A Survey on Technologies and Developments of Large Language Models

Siqi Wang
Department of Electronic Engineering
Shanghai Jiao Tong University
Shanghai, China

***Abstract*—Language modeling has been extensively explored for language understanding and generation over recent decades, evolving from early statistical language models to neural language models. More recently, pre-training language models (PLMs) have been proposed by pre-training Transformer models on large-scale corpora, exhibiting strong performance on various natural language processing tasks. Researchers discovered model scaling effects, where increasing model size leads to improved capabilities, motivating the study of scaling PLMs to even larger sizes. To differentiate based on scale, the term large language models (LLMs) was coined referring to the PLMs of significant size. Both academia and industry have rapidly advanced LLM research, with the advent of ChatGPT generating widespread interest. Given these swift technical advances, this survey reviews the recent progress on LLMs, covering key development stages and mainstream techniques. Additionally, it summarizes major LLM applications and discusses open challenges for future work. This review article aims to provide a fundamental understanding of LLMs for newcomers to the field.**

***Index Terms*—Large Language Models, LLMs, ChatGPT, Adaptation Tuning**

I. Introduction

Language, a fundamental medium of human interaction, is the cornerstone of communication and self-expression. There is an increasing demand for machines to perform complex language tasks like translation, summarization, information retrieval, and conversational interactions. This growing need is driving the development of more generalized language models that can help equip machines with the AI required to understand and communicate in natural language [1]. The advances in language models can be attributed to key technological innovations like more powerful transformer architectures [2], growth in computational resources for model training, and greater access to large datasets for models to learn from.

These developments have brought about a revolutionary transformation by enabling the creation of Large Language Models (LLMs). LLMs have emerged as cutting-edge AI systems that can process and generate text with coherent communication, while also generalizing to perform well on multiple different tasks [3], [4]. The creation of these powerful models that can understand and produce natural language has achieved near-human level [5], [6]. LLMs display some surprising emergent abilities that are key to the performance of language models on complex tasks, making AI algorithms unprecedentedly powerful and effective.

LLMs are having a profound impact on artificial intelligence research and development. Recent models like ChatGPT and GPT-4 have catalyzed renewed excitement about the possibility of artificial general intelligence. The rapid advances in LLMs are transforming many subfields of AI and may enable a host of real-world applications. The progress in LLMs is blurring the traditional distinction between research and engineering. The development of LLMs now requires extensive engineering experience with large-scale data processing and distributed parallel training systems. LLM researchers must tackle complex systems engineering challenges, either by partnering closely with engineers or by taking on engineering responsibilities themselves [7]. The fast pace of innovation in LLMs signals a new era in AI requiring tight integration of research insights and practical engineering solutions.

Despite the rapid development and profound impact, the underlying principles behind LLMs' capabilities remain poorly understood [7]. First, it is mysterious why advanced abilities emerge in LLMs. More broadly, there has been a lack of rigorous investigation into the key factors driving LLMs' superior performance. Second, the research community faces difficulties in training cutting-edge LLMs themselves, due to the enormous computational resources required. The high costs make it infeasible to run repetitive, ablation studies to fully explore different training strategies. As a result, most capable LLMs are developed within the industry, with many critical training details not revealed publicly. Third, aligning LLMs with human values and preferences remains an open challenge. Although LLMs demonstrate impressive capacities, they can still generate toxic, false, or harmful content. Developing effective and efficient techniques to control LLMs and mitigate risks is required [8].

Given the opportunities and challenges posed by LLMs, a greater focus on LLM technology and development is needed. To provide a fundamental understanding of LLMs, this survey conducts a literature review of the latest progress made in LLMs. The rest of this survey is structured as follows. Section II summarizes four key development stages of language modeling. Section III provides the relevant background and key techniques to understand the fundamentals related to LLMs. Section IV reviews the applications of LLMs in several representative fields. Section V discusses the safety risks and challenges of LLMs, along with potential mitigation strategies. Finally, Section VI summarizes the major findings and discusses the remaining issues for future work.

