Proceedings of CONF-MSS 2026 Syposium: Mechanical Control and Automation
DOI: 10.54254/2755-2721/2026.BJ32952

Status and Trends of Large Model Fine-tuning in Various Scenarios

Sihan Zhang

School of Artificial Intelligence, China University of Mining & Technology-Beijing, Beijing, China
2796747021zsh@gmail.com

Abstract. Large language models (LLMs) have achieved breakthroughs in natural language processing. LLMs such as ChatGPT-4o, Claude-3.5, and deepseek-V2 have been released one after another, ushering in a blowout development of LLMs. However, general LLMs are difficult to perfectly adapt to the needs of professional fields, and fine-tuning of large models is an important way to solve this problem. This paper reviews the development of large model fine-tuning and introduces the mainstream technologies of large model fine-tuning, such as full fine-tuning and parameter efficient fine-tuning. Summarize the application and expansion of large-scale model fine-tuning in professional fields such as finance, medicine, and law in recent years. This paper discusses the necessity of large model fine-tuning in various professional fields, the main application methods of model fine-tuning, and conducts a comparative analysis before and after fine-tuning. Finally, based on the survey of large model fine-tuning in various fields, this paper gives suggestions on the application of large model fine-tuning in various fields, provides theoretical support and practical reference for large model fine-tuning in various fields, and promotes the development of large model fine-tuning in various professional fields.

Keywords: Large model, Model fine-tuning, Fine-tuning method, Model application

1. Introduction

In 2013, Mikolov et al. proposed the word vector model (Word2Vec) to simplify the previous distributed representation learning problem into a supervised learning problem [1]. It connects the geometric relationship in the vector space with the relationship between words and successfully quantifies the semantic information. It effectively shifts people from relying on manually annotated data to predict the context of words through "self-supervised learning" (CBOW model) or learning vector representation automatically through word prediction context (Skip-gram model). This feature makes it possible to label large amounts of data, and more researchers are focusing on parameter adjustment rather than large amounts of manual data labeling to obtain better language model performance.

In 2018, Jeremy Howard and Sebastian Ruder proposed ULMFiT, which was a key milestone in the history of large model fine-tuning. It was the first time that the "pre-training-fine-tuning" paradigm was systematically established and verified in the field of Natural Language Processing (NLP) [2]. ULMFiT has a very high sample efficiency, and downstream tasks can achieve good

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.BJ32952

results with very little annotation, which greatly improves data utilization efficiency and reduces dependence on large amounts of data. It pioneered three fine-tuning techniques: Discriminative Fine-tuning, Slanted Triangular Learning Rates, and Gradual Unfreezing to solve the problem of "catastrophic forgetting". ULMFiT builds a bridge connecting Word2Vec's "pre-trained word vector" and Transformer's "pre-trained model", laying a solid foundation for the explosive development of NLP [3].

The BERT model uses the Transformer encoder structure, and its pre-training process is bidirectional, which changes the previous pre-training model that could only encode text from left to right [4]. The BERT model uses the masked language model (MLM) for model pre-training, achieving bidirectional prediction of masked words, so that each word in the model is established in the context. BERT's excellent performance proves that it does not need to design a complex solution from scratch. It only needs to download a pre-trained BERT model and fine-tune it with its own professional data to achieve good results. This greatly reduces the technical threshold and cost requirements for fine-tuning, allowing people to focus on higher-quality datasets and larger-scale models.

Building on its GPT series models (GPT-1 [5], GPT-2 [6]), OpenAI continued to use the decoder-structured Transformer and autoregressive training approach to introduce the GPT-3 model [7]. GPT-3 ignited global research and attention on large language models (LLMs). GPT-3, with its massive model size, demonstrates the power of scale [8]. GPT-3 is the first model with a scale of hundreds of billions of parameters. 175 billion parameters give GPT-3 excellent capabilities in natural language processing. It requires almost no fine-tuning and only a small number of examples, and users can complete the task by constructing a prompt in natural language. By demonstrating its powerful emergent abilities, GPT-3 shifted the focus of AI development from "specialization" to "generalization," directly promoting the emergence and development of various LLMs [9].

