The Evolution, Application, and Future Challenges of Large Language Model

Weiqi Li

{wei2143613@maricopa.edu}

Department of Architectural Engineering, Harbin University of Science and Technology, Harbin, 150000, China

Abstract. Large language models have been adopted as a fundamental advance in artificial intelligence, showing natural language understanding and natural language generation systems that are almost humanly good, through pre-training on massive quantities of text. This paper gives an attempt to review its main technologies, extensive usage, and challenges systematically. It begins with outlining the technological platforms behind its growth such as establishment of pre-training and a fine-tuning paradigm, scaling law to emergent capability development and state-of-the-art progress in secure alignment and efficient deployment. Second, its revolutionary effect on content generation, machine translation and intelligent answering of questions were discussed. And then goes to the primary issues that the model has in factual accuracy, security, bias, and energy consumption. And lastly, it has an outlook to the future on how the technology will evolve into additional reliability, security, and accessibility.

Keywords: Large language model; Pre-training; Scaling law; Security; Efficient deployment

1 Introduction

The most recent advancement in language intelligence, however, is large language models (LLMs), which have experienced a paradigm shift on the way a company should approach its language models, shifting them towards general task solvers. As the size of model parameters reach tens or even hundreds of billions they do not only produce breakthroughs in different language tasks as well as an unparalleled set of emergent behaviors including context learning and sophisticated reasoning [1].The most sophisticated models, exemplified by GPT-4, have outperformed humans at the highest level in most of the professional tests and this shows that GPT-4 can be used as a basis to general intelligence. Nonetheless, this increase in capabilities is also challenging, with the models potentially giving people an illusionary work that is against the reality [2]. It also has various security and privacy threats, including the possibility of attackers to contribute to the model output via indirect prompting injection [3]. In the meantime, its large computing requirements and implementation costs have given rise to the emergent research topics of efficient inference and model compression on edge devices [4][5]. That is why a logical overview of its technical principles, implementation in the existing

ICIAAI 2026, May 22-24, Singapore, Singapore

DOI 10.4108/eai.22-5-2026.2365155

context, and the challenges it faces is critically important both theoretically and practically to understand the leading edge of technology and develop it responsibly.

Key research projects have led to a rapid development of large language models, with three major themes being: ability enhancement, safety alignment, and efficient deployment Regarding the ability of the capabilities enhancement, the early attempts involve BERT [6]. The pre-training and fine-tuning paradigm was laid down. GPT-3 later empirically showed the line of scaling of model performance with scale, which showed an emergent ability of few-shot learning [7]. The scaling laws do not only help in the creation of the bigger models but they also indicate the diminishing returns of merely increasing the scale without simultaneous increase of data or computer efficiency. As an example, recent studies have found that most of the currently available LLM are undertrained, and that smaller models trained on larger data tend to be optimally powered by a given compute budget. This observation has led to the study of compute-optimal training recipes which have better performance with lower parameter counts through balancing between model size, and training tokens. Moreover, the emergent capabilities are not coded to perform reasoning (in terms of a chain of thought) or solving a mathematical problem, or being able to generate code, but instead, emerge spontaneously as models reach particular scale limits. These capabilities have been studied in a systematic manner, and they are impossible to extrapolate by the performance of smaller systems, which highlights the qualitative difference between a small model and the large model. The newest models, including GPT-4 and LLaMA 2, keep advancing the scope of possibilities and improve the multimodal comprehension and dialogue safety [8].

Regarding security and value alignment, the focus of value, reliability and security alignment in relation to model capabilities is enhanced as improvement of model capabilities. Ouyang et al. were the first to apply human feedback-based reinforcement learning (RLHF) to fine-tune models with a significant improvement in the ability to match the human intentions to the output [9]. The open-source models (LLaMA 2) also discuss the process of their security fine-tuning and red team testing. The study conducted by Greshake et al. demonstrated the susceptibility to indirect hint injection attacks in multifunctional applications of LLM in which an attacker may inject malicious instructions with the purpose of influencing model behavior, which indicates the insecurity of the deployment [3].

