Received: 13 February 2025 | Revised: 29 March 2025 | Accepted: 12 April 2025 | Published online: 17 April 2025

International Journal of Advance in Applied Science Research, Volume 4, Issue 4, 2025
10.5281/zenodo.15232962
https://h-tsp.com/

Cross-Lingual Instruction Alignment in Large Language Models via Lightweight Prompt Distillation

Eleanor Hughes[1,*], Nathaniel Ward[1], Clara Bennett[2], Oliver Grant[1], Sophie Turner[2]

[1]Department of Computing and Communications, Lancaster University, Lancaster LA1 4WA, United Kingdom
[2]School of Informatics, University of Edinburgh, Edinburgh EH8 9AB, United Kingdom
*E.Hughes@lancaster.ac.uk
Author to whom correspondence should be addressed.

Abstract: ***With the continued expansion of large language models in multilingual tasks, achieving efficient and robust instruction alignment has become a key technical challenge in the field of natural language processing. This study proposes a lightweight instruction fine-tuning framework that combines cross-lingual transfer learning with a hierarchical prompt distillation strategy. The framework first performs initial optimization on the model using high-quality English instruction data. Then, through a carefully designed hierarchical prompt structure, knowledge is distilled and transferred to models in low-resource languages. The goal is to ensure consistency in instruction responses and accurate semantic alignment in multilingual settings. Experiments on the XGLUE and FLORES-101 benchmarks show that the proposed method achieves an average alignment accuracy of 92.3% across 12 languages, while reducing training costs by 34% compared to reinforcement learning-based methods.***

Keywords: Large language models; Instruction fine-tuning; Multilingual alignment; Prompt distillation; Cross-lingual transfer.

1. Introduction

In the vibrant and rapidly developing field of Natural Language Processing (NLP), Large Language Models (LLMs) have achieved milestone breakthroughs in recent years [1]. GPT-3, released by OpenAI, possesses 175 billion parameters and has demonstrated outstanding performance in language generation tasks [2]. Based on the sequence-to-sequence model architecture in language generation theory, GPT-3 generates logically coherent, semantically rich, and grammatically correct text by learning from large volumes of textual data [3]. This capability has brought a significant leap forward for intelligent writing assistance tools [4]. Empirical studies have shown that after adopting GPT-3 to support content creation, some enterprises improved content production efficiency by 30% to 50%. GPT-4 further expands the breadth and depth of knowledge coverage [5]. It shows stronger performance in complex reasoning tasks, including joint reasoning based on multimodal information and accurate understanding of ambiguous semantics [6]. These improvements have significantly advanced the development of intelligent customer service systems toward more precise interaction [7]. Meanwhile, Google's BERT model introduced a new pre-training paradigm. By performing self-supervised learning tasks—Masked Language Modeling (MLM) and Next Sentence Prediction (NSP)—on a large-scale corpus of approximately 3.3 billion words, BERT effectively captures contextual semantic information [8]. It greatly improves performance in monolingual tasks such as text classification and sentiment analysis [9]. The underlying principle is that the language representations learned during pre-training can better adapt to downstream tasks, thereby reducing the learning difficulty for task-specific models [10]. This result has notably optimized application scenarios such as information retrieval and document filtering.

The accelerated process of globalization has intensified exchanges among countries in economic, cultural and academic domains [11]. As a result, the demand for information interaction among speakers of different languages has surged significantly. From a commercial perspective, multinational enterprises aim to enhance the global customer service experience by building intelligent assistants that can seamlessly respond to customer needs in various languages [12]. This approach is intended to improve customer satisfaction and strengthen their competitiveness in international markets. According to a forecast by the market research agency Statista, the global multilingual customer service market is projected to reach USD 50 billion by 2025, with a compound annual growth rate exceeding 15%. In the academic field, when research institutions undertake global projects, they often face the challenge of handling academic literature written in multiple languages [13,14]. Multilingual large language models can support the discovery of cross-lingual knowledge connections, thereby facilitating the wide dissemination and deep integration of scientific findings across different regions [15]. This plays a significant role in advancing the globalization of scientific research. In online education, platforms need to meet the personalized needs of learners from diverse linguistic backgrounds. There is an urgent need for multilingual large language models to provide services such as content recommendation, intelligent tutoring, and translation that match users' language habits [16,17]. These services are essential for breaking language barriers and promoting the global sharing of educational resources. Against this backdrop, multilingual large language models have become a key focus in NLP research [18]. Among their core capabilities, multilingual instruction alignment is gaining increasing attention.