II. Development Stages

Language modeling (LM) is one of the major approaches to improving the language intelligence of machines. LM aims to model the generative likelihood of word sequences, enabling the prediction of future or missing tokens. LM research has been an area of significant focus in the literature and can be categorized into four major development stages:

A. Statistical language models (SLM)

Based on statistical learning methods, SLMs rose in the 1990s. SLM amounts to estimating the probability distribution of various linguistic units, such as words, sentences, and whole documents [9]. The main tasks of SLMs are determining the structure of a statistical model and determining the free parameters of the model [10]. The dominant models are n-gram language models, *e.g.*, bigram and trigram language models, which have a fixed context length n.

SLMs have been widely used to improve performance on information retrieval and natural language processing tasks [11]–[13]. However, accurately estimating higher-order language models is challenging and the goal of statistical language modeling, to learn the joint probability function of sequences of words, is intrinsically difficult. This is because of the curse of dimensionality [14].

B. Neural language models (NLM)

NLMs model the probability of word sequences using neural networks, *e.g.*, multi-layer perceptron (MLP) and recurrent neural networks (RNNs) [15], [16].

Compared with SLMs which can only assist in specific tasks, NLMs can solve typical NLP tasks. The work in [14] introduced the concept of distributed representation of words and built the word prediction model conditioned on aggregated context vector representations. This idea of learning effective features for text data was extended to develop end-to-end neural network frameworks for diverse NLP tasks [17].

Additionally, word2vec [18], [19] proposed a simplified shallow network for learning distributed word representations that proved highly effective across many NLP tasks. These studies initiated the use of language models for representation learning beyond word sequence modeling, greatly impacting the field of NLP.

C. Pre-trained language models (PLM)

While conventional language modeling (LM) trains task-specific models in supervised settings, PLMs are trained in a self-supervised setting on a large corpus of text to learn generic representation shareable among various NLP tasks.

As a former approach, ELMo [20] proposed pre-training a bidirectional LSTM (biLSTM) network to learn context-aware word representations, before fine-tuning the network on downstream tasks. Different from ELMo using a shallow concatenation of independently trained left-to-right and right-to-left LMs, BERT [21] uses masked language models to enable pre-trained deep bidirectional representations. Building on the highly parallelizable Transformer architecture [22], BERT proposed pre-training bidirectional language models with customized different pre-training tasks on large-scaled unlabeled data. The resulting context-aware word representations proved highly effective as general semantic features, largely improving NLP performance. This work inspired extensive follow-up research adopting the *pre-train then fine-tune* paradigm [7]. Since then, many studies on PLMs have developed different architectures like GPT-2 and BART [23] or enhanced pre-training strategies. This paradigm usually requires fine-tuning the models on target downstream tasks.

After fine-tuning on downstream tasks, PLMs outperform traditional LMs. The larger PLMs lead to more performance gains, driving the transitioning of PLMs to LLMs by significantly increasing model parameters (tens to hundreds of billions) [24] and training dataset (many GBs and TBs) [24]. This development has led to the proposal of many new LLMs.

D. Large language models (LLM)

Scaling up PLMs can improve model capacity for downstream tasks as described by scaling laws [25]. There are performance limits to training ever-larger models (*e.g.*, 175B-parameter GPT-3 and 540B-parameter PaLM [26]). Although scaling has largely involved increasing model size with similar architectures and pre-training tasks, these large PLMs exhibit different behaviors compared to smaller predecessors like 330M-parameter BERT and 1.5B-parameter GPT-2. The larger models demonstrate surprising capabilities, called emergent abilities [27], in solving complex tasks. For example, GPT-3 can perform few-shot learning through in-context examples while GPT-2 struggles to. Thus, these large PLMs are characterized as large language models (LLMs) to reflect their distinct performance. A remarkable LLM application is ChatGPT which adapts the GPT family for conversational dialogue and interacts impressively.

The ability of LLMs to solve varied tasks at human-level performance comes with drawbacks of slow training and inference, extensive hardware demands, and high running costs. These requirements have hindered adoption and motivated research into more efficient architectures [26] and training approaches [28]. Methods like parameter efficient tuning [29], pruning, quantization, knowledge distillation, and context length interpolation have been widely explored to enable more efficient use of LLMs.

III. Key Techniques

Large language models (LLMs) typically refer to Transformer-based language models with hundreds of billions or more of parameters, trained on very large quantities of textual data [30]. Examples include GPT-3 [3], PaLM [26], LLaMA [31], and GPT-4 [8]. LLMs demonstrate strong capabilities in natural language understanding and solving complex tasks through text generation, which have come a long way to reach their current state as general and capable learners. During this evolution, several important techniques were introduced that greatly enhanced LLMs' capabilities. This section outlines the key techniques behind LLMs as follows.

