Generative AI Workshop References

Articles, Tutorials, and Courses

- Course from DeepLearning,AI on Generative AI
    - https://www.deeplearning.ai/courses/generative-ai-with-llms/
- Nice explainer: https://towardsdatascience.com/attention-and-transformer-models-fe667f958378
- Learn More about RAG
    - https://medium.com/snowflake/langchain-and-streamlit-rag-c5f53af8f6ba
- https://d2l.ai Online, free, interactive Deep Learning textbook (Zhang et al., 2023)
- https://huggingface.co/blog/rlhf
- https://medium.com/the-modern-scientist/detailed-explanations-of-transformer-step-by-step-dc32d90b3a98
- Arctic Model Blog: https://www.snowflake.com/blog/arctic-open-and-efficient-foundation-language-models-snowflake
- Arctic LLM Demo App: https://arctic.streamlit.app/
- Explore Fine Tuning Models
    - https://msuryavanshi.medium.com/large-language-model-fine-tuning-techniques-df8975396989
    - https://www.kdnuggets.com/how-to-use-hugging-face-autotrain-to-finetune-llms
    - https://www.philschmid.de/dpo-align-llms-in-2024-with-trl
    - Training with PEFT: *Memory efficient RLHF training using adapters with PEFT*
    - https://huggingface.co/docs/peft/main/en/conceptual_guides/lora
    - https://huggingface.co/docs/trl/en/sft_trainer
    - https://huggingface.co/docs/trl/en/dpo_trainer
- https://huggingface.co/blog/hf-bitsandbytes-integration
- https://paperswithcode.com/method/multi-head-attention
- https://www.promptingguide.ai/techniques/cot

- Useful article: https://towardsdatascience.com/retrieval-augmented-generation-rag-from-theory-to-langchain-implementation-4e9bd5f6a4f2
- “Snowflake Cortex: Industry-leading AI Models And LLMs In Snowflake | Snowday 2023”
  - https://www.youtube.com/watch?v=-ZagrEDUnHQ
- Good source on agents: https://www.promptingguide.ai/research/llm-agents
- Semantic search in Streamlit Community Cloud - “Seeing through the Cloud”
- Build a RAG using Cortex and Snowflake: https://quickstarts.snowflake.com/guide/asking_questions_to_your_own_documents_with_snowflake_cortex/index.html#4
- DocDocGo Streamlit App
  - Dmitriy Vasilyuk's DocDocGo chatbot ingests content from websites and uploaded docs, + lets you navigate and edit with special commands. 🤯
  - https://github.com/reasonmethis/docdocgo-core/blob/main/README.md
  - From LinkedIn: https://www.linkedin.com/posts/streamlit_ai-rag-chatgpt-activity-7189352534971486208-woYY?utm_source=share&utm_medium=member_desktop

Models

- Gemma
  - https://ai.google.dev/gemma/docs/model_card
- Mistral 7B Instruct
  - https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
- Llama3
  - https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md
  - https://ai.meta.com/blog/meta-llama-3/
- Nomic Embed
  - https://huggingface.co/nomic-ai/nomic-embed-text-v1.5
- Arctic
  - https://huggingface.co/Snowflake/snowflake-arctic-instruct
  - https://github.com/Snowflake-Labs/snowflake-arctic
- Arctic Embed
  - https://github.com/Snowflake-Labs/arctic-embed

- Zephyr-7B
    - https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

Tools and Libraries

Streamlit

- Streamlit - https://streamlit.io/
- GenAI uses - https://streamlit.io/generative-ai
- App Gallery - https://streamlit.io/gallery
- Ask Streamlit docs - https://llamaindex-chat-with-docs.streamlit.app/
- Streamlit Playground (create and run online live) - https://create.streamlit.app/
- Streamlit Community Cloud (share.streamlit.io)