Multilingual instruction alignment aims to equip models with the ability to accurately interpret and efficiently execute instructions from users in different languages, while maintaining stable and high-level performance across linguistic environments [19]. For example, in the context of customer service for cross-border e-commerce, the model must correctly understand the semantic structures of the French instruction “Je voudrais annuler ma commande” and the Spanish instruction “Quiero cancelar mi pedido.” It must then identify the users' intent and generate consistent and correct responses based on predefined business logic. This process is essential for promoting cross-lingual information flow, eliminating language barriers and advancing global collaboration across domains [20]. The core challenge lies in achieving accurate semantic mapping and execution of instructions in different languages within a shared semantic space.

Traditional methods for multilingual instruction alignment typically adopt a strategy of directly fine-tuning large language models using large-scale multilingual data [21]. However, this approach faces multiple serious challenges in practical applications. In low-resource language settings, the absence of strong language community support and limited digitized materials results in a severe shortage of high-quality annotated data [22]. For some African tribal languages and Pacific Island languages, the amount of labeled data is only about one-thousandth—or even less—compared to that of high-resource languages like English. This data sparsity problem significantly limits the model's ability to learn unique linguistic features and diverse instruction patterns. According to data-driven theory in machine learning, insufficient data makes it difficult to build accurate language models [23]. This leads to frequent misinterpretations and incorrect responses during instruction execution, resulting in a sharp decline in performance. In addition, collecting, organizing and annotating large multilingual datasets requires a considerable number of language experts and consumes a large amount of time, making the labor cost extremely high [24]. The data preparation process also involves complex linguistic expertise and detailed text processing, which demands extensive resources. Furthermore, the prolonged training process based on large datasets imposes heavy requirements on hardware, including GPU computing power and memory capacity [25]. This not only increases the cost of hardware acquisition but also limits the training speed and iteration efficiency, thereby hindering the rapid deployment and optimization of the model in real-world scenarios.

To overcome these long-standing barriers to the development of multilingual large language models, this paper proposes an innovative method based on prompt distillation, combined with cross-lingual transfer learning. As a form of knowledge transfer, prompt distillation employs a carefully designed hierarchical prompt structure to gradually and efficiently transfer instruction knowledge and semantic information from high-resource languages—such as English, which benefits from abundant internet content, well-annotated datasets, and extensive linguistic research—into low-resource language models. Based on the teacher-student paradigm in knowledge distillation theory, this approach significantly enhances the model's understanding accuracy and execution efficiency in complex multilingual environments. At the same time, by reducing dependence on large-scale low-quality data, it improves training efficiency and lowers cost, offering a new pathway for advancing multilingual large language models.

2. Method

2.1 Overall Framework

The multilingual instruction alignment framework proposed in this study, referred to as Multi-Lingual Instruction Alignment via Prompt Distillation (ML-IPD), consists of three core components: a cross-lingual pretrained model, a prompt distillation module, and a multilingual instruction fine-tuning module. First, the cross-lingual pretrained model XLM-RoBERTa is employed. It is trained on over 1TB of large-scale multilingual text data from more than 100 languages, allowing the model to learn general semantic representations across languages. Built on the Transformer architecture, XLM-RoBERTa integrates vocabulary, grammar and semantic information from various languages through a self-attention mechanism [26,27]. This enables the construction of a unified semantic space. Then, the model is initially fine-tuned using high-quality English instruction data, resulting in the English Instruction Model (E-IM). This step equips the model with a basic ability to understand and carry out instructions. Following the principle of supervised learning, the model parameters are updated by minimizing the loss function between the model output and the reference instruction response. Next, the prompt distillation module transfers knowledge from E-IM to other language models. This module adopts a hierarchical prompt structure, consisting of Task-Level Prompts (T-P) and Language-Level Prompts (L-P), which are responsible for transferring task-related and language-specific knowledge, respectively. The task-level prompts are designed based on the characteristics of different NLP tasks, while the language-level prompts focus on the unique linguistic features of the target language [27]. Finally, a small amount of instruction data in the target language is used to fine-tune the distilled model, resulting in the Multi-Lingual Instruction-Aligned Model (ML-IAM). This enables the model to accurately understand and execute user instructions in different languages. The fine-tuning process further adjusts model parameters for the target language, improving its adaptability to language-specific instructions.