A. Tokenization

As with other natural language processing systems, LLMs rely on tokenization [32] as an essential preprocessing step before training on textual data. Tokenization involves parsing text into non-decomposing units called tokens, which can be characters, subwords [33], symbols [34], or words, depending on the LLM architecture.

B. Scaling

As mentioned in the previous section, Transformer language models exhibit clear scaling effects - larger model sizes, dataset sizes, and training compute resources generally improve model performance [25], [35]. As two representative models, GPT-3 and PaLM explored the scaling limits by increasing the model size to 175B and 540B parameters, respectively. Since compute budgets are often limited, scaling laws can optimize the allocation of compute resources. For instance, Chinchilla outperformed the larger Gopher model by training on more data with the same compute budgets [35]. Moreover, data scaling requires careful cleaning, as pre-training data quality is key for model capability.

C. Training

The enormous size of LLMs makes training them successfully very challenging, so it needs distributed training algorithms to learn the parameters of LLMs, typically combining various parallel strategies. Several optimization frameworks (*e.g.*, DeepSpeed [28] and Megatron-LM [36]) help implement and deploy parallel algorithms for distributed LLM training. Additionally, employing optimization tricks (*e.g.*, re-start to mitigate training loss spikes [26] and mixed-precision training [37]) can improve training stability and model performance. More recently, GPT-4 [8] proposed specialized infrastructure and optimization techniques to predict large model performance using much smaller models reliably.

D. Fine-Tuning

There are different styles to fine-tune an LLM. This part briefly discusses fine-tuning approaches.

1) Transfer Learning: The pre-trained LLMs demonstrate strong performance on various tasks [3], [26]. However, to further improve performance on a specific downstream task, the pre-trained models are fine-tuned using task-specific data [24]. This transfer learning approach adapts the models to new tasks.

There are two major approaches to adapting pre-trained LLMs: instruction tuning and alignment tuning. The former approach mainly aims to enhance (or unlock) the abilities of LLMs, while the latter approach aims to align the behaviors of LLMs with human values or preferences. The following introduces the two kinds of adaptation tuning in a bit more detail.

2) Instruction-tuning: To enable a model to effectively respond to user queries, the pre-trained model is fine-tuned on instruction-formatted data (i.e. instructions and input-output pairs). Instructions typically consist of multi-task data in plain natural language that guides the model to respond properly to a given prompt and input. This instruction fine-tuning enhances zero-shot generalization and downstream performance [38].

3) Alignment-tuning: LLMs are likely to generate false, biased, and harmful content for humans, so it is necessary to align LLMs with human values and make them helpful, honest, and harmless. Alignment involves asking LLMs to generate unexpected responses and then updating their parameters to avoid such outputs [39], [40], ensuring LLMs behave according to human intentions and values. Researchers employ reinforcement learning with human feedback (RLHF) [41] for model alignment, where a fine-tuned model is further trained with reward modeling (RM) and reinforcement learning (RL).

E. Prompting

Prompting is a method to querying trained LLMs to generate responses. LLMs can be prompted using different setups: they may adapt to instructions without fine-tuning and in other cases require fine-tuning on data with various prompt styles [38], [42]. The following discusses widely used prompt setups.

1) Zero-Shot Prompting: LLMs are zero-shot learners and able to answer queries never seen before. This prompting style requires LLMs to answer user questions without seeing any examples in the prompt.

2) In-context Learning: Here, multiple input-output demonstration pairs are shown to the model to generate the desired response. This adaptation style is also called few-shot learning. In-context learning (ICL) was formally introduced by GPT-3 [3]: with natural language instructions and/or task demonstrations, the language model can generate expected outputs for test instances by completing the word sequence of input text, without additional training or gradient update. Among GPT-series models, 175B parameter GPT-3 exhibited strong and general ICL ability, unlike smaller GPT-1 and GPT-2. The ICL ability also depends on the specific downstream task [7] - it can emerge on arithmetic tasks like 3-digit addition/subtraction for the 13B GPT-3, yet 175B GPT-3 cannot even work well on Persian QA [27].

