Structured Multi-Stage Alignment Distillation for Semantically Consistent Lightweight Language Models

Jinxu Guo *

Posted Date: 5 February 2026

doi: 10.20944/preprints202602.0355.v1

Keywords: structured distillation; multi-stage alignment; semantic consistency; model compression

 doi:10.20944/preprints202602.0355.v1

Article

Structured Multi-Stage Alignment Distillation for Semantically Consistent Lightweight Language Models

Jinxu Guo

Dartmouth College, Hanover, USA; jinxuguo2c@gmail.com

Abstract

This study presents a multi-stage alignment fine-tuning framework based on structured knowledge distillation to address semantic mismatch, representation drift, and the loss of deep structural information during knowledge transfer in lightweight models. The method first extracts structured knowledge across layers and semantic scales from the teacher model and constructs a unified structural space through schematic modeling, which enables explicit alignment of latent relations across semantic levels. The framework then introduces a progressive optimization process composed of multiple stages. Each stage focuses on semantic absorption, structural alignment, and representation stability, while cross-stage consistency constraints ensure smooth feature evolution and reduce semantic discontinuity and distribution shifts. The model further incorporates cross-layer feature distillation and attention structure alignment, allowing the student model to inherit not only surface outputs but also reasoning paths and internal semantic logic from the teacher. By integrating structured modeling, multi-stage alignment, and cross-layer distillation, the method forms a knowledge transfer system with strong consistency and high fidelity, improving distribution alignment, hierarchical semantic understanding, and structural stability. Overall, the study shows that structured and staged design can produce a more expressive, robust, and interpretable distillation framework while maintaining model efficiency, making it suitable for model compression and efficient customization in multi-task settings.

Keywords: structured distillation; multi-stage alignment; semantic consistency; model compression

I. Introduction

Recent years have witnessed significant advances in large language models in text understanding, reasoning, generation, and multi-task transfer. However, the rapid increase in model size has also led to soaring computational costs, difficulties in deployment, and limited efficiency in task adaptation. In real applications, models must achieve strong performance while maintaining controllable resource consumption. This raises an important challenge [1]. It concerns how to compress and customize large models efficiently without sacrificing accuracy. Traditional fine-tuning methods often rely on single-stage parameter updates and lack a refined understanding of knowledge structure, semantic hierarchy, and task requirements. As a result, models may suffer from representation drift, semantic mismatch, and the loss of useful knowledge when transferred to specific scenarios. Under the tension between model scale and task complexity, a structured, interpretable, and layer-wise alignment mechanism becomes essential for improving model adaptability [2].

In research on model compression and knowledge distillation, many existing approaches focus on global distributions or output-level knowledge. They pay insufficient attention to internal knowledge structures, semantic hierarchies, and interactions across multi-scale features. The knowledge encoded within large models contains more than predictions. It includes multi-level semantic abstractions, logical relations, domain conventions, and task-specific priors. Distillation

 doi:10.20944/preprints202602.0355.v1

based only on final outputs or a single feature space often ignores these structural forms of knowledge. Student models, therefore, struggle to inherit the teacher's reasoning patterns and expressive capacity. In addition, most existing distillation strategies rely on a single alignment mechanism. They lack cross-stage stabilization of representations. This may cause representation drift or semantic collapse during training. As a result, knowledge transfer becomes incomplete and discontinuous, which limits further improvements [3].

Meanwhile, as application scenarios expand, task requirements show stronger structural, multi-dimensional, and dynamic characteristics. In real systems, models must understand complex text structures, domain rules, semantic hierarchies, and cross-sentence relations [4]. This increases the need for a framework that can absorb knowledge layer by layer, align structures across domains, and stabilize features progressively. Single-stage or single-point distillation is no longer sufficient for high-demand tasks. Recent research has shifted toward frameworks that explicitly model knowledge structure, reinforce semantic boundaries, and reconstruct representations through multiple stages. These methods help capture the rich capabilities of teacher models and improve both interpretability and generalization [5].