Although the parameter size of LLMs continues to grow, the performance of general large models in specific fields (such as finance, medicine, law, etc.) still suffer from inaccuracies. Therefore, it is usually necessary to fine-tune the general large model for domain adaptation, adding professional knowledge to make the output results more consistent with the specifications of the domain and significantly improve its accuracy and reliability in specific tasks [10].

2. Large model fine-tuning methods commonly used in various fields

In various professional fields, the main methods for fine-tuning LLMs currently include: full-parameter fine-tuning and efficient parameter fine-tuning. The ultimate goal of full parameter fine-tuning and efficient parameter fine-tuning is the same, which is to perform additional training on a pre-trained large model using task-specific labeled data without changing its core network architecture, such as the number of Transformer layers and dimensions, so that it has a stronger ability to handle downstream tasks [3]. Continuous pre-training or domain-adaptive pre-training (DAPT) is a typical scenario in which full-parameter updates are applied [10]. That is, based on the general model, full-parameter pre-training is performed using domain data to obtain domain-specific models, such as BloombergGPT and BioMedLM [11, 12]. Full parameter fine-tuning updates the pre-trained model parameters (weights and biases). Updating parameters leads to a large amount of computational and storage costs for full parameter fine-tuning. Usually, full parameter fine-tuning of a model requires updating all parameters. Therefore, in the BERT era, when the amount of data was small, this was the mainstream fine-tuning method [4]. However, facing the current large models with tens or even hundreds of billions of parameters, the cost is unacceptable. Parameter Efficient Fine-tuning achieves near-full parameter fine-tuning capabilities with a very small number of

DOI: 10.54254/2755-2721/2026.BJ32952

additional training parameters (usually less than 1% of the total number of parameters), significantly reducing computational and storage costs [13]. In 2021, Edward Hu et al. creatively proposed the LoRA technology in the field of parameter efficient fine-tuning, which greatly reduced the threshold for fine-tuning large models [14]. Developers and businesses can fine-tune their own large models, rather than only large technology companies and top research institutions having the ability to fine-tune large models. Even using LoRA technology, one only needs a consumer-grade graphics card to fine-tune own large model. Moreover, a basic model can be adapted to different LoRA adapters, and loading different LoRA adapters only takes seconds. This allows the same large model to improve its accuracy for downstream tasks in specialized fields by using different LoRA adapters. LoRA technology has greatly promoted the prosperity and development of the open source community and accelerated the application of large-scale model fine-tuning in various professional fields. It has laid the foundation for subsequent technologies such as QLoRA and AdaLoRA [15, 16].

3. Application of large model fine-tuning in various professional fields

3.1. Application of large model fine-tuning in the financial field

The financial field has a massive amount of structured and unstructured data, which is filled with a large number of professional terms that are different from the general context, reflecting a high degree of specialization. Compared with other fields, the financial field has very high requirements for data accuracy and timeliness, and even requires information extraction, logical reasoning, mathematical calculations and causal judgment of the data. This requires model fine-tuning to not only digest and understand massive amounts of high-quality data and professional terminology, but also to minimize the generation of hallucinations. Although model fine-tuning is very difficult in the financial field, its impact is enormous. In the pre-big model era, there was a Loughran-McDonald financial sentiment dictionary model based on traditional machine learning models, which pioneered a financial sentiment analysis framework and constructed a specific sentiment dictionary in the financial field [17]. In the BERT era, the FinBERT model was created by using a large amount of professional text data in the financial field to continue pre-training general models [18]. Its pre-training corpus includes financial news, analysis reports and sentiment annotation data. FinBERT also follows the same Masked Language Model(MLM) goal as BERT, but retains the Next Sentence Prediction(NSP) task in some versions. Thanks to its open-source nature, researchers can directly fine-tune the FinBERT model downstream, which greatly lowers the threshold for model fine-tuning in the financial field and promotes the vigorous development of the entire field. Then, based on the large language model, through instruction fine-tuning (SFT), a large amount of high-quality financial data was injected to create a professional large model at a controllable cost [19, 20]. For example, BloombergGPT used 3630B tokens of financial and general corpus to complete pre-training, and specially constructed financial scenario instruction dataset for implementation fine-tuning, thereby generating a large model in the financial field [11].