Regarding efficiency and accessibility, there are two directions of research to reduce the deployment threshold, one is model efficiency optimization, where Touvron et al. train smaller models with great performance (LLaMA series) on more efficient data and architecture [10]. And the designed pruning procedure named as LLM-Pruner by Ma et al [5]. The object is to compress the model without respect to the task. Secondly there are edge deployment technologies. The review by Cai et al. summarizes systematically the most important technologies created to execute LLMs on edge devices with a limited amount of resources, including speculative decoding and model offloading [4].

In the context of the given background and the status of the current research, the purpose of this paper is to offer an integrated review of technological evolution, impact thereof, and essential issues of large language models. It will begin by discussing the fundamental technological underpinnings underpinning their breakthroughs and major technology developments out of pre-training paradigms and scaling laws in order to achieve alignment.

This paper will then proceed to identify and discuss the common uses and effects of large models used in producing content, writing of code, and in the generation of intelligent questions. Moreover, it will be used to systematically examine the entire problems in question, being reliability, security, and consumption of resources, and summarize solutions, including efficient inference and model compression. Lastly, the conclusion and the anticipation of the future will be made.

2 Overview of mainstream technologies in large language models

2.1 Pre-training, fine-tuning, and alignment paradigms

Large language models have their core competency based on the pre-training-fine-tuning paradigm. The pre-training stage is a stage where the model engages in self-supervised learning on very large quantities of unlabeled text corpora. It learns to learn the common and general representations of the language and world knowledge as it takes tasks (e.g. cloze test, at BERT cloze language modeling) or as an autoregressive model (e.g. GPT) makes a prediction (e.g., the next word). This process provides the model with the capability of strong language comprehension and generation [6][7].

One of the most important procedures in the process of adjusting a general model to a particular downstream activity is fine-tuning. With additional training on labeled data to carry out particular tasks (e.g. text classification and question answering), the model can rapidly transfer general knowledge to particular tasks [6]. Nevertheless, when traditional instruction fine-tuning is done, there is still a possibility that the model would create unrealistic, harmful, or non-compliant outputs [9].

As such, there has developed the so-called alignment-technology, which is meant to bring about congruence between model behavior and human values and multifaceted intentions. Reinforcement learning has been the most mainstream method currently utilizing human feedback. This involves performing preference data, acquired by human annotators of the model output, and training a reward model to acquire the human evaluation criteria. Next, convert this reward model. It can make the output of the model very helpful, realistic, and not harmful, and is among the fundamental technologies of dialogue models including ChatGPT and InstructGPT [9]. Similar alignment processes are also commonly used in open-source models like LLaMA 2 in order to increase security [8].

Figure 1 demonstrates the safety and the human evaluation of LLaMA 2-Chat. This number is in comparison with the rates of security violations of LLaMA 2-Chat in other open-source and closed-source models but with less than 2000 adversarial hints. The findings indicate that the violation rates of all parameter versions of LLaMA 2-Chat are lower by far as compared to MPT, Vicuna, and other models and even better than ChatGPT (version 0301), which proves the success in security fine-tuning and red team testing [8].

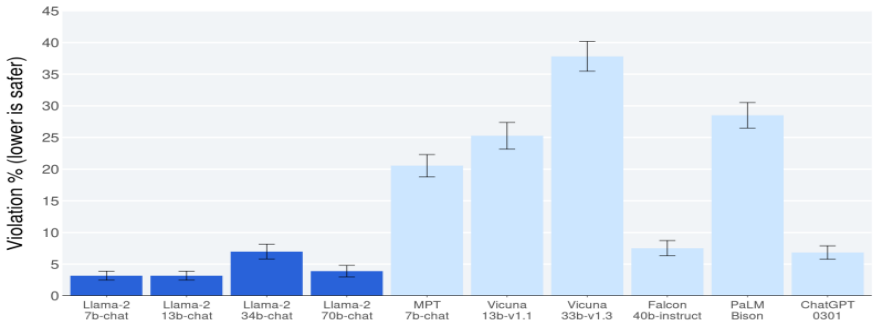

Fig. 1. Safety human evaluation results for Llama 2-Chat compared to other open-source and closed-source models. Human raters judged model generations for safety violations across ~2,000 adversarial prompts consisting of both single and multi-turn prompts [8].