Within this context, multi-stage alignment has emerged as an effective solution. It divides the learning process into several progressive stages. Each stage absorbs knowledge at different levels, such as shallow syntactic features, intermediate semantic patterns, and deeper reasoning paths or task constraints. This staged organization reduces abrupt feature shifts caused by one-step transfer. It also prevents unstable fluctuations in representations. Student models can then maintain a stable trajectory of knowledge growth. When combined with structured knowledge, this approach constrains the form of knowledge at each stage and enables controlled integration across semantic dimensions. It also aligns the structural patterns of teacher models more precisely. This provides a promising direction for inheriting complex knowledge and supports adaptation across tasks, domains, and text structures.

In addition, recent studies on multi-source feature integration indicate that unified latent space modeling and cross-modal structural alignment can significantly enhance representation consistency and information complementarity. By jointly constraining heterogeneous features within a shared structural space, such approaches improve the stability and expressiveness of complex prediction models [6]. This perspective further supports the necessity of structured modeling in multi-stage knowledge transfer.

Overall, constructing a multi-stage alignment framework based on structured knowledge distillation has strong research value and practical significance. It helps retain the structural knowledge and reasoning capability of large models while keeping the model lightweight. This creates a feasible path for low-cost deployment. It also enhances semantic consistency, feature robustness, and task generalization through staged knowledge transfer and alignment. This helps mitigate representation drift and semantic mismatch and improves reliability and interpretability in complex environments. Furthermore, the introduction of structured mechanisms makes knowledge transfer transparent and controllable rather than a black-box imitation. This is important for high-risk domains that require verifiable model behavior. Therefore, this direction holds both theoretical value and long-term impact on model compression, task adaptation, deployment, and trustworthy AI.

II. Related Work

Knowledge distillation has long served as a key technique for model compression and transfer learning [7]. It improves the performance of lightweight models through methods such as output distribution imitation, feature alignment, and attention distillation. Early approaches relied mainly on the response behavior of teacher models. They used soft labels, probability distributions, or hidden vectors to guide the learning of student models. However, these strategies often focus on point-to-point or global distribution matching. They do not explicitly model the internal structure of knowledge. As a result, student models struggle to capture hierarchical relations, semantic

 doi:10.20944/preprints202602.0355.v1

abstraction paths, and reasoning patterns in teacher models. With increasing task complexity, distillation from a single space or a single level is no longer sufficient for transferring deep knowledge. This has motivated the development of finer and more diverse knowledge transfer strategies such as cross-layer distillation, multi-scale feature distillation, and semantic contrastive distillation. Although these methods improve performance, they still lack systematic structural modeling and stage-wise training mechanisms [8].

In the area of structured knowledge modeling, research trends are shifting from representation imitation to semantic structure alignment. Existing work attempts to incorporate task priors, semantic labels, text hierarchies, or domain knowledge into models in the form of graphs, templates, or multi-dimensional latent factors. These methods aim to reduce semantic entanglement in model representations. They emphasize explicit descriptions of knowledge forms and use structural constraints to improve the organization of multi-granular semantics. However, most structured knowledge methods are designed for pre-training, inference enhancement, or single-stage fine-grained optimization. They have not yet formed a unified paradigm that integrates structured knowledge with cross-stage distillation. In addition, alignment strategies for structured knowledge are often coarse. They rely on static constraints and cannot adapt to learning needs and representation changes across different stages. This limits their usefulness in distillation scenarios.

As model size continues to increase, multi-stage training and layer-wise alignment strategies have received growing attention. Existing methods introduce staged optimization in transfer learning and parameter-efficient fine-tuning. They use phased freezing, gradual unfreezing, or local refinement to mitigate representation drift and enhance training stability. However, most of these methods focus on parameter update strategies rather than explicit modeling of knowledge forms, sources, or semantic paths. This results in a lack of logical knowledge continuity across stages. Some layer-wise or multi-stage distillation approaches consider features at different levels. Yet they still rely on surface-level alignment, such as feature similarity or attention similarity. They do not explore the semantic structures, reasoning patterns, and cross-layer relations inside teacher models. As a result, student models often gain only superficial consistency and fail to inherit deeper knowledge.