Currently, the main trends in fine-tuning large models in the financial field are: FinGPT ecosystem and AI Agent framework [21-23]. FinGPT is not a single model, but an open source framework and ecosystem that provides a complete solution from data collection, cleaning, preprocessing to model fine-tuning. The open data features, automation of processes, low-cost and efficient low-frequency fine-tuning are the direction of further development of the popularization of finance. FinGPT minimizes the cost of a financial scenario model by basing on Parameter Efficient Fine-Tuning (PEFT), and Low-Rank Adaptation(LoRA) is its primary implementation scheme. LoRA proposes very low-rank parameter-frozen trainable matrix to top of the initial general model

DOI: 10.54254/2755-2721/2026.BJ32952

parameters to enable the model to perform significant domain adaptation even under resource-restricted settings, which drastically minimizes the cost of developing financial model, and lowers the energy and resources requirements in computing. The fine-tuning module of FinGPT enables easy usage of LoRA and its variations as a default rule and the models created can fit particular financial tasks within a short period of time. The AI Agent model is indeed a part of the financial business process itself, soliciting substantial amounts of data and external applications and becoming an AI agent that will be able to carry out investment researchers, reports generation, and trading choices on its own.

The implementation of massive model fine-tuning will go a long way to enhance more efficiency and lowering labour expenses. Financial models are able to easily structure high volumes of important data, eliminating the possibility of repetitive work of financial professionals. Besides, the model is able to track the financial markets around the globe 24/7, provide early warnings and reporting timely and allow the financial experts to provide timely countermeasures. The fine-tuned model is not aimed to substitute financial professionals but make them make more scientific and correct decisions.

3.2. Application of large model fine-tuning in the medical field

Medical industry is highly specialized, risky and highly regulated. The medical information, in addition to containing the knowledge of various fields including physiology, pathology, pharmacology, etc., is also replete with numerous technical terms, abbreviations, drugs, and symbols. The terms are not common in the general model data, and many technical terms are similar that puts a very strong pressure on the model input and output accuracy. Hallucinations and bias may pose a significant risk and threat to the safety of a patient. In comparison with other disciplines, the medical models do not only require them to be able to present the predictive results, but also present a level of interpretability to the clinicians to make it easy to prove the accuracy of the results. In addition, the issue of spelling mistakes or non-conformity writing occurs in both handwritten and electronic medical records, and it is impossible to fix the information without a significant number of professional medical specialists, which is also extremely expensive. Data privacy and security are the hardest to take care of. Since medical data is the most highly secured type of personal privacy, it is treated as the object of strict rules like HIPAA and GDPR. The storage, relaying, and consumption of clinical information should have high compliance standards. Access to large-scale clinical information to fine-tune it is highly challenging and must go through complicated desensitization procedures, authorization and security steps [24]. Medical data, even with the anonymization process, may also have the risk of reversible inference, and more encryption technology is required to enhance higher protection. All these make the fine-tuning of models in the medical profession even harder. As of 2019, several early biomedical models including BioBERT were generated by further training on medical data on the BERT [25]. BioBERT is much more effective compared to general-purpose models in identifying drugs and diseases. Nonetheless, it remains limited to high-risk interactive clinical decision-making situations. Later, due to large language model maturity and instruction tuning technology, open-source projects including MedAlpaca that leverages general models to tune instruction on medical data have appeared extending the field of application of medical models to interactive-style applications like question answering, and report generation [26]. It trains on a massive medical education data both to match medical knowledge and to integrate feedback learning to enhance the reliability and consistency of the model during clinical reasoning.