2.2 Scaling Laws and Emergent Capabilities

The scaling law presents the fundamental concept of performance improvement of large language models. It has been demonstrated in the literature that analysis of the model has a predictable power-law dependence between the worth of its loss function and the expense of computation, size of data, and magnitude of parameters [7]. This implies that as the computing power, data and model size are constantly expanded this can be reasonably assured to enhance the performance of the model. This legislation offers a theoretical ground on the industry to contemplate investing constantly in the training of bigger models.

A qualitative leap takes place when the model size (in terms of parameters and data) reaches a critical size, that is, a sequence of so-called emergent capabilities revealed are no longer visible in smaller models [1][7]. Its most prominent feature is the ability of contextual learning. Without changing any parameter, the model can comprehend and execute new tasks by just giving a few examples of the tasks in the task input prompts (few-shot learning) or a simple description of the task (zero-shot learning). More features like multi-step reasoning, code generating and interpreting capabilities have also come with scale. These abilities render large language models resemble general task solvers and less so tools to execute pre-carried out tasks.

Notable instances of emergent skills are the ability to do arithmetic problems without necessarily being trained on them, comprehension of complicated instructions that involve several constraints, as well as production of meaningful code in response to a natural language description. An illustration of how scaling opens up the reasoning ability is chain-of-thought prompting, where Wei et al. introduced scaling as a way to solve reasoning problems by prompting the model to generate steps of reasoning even models that are not optimized on reasoning tasks can solve multi-step problems it previously could not. This effect is directly related to model scale; it is experimentally demonstrated that chain-of-thought reasoning can be trusted in scales of models with tens of billions of parameters or more. In addition, emergent abilities have a structure of a phase transition, involve performance being near random at low model sizes, and increasing exponentially at a low critical number of

parameters, with hints of fundamental changes in the internal representations of the model being made [13].

2.3 High-efficiency technology frontier

The high inference costs and deployment barriers of the model have been an important constraint to its use as its size of paradigms exploded tremendously. So, model efficiency methods have been a topic of active research, predominately split into two groups, model compression and efficient inference.

The objectives of model compression are to minimize the model size and computation needs without profoundly affecting the performance. Quantization transforms model weights in high precision (like FP16) to low precision (like INT8, INT4), which demands much fewer bits to store and transmit [11]. Knowledge distillation challenges a small student model to imitate the behaviour of a large teacher model [12]. Structured pruning simply eliminates unimportant entities in the model, e.g. attention heads or neurons. As an example, task-agnostic pruning of the model by LLM-Pruner detects dependencies and evaluates their importance, which is an efficient approach to revive the performance of a small number of data, making it a new method of lightweight deployment of LLM [5].

The comparative variations in the three compression paradigms are made against each other in figure 2. Fine-tuning on individual datasets followed by pruning is required in traditional task-specific compression. TinyBERT needs massive and unlabeled corpora and long distillation. By contrast, one hour of task-independent structured pruning only needs 50MB of data and a single GPU in LLM-Pruner. Besides, the smaller model (i.e. compressed LLaMA-5.4B) similarly could retain the same quality of generation as the original LLaMA-7B [5].