Overall, existing research has made progress in knowledge distillation, structured modeling, and multi-stage transfer. However, there remains a clear gap in combining structured knowledge with multi-stage alignment. On one hand, knowledge distillation requires stronger semantic organization to separate different knowledge dimensions and to build interpretable hierarchical relations. On the other hand, multi-stage training requires more precise alignment targets to ensure continuous knowledge transfer and stable representations across stages. Therefore, integrating both directions and building a new framework that explicitly models knowledge structure and incorporates progressive multi-stage alignment is of significant research value. Such a framework can address knowledge loss, semantic mismatch, and representation drift in model compression and efficient fine-tuning. It can also enhance the effectiveness and robustness of knowledge transfer. Moreover, it provides a theoretical and practical foundation for developing lightweight models that are interpretable, controllable, and adaptable to various domains.

III. Proposed Framework

This study constructs a multi-stage alignment fine-tuning framework centered on structured knowledge distillation. Explicitly modeling the semantic structure, hierarchical relationships, and cross-scale features of the teacher model enables the student model to achieve more stable and interpretable knowledge transfer during the progressively layered training process. Its model architecture is shown in Figure 1.

 doi:10.20944/preprints202602.0355.v1

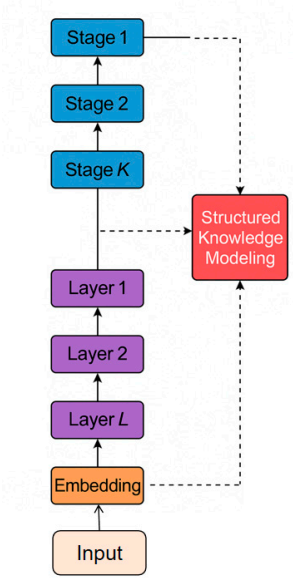

Figure 1. Overall model architecture diagram.

First, the original representation of the input sequence is defined as:

$$H^{(0)} = Embed(X) \quad (1)$$

Here, $H^{(0)}$ serves as the foundation for the construction of representations in subsequent stages. To capture the structural relationships of multi-layered knowledge within the teacher model, a structured graph modeling approach is used to map the semantic features of each layer to a unified structural space, resulting in a structured knowledge tensor:

$$S = \phi(\{H_T^{(1)}, H_T^{(2)}, \ldots, H_T^{(L)}\}) \quad (2)$$

Here, $H_T^{(l)}$ represents the latent space representation of the teacher model at layer l, and S is used to guide the subsequent multi-stage alignment process. The purpose of structured modeling is to provide a cross-layer, cross-semantic scale reference structure, so that distillation no longer depends on a single representation, but establishes consistency across multiple knowledge dimensions.

In the multi-stage optimization process of the student model, each stage progresses progressively around three objectives: semantic absorption, structural alignment, and representation stability. For the student feature $H_S^{(k)}$ in the k-th stage, a structured semantic alignment loss is first used to constrain its representation mapped to the teacher model in the structural space:

$$L_{struct}^{(k)} = ||\psi(H_S^k) - S^{(k)}||_2^2 \quad (3)$$

Where $S^{(k)}$ represents the semantic part of the structured knowledge tensor corresponding to this stage, and $\psi(\cdot)$ is the structure mapping function. To ensure the stability of cross-stage representation, the framework further incorporates inter-stage consistency constraints, enabling the model to maintain a smooth evolution of knowledge form across consecutive stages:

$$L_{smooth}^{(k)} = ||H_S^k - H_S^{(k-1)}||_2 \quad (4)$$

This design can effectively suppress representation drift and improve the continuity of multi-stage learning.

