2023 13th International Conference on Information Technology in Medicine and Education (ITME) | 979-8-3503-1915-6/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITME60234.2023.00148

Fine-tuning Large Language Models in Education

Yong Chen, Hongpeng Chen, Songzhi Su*
Institute of Artificial Intelligence
Xiamen University
Xiamen, China
e-mail: ssz@xmu.edu.cn

***Abstract*—In recent years, large language models (LLMs) have been a hot topic in artificial intelligence research, profoundly impacting many fields, including education. LLMs showcase powerful capabilities in natural language comprehension and generation, demonstrating impressive domain generalization after downstream tasks fine-tuning. Therefore, fine-tuning LLMs have attracted much attention from education, and research on their educational applications has begun to enter the public view. This paper undertook a comprehensive examination and analysis of the fine-tuning LLMs and their potential applications in the field of education. On this basis, we explore the common issues currently encountered in fine-tuning LLMs, considering both the advantages and limitations of fine-tuning LLMs and their potential to enhance the efficiency and effectiveness of education. Finally, to help educators make informed decisions and foster innovation in education for the better service of human education, we look into the future trends and applications of fine-tuning LLMs in education.**

Keywords-large language models; fine-tuning; survey; educational technology; educational application

I. INTRODUCTION

With the rapid development of AI technology, the impact of large language models (LLMs), which is epoch-making in AI, has gradually emerged in various industries due to its breakthroughs in natural language processing, computer vision, robotics, and other technical fields. Currently, the digital transformation and intelligent upgrading of education are accelerating in education. As a result, LLMs are gradually being applied to critical aspects such as constructing educational environments, supporting teaching processes, accurately evaluating teaching, and efficiently managing education. LLMs have begun a comprehensive and deep integration with all aspects of education and teaching, creating a new form of intelligent education with intelligent technology covering all aspects of education and promoting the development of ubiquitous learning and personalized learning.

LLMs are AI systems trained on massive data with large-scale parameters for natural language processing tasks. They are usually trained on large-scale corpora such as books, articles, and Internet content. LMMs, also known as Foundation Models [1], can be applied to various AI models for different tasks, such as GPT-3 [2] and LLaMa2 [3], which have achieved great success in the field of natural language processing. In addition, multimodal LLMs such as Stable Diffusion [4] and GPT-4 [5] based on the fusion of massive text and image data have further achieved cross-modal comprehension and generation, enhancing their generalization capabilities across multiple domains and tasks. However, due to the massive training data, relatively complex internal structure, and numerous parameters in LLMs, the training and usage costs are very high for a single or a few tasks, and the delay caused by model calculation is also considerable. With the pre-training and fine-tuning method, significant performance improvements can be achieved by fine-tuning the pre-trained models using task-specific labeled data, significantly reducing the cost of training specialized LLMs in specific domains. Therefore, this more efficient pre-training and fine-tuning method has gradually become the mainstream paradigm in applying LLMs in various fields.

The research and application of fine-tuning LLMs have made significant progress in many verticals, including healthcare, law, finance, and the arts. However, LLMs are still in their infancy in education, and there is an urgent need for relevant basic research and applied innovation. The most significant advantage of fine-tuning LLMs is their conversational interactivity and performance comparable to the human level in cognitive tasks in various fields, including education. The rise of fine-tuning LLMs has great potential to improve the efficiency and effectiveness of educational work, providing new development ideas for upgrading educational intelligence.

II. BASIC CONCEPTS OF FINE-TUNING LLMS

A. Development of LLMs

A language model is a modeling of the probability distribution of natural language. In a given context, language models can estimate the probability of a sentence appearing, which is used to measure the linguistic rationality of a sentence.

The development of language models has gone through grammar rule language models, statistical language models, and neural language models. Grammar rule language models are based on linguistic and domain knowledge by manually designing linguistic grammars, but they are difficult to deal with large-scale texts [6]. Among statistical language models, the most representative one is n-gram [7], whose main idea is to use statistical methods to predict the probability of the next word appearing in the text.