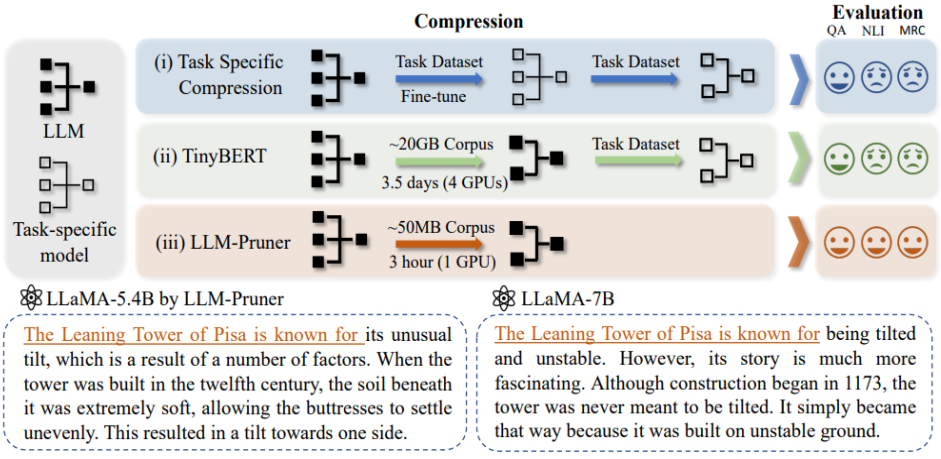

Fig. 2. Comparison of different compression paradigms. LLM-Pruner enables task-agnostic structured pruning with only 50K samples and 3 hours of recovery [5].

The most recent development in quantization techniques is the LLM. int8() method that deals with the performance loss to outlier features in 8-bit quantization by means of mixed precision

decomposition. Figure 3 shows the process of two-component quantization of LLM. int8(). To achieve 8-bit quantization of most normal values, this approach uses a preliminary step of vector-based quantization of 8-bit quantities, and then segregates the characteristics of some outliers and carries out 16-bit matrix multiplication. Lastly, the two outcomes are added resulting in a halved memory usage but the same model performance [11].

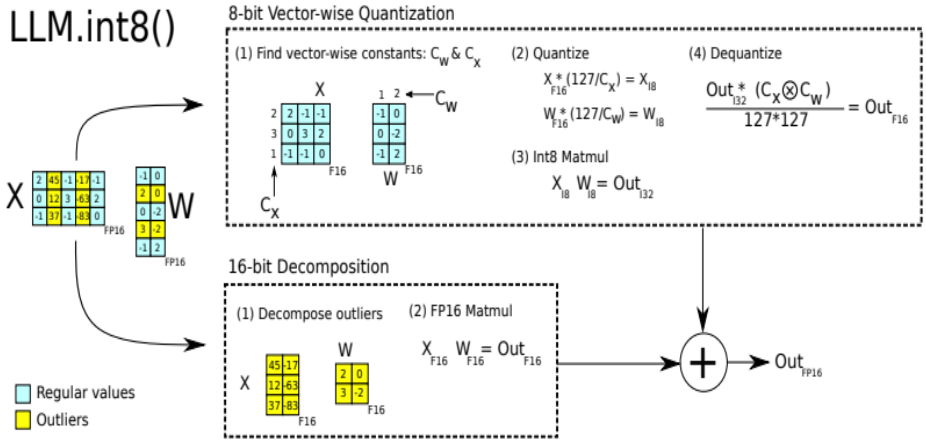

Fig. 3. Two-part quantization procedure of LLM.int8(): vector-wise quantization for most values, and 16-bit decomposition for outlier features [11]

Inference technology High efficiency Inference technology is devoted to the optimization of the inference process itself, particularly on resource constrained edge devices. Inference decoding builds upon a small, fast draft model, which then generates candidate word sequences which are validated in parallel using the original large model, thus speeding up the autoregressive generation process. Let the model offloading strategy sees the inactive model portions stored, in external slow storage (i.e. disk) and loaded into GPU memory on demand thereby working around the device memory limitations [4]. Such technologies are essential to facilitate the localization of large language models to edge devices like a mobile phone and internet of things devices.

3 Analysis of the main applications of large language models

3.1 Content generation and creation

Huge language models have already shown sturdy productivity when it comes to content production. They are able to produce high quality articles, reports, poems, scripts, and marketing copy after having brief prompts or outlines. It is not only applicable to a formatted text but also applies to situations that involve the use of creativity and style. As an illustration, GPT-4 is capable of writing coherent and logically structured long essays, and even replicate the style of individual authors in the literary work [2]. Also, with the emergence of the multimodal technologies, text-based image and text-based video applications are also achieved

that have broadened the frontiers of creative industries. But this, also has led to similar ethical and legal debates concerning copyright and the authenticity of content, and the labelling of AI-generated content.