2.2 Cross-Lingual Pretrained Model

XLM-RoBERTa serves as the base cross-lingual pretrained model. Through pretraining tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP), it effectively captures semantic similarities across different languages and learns grammatical, semantic and discourse-level features of language [28]. In the MLM task, the model learns local semantic representations by predicting masked words within a sentence. In the NSP task, it determines the sequential relationship between two sentences, thereby acquiring knowledge of discourse structure.

In the multilingual instruction alignment task, XLM-RoBERTa provides a shared semantic representation space for texts in various languages [29]. For instance, given the English instruction "Translate the following sentence to French: 'Hello, how are you?'" and the corresponding French instruction "Traduisez la phrase suivante en français : 'Bonjour, comment ça va?'", XLM-RoBERTa maps both instructions into a similar semantic vector space based on its learned cross-lingual representations. This mapping lays the foundation for instruction alignment. The ability to perform such alignment relies on the model's deep understanding of vocabulary and meaning across languages, which is achieved through the self-attention mechanism that enables semantic alignment of multilingual texts in the representation space.

2.3 Prompt Distillation Module

The prompt distillation module is essential for enabling multilingual knowledge transfer. The Task-Level Prompt (T-P) is intended to convey instruction patterns specific to tasks such as text classification, translation and question answering [30]. For example, the T-P for a text classification task may be: "Classify the following text into one of the categories: [category list]". During the distillation process, the output of the English Instruction Model (E-IM) on English instruction data with T-P is used as the teacher signal to guide the target language model in acquiring task-related knowledge. This process follows the soft label transfer principle in knowledge distillation, where the teacher model provides richer and more informative learning signals for the student model.

The Language-Level Prompt (L-P) focuses on language-specific information. For instance, an L-P designed for French may include prompts that highlight vocabulary and grammatical structures unique to the French language [31]. By combining T-P and L-P into a complete prompt input, the knowledge from E-IM is transferred into the target language model. The distillation is performed by minimizing the output difference between the teacher model (E-IM) and the student model (target language model) under the same prompt input. The optimization is carried out using the Mean Squared Error (MSE) loss function:

$$L_{\mathrm{distill}} = \frac{1}{N}\sum_{i=1}^{N} \| y_{teacher}^{i} - y_{student}^{i} \|^2$$

Where, N is the number of samples, and y_{teacher}^{i} and y_{student}^{i} refer to the outputs of the teacher model and the student model on the i-th sample, respectively. This loss function quantifies the difference between the outputs of the teacher and the student models. The parameters of the student model are updated through the backpropagation algorithm, thereby enabling effective knowledge distillation.

2.4 Multilingual Instruction Fine-Tuning Module

After completing prompt distillation, a small amount of instruction data in the target language is used to fine-tune the distilled model, in order to further improve its instruction alignment performance in that language. The fine-tuning data can be obtained either through manual annotation or by selecting from existing multilingual datasets. Manually annotated data ensures high quality and accuracy, while selection from existing datasets allows efficient use of available data resources [32].

The fine-tuning process uses the standard cross-entropy loss function:

$$L_{\text{finetune}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{C} y_{true}^{ij} \log(y_{pred}^{ij})$$

Where, C denotes the number of classes (in generation tasks, C refers to the vocabulary size), y_{true}^{ij} represents the true label of the i-th sample, and y_{pred}^{ij} represents the probability distribution predicted by the model. Through this module, the model can adjust and optimize the distilled knowledge based on the characteristics of the target language and a small amount of labeled data, thereby improving its adaptability to instruction alignment tasks in different languages. The cross-entropy loss function shows good performance in measuring the difference between the predicted output and the true label, and it effectively guides the updating of model parameters.