3) Reasoning in LLMs: For small language models, it is usually difficult to solve complex tasks that involve multiple reasoning steps, while LLMs are zero-shot reasoners and can generate answers to reasoning tasks. Generating reasons is possible only by using different prompting styles, whereas to improve LLMs further on reasoning tasks many methods train them on reasoning datasets [38]. With the chain-of-thought (CoT) prompting strategy [43] (a special case of prompting where demonstrations contain reasoning information aggregated with inputs and outputs), LLMs can solve such tasks by utilizing the prompting mechanism that involves intermediate reasoning steps for deriving the final answer. CoT prompting does not positively impact performance for small models and only yields performance gains when the model size exceeds 100B. The models of smaller scale produced fluent but illogical chains of thought, leading to lower performance than standard prompting [43].

IV. Application

This section briefly reviews the recent progress on the applications of LLMs in four aspects, including the impact on research community and representative domains.

A. LLMs for Classic NLP Tasks

LLMs have catalyzed rapid progress in NLP by advancing performance on fundamental tasks. As pre-trained models like BERT originated from NLP research, their capabilities have significantly impacted the research fields of NLP. This part discusses LLMs' applications for classic NLP tasks in three examples: word-level, sentence-level, and text generation.

1) Word/Sentence-level Tasks: Word-level tasks (*e.g.*, word sense disambiguation [44] and word clustering [45]), along with sentence-level tasks (*e.g.*, sentence matching [46] and sentiment classification [47]), are long-standing tasks in NLP that have seen extensive research and real-world application. The key to solving these tasks lies in accurately understanding the semantic information of words or sentences. Rich high-quality labeled data has been accumulated for these tasks, allowing small language models to achieve strong performance by fine-tuning as shown by existing work [21].

Recent studies [3] have also tested LLMs on these classic tasks using in-context learning, showing that LLMs perform well. However, small models can mostly outperform LLMs using in-context learning on several classic tasks, because they can be specially optimized on these tasks to learn the specific task requirement and domain knowledge. For instance, full-data fine-tuned small models tend to outperform LLMs on semantic matching and sentiment analysis [48].

2) Text Generation: Text generation tasks (*e.g.*, machine translation and automatic summarization) are long-standing NLP tasks that have also been widely researched with many real-world deployments relying on fine-tuned small models. Since pre-training focuses on text prediction, LLMs display strong language generation abilities as commercial products and humans, given proper prompts. LLMs also offer flexibility to handle specialized requirements in real-world application scenarios and enable natural language interaction with users to further improve generation quality. However, LLMs are challenging to generate high-quality text in low-resource scenarios, *e.g.*, Marathi-to-English translation [49], due to imbalanced pre-training data.

B. Multi-modal LLMs

Multi-modal models refer to systems that can process and integrate multiple modes of input data (*e.g.*, text, images, and audio) to produce corresponding outputs in certain modalities. This part focuses specifically on extending LLMs to handle non-textual data, especially the visual modality. These models are called multi-modal large language models (MLLMs) [50]. For better discussion, we specify the input to be text-image pairs and the output to be text responses.

The key idea is adapting the information from other modalities to the text modality, in order to leverage the excellent model capacities of LLMs. A typical MLLM includes an image encoder for image encoding and a LLM for text generation, associated with a connection module to align vision and language representations. During generation, the image is split into patches and encoded into patch embeddings by the image encoder and connection module. This produces a visual representation digestible by the LLM. Then, the patch embeddings are concatenated with text embeddings and fed into the MLLM to generate the response autoregressively [7].

C. LLMs for Specific Domains

This part discusses the applications of LLMs in two representative domains: healthcare and finance.

Healthcare is a vital application field directly impacting human lives. Since ChatGPT's release, researchers have explored applying ChatGPT or other LLMs to various medical tasks including biology information extraction [51], medical advice consultation [52], mental health analysis [53], and report simplification [54]. As the major technical approach, researchers typically design specific prompts or instructions to guide LLMs in handling these diverse healthcare challenges. To further utilize LLMs' potential for healthcare, researchers propose domain-specific medical LLMs. To be specific, the Med-PaLM models [55], [56] achieve expert-level performance on the United States Medical Licensing Examination (USMLE) and gain increased physician approval for answering consumer medical questions. However, LLMs may fabricate misinformation [54], [57] by misinterpreting medical terms or suggesting advice inconsistent with medical guidelines. Another concern is the privacy risks of uploading patient data to commercial LLM servers.

