Transactions on Computational and Scientific Methods | Vo. 4, No. 5, 2024
ISSN: 2998-8780
https://pspress.org/index.php/tcsm
Pinnacle Science Press

Layer-Wise Structural Mapping for Efficient Domain Transfer in Language Model Distillation

Xuehui Quan
University of Washington, Seattle, USA
quanxh1228@gmail.com

Abstract: This paper addresses the challenges of high computational cost and low semantic transfer efficiency in adapting large language models to specific domains. It proposes a domain-oriented knowledge distillation framework for large language models. The framework employs a teacher-student architecture to enable model compression and knowledge transfer. On this basis, it incorporates a structural alignment mechanism and a domain-aware module to enhance the student model's ability to represent domain-specific semantic structures. Specifically, the teacher model first constructs a domain representation based on the raw input. This representation is then projected into a unified semantic space through structural mapping. At the same time, the student model is guided to learn semantic representations and domain features layer by layer. To improve semantic compression efficiency, the student model integrates a multi-granularity aggregation mechanism. This component structurally fuses semantic information, enhancing the compactness and consistency of representations. In the experimental section, multiple sensitivity experiments are designed to evaluate the impact of distillation depth, projection dimension, and sampling strategy. The evaluation focuses on the student model's ability to align semantics and model domain features. Comparative analysis based on real-world domain datasets shows that the proposed method outperforms several mainstream distillation baselines. It achieves better performance in semantic retention, structural consistency, and model efficiency. These results confirm the effectiveness and robustness of the proposed approach in domain adaptation tasks.

Keywords: Knowledge distillation, semantic compression, knowledge transfer, domain modeling

1. Introduction

With the continuous advancement of artificial intelligence, large language models (LLMs) have become a key technology driving innovation in intelligent systems due to their powerful capabilities in natural language understanding and generation[1]. However, general-purpose LLMs are usually trained on large-scale general corpora and lack deep adaptation to the linguistic features and knowledge systems of specific domains[2]. As a result, they often fail to meet the high demands for precision and professionalism in vertical domains such as law, finance, and biomedical fields. Therefore, how to enable effective domain adaptation for LLMs has become an urgent issue in natural language processing. Under constraints such as data privacy, training cost, and model size, developing an efficient, controllable, and generalizable domain adaptation mechanism is particularly critical.

Although mainstream LLMs offer broad language understanding abilities, their large parameter size leads to high deployment costs and significant resource consumption. This limits their practical application in resource-constrained scenarios. In contrast, small models are more deployable and efficient but often underperform in complex tasks and domain-specific processing. Against this backdrop, knowledge distillation has shown great potential. It transfers the language knowledge, reasoning abilities, and domain-specific insights from a large teacher model to a smaller student model. This approach maintains performance while reducing computational demands and supports fast domain adaptation[3].

However, general knowledge distillation methods often ignore domain-specific differences in representational structures, semantic distribution, and conceptual expression. This can cause the student model to lose critical semantic capabilities during transfer. Furthermore, domain corpora are often sparse, imbalanced, and fragmented. These challenges make it difficult for general distillation strategies to capture domain-specific knowledge patterns[4]. A domain-oriented knowledge distillation framework should model domain features at the representation level and integrate domain-sensitive mechanisms into the distillation process. This ensures accurate semantic transfer and a deeper understanding of professional content.

To address these challenges, researchers need to explore a distillation framework that integrates structural guidance, knowledge fusion, and dynamic control. This framework should support fine-grained knowledge alignment across models at different scales and representation levels. It should also incorporate external knowledge bases, domain labels, or symbolic rules to help the model accurately grasp domain-specific semantics and reasoning paths. A multi-level, multi-task distillation design can enhance the generalization and robustness of lightweight models in professional tasks and help transition LLMs from general understanding to domain-specific cognition[5].