2474-3828/23/$31.00 ©2023 IEEE
DOI 10.1109/ITME60234.2023.00148

With advanced deep learning, neural networks have gradually been introduced into language model modeling, opening a new phase known as Neural Language Models. The Word Embedding method proposed by Bengio [8], which maps the unique one-hot encoding of words into a low dimensional dense real number vector by constructing a shallow neural network, has profoundly impacted the development of language models. Since then, neural network approaches such as RNN and LSTM [9], which use distributed word vectors to model contextual relationships in language, have begun to emerge. In recent years, the explosive growth of data, improved performance of hardware devices, and the development of self-supervised learning [10] techniques have made it possible to train super-large scale neural network-based language models. ELMo [11] opens the door to pre-trained methods for language models. The advent of large-scale pre-trained language models such as BERT [12] and GPT-4 based on the Transformer architecture [13] has made the pre-training and fine-tuning paradigm a mainstream research direction. The pre-training approach involves training on a vast amount of data to help the model learn how to extract features. Then, the model is fine-tuned based on the specific objectives of the task. This means that the pre-trained model is trained with labeled data that is specific to the task, which allows for the transfer of knowledge from the pre-trained model to the downstream task in an effective manner.

B. Different Approaches to Fine-tuning

Supervised fine-tuning, also called instruction tuning, involves using task-specific labeled data to fine-tune a pre-trained model, which allows the model to follow instructions accurately. Instruction tuning is a common method for adapting pre-trained LLMs to downstream tasks.

The surging number of parameters in LLMs significantly increases the model's performance and generates an enormous demand for computational resources. For example, GPT-3 proposed by OpenAI has 175B parameters and requires at least 3.14e23 FLOPs of computation and 700GB of graphics memory space for training under single precision [2]. The training of GPT-3 requires parallelization on thousands of high-performance GPUs, costing millions of dollars [2]. It can be seen that the high expense of fully training LLMs is difficult to undertake for most enterprises and laboratories.

Traditional transfer learning methods require fine-tuning all parameters of the pre-trained model, called full fine-tuning. However, this method's computational cost is exceedingly high for the large number of LLMs parameters. In contrast, parameter-efficient fine-tuning (PEFT) only fine-tunes a small or additional number of model parameters, fixes most pre-trained parameters, greatly reduces computational and storage costs, and expands the application range of pre-trained LLMs. In addition, the advanced PEFT technology can also achieve performance comparable to full fine-tuning.

The commonly used PEFT methods include adapters, soft prompts, and low-rank adaptation (LoRA). Adapter tuning was proposed by Houlsby et al. in 2019 [14], which focuses on adapting downstream tasks by adding a learnable network module, adapter, to the pre-trained model based on the Transformer architecture. The adapter network is structured as a bottleneck, which decreases the input's dimensionality, applies a nonlinear transformation, and subsequently reconstructs the input to its original high dimensionality. Finally, the residuals are added to the ultimate output via a residual connection. The adapter is inserted twice into every Transformer layer: once after the multi-head attentional mapping and again following the two-layer feedforward neural network. During fine-tuning on specific data, only the parameters within the adapter are updated, while the parameters of the pre-trained model remain unchanged. The adapter tuning method is widely used, and several variants have been proposed. Wang et al. applied the adapter to transfer learning and proposed the K-adapter method to solve the problem of catastrophic forgetting during new knowledge injection [15]. Pfeiffer et al. proposed the AdapterFusion method to achieve maximum task transfer between multiple adapter modules [16].

Prompt-tuning [17], proposed by Lester, Al-Rfou, and Constant for learning soft prompts, involves splicing a task-specific, continuous, and learnable prefix tensor at the front end of the input embedding of a pre-trained LLM. This approach enables adaptation for downstream tasks. During training, the gradient descent algorithm optimizes the learnable prefix tensor on downstream tasks while maintaining the pre-trained LLMs' parameters in a frozen state. Prefix-tuning, designed by Li and Liang [18], is an improved method of prompt-tuning, which further splices the learnable prefix tensor at the front end of all hidden states in the model to improve the stability of training.