In the cross-layer knowledge distillation part, the framework aligns not only surface features and attention patterns, but also inference paths and semantic abstraction logic. For the corresponding layer l of the teacher model and student model, a cross-layer feature distillation loss is used:

$$L_{attn}^{(l)} = ||A_T^{(l)} - A_S^{(l)}||_2^2 \quad (5)$$

Where $A_T^{(l)}$ and $A_S^{(l)}$ are the attention matrices of the teacher and student at layer l, respectively. Through joint modeling of feature alignment and attention alignment, the student model can acquire a more complete reasoning pattern and information flow structure, rather than just surface features.

 doi:10.20944/preprints202602.0355.v1

Ultimately, the framework unifies multi-stage loss, multi-level distillation loss, and structured constraints into a single comprehensive optimization objective, enabling the model to progressively absorb knowledge at different levels at each stage and maintain a stable structural evolution process. The overall optimization objective is:

$$L = \sum_{k=1}^{K}(L_{struct}^{(k)} - \lambda L_{struct}^{(k)}) + \sum_{l=1}^{L} \alpha L_{struct}^{(l)} - \beta L_{struct}^{(l)} \quad (6)$$

Here, λ, α, β represents the importance control factor for balancing different knowledge sources. Through this comprehensive optimization approach, the model can achieve more robust, interpretable, and efficient knowledge transfer through the combined effects of layer-by-layer progression, cross-stage continuity, and multi-dimensional structural alignment, providing a systematic distillation and fine-tuning mechanism for building lightweight large models.

IV. Experimental Analysis

A. Dataset

This study adopts the GLUE dataset as the primary data source. GLUE is a multi-task benchmark for general language understanding. It includes sentiment classification, natural language inference, semantic similarity evaluation, and question-answering inference. It provides a comprehensive assessment of a model's semantic processing ability across different text characteristics and task structures. Due to its diverse task types, rich label formats, and complex semantic relations, it has long been regarded as an important reference for evaluating model generalization and structured semantic understanding.

GLUE contains data from multiple text domains, such as social media posts, news articles, and encyclopedia-style sentence pairs. This leads to clear cross-domain variation in data distribution. The combination of multi-source texts not only increases opportunities for learning structured semantics but also reflects the diversity and complexity of real-world language tasks. For research on knowledge distillation and multi-stage alignment, such multi-domain text structures provide a rich context for examining how models adapt to different semantic levels.

In the framework of this study, the dataset enables systematic exploration of structured semantic modeling, cross-layer representation alignment, and stage-wise convergence behavior. The differences across tasks, the diversity of label formats, and the complex cross-sentence relations create a suitable test environment for multi-stage alignment mechanisms. Through its multi-task and multi-style characteristics, GLUE offers an appropriate data foundation for studying structured knowledge transfer and model robustness in depth.

B. Experimental Results

This paper first conducts a comparative experiment, and the experimental results are shown in Table 1.

Table 1. Comparative experimental results.

Method	KL Divergence	RCS	Pearson	Spearman
LLM-adapter [9]	0.214	0.128	0.864	0.851
Lofit [10]	0.187	0.103	0.881	0.872
Fedbiot [11]	0.176	0.097	0.889	0.878
Memento [12]	0.163	0.091	0.895	0.883
Ours	0.142	0.067	0.913	0.901

In the overall comparison, the methods show clear differences in distillation consistency, structural stability, and semantic relevance. These differences reflect the influence of knowledge transfer mechanisms on internal representations and semantic understanding. Traditional adaptation

 doi:10.20944/preprints202602.0355.v1

methods show higher KL divergence values. This indicates that student models have a limited ability to fit the output distribution of teacher models. Their knowledge transfer depends more on local features or task prompts and cannot fully absorb deeper semantic structures. In contrast, the multi-stage structured alignment framework performs better on this metric. It shows that the framework can converge more stably to the probability distribution of the teacher across stages and achieve more complete knowledge inheritance.