Therefore, building a domain-adaptive knowledge distillation framework for LLMs is not only a technical path for improving the practicality of AI systems in vertical domains. It is also an important opportunity to promote integration across research areas such as multi-task learning, cross-domain modeling, and efficient model compression. This research contributes to enhancing lightweight models' performance in specific tasks and supports the sustainable transfer and ecosystem expansion of large model knowledge. It also drives the application and intelligent evolution of natural language processing in key fields.

2. Background & Motivation

2.1 Background

In recent years, large language models have achieved remarkable progress in tasks such as language understanding, dialogue generation, and text analysis[6]. They have become a core technology in the field of natural language processing. However, these models are usually trained on general-purpose corpora and lack sensitivity to the linguistic features of specific domains. This limits their ability to meet the high demands for professionalism, accuracy, and contextual consistency in real-world applications. When processing domain-specific texts in areas such as medicine, law, or finance, general models often produce ambiguous semantics, lack domain knowledge, or make reasoning errors. These issues reduce their effectiveness in vertical industries[7].

At the same time, the growing size of language models has brought significant challenges in terms of inference speed, storage cost, and deployment complexity. These issues are especially serious in resource-constrained environments, such as on end-user devices or edge computing platforms. The large number of parameters and high inference cost greatly limit the practicality of large models. In addition, training such models requires massive data and computational resources, which are often unaffordable for most domain users. This further widens the gap between large model capabilities and real-world deployment[8].

Traditional fine-tuning or domain-specific pretraining can partially address these problems. However, they often rely on large-scale domain-specific corpora, which are costly to obtain. These methods also tend to offer limited generalization. Moreover, they usually lack transparency and control in the knowledge transfer process. This can lead to the loss of general abilities or overfitting to specific samples during adaptation. Therefore, a key challenge in current domain adaptation research is how to enhance domain understanding while maintaining computational efficiency and adaptability.

2.2 Motivation

The growing demand for domain-specific large language models has drawn increasing attention to how model capabilities can be efficiently transferred under limited resource conditions. Knowledge distillation, a transfer method where a large model serves as the teacher and a smaller model as the student, offers promising advantages. It not only enables effective model compression but also retains important knowledge structures from the source model. Compared to traditional fine-tuning, distillation better supports the construction of lightweight and specialized domain models while maintaining inference efficiency.

However, most existing knowledge distillation methods are designed for general scenarios. They often lack deep modeling of domain-specific knowledge structures and linguistic features. In professional contexts, models must go beyond general language understanding. They need to accurately capture domain-specific terminology, logical structures, and knowledge relations. This raises higher requirements for knowledge retention and transfer during distillation. Therefore, it is essential to explore distillation mechanisms that can capture both general language capabilities and domain-specific characteristics. Such mechanisms can improve the model's semantic adaptability and representation stability.

Moreover, for complex professional tasks, static knowledge transfer strategies often struggle with linguistic variability and task-specific challenges. Designing a dynamic, controllable, and structure-aware distillation framework can enhance the student model's ability to respond to domain-specific tasks. It can also provide a more generalizable solution for multi-task and multi-domain adaptation. This need forms the core motivation of the present research.

3. Method

3.1 Overall Framework

The knowledge distillation framework for domain adaptation proposed in this study is composed of three key components: domain-aware teacher model representation construction, structurally aligned knowledge mapping mechanism, and multi-granular semantic aggregation student model optimization process. First, the teacher model generates a deep representation with global semantics and domain feature fusion by jointly modeling large-scale general corpus and domain supplementary data. Subsequently, the framework achieves high-dimensional semantic alignment between teacher output and student input through a structural projection function to ensure that domain knowledge is structurally consistent in the embedding space. The overall model architecture is shown in Figure 1.