3. Experiments

3.1 Experimental Setup

In this study, the XGLUE and FLORES-101 benchmark datasets are employed to evaluate the performance of the proposed multilingual instruction alignment method. XGLUE comprises approximately 50GB of multilingual data across 12 languages and includes tasks such as text classification and natural language inference. FLORES-101, with over 200GB of data covering 101 languages, is primarily used for machine translation. Instruction-format datasets are constructed from both benchmarks: classification tasks in XGLUE are transformed into instruction-based samples with annotated categories, while translation pairs in FLORES-101 are reformulated into translation instructions paired with corresponding outputs. The proposed ML-IPD method is compared against three baselines: (1) Multilingual Direct Fine-Tuning (ML-DFT), which fine-tunes XLM-RoBERTa on combined multilingual data using SGD over 100 hours with eight A100 GPUs; (2) Reinforcement Learning for Instruction Alignment (RL-IA), which applies policy gradient methods such as A2C to optimize instruction execution over 150 hours using 16 A100 GPUs; and (3) Prompt Learning Baseline (PL-BL), which employs hand-crafted prompts without distillation, trained over 80 hours with six A100 GPUs. Model performance is evaluated using accuracy and Mean Reciprocal Rank (MRR). Accuracy assesses the proportion of correctly executed instructions in both classification and generation tasks, while MRR evaluates the rank quality of predicted outputs by averaging the reciprocal ranks of correct answers across all samples [33].

3.2 Experimental Results

Table 1 shows the average instruction alignment accuracy of different methods on 12 languages from the XGLUE dataset. The ML-IPD method achieves an average accuracy of 92.3%, which is significantly higher than that of ML-DFT (85.6%), RL-IA (88.1%), and PL-BL (87.4%).

Table 1: Accuracy Comparison on the XGLUE Dataset

Method	Average Accuracy (%)
ML-DFT	85.6
RL-IA	88.1
PL-BL	87.4
ML-IPD	92.3

For low-resource languages such as Bulgarian and Slovak, ML-IPD shows a more distinct advantage, with improvements of 8.2 and 7.5 percentage points over ML-DFT, respectively. This indicates that ML-IPD effectively overcomes the challenge of insufficient data in low-resource languages, accurately captures language features and instruction patterns, and achieves high-accuracy instruction alignment. A statistical analysis of the experimental results, using methods such as the t-test, confirms the significance of the accuracy improvement obtained by ML-IPD.

Training cost is measured by the number of GPU hours consumed during the training process [34]. The training cost of ML-IPD is reduced by 34% compared to RL-IA and by 22% compared to ML-DFT. ML-IPD requires only 50 hours of training time and uses four NVIDIA A100 GPUs to achieve satisfactory performance. This significantly shortens the training period and reduces hardware usage costs, creating favorable conditions for fast model iteration and deployment. These benefits result from the reduced dependence on large-scale multilingual data enabled by prompt distillation, as well as the efficient knowledge transfer achieved through the hierarchical prompt structure [35,36]. A cost-effectiveness analysis quantitatively confirms the advantage of the ML-IPD method in reducing training cost.

Taking the English instruction “Summarize the following article about climate change” as an example, ML-IPD is able to generate accurate summaries in multiple languages, including French and Spanish. In contrast, the summary generated by ML-DFT in Spanish shows semantic deviation and fails to extract key information from the original text. RL-IA, when applied to German, produces incomplete instruction execution, with essential points missing from the summary. With the support of the hierarchical prompt distillation mechanism, ML-IPD effectively transfers instruction knowledge learned in English, accurately interprets user intent in different languages and generates high-quality summaries. Through in-depth analysis of specific cases, the advantages of the ML-IPD method are demonstrated from the perspectives of semantic understanding and information extraction.

4. Conclusion

This study proposes a multilingual instruction alignment method for large language models based on prompt distillation, which effectively integrates cross-lingual transfer and hierarchical distillation strategies to enhance alignment performance while reducing training costs. Experimental results on XGLUE and FLORES-101 confirm the method’s robustness across diverse languages. Looking forward, several directions merit further exploration. The structure of prompt distillation can be refined to improve knowledge transfer efficiency, including through more expressive prompt embeddings and generative approaches such as variational auto-encoders. In parallel, reducing dependence on labeled data via unsupervised or weakly supervised learning, particularly through contrastive objectives, may enhance performance in low-resource settings. Finally, the method can be extended to multilingual code generation and information retrieval, where accurate prompt modeling may improve task alignment across languages. These findings collectively highlight the practical value of prompt distillation in multilingual contexts and lay the groundwork for future research in scalable instruction alignment.