However, all of the above methods have some drawbacks. Adapter tuning adds additional model parameters that lead to the inference latency problem in the inference phase. Prompt tuning is difficult to optimize; its performance varies nonlinearly with the size of the training parameters [19]. Fundamentally, prefixes reduce the length of the sequences used to process downstream tasks. Hu et al. hypothesized that the updated values of model weights actually have a lower intrinsic rank and thus proposed a method called LoRA [19], which adds a branch network to all fully connected layers in the pre-trained LLMs. This network utilizes the product of two rank decomposition matrices to approximate the updated values of fully connected layer weights in domain adaptation. Therefore, only the parameters in this branch network need to be updated during training. The proposal of LoRA attracted much attention, according to which researchers have successively proposed methods such as AdaLoRA [20], QLoRA [21], and IncreLoRA [22].

Based on the analysis above, PEFT can significantly reduce the training cost of LLMs, achieving performance comparable to full fine-tuning while only requiring a small number of additional parameters to achieve domain adaptation. In addition, PEFT can alleviate the catastrophic forgetting issue of knowledge resulting from full fine-tuning, thereby improving generalization. Notably, PEFT is a

flexible and general purpose, which can adapt well to different downstream tasks.

III. Educational Application of Fine-tuning LLMs

To meet various application requirements in education, it is necessary first to construct general-purpose LLMs. These LLMs allow fine-tuning on downstream tasks to form three typical applications: automatic generation of teaching resources, human-AI interactive learning, and teaching intelligent assistance. The process involves collecting massive amounts of data and knowledge from both general and educational fields, such as subject knowledge, assignments, exam papers, MOOCs, teaching theories, etc., and then using self-supervised learning to pre-train on these data. The fine-tuning LLMs obtained in this way can deeply understand the three educational elements of teaching resources, teaching objects, and teaching processes to serve and support educational participants better.

A. Automatic Generation of Teaching Resources

The concept of using artificial intelligence systems to automatically generate educational resources is not new, and discussions on algorithm-generated learning materials can be traced back to the 1970s [23]. In recent years, the rapid development of LLMs has brought powerful generative abilities, making it a tool for supplementing teaching resources and broadening the application of related fields.

Regarding the automatic generation of teaching resources, innovative technologies based on LLMs have demonstrated certain capabilities. Image generation models, such as Stable Diffusion, can generate art teaching resources with various styles, novelty, uniqueness, and aesthetics by inputting text descriptions of images based on teaching needs. In terms of text resource generation, there is little difference between abstracts generated by LLMs and those produced by people. DeepMind and Stanford University proposed Dramatron [24], a text-generation model which can generate specific and vivid script content. In addition, Google's MusicLM [25] can generate high-quality music clips directly based on natural language descriptions. Fine-tuning LLMs are also making continuous progress in applications that require strong logical reasoning abilities. Researchers at Rice University propose generating multidisciplinary and high-quality questions that can be directly applied to teaching, utilizing GPT-3 and prompt fine-tuning [26]. In the programming course, the most advanced LLM, Codex, can generate reasonable programming exercises for students and provide examples and accurate code explanations [27]. The educational technology developed through fine-tuning Codex also demonstrates the ability to solve 81% of advanced mathematics problems [28].

In summary, based on existing LLMs, fine-tuning for automatic generation of teaching resources is expected to make continuous progress in functionality and performance. Especially in terms of personalization, inspiration, multimodality, and interdisciplinarity, it will provide more possibilities for promoting the digitization and intelligence of education.

B. Human-AI Interactive Learning

Human-AI Interactive Learning is gradually becoming an important form and component of teaching activities, and the biggest advantage of LLMs lies in their direct interaction capability. An important advantage of combining interaction with teaching is that personalized elements can be better integrated into the learning process.

Pataranautaporn et al., based on the GAN architecture, used AI-generated animated characters to interact with learners [29]. Dong et al. used LLM as a partner to tackle complex scientific challenges, introducing a framework called Socratic reasoning and proposing a paradigm named LLM for Science [30]. In education, fine-tuning LLMs with a Socratic tutoring mode can encourage students to think independently and guide them to find answers by asking appropriate questions. Although appropriate questions are essential, evaluating how students respond to and interact with these questions is also necessary. Abdelghani et al. studied the impact of prompt learning on the question-asking behavior of students. They found that such automatic prompts generally have positive effects, such as enhancing students' curiosity and learning initiative [31].