Table 1 presents the results of GPT-4 and GPT-3.5 in a number of professional tests. GPT-4 scored on the top 10 percentile of candidates in the mock bar exam whereas GPT-3.5 scored on the bottom 10 percentile of candidates. Its GPT-4 in the graduate school entrance exam, including the LSAT and GRE, also have shown very high performance over its older versions, with its verbal reasoning score in the 99 th percentile [2]. These metrics indicate the ability of large language models to perform knowledge-intensive tasks with great power.

Table 1. Performance comparison of GPT-4 and GPT-3.5 on representative professional exams.

Exam	GPT-4	GPT-3.5
Uniform Bar Exam (simulated)	298/400 (top 10%)	213/400 (bottom 10%)
LSAT	163 (88th percentile)	149 (40th percentile)
GRE Verbal	169/170 (99th percentile)	154/170 (63rd percentile)
USABO Semifinal Exam 2020	87/150 (99-100th percentile)	43/150 (31-33rd percentile)
Codeforces Rating	392 (below 50th percentile)	260 (below 50th percentile)

3.2 Code generation and program assistance

Big language models have been very promising in machine translation and cross-linguistic. The classic neural machine translators use huge parallel corpora, and with the use of large language models, by pre-training them through huge volumes of multilingual text, can perform translation operations under zero-sample or few-sample specifications. According to a study conducted by Brown et al., GPT-3 can also learn to perform nearly supervised system performance on many language pairs, with few input translation examples, and do not require gradient updates [7]. Moreover, a model of a larger size gains far better cross-linguistic generalization capabilities, most notably in the tests of 54 languages of MMLU, GPT-4 done better on 24 languages than the best-then English model [2]. This is quite useful when it comes to the translation task of low-resource languages. The high cross-lingual performance of LLMs is due to the pre-training the model at scale of huge multilingual data corpus, consisting of web crawls, books, and articles in dozens of languages. This provides the ability to translate between language pairs that were not often observed together to provide training feedback, like Thai to Romanian, without knowing any parallel data. Nevertheless, there are still differences in performance: languages with rich training data (e.g., English, Chinese, Spanish) can be translated with almost human accuracy, whereas with low-resource languages, the error rates are still higher and language experiences a propensity to hallucinations. To counter this imbalance, methods of targeted data augmentation, cross-lingual transfer learning and creating more diverse evaluation benchmarks, beyond English-centric ones are necessary. Nevertheless, existing models continue to be weak in handling the literary texts, idioms, and culturally-specific expressions, and the issue of English-centric evaluation standards has not been properly tackled.

3.3 Intelligent question answering and complex reasoning

One of the most persuasive uses of large language models is the capability of question answering and knowledge reasoning engines, which can be intelligent. The model can resolve open-ended questions, engage in multi-turn conversations, summarize long texts, and identify important points through the integration of massive amounts of knowledge. On a more advanced level, with the support of the ideas of prompting engineering, like the thought chains, the model is capable of proving step-by-step reasoning, thus solving complex problems that can have mathematical, logic, and common sense solutions [13]. This renders them very promising in educational tutoring, intelligent customer service, research assistance, and legal documents analysis. Nonetheless, the illusion issue about their responses that is, making sure to come up with apparently sensible yet really false or artificial information is a fundamental reliability issue in this application field [2][3].

4 Current Challenges and Future Prospects

4.1 The main challenges

The further adhesion and the proliferation of large language models remains a subject of several problems despite remarkable outcomes of the models described in the preceding sections. Reliability-wise, the most noticeable issue that the model is likely to reproduce is the phenomenon of the illusion: the resulting content can be counter to the facts or utterly unsubstantiated, in the cases when there are no sources of outside knowledge to back them up [2][3]. This does not only compromise user trust, but it also reduces the use of the same in high risk environments like healthcare and law. Security-wise, a model can replicate social biases of the training data, creating discriminatory material; in addition, malicious people can use a model to make fake information, launch a phishing attack, or even automate an attack. Moreover, one should not overlook the privacy threats of the model; personal information in the training data is likely to be inadvertently pulled out or it can be leaked during early injection [3]. Resources and deployment Hundreds of billions of models can be trained and inferred with very large electricity demands, which underscores the increasing carbon footprint problem. Moreover, the technology is expensive and is a significant barrier to the use of the technology [5][11][12]. And, last but not least, the evaluation system remains subpar: the current standards are not sufficient to fully assess the real-life performance of the model in terms of safety, fairness, and robustness, and not a unified evaluation framework to do it across languages and cultures [1].