In terms of representation stability, the RCS scores reveal a consistent advantage for structured alignment. Single-stage or shallow alignment methods lack mechanisms for cross-stage stability. They tend to produce feature drift during training, which leads to large differences between stages. The multi-stage structured distillation framework reduces this instability through progressive alignment across layers. It allows student models to approach the teacher structure along a coherent representational path at each stage. The lower RCS values highlight this stable evolution and provide a foundation for robustness in complex tasks.

For semantic relevance, the methods show significant differences in Pearson and Spearman correlation scores. This reflects the importance of structural modeling for understanding inter-sentence relations. Traditional distillation methods rely mainly on surface-level feature alignment. They produce only limited improvements in shallow semantic similarity. In contrast, structured student models incorporate cross-layer semantic structures. They capture both high-level sentence features and deeper semantic composition patterns. This leads to higher quality in learning sentence-level correlations and results in better performance on both correlation metrics.

Overall, the multi-stage structured distillation framework demonstrates comprehensive advantages across all four core metrics. This confirms the effectiveness of combining structured knowledge modeling with multi-stage semantic alignment. The framework addresses representation instability and knowledge loss commonly seen in traditional distillation. It also strengthens the hierarchical semantic representations of student models. From the perspective of capability building, this progressive and structurally aligned mechanism preserves deeper reasoning patterns and semantic structures of the teacher model. It provides a lightweight solution that is better suited for multi-task and multi-domain applications.

This paper also presents the experimental results for the learning rate, as shown in Table 2.

Table 2. Experimental results of the learning rate.

Learning Rate	KL Divergence	RCS	Pearson	Spearman
0.0004	0.198	0.109	0.874	0.861
0.0003	0.176	0.091	0.889	0.876
0.0002	0.158	0.078	0.902	0.888
0.0001	0.142	0.067	0.913	0.901

Under different learning rate settings, the model shows a clear downward trend in distillation consistency. A larger learning rate leads to higher KL divergence. This indicates that parameter updates introduce large fluctuations, making it difficult for the student model to approach the teacher's output distribution stably. As the learning rate decreases, updates become smoother. The student model can then capture deeper semantic structures from the teacher more accurately. At a learning rate of 0.0001, the KL value reaches its lowest point. This shows that structured knowledge is transferred more effectively and highlights the importance of multi-stage alignment in maintaining a stable optimization path.

In terms of representation stability, the RCS score decreases as the learning rate becomes smaller. This reflects a consistent reduction in representation differences across stages. A high learning rate often causes representation drift during multi-stage training. It becomes difficult for the student model to maintain continuous semantic structures across stages, which harms the effect of progressive alignment. A lower learning rate strengthens the progressive absorption of knowledge.

 doi:10.20944/preprints202602.0355.v1

Representations from each stage connect naturally to those of the next stage, resulting in smoother structural alignment. This trend further confirms the advantage of the multi-stage framework in stabilizing the knowledge transfer process.

For semantic relevance, both Pearson and Spearman correlations increase steadily as the learning rate decreases. This indicates that the student model achieves better alignment in sentence-level semantic relations and semantic ranking. This outcome is consistent with the goal of structured knowledge distillation. The framework emphasizes the preservation of cross-layer semantic structures. It enables the student model to inherit deeper semantic logic rather than relying only on surface-level feature imitation. At the optimal learning rate, both correlation scores reach their highest values. This improvement reflects the enhanced structural consistency achieved through fine-grained updates and confirms that combining multi-stage semantic alignment with structured knowledge modeling significantly strengthens the student model's expressive ability in complex semantic tasks.

This paper also presents experimental results for different optimizers, as shown in Table 3.

Table 3. Experimental results of the optimizer.

Optimizer	KL Divergence	RCS	Pearson	Spearman
AdaGrad	0.194	0.112	0.871	0.858
Adam	0.167	0.089	0.891	0.879
SGD	0.183	0.097	0.882	0.871
AdamW	0.142	0.067	0.913	0.901

From the overall trend, the optimizers show clear differences in distillation consistency and structured alignment ability. This reflects the important role of optimization strategies in structured knowledge transfer. AdaGrad performs relatively weakly on KL divergence and Pearson or Spearman correlations. Its rapidly decaying learning rate suits sparse gradient settings but does not support stable and continuous approximation of the teacher's semantic distribution in multi-stage distillation. The higher KL value indicates that the student model struggles to absorb deep structured knowledge and that the efficiency of distillation is limited.