Med-PaLM2 is now based on large-scale instruction fine-tuning, the new method of instruction fine-tuning that is exclusive to the medical domain, which aligns medical knowledge, and integrates

DOI: 10.54254/2755-2721/2026.BJ32952

it with feedback learning techniques [27]. The first large scale model won the United States Medical Licensing Examination(USMLE) style question answering test was Med-PaLM2. In answer justification, reasoning chains and interpretability,Med-PaLM2 stresses the factors more in training, making it more applicable in high-risk medical contexts. Several hospitals and businesses also started to utilize services like the Azure OpenAI to refine the interface and integrate it with their own clinical data to create their own medical assistant programs [28]. To address the paradox of the necessity of large-volume high-quality data, as well as necessities of data privacy and security in fine-tuning large medical models, Federated Learning offers a compliant and viable technical pathway of large medical models [29,30]. Every data is localized in every hospital and is not transferred to the central server. Rather, the global model is distributed by the central server to every hospital. Once the hospital receives a new model, then it encrypts the new model parameters (weights and biases, etc.) and uploads them to the central server so that these are aggregated.

The model fine-tuning has proven the opportunity to apply Guanfan in the medical area in a significant way that can enhance the efficiency of medical facilities and also serve a wide range of functionalities, including, but not limited to, first-triage, clinical decision support, question and answer in medical knowledge, as well as, personalized diagnosis and treatment. Nevertheless, medical information is very sensitive, clinical judgment is very risky among them, needs high interpretability, data annotation expenses are very much high and regulatory compliance is very rigorous. All these variables have determined that industrialization of huge medical models has to create a compromise between technological provisions, safety requirement, and ethical principles.

3.3. Application of large model fine-tuning in the legal field

Although there is a huge amount of legal text data in the legal field, such as legal provisions and judicial documents, there is relatively little high-quality data suitable for fine-tuning instructions [31]. It requires manual writing and annotation by legal experts, which has a high threshold and cost. Different judges will have different sentencing styles and reasoning logic, so the model must be aligned with the core values of the legal system, such as fairness, equality, and the presumption of innocence. The development of model fine-tuning in the legal field has promoted the continuous improvement of technology efficiency, specialization and human-machine collaboration. Between 2016 and 2018, deep learning began to be applied to the legal field. At this time, the research was mostly exploratory, and the models were often named after tasks or techniques. These models predict case verdicts using Word2Vec word vectors and convolutional neural network models (CNN). From 2018 to 2020, Legal-BERT was released by research teams from institutions such as University College London (UCL) and the University of Sheffield [32]. It used large-scale legal texts from the United States, the United Kingdom, and the European Union for further pre-training and fine-tuning based on the general model (BERT), becoming one of the most influential models in the legal field in the early days.

Currently, GPT-4 itself ranks in the top 90% of all test takers on the UBE without any specific fine-tuning [33,34] . Many legal applications achieve output by calling GPT-4's API. But there are still many companies and teams working on fine-tuning large models in the legal field. For example, Harvey AI, developed by Harvey, not only uses a large amount of high-quality data to fine-tune instructions, but also uses feedback reinforcement learning by a team of top lawyers to develop an AI assistant for workers in the legal field [35]. CoCounsel model, created by CaseText has been streamlined and refined in different aspects like legal document review and study of a contract which makes it able to have e-Discovery and readability features.