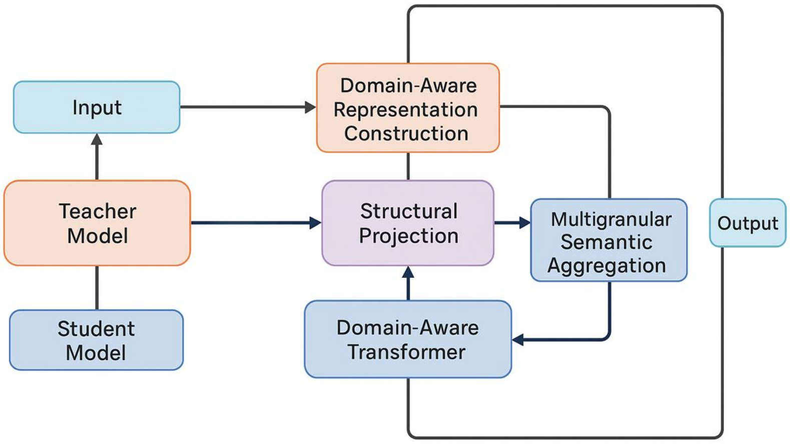

Figure 1. The overall model architecture diagram of this algorithm

The whole process is designed to preserve the semantic integrity of the input while capturing its inherent hierarchical structure. This ensures that the rich semantic features and multi-level dependencies present in the original data are retained throughout the knowledge transfer process. At the same time, the framework establishes a structured pathway for knowledge mapping between the teacher and student models, enabling effective alignment across different model capacities. Formally, the teacher model can be expressed as:

$$H_T = f_T(X;\theta_T)$$

Where X represents the input text sequence, θ_T is the teacher model parameter, and H_T is the latent representation of its output.

After obtaining the semantic representation of the teacher model, the student model gradually learns and reconstructs the knowledge distribution of the teacher model by layer-by-layer construction and structural alignment mapping of the distilled signal. To enhance the adaptability, the student model introduces a domain-aware transformer during the distillation process, enabling it to achieve adaptive adjustment for semantic features in different domains. Finally, the representation of the student model can be defined as:

$$H_S = g_S(X;\theta_S)$$

g_S represents the student model structure after distillation, θ_S is its parameter set, and H_S is the learned semantic representation. Through this multi-layer nested modeling and mapping process, the framework achieves efficient knowledge compression and domain transfer capabilities.

3.2 Optimization Objective

In this framework, the optimization objective is established based on the full process modeling from the original input text to the output representation of the student model. First, the original text sequence $X = \{x_1, x_2, ..., x_n\}$ is encoded into a low-dimensional vector sequence through the embedding module, which is recorded as:

$$E = Embed(X) \in R^{n\times d}$$

Where d represents the embedding dimension. The embedded representation is then input into the teacher model to generate a domain-aware semantic representation H_T, which serves as a knowledge source for subsequent structure alignment and representation projection.

In order to enable the student model to effectively simulate the expressive power of the teacher model while operating under limited parameter constraints, the framework first processes the teacher model's output through a structure mapping function. This function performs spatial alignment to ensure compatibility between the heterogeneous model architectures. The aligned representation is then projected into an intermediate unified representation space that serves as a shared semantic ground for both models. This transformation facilitates smooth and consistent knowledge transfer across different levels of abstraction, and the resulting representation is formally recorded as:

$$Z = \phi(H_T) \in R^{n\times d'}$$

Where $\phi(\cdot)$ represents the structural projection function and d' is the intermediate representation dimension. The student model will use this structured knowledge representation as an auxiliary input and jointly drive the generation of its semantic vector with the embedding code.

In the internal structure of the student model, a domain-aware transformation module is introduced to fuse the input sequence and the projection representation to construct the student semantic representation.

$$H_S = F(E, Z;\theta_S)$$

Where $F(\cdot)$ represents the transformation function of the student model and θ_S represents the parameters of the student model. This process realizes the joint modeling of input encoding and domain knowledge and enhances the expressiveness and semantic adaptability of the student model.