In sum, fine-tuning LLMs in specific fields through human-AI interaction can provide personalized learning experiences. In this regard, role-playing, automatic feedback, social interaction, and other methods can encourage exploration and questioning, focusing on learners' cognitive state, intentions, and teaching-oriented interactions. This will help learners enhance their knowledge and achieve an efficient human-AI learning process.

C. Teaching Intelligent Assistance

The current basic LLMs already possess strong problem-solving capabilities and can be further fine-tuned in education to expand their ability to assist in teaching. Fine-tuning LLMs can play a significant role, especially in applying educational theory, automatically grading test questions, and managing teaching processes.

Ziyue is an educational LLM released by NYSE: DAO, equipped with various functions such as LLMs-based translation, virtual oral coaching, and AI essay guidance to assist learning. Khan Academy is a non-profit educational institution actively researching how to apply LLMs in Khanmigo to optimize online teaching. EduChat is an LLMs-based educational chat robot system developed by Dan et al. [32]. This chat robot system is guided by psychology and educational theories. It learns domain-specific knowledge through pre-training on educational corpora and fine-tuning on designed system prompts and instructions. This further enhances educational functions such as open question answering, essay assessment, Socratic teaching, and emotional support.

Fine-tuning LLMs has shown great potential in teaching intelligent assistance, and there is still much room for exploration in both technology and application domains.

IV. Challenges of Fine-tuning LLMs in Education

A. Technical Challenges

In the process of gradually applying fine-tuning LLMs to the education field, there are some corresponding technical challenges. Due to the training mechanism of LLMs, they rarely involve comprehension-level content and exhibit significant limitations in some respects, such as reasoning, self-awareness, emotions, intuition, responsibility, and morality. Furthermore, knowledge in the field of education is constantly evolving. Since LLMs do not have real-time internet access, they cannot learn the most up-to-date information. Consequently, their knowledge base remains fundamentally limited. LLMs are mainly pre-trained on massive unlabeled data, and even after fine-tuning, it is difficult to avoid inherent issues such as data bias, intellectual property, and knowledge accuracy. In fact, diverse and high-quality training data is relatively scarce in education, and the accuracy of training data cannot be effectively ensured. Consequently, fine-tuning LLMs may provide incorrect answers. Learners lacking specialized knowledge may be unable to find and correct these problems, leading to potential misguidance. Worse, the hallucination problem has been widely observed in fine-tuning LLMs, where inaccurate information is fabricated and espoused lucidly. These errors, biases, and hallucinations will make it difficult for learners to discern whether they have been accurately imparted knowledge.

LLMs are typically black box models whose internal decision-making processes are difficult to explain, making it challenging to gain people's trust truly. Consequently, none of the existing fine-tuning LLMs can be considered AI systems that are totally transparent to educational stakeholders. Additionally, real-time interaction and feedback greatly impact learners' initiative and learning effect in the teaching process. However, due to reasoning delay, fine-tuning LLMs still have certain defects in real-time interaction. The existing fine-tuning LLMs are difficult to effectively help learners in learning motivation and emotional support. More fundamentally, in the interactive teaching between teachers and students, teachers and students will also share an emotional experience and moral resonance while imparting knowledge. In this regard, the general intelligence embodied by LLMs is still far from genuine human intelligence.

B. Ethical Issues

Powerful technology is often a double-edged sword. Improper or misuse of LLMs will lead to negative application effects. When fine-tuning LLMs are applied in education, it is essential to conduct risk assessments from various dimensions such as scientificity, fairness, accuracy, and values. Fine-tuning LLMs can quickly answer questions, generate papers, write code, and compose scripts, and educators may be unable to identify them. If misused, they can easily become a tool for cheating in education. On the one hand, it will cause unfairness, integrity crisis, and other problems. On the other hand, it may prevent teachers from truly understanding students' learning statuses, according to which they can objectively adjust the teaching plan. These factors will diminish the quality of teaching in numerous ways and directly impact the conventional educational process and system.

Various factors, including training data, influence the bias of LLMs. Gebru, a renowned expert in artificial intelligence ethics, has pointed out that it is almost impossible to completely eliminate certain social biases, such as those related to politics, ethnicity, gender, class, and age, from training data and models [33]. Therefore, in the application of education, it is crucial for educators to supervise and define the scope of the model's usage to prevent any adverse impact on learners' independent thinking and cognitive processes. The problems in generating teaching resources by fine-tuning LLMs may harm students' social cognition and ethics. Teachers who use these resources will also encounter new pressures and risks.