DOI: 10.54254/2755-2721/2026.BJ32952

Out of the largest issues with scaling large models in legal domain, model hallucination is the biggest problem [36]. The most recent key direction is to apply retrieval-augmented generation (RAG) in order to solve this issue [37]. Facilitating large models to search external professional knowledge bases is an effective way of reducing the model hallucinations and enhancing the accuracy and credibility of output answers. Self-verification and tracing of the big model is possible. The origin of any answer of the big model is automatically indicated, and it becomes easier to search and confirm by the users. To work out the response strategy of the model, the strength of the constraint of the must answer of the reasoning stage can be minimized, by which the model can directly produce strong responses to the format of cannot answer in case of lack of knowledge or under the condition of making no judgment without necessarily producing potentially false response.

Large-scale model fine-tuning will help legal workers free their time and resources on the time-consuming and boring activities like text review and legal search, significantly enhancing their work performance. It may also help the judges to make a decision and reference to the rulings in other related cases, which will aid in enhancing uniformity of judicial principles and bringing about social fairness and justice.

4. Future research directions

4.1. Multimodal data

The data is not a text only, but contains other types of data like images, audio and video in the process of fine-tuning of a model and its use in other areas (such as in visual arts, a visual channel) [38]. It is possible to enhance and deepen the merging and calibration of the multimodal data to a considerable extent to enhance the knowledge and creation of the multimodal data by the model. One of the beneficial research directions to the application of model fine-tuning in different professional domains is the improvement of the capacity to process and integrate multimodal information.

4.2. Federated learning

In federated learning, various parties are able to learn jointly without the exchange of raw data [29]. The information in the field of work is very secretive and as the new wave of data privacy gains more exposure, it is starting to get all the more difficult to access the data directly. The issue of raw data leakage can be prevented through federated learning. This preserves the privacy of the data and also the model possesses sufficient data to train and fine-tune.

4.3. AI agent

The long-term target of implementation model fine-tuning in multiple sectors is to emerge as a professional intelligent being such as an AI agent that should be able to break down work, formulate plans and external tools. It possesses some level of autonomy and can help the professional personnel to accomplish certain tasks as a professional assistant, therefore enhancing the work efficiency and the productivity of professional personnel [22]. The studies of AI agents are very relevant to the usage of model fine-tuning in different spheres.

DOI: 10.54254/2755-2721/2026.BJ32952

4.4. Retrieval-augmented generation

Through the RAG technology, the model obtains the latest and accurate data by calling an external professional knowledge base, thereby effectively suppressing the model's hallucination [36,37] . The RAG system increases the transparency and credibility of answers by marking the sources of the generated answers. In addition, it is costly to fine-tune the model to obtain new data. By updating external professional databases through RAG, data can be updated without updating the model. RAG has become one of the key technologies for model fine-tuning in various fields.

5. Conclusion

This paper reviews the development of model fine-tuning and conducts an in-depth analysis of the three high-value and high-barrier fields of finance, healthcare, and law. It demonstrates that the development paths of large-scale model fine-tuning in various professional fields are basically similar: from the foundational "pre-training-fine-tuning" paradigm, to efficient parameter fine-tuning and instruction fine-tuning for large models, and then to reinforcement learning fine-tuning and retrieval-enhanced generation. As professional field models are fine-tuned and updated, the problems that arise are basically similar, such as high-quality data acquisition, data security, model "hallucinations", and the construction of professional field ecosystems. Although the development of big models has brought challenges in terms of data security and responsibility determination, the ultimate goal of big models is not to replace humans, but to become a tool for humans to help them make more scientific decisions. The development of large-scale model fine-tuning in various professional fields is an arduous and lengthy process. People should objectively understand the development laws of large model fine-tuning in various professional fields, and actively and steadily promote the application of large model fine-tuning technology in various professional fields. Ultimately, it will empower all industries and unleash the tremendous productivity of the intelligent era.

References

[1] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv: 1301.3781.
[2] Howard, J., & Ruder, S. (2018). Universal language model fine-tuning for text classification. arXiv preprint arXiv: 1801.06146.
[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). 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) (pp. 4171-4186).
[5] Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.
[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. OpenAI blog, 1(8), 9.
[7] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.
[8] Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... & Amodei, D. (2020). Scaling laws for neural language models. arXiv preprint arXiv: 2001.08361.
[9] Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv: 2206.07682.