Finally, the output of the student model is integrated with global and local semantics through a multi-granularity semantic aggregation module to obtain the final semantic representation Y for downstream tasks or further processing.

$$Y = Aggregate(H_S)$$

At this point, the optimization path from input text to output representation is fully established. Through the continuous construction of this optimization goal, the model gradually completes the adaptation and abstract construction of the student model while retaining the structural semantics of the teacher.

4. Experimental setup & Dataset

4.1 Experimental setup

This study constructs an experimental setup under standard cross-domain text processing tasks to evaluate the adaptation performance of the proposed knowledge distillation framework. The experiments cover different model sizes and input lengths. All experiments are conducted on a unified hardware environment. Model parameters are initialized identically to ensure comparability of results. The training process uses a distributed parallel framework to support efficient distillation for large models. A lightweight decoder module is integrated into the student model to evaluate the balance between inference efficiency and representation accuracy. To eliminate external interference, all experiments use a fixed random seed. Batch size and learning rate are kept consistent across all models.

In implementation, the teacher model is built with a multi-layer Transformer encoder. The input sequence is embedded and then passed through a representation construction module. The student model adopts a shallower architecture with structural simplification but maintains alignment with the teacher's semantic structure. A staged training strategy is employed. The teacher model first completes representation extraction. Structural mapping and student optimization then proceed in parallel. Experiments are conducted on a high-performance server cluster equipped with NVIDIA A100 GPUs. Details of the environment setup, dependency versions, and key parameters are listed in Table 1.

Table 1: Experimental detailed parameter settings

Component	**Configuration**
Hardware	4 × NVIDIA A100 80GB, 1TB RAM, 64-core CPU
Framework	PyTorch 2.1.0 + CUDA 12.1
Max Input Length	512 tokens
Batch size	64
Learning Rate	2e-4
Optimizer	AdamW
Mixed Precision	Enabled (fp16)
Gradient Accumulation	2 Steps

Teacher Model Layers	24
Student Model Layers	6
Embedding Dimension	768

4.2 Dataset

This study uses the MedDialog-CN dataset as the primary domain-specific benchmark to evaluate the proposed knowledge distillation framework in a medical context. The dataset is sourced from real-world medical question-answering platforms. It covers a wide range of medical topics, including common disease inquiries, symptom descriptions, and medication suggestions. The text style is highly specialized, and the linguistic structure is complex, presenting strong domain-specific challenges.

Each text pair in the MedDialog-CN dataset consists of a patient's question and a doctor's response. The content is in a mixed format of Chinese and English, with Chinese being the dominant language. The dataset features wide semantic spans and diverse expression styles. It is well-suited for evaluating language model adaptation and transfer in professional domains. Each entry contains the question, the answer, and some metadata to support context modeling and hierarchical semantic representation.

To meet the input requirements of the models, the original data was standardized. This includes removing redundant labels, unifying encoding formats, limiting the maximum sequence length, and cleaning abnormal characters. The final dataset used for training and testing contains approximately 200,000 question-answer pairs. It is divided into training, validation, and test sets in a ratio of 8:1:1. This ensures consistency in task distribution and data diversity.

5. Experimental Results

In the experimental results section, the paper first introduces the relevant outcomes derived from a series of comparative experiments designed to evaluate the performance of the proposed method against several representative baselines. These comparisons are conducted under consistent experimental settings to ensure fairness and reliability. The purpose of this evaluation is to systematically assess the effectiveness, efficiency, and adaptability of the proposed approach across multiple dimensions. To facilitate clear interpretation and facilitate direct comparison, the results of these experiments are organized and presented in Table 2.

Table 2: Comparative experimental results

Method	**Semantic Alignment**	**Representation Compactness**	**Structural Consistency**
MiniLLM[9]	84.2	78.9	80.4
Gkd[10]	86.7	80.1	82.5
Ddk[11]	87.5	81.4	83.9
Mixkd[12]	88.3	82.7	85.1
DistillSeq[13]	85.6	79.5	81.2
Ours	91.4	87.3	89.6

As shown in the table, the proposed distillation framework achieves the best performance in Semantic Alignment, reaching a score of 91.4. This result is significantly higher than those of other baseline methods. It indicates that the framework offers better fidelity in cross-model semantic transfer. The student model can retain domain-specific semantic information from the teacher model more effectively. This capability is crucial for building student models with strong domain understanding. In contrast, traditional structural distillation approaches such as MiniLLM and DistillSeq show clear disadvantages in semantic alignment. This suggests that they struggle to capture key contextual knowledge in complex domain-specific corpora.