Finance is a crucial domain where LLMs show promise for automating tasks like numerical claim detection [58], financial named entity recognition [59], and financial reasoning [60]. Despite general-purpose LLMs exhibiting competitive zero-shot performance on finance tasks, they still underperform domain-specific PLMs containing million-scale parameters [58]. To leverage the scaling benefits of LLMs, researchers collect large-scale finance corpora for continued pre-training (*e.g.*, BloombergGPT [61], XuanYuan 2.0 [62], FinGPT [63]). BloombergGPT demonstrates strong performance on diverse finance tasks while maintaining competitive proficiency on general tasks [61]. However, potential risks exist in applying LLMs to finance since inaccurate or harmful content could be generated, which will significantly impact markets [61]. Therefore, more rigorous reviewing and monitoring of the uses of LLMs is imperative to prevent adverse outcomes in the financial field.

V. Safety and Challenges

This section discusses ethical considerations related to LLMs about their safety and discusses potential mitigation strategies.

Despite their capabilities, LLMs face great safety issues in real-world use. Their probabilistic modeling nature makes them tend to generate hallucinations [64], meaning plausible but incorrect texts [8]. What is worse, malicious instructions could elicit harmful, biased, or toxic generation, resulting in the potential risks of misuse [3]. LLM safety risks like privacy leaks, overreliance, disinformation, and influence operations require careful handling.

To avert these risks, alignment methods (*e.g.*, RLHF) [40] leverage human feedback to develop well-aligned LLMs. However, RLHF heavily relies on high-quality human feedback data from professional labelers, which is costly and time-consuming to recruit qualified human annotators. Consequently, it is necessary to improve the RLHF framework to reduce the efforts of human labelers and seek a more efficient annotation approach with guaranteed data quality, *e.g.*, LLMs can be employed to assist the labeling work [7]. Furthermore, developing simplified optimization algorithms for alignment could reduce the training difficulty and instability of RLHF. Red teaming [65], [66] is another approach used to improve model safety, using the collected adversarial prompts to make LLMs more robust (*i.e.*, avoiding the attacks from red teaming). Additionally, privacy is a concern when fine-tuning LLMs on sensitive data, for which federated learning [67] helps avoid centralized data in privacy-restricted scenarios.

VI. Conclusion

This survey has reviewed the major developments of LLMs. It contributes to summarizing significant findings in the existing literature and provides a detailed analysis of LLMs, including the evolution of language modeling research through four key stages, the core techniques powering LLMs, and diverse applications of LLMs. It also reveals that continued advances in model scale, architecture design, training techniques, and prompting strategies have progressively enhanced LLMs' language understanding and generation capacities. Moreover, there remain open challenges around training extreme-scale models efficiently, aligning LLMs to human preferences, and monitoring potential risks during deployment.

Acknowledgment

The author would like to express sincere gratitude to Professor Song and the teaching assistants of the Multimedia Communication Systems and Applications course for their tremendous help and support. Their willingness to explain complex concepts and provide assistance on assignments has been invaluable. I am deeply thankful for the knowledge and skills I have developed under their instruction. It has been a privilege to take this course and engage with such excellent instructors.

References

[1] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *arXiv preprint arXiv:2307.06435*, 2023.

[2] A. Chernyavskiy, D. Ilvovsky, and P. Nakov, "Transformers: "the end of history" for natural language processing?," in *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21*, pp. 677–693, Springer, 2021.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[4] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa, "Large language models are zero-shot reasoners," *Advances in neural information processing systems*, vol. 35, pp. 22199–22213, 2022.

[5] A. Wang, Y. Pruksachatkun, N. Nangia, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman, "Superglue: A stickier benchmark for general-purpose language understanding systems," *Advances in neural information processing systems*, vol. 32, 2019.

[6] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," *arXiv preprint arXiv:2309.05519*, 2023.

[7] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[8] OpenAI, "Gpt-4 technical report," 2023.

[9] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?," *Proceedings of the IEEE*, vol. 88, no. 8, pp. 1270–1278, 2000.