DOI: 10.54254/2755-2721/2026.BJ32952

[10] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't stop pretraining: Adapt language models to domains and tasks. arXiv preprint arXiv: 2004.10964.
[11] Wu, S., Irsoy, O., Lu, S., Dabravolski, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). Bloomberggpt: A large language model for finance. arXiv preprint arXiv: 2303.17564.
[12] Bolton, E., Venigalla, A., Yasunaga, M., Hall, D., Xiong, B., Lee, T., ... & Manning, C. D. (2024). Biomedlm: A 2.7 b parameter language model trained on biomedical text. arXiv preprint arXiv: 2403.18421.
[13] Ding, N., Qin, Y., Yang, G., Wei, F., Yang, Z., Su, Y., ... & Sun, M. (2023). Parameter-efficient fine-tuning of large-scale pre-trained language models. Nature machine intelligence, 5(3), 220-235.
[14] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.
[15] Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115.
[16] Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., ... & Zhao, T. (2023). Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv: 2303.10512.
[17] Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of finance, 66(1), 35-65.
[18] Liu, Z., Huang, D., Huang, K., Li, Z., & Zhao, J. (2021, January). Finbert: A pre-trained financial language representation model for financial text mining. In Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence (pp. 4513-4519).
[19] Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... & Le, Q. V. (2021). Finetuned language models are zero-shot learners. arXiv preprint arXiv: 2109.01652.
[20] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744.
[21] Liu, X. Y., Wang, G., Yang, H., & Zha, D. (2023). Fingpt: Democratizing internet-scale data for financial large language models. arXiv preprint arXiv: 2307.10485.
[22] Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., ... & Gui, T. (2025). The rise and potential of large language model based agents: A survey. Science China Information Sciences, 68(2), 121101.
[23] Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Wang, J., ... & Schmidhuber, J. (2023, August). MetaGPT: Meta programming for a multi-agent collaborative framework. In The Twelfth International Conference on Learning Representations.
[24] Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... & Liu, H. (2018). Clinical information extraction applications: a literature review. Journal of biomedical informatics, 77, 34-49.
[25] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36(4), 1234-1240.
[26] Han, T., Adams, L. C., Papaioannou, J. M., Grundmann, P., Oberhauser, T., Löser, A., ... & Bressem, K. K. (2023). MedAlpaca--an open-source collection of medical conversational AI models and training data. arXiv preprint arXiv: 2304.08247.
[27] Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Amin, M., ... & Natarajan, V. (2025). Toward expert-level medical question answering with large language models. Nature Medicine, 31(3), 943-950.
[28] Microsoft Corporation. (2025). Azure OpenAI Service. Retrieved September 18, 2025, from Azure Machine Learning - ML as a Service |Microsoft Azure
[29] McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In Artificial intelligence and statistics (pp. 1273-1282). PMLR.
[30] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ digital medicine, 3(1), 119.
[31] Sulea, O. M., Zampieri, M., Malmasi, S., Vela, M., Dinu, L. P., & Van Genabith, J. (2017). Exploring the use of text classification in the legal domain. arXiv preprint arXiv: 1710.09306.
[32] Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). LEGAL-BERT: The muppets straight out of law school. arXiv preprint arXiv: 2010.02559.
[33] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ... & McGrew, B. (2023). Gpt-4 technical report. arXiv preprint arXiv: 2303.08774.
[34] Martínez, E. (2024). Re-evaluating GPT-4's bar exam performance. Artificial intelligence and law, 1-24.
[35] Harvey Company. (2025). Harvey AI. Retrieved September 18, 2025, from https: //www.harvey.ai
[36] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM computing surveys, 55(12), 1-38.

DOI: 10.54254/2755-2721/2026.BJ32952

[37] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing systems, 33, 9459-9474.
[38] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.