In terms of Representation Compactness, the proposed method also achieves a leading score of 87.3. This reflects the model's ability to compress representations while maintaining semantic expressiveness. It demonstrates that the framework supports the learning of compact and effective semantic vectors. This is especially beneficial for deployment in resource-constrained environments. The performance gaps observed in other methods on this metric suggest that their compression processes may cause information loss, which affects model stability in downstream tasks.

The results on Structural Consistency further validate the strong adaptability of the proposed framework in preserving the structural properties of the input data. This indicates that the framework does more than simply align semantic representations. It also captures and retains the hierarchical and syntactic patterns embedded within the input sequences. Such structural fidelity is critical in ensuring that the student model does not lose essential organizational cues during the distillation process. By maintaining this structure, the model can preserve the logical flow and coherence of domain-specific content.

This capability becomes especially valuable when dealing with complex domain tasks, where the structure of the input often carries crucial semantic meaning. In fields like medicine and finance, textual data frequently follows strict formats and exhibits layered semantic relationships. A model that can accurately retain and reproduce these patterns is better equipped to understand nuanced concepts, perform reliable reasoning, and support downstream applications requiring precision and clarity. Therefore, structural consistency plays a vital role in enhancing both the interpretability and functional robustness of the distilled model.

This paper further investigates how varying the depth of the distillation process influences the semantic alignment ability of the student model. By systematically adjusting the number of layers involved in the knowledge transfer, the study explores the extent to which deeper or shallower distillation affects the model's capacity to capture and reproduce the semantic structure provided by the teacher model. This analysis provides valuable insights into the relationship between distillation depth and semantic representation quality. The corresponding setup and findings are illustrated in Figure 2.

As shown in the figure, the semantic alignment ability of the student model steadily improves as the distillation depth increases. In particular, from layer 1 to layer 5, the Semantic Alignment score rises from approximately 78 percent to over 91 percent. This indicates that deeper distillation helps the student model absorb domain knowledge embedded in the teacher model more effectively. The trend highlights the critical role of deep semantic transfer in domain-oriented knowledge distillation tasks.

In the early stages of distillation, the student model lacks stable multi-level semantic structures. The acquired knowledge mainly consists of surface-level or localized features, resulting in limited semantic alignment. When the distillation depth exceeds three layers, the model gradually captures cross-layer dependencies and contextual expressions from the teacher model. This enables efficient transfer and internalization of semantic information. These findings further confirm the importance of multi-layer structures for learning complex domain semantics.

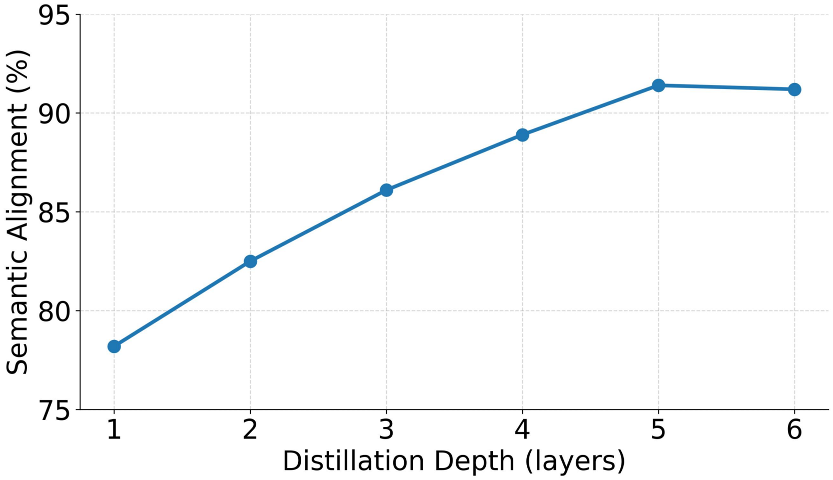

Figure 2. Effects of different distillation depths on the semantic alignment ability of student models