[10] J. Gao and C.-Y. Lin, "Introduction to the special issue on statistical language modeling," 2004.

[11] C. Zhai *et al.*, "Statistical language models for information retrieval a critical review," *Foundations and Trends® in Information Retrieval*, vol. 2, no. 3, pp. 137–213, 2008.

[12] X. Liu and W. B. Croft, "Statistical language modeling for information retrieval.," *Annu. Rev. Inf. Sci. Technol.*, vol. 39, no. 1, pp. 1–31, 2005.

[13] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the eighth international conference on Information and knowledge management*, pp. 316–321, 1999.

[14] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in neural information processing systems*, vol. 13, 2000.

[15] T. Mikolov, M. Karafiát, L. Burget, J. Cernockỳ, and S. Khudanpur, "Recurrent neural network based language model.," in *Interspeech*, vol. 2, pp. 1045–1048, Makuhari, 2010.

[16] S. Kombrink, T. Mikolov, M. Karafiát, and L. Burget, "Recurrent neural network based language modeling in meeting recognition.," in *Interspeech*, vol. 11, pp. 2877–2880, 2011.

[17] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[18] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[19] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[20] 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)* (M. Walker, H. Ji, and A. Stent, eds.), (New Orleans, Louisiana), pp. 2227–2237, Association for Computational Linguistics, June 2018.

[21] J. Devlin, M.-W. 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, Volume 1 (Long and Short Papers)* (J. Burstein, C. Doran, and T. Solorio, eds.), (Minneapolis, Minnesota), pp. 4171–4186, Association for Computational Linguistics, June 2019.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.

[23] 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* (D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, eds.), (Online), pp. 7871–7880, Association for Computational Linguistics, July 2020.

[24] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *The Journal of Machine Learning Research*, vol. 21, no. 1, pp. 5485–5551, 2020.

[25] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.

[26] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann, *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[27] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, *et al.*, "Emergent abilities of large language models," *arXiv preprint arXiv:2206.07682*, 2022.

[28] J. Rasley, S. Rajbhandari, O. Ruwase, and Y. He, "Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, 2020.

[29] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv preprint arXiv:2104.08691*, 2021.

[30] M. Shanahan, "Talking about large language models," *arXiv preprint arXiv:2212.03551*, 2022.

[31] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[32] J. J. Webster and C. Kit, "Tokenization as the initial phase in nlp," in *COLING 1992 volume 4: The 14th international conference on computational linguistics*, 1992.

[33] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (I. Gurevych and Y. Miyao, eds.), (Melbourne, Australia), pp. 66–75, Association for Computational Linguistics, July 2018.

[34] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (K. Erk and N. A. Smith, eds.), (Berlin, Germany), pp. 1715–1725, Association for Computational Linguistics, Aug. 2016.

[35] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, *et al.*, "Training compute-optimal large language models," *arXiv preprint arXiv:2203.15556*, 2022.

[36] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-lm: Training multi-billion parameter language models using model parallelism," *arXiv preprint arXiv:1909.08053*, 2019.

[37] B. Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, *et al.*, "Bloom: A 176b-parameter open-access multilingual language model," *arXiv preprint arXiv:2211.05100*, 2022.

[38] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, *et al.*, "Scaling instruction-finetuned language models," *arXiv preprint arXiv:2210.11416*, 2022.

[39] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, *et al.*, "Llama 2: Open foundation and fine-tuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[40] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, *et al.*, "Training language models to follow instructions with human feedback," *Advances in Neural Information Processing Systems*, vol. 35, pp. 27730–27744, 2022.

[41] D. M. Ziegler, N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving, "Fine-tuning language models from human preferences," *arXiv preprint arXiv:1909.08593*, 2019.

[42] Q. Liu, F. Zhou, Z. Jiang, L. Dou, and M. Lin, "From zero to hero: Examining the power of symbolic tasks in instruction tuning," *arXiv preprint arXiv:2304.07995*, 2023.

[43] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24824–24837, 2022.

[44] R. Navigli, "Word sense disambiguation: A survey," *ACM computing surveys (CSUR)*, vol. 41, no. 2, pp. 1–69, 2009.

[45] S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," *Speech communication*, vol. 24, no. 1, pp. 19–37, 1998.

