A Review of Transformer Models

Preprint · September 2023

DOI: 10.48366/r640001

Jennifer D'Souza
Leibniz Information Centre for Science and Technology University Library

A Review of Transformer Models

Read the full and interactive version of this article on the ORKG website: https://orkg.org/review/R609546

Artificial Intelligence   Jennifer D’Souza

Introduction

Large language models (LLMs) have emerged as a transformative force in the field of natural language processing (NLP). These models, powered by deep learning techniques, have demonstrated remarkable capabilities in understanding and generating human-like text. The development of LLMs gained momentum with the advent of GPT (Generative Pre-trained Transformer) introduced by OpenAI, with its pioneering model GPT-3 being one of the most well-known and influential language models to date. With billions of parameters and extensive pre-training on vast amounts of text data from the internet, ChatGPT, based on the GPT-3 architecture, has become a leading example of LLMs. It showcases extraordinary language comprehension, fluency, and contextual understanding, enabling it to excel across a wide range of NLP tasks. As these models continue to evolve and grow, they hold tremendous potential to revolutionize numerous applications in human-computer interaction, content generation, and information retrieval (Dale, 2021), (Radford, 2019).

The development of LLMs which have revolutionized NLP are directly facilitated by the introduction of the Transformer architecture in "Attention Is All You Need" by (Vaswani, 2017) on which they are based. Before Transformers, recurrent neural networks (RNNs) and variants like LSTMs dominated sequential data processing tasks, but struggled with long-range dependencies and parallelization. The Transformer's attention mechanism enables it to efficiently capture long-range dependencies, making it highly effective for NLP tasks. Notably, LLMs have been developed both by major research institutions and commercial organizations. On the one hand, models like GPT-3, with over 175 billion parameters, developed by OpenAI, are representative of the most well-known commercial LLMs, demonstrating human-like language understanding and generation capabilities. On the other hand, the research community has contributed several open-source large language models, such as T5 (Text-to-Text Transformer) by Google (Raffel, 2020), and LLaMA (Large Language Model Meta AI) by Meta AI (Touvron, 2023). These models have also made significant impacts and advancements in various NLP applications, ranging from sentiment analysis to machine translation. The pre-training of large language models on vast amounts of internet text data is a common practice, followed by fine-tuning on specific tasks with labeled data to adapt their knowledge for targeted applications. Consequently, both paywalled and open-source large language models have played crucial roles in advancing the state of NLP, and they continue to pave the way for exciting developments in human-machine interaction and natural language understanding.

This article reviews various LLM model series and concludes with a comprehensive comparison of all models released to date.

Google's T5

Google's T5 was one of the earliest models that implemented the text-to-text as a sequence-to-sequence generation objective where a where the model is fed some text for context or conditioning on a range of diverse NLP tasks and is then asked to produce some output text. To specify which task the model should perform, they added a task-specific (text) prefix to the original input sequence before feeding it to the model. Consequently, it was found that a huge variety of NLP tasks can be cast in this format, including translation, summarization, and even classification and regression tasks. As an example, to ask T5 to translate the sentence “That is good.” from English to German, the model would be fed the sequence “translate English to German: That is good.” and would be trained to output “Das ist gut.” For text classification tasks, the model simply predicts a single word corresponding to the target label. For example, on the MNLI benchmark (Williams, 2018) the goal is to predict whether a premise implies (“entailment”), contradicts (“contradiction”), or neither (“neutral”) a hypothesis. With our preprocessing, the input sequence becomes “mnli premise: I hate pigeons. hypothesis: My feelings towards pigeons are filled with animosity.” with the corresponding target word “entailment”. Their subsequent work FLAN advocated for a zero-shot task learning approach via instruction tuning. The idea is that by using supervision to teach an LM to perform tasks described via instructions, the LM will learn to follow instructions and do so even for unseen tasks. Distinguishing between T5 and FLAN prompts, T5 prompts were mostly just a tag for the dataset, which would not work in the zero-shot setting. In contrast, the prompts that are used for FLAN are similar to what would be used to ask a human to perform the task. As such just as the dessert that follows the end of meals, the chosen "FLAN" strategy could be applied over any pretrained language models. Specifically Google applied FLAN to their pretrained LAMDA dialog model and their pretrained T5.