Snowflake

- Streamlit in Snowflake - https://www.snowflake.com/en/data-cloud/overview/streamlit-in-snowflake/
- Snowflake GenAI: https://www.snowflake.com/dca-thankyou/generative-ai-llm-school-thank-you/
- Snowflake Cortex LLM functions: blog docs
- Snowflake Notebooks - [EXTERNAL] Notebooks Private Preview

Python - https://www.python.org/

Miniconda - https://docs.anaconda.com/free/miniconda/

Chromadb - https://www.trychroma.com/

Jupyter - https://jupyter.org/

LangChain- http://langchain.com/

HuggingFace - https://huggingface.co/

Ollama - https://ollama.com/

DeepEval - docs.confident-ai.com/

TruLens - https://www.trulens.org/trulens_eval/getting_started/quickstarts/quickstart/

References

AI@Meta. (2024). *Llama 3 Model Card*. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., & Amodei, D. (2020). Language Models are Few-Shot Learners. *ArXiv, abs/2005.14165*.

Chang, Y., Wang, X., Wang, J., Wu, Y., Zhu, K., Chen, H., Yang, L., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P.S., Yang, Q., & Xie, X. (2023). A Survey on Evaluation of Large Language Models. *ArXiv, abs/2307.03109*.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *North American Chapter of the Association for Computational Linguistics*. https://arxiv.org/abs/1810.04805

Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L. (2022). LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale. *ArXiv, abs/2208.07339.*

Gemma Team, T. M., Hardin, C., Dadashi, R., Bhupatiraju, S., Sifre, L., Rivière, M., ... al., E. (2024). *Gemma*. doi:10.34740/KAGGLE/M/3301

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., Laroussilhe, Q.D., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-Efficient Transfer Learning for NLP. *ArXiv, abs/1902.00751*.

Hu, J.E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., & Chen, W. (2021). LoRA: Low-Rank Adaptation of Large Language Models. *ArXiv, abs/2106.09685*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. D. L., ... & Sayed, W. E. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large Language Models are Zero-Shot Reasoners. *ArXiv, abs/2205.11916*.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., & Rastogi, A. (2023). RLAIF: Scaling Reinforcement Learning from Human Feedback with AI Feedback. *ArXiv, abs/2309.00267*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. *Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1%2F2020.acl-main.703

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459-9474. https://arxiv.org/pdf/2005.11401.pdf

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692*. https://arxiv.org/abs/1907.11692

Radford, A., & Narasimhan, K. (2018). Improving Language Understanding by Generative Pre-Training. https://openai.com/index/language-unsupervised/

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C.D., & Finn, C. (2023). Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *ArXiv, abs/2305.18290*.

Raffel, C., Shazeer, N.M., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P.J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *J. Mach. Learn. Res., 21*, 140:1-140:67. https://arxiv.org/abs/1910.10683

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., … Wolf, T. (2023). Zephyr: Direct Distillation of LM Alignment. *arXiv [Cs.LG]*. Retrieved from http://arxiv.org/abs/2310.16944

Vaswani, A., Shazeer, N.M., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. *Neural Information Processing Systems*. https://arxiv.org/abs/1706.03762

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Chi, E.H., Xia, F., Le, Q., & Zhou, D. (2022). Chain of Thought Prompting Elicits Reasoning in Large Language Models. *ArXiv, abs/2201.11903*.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J.G., Salakhutdinov, R., & Le, Q.V. (2019). XLNet: Generalized Autoregressive Pretraining for Language Understanding. *Neural Information Processing Systems*. https://arxiv.org/abs/1906.08237

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. *arXiv preprint arXiv:2210.03629*.

Zhang, Z., Zhang, A., Li, M., & Smola, A.J. (2022). Automatic Chain of Thought Prompting in Large Language Models. *ArXiv, abs/2210.03493*.

https://d2l.ai - Zhang, A., Lipton, Z. C., Li, M., & Smola, A. J. (2023, December 7). Dive into Deep Learning. Cambridge University Press.