Natural language data used by LLMs for fine-tuning in educational applications may contain personal and sensitive information about individuals' private lives and identities. This will easily lead to privacy and data security issues such as leakage, unauthorized access, and data abuse. Therefore, effective measures must be taken to ensure security.

V. Directions for Future LLMs in Education

The scaling law of language models suggests that the model's performance demonstrates a linear improvement with exponential increases in the number of parameters, the amount of data, and the training time [34]. According to this law, increasing the amount of training data and expanding the scale of large models is a straightforward approach to improving LLMs' performance. Currently, the training data for LLMs is typically at the TB level and may progress to the PB level in the future. It is anticipated that future LLMs will achieve breakthroughs in capabilities through the surge in training data. In the fine-tuning process of educational applications, data quality, structure, and diversity are important factors affecting the model's performance. In this regard, we can continue to explore data mining techniques and extract meaningful and valuable information for fine-tuning LLMs from educational big data.

Current LLMs are primarily based on the Transformer regarding algorithm and model architecture. In the future, improvements to the Transformer or the emergence of other superior architectures will further enhance the capabilities of LLMs. During training, innovations and breakthroughs in self-supervised learning and fine-tuning algorithms will also imbue LLMs with new capabilities. A possible research direction is exploring ways to effectively reduce the scale of LLMs while maintaining their performance on specific tasks, by leveraging their capabilities in few-shot learning, domain generalization, and task generalization. Additionally, designing algorithms that consume less on computility to achieve breakthroughs in computility is a path for enhancing real-time interaction. About reinforcement learning from human feedback, LLMs in education can use the interaction and feedback information with learners during the learning process to further fine-tune the model based on personalized instructions from learners. This enables LLMs to improve

their capabilities continuously and provides more personalized teaching services.

In educational applications, LLMs can be connected to the Internet to access the latest teaching resources and information, ensuring the accuracy and credibility of the generated teaching content. Moreover, precise user profiles can be created based on education-related information and recommendation algorithms to generate personalized learning resources, enabling LLMs to showcase different teaching styles for learners. In the teaching process, the accurate comprehension of multimodal data and autonomous utilization of teaching tools are crucial capabilities that fine-tuning LLMs in education are expected to achieve breakthroughs in the future. This will render educational LLMs more intelligent and have greater development potential.

VI. Conclusion

Motivated by the ongoing digital transformation in education, this paper explores the application of advanced fine-tuning LLMs in education. Firstly, this paper introduces the technical background knowledge of fine-tuning LLMs, providing an overview of the model architecture and training process. Subsequently, the application of fine-tuning LLMs technology in education was exemplified, primarily focusing on published cases. Finally, we summarize the main challenges in fine-tuning LLMs in education and offer prospects for their future development. Fine-tuning LLMs, which are currently one of the cutting-edge research domains in artificial intelligence, will continue to develop rapidly and profoundly influence education in the future. However, we should also remain vigilant in addressing ethical, legal, and social issues related to the application of fine-tuning LLMs in education, ensuring that their applications always comply with ethical standards and regulatory requirements.

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, and E. Brunskill, "On the opportunities and risks of foundation models," arXiv preprint arXiv:2108.07258, 2021.

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

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

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models." pp. 10684-10695.

[5] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, and S. Lundberg, "Sparks of artificial general intelligence: Early experiments with gpt-4," arXiv preprint arXiv:2303.12712, 2023.

[6] J. Gillett, and W. Ward, "A language model combining trigrams and stochastic context-free grammars."

[7] P. F. Brown, V. J. Della Pietra, P. V. Desouza, J. C. Lai, and R. L. Mercer, "Class-based n-gram models of natural language," Computational linguistics, vol. 18, no. 4, pp. 467-480, 1992.

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

[9] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," Physica D: Nonlinear Phenomena, vol. 404, pp. 132306, 2020.

[10] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," Technologies, vol. 9, no. 1, pp. 2, 2020.