Comparison available via https://orkg.org/comparison/R605894/

EleutherAI's GPT-J, GPT-NeoX, and Pythia series

With the GPT-NeoX-20B model, EleutherAI sought to offer a strong performing open-source alternative mainly to GPT-3, which in its empirical evaluations proved an effective model--it was fairly close to GPT-3 performance. At the time of its release i.e. April 2022, it was the largest open-source model available. The earlier released GPT-J-6B model had a similar objective and outperformed GPT-3 Babbage and underperformed GPT-3 175B model by 10 points in the lowest results, still proving a strong open-source contender.

The Pythia series from EleutherAI mainly focused on ground research on LLMs to answer a central research question: "How do large language models (LLMs) develop and evolve over the course of training? How do these patterns change as models scale?" It is well established that there are regular and predictable patterns in the behavior of trained language models as they scale (Kaplan, 2020) (Henighan, 2020) (Ghorbani, 2021) (Pu, 2021) (Mikami, 2021) (Hernandez, 2021). But prior work connecting these “Scaling Laws” to the learning dynamics of language models is minimal. One of the driving reasons for this gap in research is a lack of access to appropriate model suites to test theories. This is addressed by EleutherAI in their release of the Pythia-suite of 8 main models at parameter sizes of 70M, 160M, 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B, which are respectively released in two variants of the underlying Pile training dataset, thus resulting in 16 total models. Furthermore, public access to 154 checkpoints for each one of the 16 models, alongside tools to download and reconstruct their exact training dataloaders is provided for further study.

Comparison available via https://orkg.org/comparison/R609226/

MosaicML's Mosaic Pretrained Transformer (MPT)

MosaicML based on San Francisco, California, is a leading generative AI platform that empowers enterprises to build their own AI. As of July 19, 2023, Databricks, the Data and AI company, announced it has completed its acquisition of MosaicML. MosaicML is widely known for its state-of-the-art MPT large language models (LLMs).

Its first model released in early May 2023 was MPT-7B a transformer trained from scratch on 1T tokens of text and code. It is open source, available for commercial use, and matches the quality of LLaMA-7B. MPT-7B was trained on the MosaicML platform in 9.5 days with zero human intervention at a cost of ~$200k. The base model is accompanied with three different finetuned models for practical application purposes: 1) an instruction-tuned model called MPT-7B-Instruct (licensed for commercial use), 2) a non-commercially licensed MPT-7B-Chat as a conversational version of MPT-7B. MPT-7B-Chat has been finetuned using ShareGPT-Vicuna, HC3, Alpaca, Helpful and Harmless, and Evol-Instruct, ensuring that it is well-equipped for a wide array of conversational tasks and applications. While MPT-7B-Instruct focuses on delivering a more natural and intuitive interface for instruction-following, MPT-7B-Chat aims to provide seamless, engaging multi-turn interactions for users. And finally, 3) they also announced MPT-7B-StoryWriter-65k+—“a model designed to read and write stories with super long context lengths”—with a previously unheard of 65,000 token context length. MPT-7B matches the quality of LLaMA-1-7B and outperforms other open source 7B - 20B models such as Pythia on standard academic tasks. The FALCON, Llama-2, or OpenLAMA models were not part of MPT-7B evaluations.

Subsequently in June, MosaicML expanded the MPT series with the bigger brother of the 7B model as MPT-30B -- a new, completely open-source model licensed for commercial use. This model is significantly more powerful than 7B and outperforms GPT-3 on many benchmarks. This model has been released in 2 fine-tuned variants too: MPT-30B-Instruct and MPT-30B-Chat licensed for non-commercial use only. All models in this series demonstrate strong abilities to generate code. MPT-30B models outperform TII's Falcon-40B.

