. A., & Zhou, L., 2022, GLAM: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning* (pp. 2790-2800). PMLR.

24. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D., 2020, Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

25. Hoffmann, Jordan, et al., 2022, Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

26. Shoeybi, Mohammad, et al., 2019, Megatron-LM: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

27. Muennighoff, N. et al., 2023, Scaling data-constrained language models. *arXiv preprint arXiv:2305.16264*.

28. Chowdhery, Aakanksha, et al., 2023, Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240), 1-113.

29. Wang, Alex, et al.,2019, SuperGLUE: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

30. Anil, Rohan, et al., 2023, Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

31. DeepMind, 2023, Gemini: A family of highly capable multimodal models. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_1_report.pdf.

32. DeepMind, 2024, Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Available at: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

33. Google Developers, 2024, Introducing PaLi-Gemma, Gemma 2, and an upgraded responsible AI toolkit. Available at: https://developers.googleblog.com/en/gemma-family-and-toolkit-expansion-io-2024/.

34. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M., Lacroix, T., ... & Jegou, H., 2023, Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

35. Jiang, A. Q., 2024, Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

36. Qwen, 2024, Introducing Qwen1.5. Available at: https://qwenlm.github.io/blog/qwen1.5/.

37. Young, A., 2024, Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*.

38. Grok-1, 2024, Available at: https://github.com/xai-org/grok-1.

39. Duan, Haodong, et al., 2023, BotChat: Evaluating LLMs' capabilities of having multi-turn dialogues. *arXiv preprint arXiv:2310.13650*.

40. Google Cloud, 2024, *Tune text models with reinforcement learning from human feedback*. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/models/tune-text-models-rlhf.

41. Bai, Yuntao, et al., 2022, Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

42. Wikipedia, 2024, Likert scale. Available at: https://en.wikipedia.org/wiki/Likert_scale.

43. Sutton, R. S., & Barto, A. G., 2018, *Reinforcement learning: An introduction*. MIT Press.

44. Bai, Yuntao, et al, 2022, Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

45. Rafailov, Rafael, et al., 2023, Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

46. Houlsby, Neil, et al., 2019, Parameter-efficient transfer learning for NLP. In *International Conference on Machine Learning* (pp. 2790-2799). PMLR.

47. Hu, Edward J., et al., 2021, LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

48. Dettmers, Tim, et al., 2023, QLoRA: Efficient finetuning of quantized LLMs. *arXiv preprint arXiv:2305.14314*.

49. Lester, B., Al-Rfou, R., & Constant, N., 2021, The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

50. HuggingFace., 2020, How to generate text? Available at: https://huggingface.co/blog/how-to-generate.

51. Google AI Studio Context caching. Available at: https://ai.google.dev/gemini-api/docs/caching?lang=python.

52. Vertex AI Context caching overview. Available at: https://cloud.google.com/vertex-ai/generative-ai/docs/context-cache/context-cache-overview.

53. Gu, A., Goel, K., & Ré, C., 2021, Efficiently modeling long sequences with structured state spaces. Available at: https://arxiv.org/abs/2111.00396.

54. Hubara et al., 2016, Quantized neural networks: Training neural networks with low precision weights and activations. Available at: https://arxiv.org/abs/1609.07061.

55. Benoit Jacob et al., 2017, Quantization and training of neural networks for efficient integer-arithmetic-only inference. Available at: https://arxiv.org/abs/1712.05877.

56. Bucila, C., Caruana, R., & Niculescu-Mizil, A., 2006, Model compression. *Knowledge Discovery and Data Mining*. Available at: https://www.cs.cornell.edu/~caruana/compression.kdd06.pdf.

57. Hinton, G., Vinyals, O., & Dean, J., 2015, Distilling the knowledge in a neural network. Available at: https://arxiv.org/abs/1503.02531.

58. Zhang, L., Fei, W., Wu w., He Y., Lou Z., Zhou H., 2023, Dual Grained Quantisation: Efficient Finegrained Quantisation for LLM. Available at: https://arxiv.org/abs/2310.04836.

59. Agarwal, R., Vieillard, N., Zhou, Y., Stanczyk, P., Ramos, S., Geist, M., Bachem, O., 2024, On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. Available at: https://arxiv.org/abs/2306.13649.

60. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., & Dean, J., 2017, Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. Available at: https://arxiv.org/abs/1701.06538.

61. Schuster, T., Fried, D., & Jurafsky, D., 2022, Confident adaptive language modeling. Available at: https://arxiv.org/abs/2207.07061.

62. Tri Dao et al. "FlashAttention. Available at: https://arxiv.org/abs/2205.14135.

63. Leviathan, Y., Ram, O., Desbordes, T., & Haussmann, E., 2022, Fast inference from transformers via speculative decoding. Available at: https://arxiv.org/abs/2211.17192.

64. Li, Y., Humphreys, P., Sun, T., Carr, A., Cass, S., Hawkins, P., ... & Bortolussi, L., 2022, Competition-level code generation with AlphaCode. *Science*, 378(1092-1097). DOI: 10.1126/science.abq1158.

65. Romera-Paredes, B., Barekatain, M., Novikov, A., Novikov, A., Rashed, S., & Yang, J., 2023, Mathematical discoveries from program search with large language models. *Nature*. DOI: 10.1038/s41586-023-06924-6.

66. Wikipedia., 2024, Cap set. Available at: https://en.wikipedia.org/wiki/Cap_set.

67. Trinh, T. H., Wu, Y., & Le, Q. V. et al., 2024, Solving olympiad geometry without human demonstrations. *Nature,* 625, 476–482. DOI: 10.1038/s41586-023-06747-5.

68. Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013, Efficient Estimation of Word Representations in Vector Space. Available at: https://arxiv.org/pdf/1301.3781.

69. Shi, L., Ma, C., Liang, W., Ma, W., Vosoughi, S., 2024, Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. Available at: https://arxiv.org/abs/2406.07791

70. Pandit, B., 2024, What Is Mixture of Experts (MoE)? How It Works, Use Cases & More. Available at: https://www.datacamp.com/blog/mixture-of-experts-moe