[11] J. Sarzynska-Wawer, A. Wawer, A. Pawlak, J. Szymanowska, I. Stefaniak, M. Jarkiewicz, and L. Okruszek, "Detecting formal thought disorder by deep contextualized word representations," Psychiatry Research, vol. 304, pp. 114135, 2021.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," arXiv preprint arXiv:1810.04805, 2018.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[14] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for NLP." pp. 2790-2799.

[15] R. Wang, D. Tang, N. Duan, Z. Wei, X. Huang, G. Cao, D. Jiang, and M. Zhou, "K-adapter: Infusing knowledge into pre-trained models with adapters," arXiv preprint arXiv:2002.01808, 2020.

[16] J. Pfeiffer, A. Kamath, A. Rücklé, K. Cho, and I. Gurevych, "AdapterFusion: Non-destructive task composition for transfer learning," arXiv preprint arXiv:2005.00247, 2020.

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

[18] X. L. Li, and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," arXiv preprint arXiv:2101.00190, 2021.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," arXiv preprint arXiv:2106.09685, 2021.

[20] Q. Zhang, M. Chen, A. Bukharin, P. He, Y. Cheng, W. Chen, and T. Zhao, "Adaptive budget allocation for parameter-efficient fine-tuning," arXiv preprint arXiv:2303.10512, 2023.

[21] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "Qlora: Efficient finetuning of quantized llms," arXiv preprint arXiv:2305.14314, 2023.

[22] F. Zhang, L. Li, J. Chen, Z. Jiang, B. Wang, and Y. Qian, "IncreLoRA: Incremental Parameter Allocation Method for Parameter-Efficient Fine-tuning," arXiv preprint arXiv:2308.12043, 2023.

[23] C. Guan, J. Mou, and Z. Jiang, "Artificial intelligence innovation in education: A twenty-year data-driven historical analysis," International Journal of Innovation Studies, vol. 4, no. 4, pp. 134-147, 2020.

[24] P. Mirowski, K. W. Mathewson, J. Pittman, and R. Evans, "Co-Writing Screenplays and Theatre Scripts with Language Models: Evaluation by Industry Professionals." pp. 1-34.

[25] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, and M. Tagliasacchi, "Musiclm: Generating music from text," arXiv preprint arXiv:2301.11325, 2023.

[26] Z. Wang, J. Valdez, D. Basu Mallick, and R. G. Baraniuk, "Towards human-like educational question generation with large language models." pp. 153-166.

[27] S. Sarsa, P. Denny, A. Hellas, and J. Leinonen, "Automatic generation of programming exercises and code explanations using large language models." pp. 27-43.

[28] I. Drori, S. Zhang, R. Shuttleworth, L. Tang, A. Lu, E. Ke, K. Liu, L. Chen, S. Tran, and N. Cheng, "A neural network solves, explains, and generates university math problems by program synthesis and few-shot learning at human level," Proceedings of the National Academy of Sciences, vol. 119, no. 32, pp. e2123433119, 2022.

[29] P. Pataranutaporn, V. Danry, J. Leong, P. Punpongsanon, D. Novy, P. Maes, and M. Sra, "AI-generated characters for supporting personalized learning and well-being," Nature Machine Intelligence, vol. 3, no. 12, pp. 1013-1022, 2021.

[30] Q. Dong, L. Dong, K. Xu, G. Zhou, Y. Hao, Z. Sui, and F. Wei, "Large Language Model for Science: A Study on P vs. NP," arXiv preprint arXiv:2309.05689, 2023.

[31] R. Abdelghani, Y.-H. Wang, X. Yuan, T. Wang, P. Lucas, H. Sauzéon, and P.-Y. Oudeyer, "GPT-3-driven pedagogical agents to train children's curious question-asking skills," International Journal of Artificial Intelligence in Education, pp. 1-36, 2023.

[32] Y. Dan, Z. Lei, Y. Gu, Y. Li, J. Yin, J. Lin, L. Ye, Z. Tie, Y. Zhou, and Y. Wang, "EduChat: A Large-Scale Language Model-based Chatbot System for Intelligent Education," arXiv preprint arXiv:2308.02773, 2023.

[33] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big??." pp. 610-623.

[34] 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.