Both MPT-7B and 30B are trained on a large amount of data (1T tokens like LLaMA vs. 300B for Pythia, and 800B for StableLM).

With its models, MosiacML aims to build on and furthermore set a strong precedent to the development of open-source LLMs pioneered in afore-mentioned efforts such such as the LLaMA series from Meta, the Pythia series from EleutherAI, the Falcon model from Technology Innovation Institute (TII) and the OpenLLaMA model from Berkeley AI Research.

Comparison available via https://orkg.org/comparison/R604254/

DeepMind's compute-optimal Chinchilla and other models

DeepMind In a seminal paper called "Training Compute-Optimal Large Language Models" published in 2022, carried out a detailed study of the performance of language models of various sizes and quantities of training data. The goal was to find the optimal number of parameters and volume of training data for a given compute budget. This paper is often referred to as the Chinchilla paper. Some of the findings of the paper: 1) it hints that many of the 100B parameter LLMs like GPT-3 may actually be over parameterized, meaning they have more parameters than they need to achieve a good understanding of language and under trained so that they would benefit from seeing more training data. The authors hypothesized that smaller models may be able to achieve the same performance as much larger ones if they are trained on larger datasets.

One important takeaway from the Chinchilla paper is that the optimal training dataset size for a given model is about 20 times larger than the number of parameters in the model. Chinchilla was determined to be compute optimal. For a 70 billion parameter model, the ideal training dataset contains 1.4 trillion tokens or 20 times the number of parameters. Per this, 175B parameter models like GPT-3, OPT, and BLOOM were trained on datasets that are smaller than the Chinchilla optimal size. These models may actually be under trained. In contrast, LLaMA 65B was trained on a dataset size of 1.4 trillion tokens, which is close to the Chinchilla recommended number. Another important result from the paper is that the compute optimal Chinchilla model outperforms non compute optimal models such as GPT-3 on a large range of downstream evaluation tasks. With the results of the Chinchilla paper in hand teams have recently started to develop smaller models that achieved similar, if not better results than larger models that were trained in a non-optimal way. Moving forward, you can probably expect to see a deviation from the bigger is always better trends of the last few years as more teams or developers like you start to optimize their model design.

This comparison characteristically describes Chinchilla as well as other models from DeepMind on certain salient properties of LLMs.

Comparison available via https://orkg.org/comparison/R609271/

MetaAI's LLaMA

In the spirit of building LLMs, Meta AI first introduced LLaMA-1 which was released under a non-commercial license. In subsequent work, Meta AI introduced open-sourced Llama 2 and Llama 2-Chat. Crucially, both these models support community development, marking a significant stride towards fostering transparency and promoting the development of more responsible, replicable LLMs. Llama 2 models are trained on 2 trillion tokens and have double the context length of Llama 1. Llama-2-chat models have additionally been trained on over 1 million new human annotations. Furthermore, compared to Llama 1, in Llama 2 they performed more robust data cleaning, updated the data mixes, trained on 40% more total tokens, doubled the context length, as well as leveraged grouped-query attention (GQA) for inference scalability improving. Evaluation comparisons of open-source models, including Llama 1, Llama 2 base models, MPT (MosaicML), and Falcon on standard academic benchmarks for Commonsense Reasoning, World Knowledge, Reading Comprehension, Math, MMLU, BBH, and AGI Eval indicates that Llama 2 outperforms other models including Llama 1. The following comparison showcases Llama 1 versus 2 on their essential characteristics.

Comparison available via https://orkg.org/comparison/R604183/

Llama-1, Alpaca, and Vicuna

With major industry players introducing open-source models, the academic research world was not meant to be far left behind. Before going into further details, a general difference between academic and industry players is that, more often than not, academics have access to far lesser compute facilities compared to their industry counterparts.