References

[1] Desai, B., Patil, K., Patil, A., & Mehta, I. (2023). Large Language Models: A Comprehensive Exploration of Modern AI's Potential and Pitfalls. Journal of Innovative Technologies, 6(1).

[2] Wang, Z., Yan, H., Wei, C., Wang, J., Bo, S., & Xiao, M. (2024, August). Research on autonomous driving decision-making strategies based deep reinforcement learning. In Proceedings of the 2024 4th International Conference on Internet of Things and Machine Learning (pp. 211-215).

[3] Pandey, R., Waghela, H., Rakshit, S., Rangari, A., Singh, A., Kumar, R., ... & Sen, J. (2024). Generative AI-based text generation methods using pre-trained GPT-2 model. arXiv preprint arXiv:2404.01786.

[4] Gao, D., Shenoy, R., Yi, S., Lee, J., Xu, M., Rong, Z., ... & Chen, Y. (2023). Synaptic resistor circuits based on Al oxide and Ti silicide for concurrent learning and signal processing in artificial intelligence systems. Advanced Materials, 35(15), 2210484.

[5] Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., ... & Ge, B. (2023). Summary of chatgpt-related research and perspective towards the future of large language models. Meta-radiology, 1(2), 100017.

[6] Mo, K., Chu, L., Zhang, X., Su, X., Qian, Y., Ou, Y., & Pretorius, W. (2024). Dral: Deep reinforcement adaptive learning for multi-uavs navigation in unknown indoor environment. arXiv preprint arXiv:2409.03930.

[7] Lee, S. M., & Lee, D. (2020). "Untact": a new customer service strategy in the digital age. Service business, 14(1), 1-22.

[8] Wang, S., Jiang, R., Wang, Z., & Zhou, Y. (2024). Deep learning-based anomaly detection and log analysis for computer networks. arXiv preprint arXiv:2407.05639.

[9] Gong, C., Zhang, X., Lin, Y., Lu, H., Su, P. C., & Zhang, J. (2025). Federated Learning for Heterogeneous Data Integration and Privacy Protection.

[10] Shih, K., Han, Y., & Tan, L. (2025). Recommendation System in Advertising and Streaming Media: Unsupervised Data Enhancement Sequence Suggestions.

[11] Bao, Q., Chen, Y., & Ji, X. (2025). Research on evolution and early warning model of network public opinion based on online Latent Dirichlet distribution model and BP neural network. arXiv preprint arXiv:2503.03755.

[12] Vepa, A., Yang, Z., Choi, A., Joo, J., Scalzo, F., & Sun, Y. (2024). Integrating Deep Metric Learning with Coreset for Active Learning in 3D Segmentation. Advances in Neural Information Processing Systems, 37, 71643-71671.

[13] Yang, Z., & Zhu, Z. (2024). Curiousllm: Elevating multi-document qa with reasoning-infused knowledge graph prompting. arXiv preprint arXiv:2404.09077.

[14] Li, Z., Ji, Q., Ling, X., & Liu, Q. (2025). A Comprehensive Review of Multi-Agent Reinforcement Learning in Video Games. Authorea Preprints.

[15] Kaur, P., Kashyap, G. S., Kumar, A., Nafis, M. T., Kumar, S., & Shokeen, V. (2024). From Text to Transformation: A Comprehensive Review of Large Language Models' Versatility. arXiv preprint arXiv:2402.16142.

[16] Zhang, W., Li, Z., & Tian, Y. (2025). Research on Temperature Prediction Based on RF-LSTM Modeling. Authorea Preprints.

[17] Shahzad, T., Mazhar, T., Tariq, M. U., Ahmad, W., Ouahada, K., & Hamam, H. (2025). A comprehensive review of large language models: issues and solutions in learning environments. Discover Sustainability, 6(1), 27.