[46] W. H. Gomaa, A. A. Fahmy, *et al.*, "A survey of text similarity approaches," *international journal of Computer Applications*, vol. 68, no. 13, pp. 13–18, 2013.

[47] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: a comprehensive review," *ACM computing surveys (CSUR)*, vol. 54, no. 3, pp. 1–40, 2021.

[48] X. Chen, J. Ye, C. Zu, N. Xu, R. Zheng, M. Peng, J. Zhou, T. Gui, Q. Zhang, and X. Huang, "How robust is gpt-3.5 to predecessors? a comprehensive study on language understanding tasks," *arXiv preprint arXiv:2303.00293*, 2023.

[49] W. Yang, C. Li, J. Zhang, and C. Zong, "Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages," *arXiv preprint arXiv:2305.18098*, 2023.

[50] C. Li, Z. Gan, Z. Yang, J. Yang, L. Li, L. Wang, and J. Gao, "Multimodal foundation models: From specialists to general-purpose assistants," *arXiv preprint arXiv:2309.10020*, vol. 1, no. 2, p. 2, 2023.

[51] R. Tang, X. Han, X. Jiang, and X. Hu, "Does synthetic data generation of llms help clinical text mining?," *arXiv preprint arXiv:2303.04360*, 2023.

[52] O. Nov, N. Singh, and D. Mann, "Putting chatgpt's medical advice to the (turing) test: survey study," *JMIR Medical Education*, vol. 9, p. e46939, 2023.

[53] K. Yang, S. Ji, T. Zhang, Q. Xie, and S. Ananiadou, "On the evaluations of chatgpt and emotion-enhanced prompting for mental health analysis," *arXiv preprint arXiv:2304.03347*, 2023.

[54] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. O. Sabel, J. Ricke, *et al.*, "Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports," *European radiology*, pp. 1–9, 2023.

[55] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[56] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, K. Clark, S. Pfohl, H. Cole-Lewis, D. Neal, *et al.*, "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[57] S. Chen, B. H. Kann, M. B. Foote, H. J. Aerts, G. K. Savova, R. H. Mak, and D. S. Bitterman, "The utility of chatgpt for cancer treatment information," *medRxiv*, pp. 2023–03, 2023.

[58] A. Shah and S. Chava, "Zero is not hero yet: Benchmarking zero-shot performance of llms for financial tasks," *arXiv preprint arXiv:2305.16633*, 2023.

[59] J. C. S. Alvarado, K. Verspoor, and T. Baldwin, "Domain adaption of named entity recognition to support credit risk assessment," in *Proceedings of the Australasian Language Technology Association Workshop 2015*, pp. 84–90, 2015.

[60] G. Son, H. Jung, M. Hahm, K. Na, and S. Jin, "Beyond classification: Financial reasoning in state-of-the-art language models," *arXiv preprint arXiv:2305.01505*, 2023.

[61] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," *arXiv preprint arXiv:2303.17564*, 2023.

[62] X. Zhang and Q. Yang, "Xuanyuan 2.0: A large chinese financial chat model with hundreds of billions parameters," in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pp. 4435–4439, 2023.

[63] H. Yang, X.-Y. Liu, and C. D. Wang, “Fingpt: Open-source financial large language models,” *arXiv preprint arXiv:2306.06031*, 2023.

[64] Y. Bang, S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, *et al.*, “A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity,” *arXiv preprint arXiv:2302.04023*, 2023.

[65] D. Ganguli, L. Lovitt, J. Kernion, A. Askell, Y. Bai, S. Kadavath, B. Mann, E. Perez, N. Schiefer, K. Ndousse, *et al.*, “Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned,” *arXiv preprint arXiv:2209.07858*, 2022.

[66] E. Perez, S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving, “Red teaming language models with language models,” in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (Y. Goldberg, Z. Kozareva, and Y. Zhang, eds.), (Abu Dhabi, United Arab Emirates), pp. 3419–3448, Association for Computational Linguistics, Dec. 2022.

[67] W. Kuang, B. Qian, Z. Li, D. Chen, D. Gao, X. Pan, Y. Xie, Y. Li, B. Ding, and J. Zhou, “Federatedscope-llm: A comprehensive package for fine-tuning large language models in federated learning,” *arXiv preprint arXiv:2309.00363*, 2023.