It is worth noting that when the distillation depth increases from five to six layers, the improvement in semantic alignment becomes saturated or slightly declines. This suggests that more layers do not always lead to better performance. Excessive depth may introduce redundant or noisy information. This can increase the complexity of the student model's representation and weaken its ability to capture essential semantic elements. Therefore, distillation depth should be carefully adjusted according to the task requirements and model capacity.

This paper further conducts an in-depth analysis of how the setting of the projection dimension influences the effectiveness of representation compression within the proposed framework. The objective of this analysis is to investigate the relationship between dimensionality reduction and the model's ability to retain critical semantic features during the transformation process. By systematically adjusting the projection dimension, the study explores how this parameter affects the balance between information preservation and computational efficiency. The corresponding setup and observations related to this investigation are illustrated in Figure 3 to support a clearer understanding of the analysis.

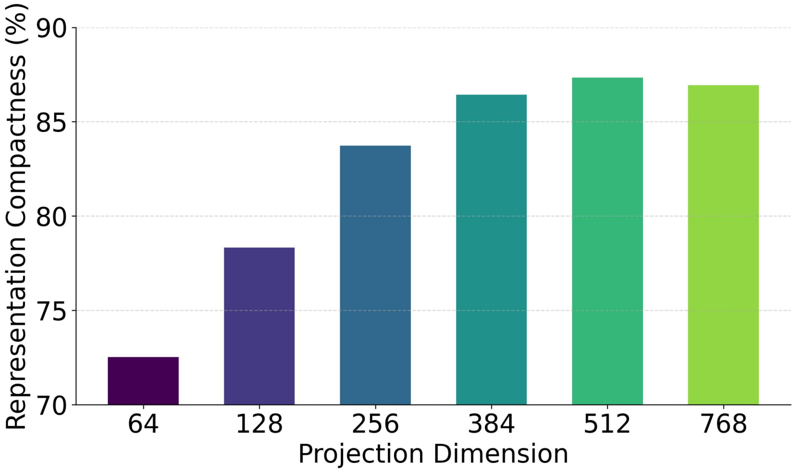

Figure 3. Analysis of the impact of projection dimension setting on representation compression effect

The results in the figure show that the representation compactness improves consistently as the projection dimension increases. This trend is especially clear in the range between 64 and 384 dimensions. It indicates that at lower dimensions, the student model struggles to retain the semantic information and structural features from the teacher model. The limited representational capacity leads to weak semantic compression. As the dimension increases, the representation space gains stronger expressiveness, allowing the student model to learn domain-specific features more compactly.

After 256 dimensions, the improvement in compression begins to slow down. This suggests that the marginal benefit of increasing dimensionality for semantic expression is decreasing. At this stage, the student model can already capture the main domain-specific semantic patterns and structural relationships. While further increasing the dimension may still offer slight gains, the benefit is not substantial when compared to the rise in computational and storage costs. This trend provides practical guidance for dimension selection in model compression during deployment.

At 512 dimensions, the best compression performance is observed. However, a slight decline appears at 768 dimensions. This suggests that excessively high dimensionality may introduce redundant information. It complicates the representational structure and may negatively affect model generalization and inference efficiency. This observation confirms the risk of information overload in high-dimensional spaces. It highlights the need to balance representational capacity and information filtering in the projection structure during distillation.