In the spirit of fostering academic research on LLMs, a group of researchers at Stanford university released the first finetuned LLM of its kind called Alpaca. The identify two important parts to creating an instruction following model in the spirit of OpenAI's ChatGPT. They are: 1) a strong baseline pretrained model to finetune. They select Llama-1-7B as their deemed most effective at the time. 2) A highly optimized and effective finetuning dataset with instruction following demonstrations. For this, they generated 52K instruction-following demonstrations by building upon the self-instruct method generated from OpenAI’s text-davinci-003. Thus they created Alpaca as a language model fine-tuned using supervised learning from a LLaMA 7B model on 52K instruction-following demonstrations generated from OpenAI’s text-davinci-003.

Inspired from Llama-1 and Alpaca, a team of students from UC Berkeley in collaboration with UC San Diego and CMU released the Vicuna chat model at a small price of 300 dollars. Still based on the Llama models in several parameter sizes, Vicuna was finetuned on user conversations from the ShareGPT platform. The team collected 70K conversations from ShareGPT.com, a website where users can share their ChatGPT conversations. The team also made several improvements to the training recipe, including memory optimizations to enable Vicuna's understanding of long context. They also adjusted the training loss to account for multi-round conversations and compute the fine-tuning loss solely on the chatbot's output. Thus, after fine-tuning Vicuna with 70K user-shared ChatGPT conversations, it was discovered that Vicuna was capable of generating more detailed and well-structured answers compared to Alpaca, with the quality on par with ChatGPT.

In a nutshell, the Meta's Llama models are proving very effective in influencing academic research on optimizing LLMs via finetuning.

Comparison available via https://orkg.org/comparison/R604319/

Berkeley AI Research's OpenLLaMA

The gap between commercial and non-commercial LLMs is rapidly narrowing. This is owing to open-sourced LLM development research efforts from Google with the T5 series, Meta with the LLaMA series, and now OpenLLaMA developed by researchers Xinyang Geng and Hao Liu from Berkeley AI Research.

With Google and Meta's models discussed above, this section presents the OpenLLaMA series engineered as a fully open-source model based on LLaMA-1. OpenLLaMA exhibits comparable performance to the original LLaMA and GPT-J across a majority of tasks, and outperforms them in some tasks.

Comparison available via https://orkg.org/comparison/R605927/

OpenAI's GPT

OpenAI's GPT (Generative Pre-trained Transformer) models have undergone remarkable advancements, evolving from GPT-1's foundational contextual text generation to the groundbreaking capabilities of GPT-4. Beginning with GPT-1, these models harnessed the power of the transformer architecture and large-scale pre-training on diverse text sources to generate coherent and contextually relevant human-like text. GPT-2 pushed boundaries by demonstrating more coherent long-form text generation, albeit raising concerns about misuse. GPT-3 further amplified this prowess, showcasing its capacity for a wide array of tasks and its uncanny ability to understand and produce human-like language. GPT-4 represents a culmination of these developments, featuring handling multimodal context, even more refined context comprehension, nuanced responses, and also enhanced handling harm, safety, and helpfulness criteria for LLMs.

Comparison available via https://orkg.org/comparison/R606160/

Microsoft's Wizard LLM series

WizardLM is a LLM based on LLaMA trained using a new method, called Evol-Instruct, on complex instruction data. By using AI to "evolve" instructions, WizardLM outperforms similar LLaMA-based LLMs trained on simpler instruction data. On the 6th of July, 2023, WizardLM V1.1 was released with significantly improved performance. Apart from a generalist LLM, WizardLM is also released as coding (WizardCoder) and math solver (WizardMath) variants able to handle more complex scenarios given their finetuning complex instructions setups than prior released specialized models.

Comparison available via https://orkg.org/comparison/R609288/

A Comprehensive Catalog of Large Language Models (LLMs)

This last section brings together 87 different transformer-based language models in a comprehensive comparison of their characteristic properties.