4.2 Future development trends

In the future, the trend of studies on large language models will move to a more efficient and quality-focused approach, as opposed to a scale-first one. At the model level, effective architectural design (e.g. hybrid expert models), quantization and pruning will also keep on reducing the threshold of deployment [5][11]. Of the many architectural new ideas, mixture-of-experts (MoE) models have also become a new way forward in terms of scaling model capacity without the incurred cost in inference being disproportionately high. MoE models like GShard and Switch Transformer use smaller proportions of their parameters with each token, which allows them to be more efficient, reducing the importance of

billion-parameter models, allowing them to be deployed. The improvement of routing, to achieve expert balancing and enhance training stability, is something that can be done in the future. Instead, knowledge distillation assists in everyday transfers of the abilities of great models into the lightweight models [12]. On the capability level, methods of cueing engineering like thought chains are also known to contribute considerably to the complex reasoning abilities of models [13]. Elucidation mechanisms of reasoning that are more explainable and verifiable will be developed in the future. Human feedback based reinforcement learning has been successful in the first steps in terms of safety and alignment [9]. Future studies should come up with more effective preference gathering techniques, stronger red team testing systems and cross cultural value alignment techniques [3][8]. Multimodal fusion and embodied intelligence will emerge as important direction at the application level and models will go beyond pure text interaction level to the interaction of vision, hearing and the physical world. After all, it is a common cause of academia and industry to establish an open, transparent, and responsible AI governance system to make sure that technological development never fails to benefit the entire human society on the whole.

5 Conclusion

This paper in systematic review covers the development of technology, practices of use, and fundamental issues of large language models. On the technical level, the new basic paradigm of pre-training-fine-tuning up to scale and expansion accelerated by the scaling law to the secure alignment technology with RLHF as its core part, all these aspects feature in its capabilities leap. Meanwhile, other efficient inference technologies involving model compression and suitable technologies to deploy the technology in real-life situations have paved the way to the implementation of the technology. On the application level, large language models have progressed beyond text generation systems to become a general purpose content creation assistant, programming assistant, and general purpose reasoning assistant, and have the potential to redefine many industries. Nonetheless, issues like hallucinations, security threats and excessive costs are still eminent and its prudent large scale usage has been hampered. In the future, large language model studies will also stop being just interested in scaling their parameters to consider alternative approaches to enhancing architectural efficiency, improving inference reliability, enhancing security governance, and lowering the cost of deployment. Large language models can only be realized as a viable intelligent infrastructure that gains human advantage and grabs the societies through shared innovation of technology, ethics, and governance.

References

[1] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361. (2020)
[2] OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., … Zoph, B.: GPT-4 technical report. arXiv preprint arXiv:2303.08774. (2023)
[3] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M.: Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (pp. 79–90). (2023)

[4] Cai, G., Tian, R., Yang, L., Jia, Y., Li, L., & Wang, J. (2026). Efficient inference for edge large language models: A survey. Tsinghua Science and Technology, 31(3), 1365-1380. https://doi.org/10.26599/TST.2025.9010166
[5] Ma, X., Fang, G., & Wang, X.: LLM-Pruner: On the structural pruning of large language models. arXiv preprint arXiv:2305.11627. (2023)
[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4171–4186). Association for Computational Linguistics. (2019)
[7] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D.: Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877–1901. (2020)
[8] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., … Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288. (2023)
[9] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., … Lowe, R.: Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155. (2022)
[10] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., … Lample, G. LLaMA: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971. (2023)
[11] Dettmers, T., Lewis, M., Belkada, Y., & Zettlemoyer, L.: LLM.int8(): 8-bit matrix multiplication for transformers at scale. arXiv preprint arXiv:2208.07339. (2022)
[12] Hinton, G., Vinyals, O., & Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531. (2015)
[13] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., … Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems, 35, 24824–24837. (2022)