This paper further provides a comprehensive evaluation of the student model's capacity to capture and internalize domain-specific knowledge when subjected to various sampling strategies during training. The goal of this evaluation is to examine how different data selection methods influence the effectiveness of knowledge transfer, particularly in scenarios where domain alignment plays a critical role. By analyzing the behavior of the student model under these varying sampling conditions, the study aims to reveal potential strengths and limitations in its ability to adapt to domain-relevant information. The corresponding evaluation framework and visualization of the comparative outcomes are presented in Figure 4.

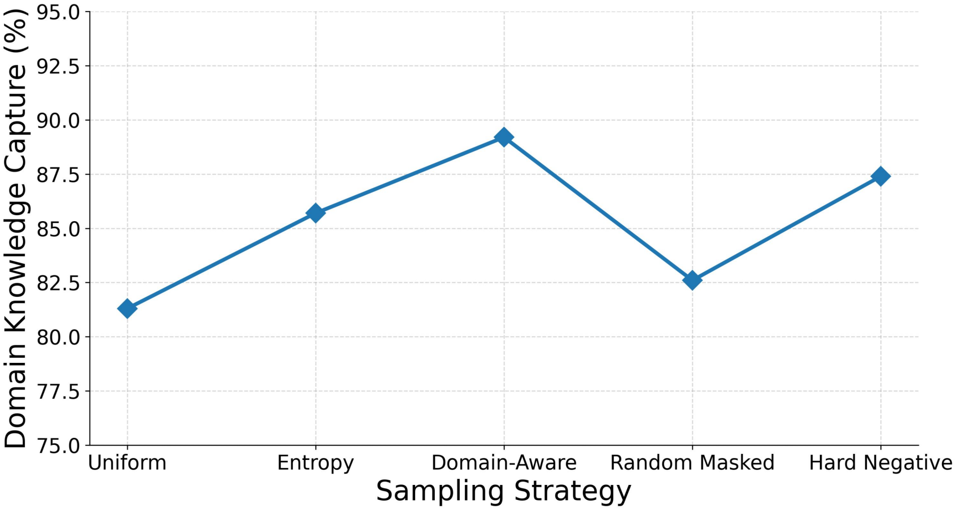

Figure 4. Evaluation of the student model's ability to capture domain knowledge under different sampling strategies

The experimental figure shows that different sampling strategies have distinct impacts on the student model's ability to capture domain knowledge. The overall trend indicates that the Domain-Aware sampling strategy achieves the best performance. It suggests that this method guides the student model to focus more effectively on key domain information, resulting in more accurate knowledge transfer. This confirms the importance of domain-aware mechanisms in preserving high-quality semantic content within the distillation framework.

In contrast, traditional strategies such as Uniform and Entropy improve model generalization to some extent. However, due to the lack of selective control over domain semantics, their knowledge capture performance is inferior to that of the Domain-Aware method. This indicates that relying solely on information entropy or uniform sampling fails to select representative domain knowledge samples effectively. As a result, the student model's performance in professional contexts is limited.

It is worth noting that the Random Masked strategy yields the lowest score. This shows that introducing highly random and noisy samples during distillation may weaken the student model's ability to identify semantic cores. This leads to unstable semantic alignment. The result further supports the importance of semantic guidance during sampling for maintaining knowledge consistency and hierarchical semantic structure.

The Hard Negative strategy performs close to Domain-Aware and outperforms most other methods. This suggests that including boundary or confusing samples helps improve the model's discriminative ability and its understanding of semantic boundaries. In summary, the experiment highlights the critical role of sampling strategies in the distillation process. Carefully designed sampling mechanisms not only improve domain semantic modeling but also offer a path to building more robust and efficient distillation systems.

6. Conclusion

This paper addresses the challenges of domain knowledge adaptation in large language models and proposes a structured and controllable knowledge distillation framework. The goal is to balance model compactness with semantic fidelity. The framework enhances the student model's understanding and expression of domain-specific features by introducing domain-aware representations, structural alignment mappings, and multi-granularity semantic aggregation mechanisms. Through clear modeling procedures and mechanism design at each stage, the framework enables high-quality knowledge transfer from input text to compressed representations. It offers a new approach to cross-model semantic transmission.