Comparison available via https://orkg.org/comparison/R609337/

Acknowledgements

This review article was created using Open Research Knowledge Graph and the SmartReview methodology

References

- *A General Language Assistant as a Laboratory for Alignment*. (n.d.).
- *A Generalist Agent*. (n.d.).
- *ALBERT: A Lite BERT for Self-supervised Learning of Language Representations*. (n.d.).
- *Alexatm 20b: Few-shot learning using a large-scale multilingual seq2seq model*. (n.d.).
- *Alpaca: A strong, replicable instruction-following model*. (n.d.).
- *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. (n.d.).
- *BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension*. (n.d.).
- *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. (n.d.).
- *Big Bird: Transformers for Longer Sequences*. (n.d.).
- *BlenderBot 3: a deployed conversational agent that continually learns to responsibly engage*. (n.d.).
- *BLOOM: A 176B-Parameter Open-Access Multilingual Language Model*. (n.d.).
- *CM3: A Causal Masked Multimodal Model of the Internet*. (n.d.).
- *CTRL: A Conditional Transformer Language Model for Controllable Generation*. (n.d.).
- Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering*, *27*(1), 113–118.
- *Deberta: Decoding-enhanced bert with disentangled attention*. (n.d.).
- *Decision Transformer: Reinforcement Learning via Sequence Modeling*. (n.d.).
- *DialoGPT: Large-Scale Generative Pre-training for Conversational Response Generation*. (n.d.).
- *DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter*. (n.d.).
- *DQ-BART: Efficient Sequence-to-Sequence Model via Joint Distillation and Quantization*. (n.d.).
- *ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators*. (n.d.).
- *ERNIE: Enhanced Language Representation with Informative Entities*. (n.d.).
- *Exploring the limits of transfer learning with a unified text-to-text transformer*. (n.d.).
- *Falcon-40B: an open large language model with state-of-the-art performance*. (n.d.).
- *Finetuned language models are zero-shot learners*. (n.d.).

- *Flamingo: a Visual Language Model for Few-Shot Learning*. (n.d.).
- *Galactica: A large language model for science*. (n.d.).
- Ghorbani, B., Firat, O., Freitag, M., Bapna, A., Krikun, M., Garcia, X., Chelba, C., & Cherry, C. (2021). Scaling laws for neural machine translation. *arXiv Preprint arXiv:2109.07740*.
- *GLaM: Efficient Scaling of Language Models with Mixture-of-Experts*. (n.d.).
- *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models*. (n.d.).
- *GLM: General language model pretraining with autoregressive blank infilling*. (n.d.).
- *Global Context Vision Transformers*. (n.d.).
- *Godel: Large-scale pre-training for goal-directed dialog*. (n.d.).
- *GPT-4 Technical Report*. (n.d.).
- *GPT-J-6B: A 6 billion parameter autoregressive language model*. (n.d.).
- *GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow*. (n.d.).
- *GPT-NeoX-20B: An Open-Source Autoregressive Language Model*. (n.d.).
- Henighan, T., Kaplan, J., Katz, M., Chen, M., Hesse, C., Jackson, J., Jun, H., Brown, T. B., Dhariwal, P., Gray, S., & others. (2020). Scaling laws for autoregressive generative modeling. *arXiv Preprint arXiv:2010.14701*.
- Hernandez, D., Kaplan, J., Henighan, T., & McCandlish, S. (2021). Scaling laws for transfer. *arXiv Preprint arXiv:2102.01293*.
- *Hierarchical Text-Conditional Image Generation with CLIP Latents*. (n.d.).
- *Highly accurate protein structure prediction with AlphaFold*. (n.d.).
- *High-Resolution Image Synthesis with Latent Diffusion Models*. (n.d.).
- *HTLM: Hyper-Text Pre-Training and Prompting of Language Models*. (n.d.).
- *Improving alignment of dialogue agents via targeted human judgements*. (n.d.).
- *Improving Language Understanding by Generative Pre-Training*. (n.d.).
- *Introducing ChatGPT*. (n.d.).
- *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs*. (n.d.).
- *Introducing MPT-30B: Raising the bar for open-source foundation models*. (n.d.).
- *Jais and Jais-chat: Arabic-Centric Foundation and Instruction-Tuned Open Generative Large Language Models*. (n.d.).
- *Jurassic-1: Technical details and evaluation*. (n.d.).
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv Preprint arXiv:2001.08361*.
- *LaMDA: Language Models for Dialog Applications*. (n.d.).
- *Language Models are Few-Shot Learners*. (n.d.).
- *Language models are unsupervised multitask learners*. (n.d.).
- *Language Models that Seek for Knowledge: Modular Search & Generation for Dialogue and Prompt Completion*. (n.d.).
- *Learning Transferable Visual Models From Natural Language Supervision*. (n.d.).
- *Llama 2: Open Foundation and Fine-Tuned Chat Models*. (n.d.).
- *LLaMA: Open and Efficient Foundation Language Models*. (n.d.).
- *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. (n.d.).
- Mikami, H., Fukumizu, K., Murai, S., Suzuki, S., Kikuchi, Y., Suzuki, T., Maeda, S., & Hayashi, K.