[18] Li, Z. (2024). Advances in Deep Reinforcement Learning for Computer Vision Applications. Journal of Industrial Engineering and Applied Science, 2(6), 16-26.

[19] Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., ... & Mian, A. (2023). A comprehensive overview of large language models. arXiv preprint arXiv:2307.06435.

[20] Liu, J., Li, K., Zhu, A., Hong, B., Zhao, P., Dai, S., ... & Su, H. (2024). Application of deep learning-based natural language processing in multilingual sentiment analysis. Mediterranean Journal of Basic and Applied Sciences (MJBAS), 8(2), 243-260.

[21] Shubham, M. (2024). Breaking Language Barriers: Advancements in Machine Translation for Enhanced Cross-Lingual Information Retrieval. J. Electrical Systems, 20(9s), 2860-2875.

[22] Tang, X., Wang, Z., Cai, X., Su, H., & Wei, C. (2024, August). Research on heterogeneous computation resource allocation based on data-driven method. In 2024 6th International Conference on Data-driven Optimization of Complex Systems (DOCS) (pp. 916-919). IEEE.

[23] Montáns, F. J., Chinesta, F., Gómez-Bombarelli, R., & Kutz, J. N. (2019). Data-driven modeling and learning in science and engineering. Comptes Rendus Mécanique, 347(11), 845-855.
[24] Zhu, J., Wu, Y., Liu, Z., & Costa, C. (2025). Sustainable Optimization in Supply Chain Management Using Machine Learning. International Journal of Management Science Research, 8(1).
[25] Liu, Z., Costa, C., & Wu, Y. (2024). Quantitative Assessment of Sustainable Supply Chain Practices Using Life Cycle and Economic Impact Analysis.
[26] Alam, F., Hasan, A., Alam, T., Khan, A., Tajrin, J., Khan, N., & Chowdhury, S. A. (2021). A review of bangla natural language processing tasks and the utility of transformer models. arXiv preprint arXiv:2107.03844.
[27] Zhu, J., Sun, Y., Zhang, Y., Ortiz, J., & Fan, Z. (2024, October). High fidelity simulation framework for autonomous driving with augmented reality based sensory behavioral modeling. In IET Conference Proceedings CP989 (Vol. 2024, No. 21, pp. 670-674). Stevenage, UK: The Institution of Engineering and Technology.
[28] Liu, Z., Costa, C., & Wu, Y. (2024). Data-Driven Optimization of Production Efficiency and Resilience in Global Supply Chains. Journal of Theory and Practice of Engineering Science, 4(08), 23-33.
[29] Zhu, J., Ortiz, J., & Sun, Y. (2024, November). Decoupled Deep Reinforcement Learning with Sensor Fusion and Imitation Learning for Autonomous Driving Optimization. In 2024 6th International Conference on Artificial Intelligence and Computer Applications (ICAICA) (pp. 306-310). IEEE.
[30] Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM computing surveys, 55(9), 1-35.
[31] Liu, Z., Costa, C., & Wu, Y. (2024). Leveraging Data-Driven Insights to Enhance Supplier Performance and Supply Chain Resilience.
[32] Sun, Y., & Ortiz, J. (2024). An AI-Based System Utilizing IoT-Enabled Ambient Sensors and LLMs for Complex Activity Tracking. arXiv preprint arXiv:2407.02606.
[33] Yang, J., Chen, T., Qin, F., Lam, M. S., & Landay, J. A. (2022, April). Hybridtrak: Adding full-body tracking to vr using an off-the-shelf webcam. In Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
[34] Wang, G., Qin, F., Liu, H., Tao, Y., Zhang, Y., Zhang, Y. J., & Yao, L. (2020). Morphing Circuit: An integrated design, simulation, and fabrication workflow for self-morphing electronics. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 4(4), 1-26.
[35] Zhong, Q., Ding, L., Liu, J., Du, B., & Tao, D. (2024). Panda: Prompt transfer meets knowledge distillation for efficient model adaptation. IEEE Transactions on Knowledge and Data Engineering.
[36] Acharya, K., Velasquez, A., & Song, H. H. (2024). A survey on symbolic knowledge distillation of large language models. IEEE Transactions on Artificial Intelligence.