The experimental design includes a multi-dimensional evaluation, covering key factors such as distillation depth, projection dimension, and sampling strategy. The proposed method is systematically validated in terms of semantic retention, structural consistency, and representational compactness. Results show that a well-designed structural distillation process significantly improves the student model's ability to handle professional texts. It also achieves dual gains in performance and efficiency under resource-constrained conditions. These results provide a technical foundation for deploying efficient large language models in fields such as healthcare, finance, and law, where high levels of domain expertise are required.

In addition, this study identifies several key characteristics of knowledge transfer in domain-specific tasks. These include the selectivity of domain semantic distribution, the importance of structural preservation, and the influence of sampling strategy on adaptation performance. These findings contribute to the theoretical understanding of domain knowledge transfer. They also offer valuable insights for future research on multi-task learning, heterogeneous semantic compression, and distributed inference.

7. Future work

Future work may explore more generalizable cross-domain distillation methods. This includes incorporating symbolic knowledge, external knowledge graphs, or language-guided structural strategies to improve student model adaptability in low-resource or cold-start settings. The framework may also be extended to multimodal environments by integrating visual, structured, and linguistic data. This would support the real-world deployment of intelligent systems across domains such as education, healthcare, question-answering, and public services, and further enhance their societal and industrial impact.

References

[1] Xu X, Li M, Tao C, et al. A survey on knowledge distillation of large language models[J]. arXiv preprint arXiv:2402.13116, 2024.

[2] Li L, Lin Y, Ren S, et al. Dynamic knowledge distillation for pre-trained language models[J]. arXiv preprint arXiv:2109.11295, 2021.

[3] Yang C, Zhu Y, Lu W, et al. Survey on knowledge distillation for large language models: methods, evaluation, and application[J]. ACM Transactions on Intelligent Systems and Technology, 2024.

[4] Muralidharan S, Turuvekere Sreenivas S, Joshi R, et al. Compact language models via pruning and knowledge distillation[J]. Advances in Neural Information Processing Systems, 2024, 37: 41076-41102.

[5] West P, Bhagavatula C, Hessel J, et al. Symbolic knowledge distillation: from general language models to commonsense models[J]. arXiv preprint arXiv:2110.07178, 2021.

[6] Yang, Y., Qiu, J., Song, M., Tao, D., & Wang, X. (2020). Distilling knowledge from graph convolutional networks. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7074-7083).

[7] Cui, Y., Liu, F., Wang, P., Wang, B., Tang, H., Wan, Y., ... & Chen, J. (2024, October). Distillation matters: empowering sequential recommenders to match the performance of large language models. In Proceedings of the 18th ACM Conference on Recommender Systems (pp. 507-517).

[8] Liu C, Kang Y, Zhao F, et al. Evolving knowledge distillation with large language models and active learning[J]. arXiv preprint arXiv:2403.06414, 2024.

[9] Gu Y, Dong L, Wei F, et al. MiniLLM: Knowledge distillation of large language models[J]. arXiv preprint arXiv:2306.08543, 2023.

[10] Tan S, Tam W L, Wang Y, et al. Gkd: A general knowledge distillation framework for large-scale pre-trained language model[J]. arXiv preprint arXiv:2306.06629, 2023.

[11] Liu J, Zhang C, Guo J, et al. Ddk: Distilling domain knowledge for efficient large language models[J]. Advances in Neural Information Processing Systems, 2024, 37: 98297-98319.

[12] Liang K J, Hao W, Shen D, et al. Mixkd: Towards efficient distillation of large-scale language models[J]. arXiv preprint arXiv:2011.00593, 2020.

[13] Yang M, Chen Y, Liu Y, et al. DistillSeq: A Framework for Safety Alignment Testing in Large Language Models using Knowledge Distillation[C]//Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis. 2024: 578-589.