(2021). A scaling law for synthetic-to-real transfer: How much is your pre-training effective? *arXiv Preprint arXiv:2108.11018*.

- *Multilingual Denoising Pre-training for Neural Machine Translation*. (n.d.).
- *Multitask prompted training enables zero-shot task generalization*. (n.d.).
- *Offline Reinforcement Learning as One Big Sequence Modeling Problem*. (n.d.).
- *One Embedder, Any Task: Instruction-Finetuned Text Embeddings*. (n.d.).
- *OpenLLaMA: An Open Reproduction of LLaMA*. (n.d.).
- *OPT: Open Pre-trained Transformer Language Models*. (n.d.).
- *Orca: Progressive learning from complex explanation traces of gpt-4*. (n.d.).
- *PaLM: Scaling Language Modeling with Pathways*. (n.d.).
- *PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization*. (n.d.).
- *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*. (n.d.).
- Pu, J., Yang, Y., Li, R., Elibol, O., & Droppo, J. (2021). *Scaling effect of self-supervised speech models*.
- *Pythia: A suite for analyzing large language models across training and scaling*. (n.d.).
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., & others. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8), 9.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, *21*(1), 5485–5551.
- *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. (n.d.).
- *Scaling instruction-finetuned language models*. (n.d.).
- *Scaling Language Models: Methods, Analysis &amp; Insights from Training Gopher*. (n.d.).
- *Solving Quantitative Reasoning Problems with Language Models*. (n.d.).
- *StarCoder: may the source be with you!* (n.d.).
- *Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows*. (n.d.).
- *Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity*. (n.d.).
- *Teaching language models to support answers with verified quotes*. (n.d.).
- *Text Embeddings by Weakly-Supervised Contrastive Pre-training*. (n.d.).
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., & others. (2023). Llama: Open and efficient foundation language models. *arXiv Preprint arXiv:2302.13971*.
- *Training Compute-Optimal Large Language Models*. (n.d.).
- *Training language models to follow instructions with human feedback*. (n.d.).
- *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. (n.d.).
- *Ul2: Unifying language learning paradigms*. (n.d.).
- *Unsupervised Cross-lingual Representation Learning at Scale*. (n.d.).
- *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*. (n.d.).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.
- *Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality*. (n.d.).
- Williams, A., Nangia, N., & Bowman, S. (2018). A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 1112–1122.

- *WizardCoder: Empowering Code Large Language Models with Evol-Instruct*. (n.d.).
- *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. (n.d.).
- *WizardMath: Empowering Mathematical Reasoning for Large Language Models via Reinforced Evol-Instruct*. (n.d.).
- *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. (n.d.).
- *Zero-Shot Text-to-Image Generation*. (n.d.).