In contrast, Adam achieves stronger performance on most metrics. Its adaptive update mechanism maintains training speed while improving the stability of parameter convergence. This allows the student model to capture long-range dependencies and semantic structures from the teacher more accurately. However, Adam's weight decay is not strictly equivalent to regularization. Representation fluctuations remain across stages. This is reflected in its moderate RCS score, which shows that stage-to-stage alignment has not reached optimal stability.

For the momentum-driven SGD optimizer, the update direction is relatively coarse. This leads to representation drift across stages during structural alignment. As a result, both RCS and correlation metrics are lower than those of Adam. Although SGD follows a more direct optimization path, its coarse-grained updates are insufficient for structured distillation tasks that require fine semantic alignment. This limits its performance on Pearson and Spearman correlations.

Considering the characteristics of all optimizers, AdamW achieves the best results across all four core metrics. This indicates that it is particularly suitable for structured knowledge distillation. The decoupling of weight decay from gradient updates helps control parameter norms and reduces noise-induced disturbances. The student model maintains higher representation stability and stronger semantic consistency during multi-stage alignment. The lower KL divergence and RCS, together with the highest semantic correlation scores, show that this optimization strategy enhances the efficiency of structured knowledge absorption. It allows the student model to approach the teacher's semantic space in a more stable manner and fully exploits the advantages of the multi-stage alignment framework.

 doi:10.20944/preprints202602.0355.v1

This paper also presents an experiment on the sensitivity of the training data scale ratio to the KL Divergence metric, and the experimental results are shown in Figure 2.

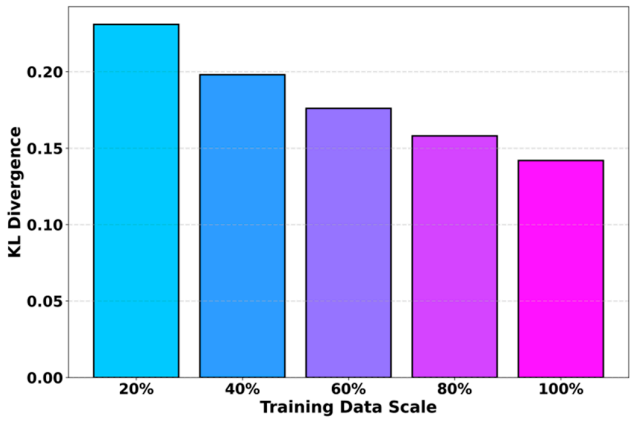

Figure 2. Experiment on the proportional sensitivity of training data size to the KL Divergence metric.

From the overall trend, KL divergence decreases steadily as the training data size increases. This indicates that the student model can approach the teacher's probability distribution more stably when trained with richer data. When the data size is small, the model lacks sufficient ability to fit the teacher's semantic space. The distillation process is affected by strong gradient fluctuations, and the output distribution remains far from that of the teacher. When the data size reaches 80 percent and 100 percent, the KL value drops markedly. This shows that the multi-stage alignment framework can fully exploit additional samples to complete finer semantic absorption and improve the overall quality of distillation.

Under medium-scale data conditions, such as 40 percent to 60 percent, the decline in KL divergence becomes smoother. This reflects that structured knowledge distillation can form a stable semantic transfer path once a certain amount of data is available. At this stage, the model can already capture the main semantic structures of the teacher. Through structured modeling and cross-stage representation alignment, the student model gradually gains a stronger knowledge expression ability. This phenomenon shows that the core advantage of the multi-stage alignment mechanism lies in its capacity to maintain strong knowledge absorption and structural stability even with limited data.

When the training data size reaches its highest level, the KL divergence reaches its minimum. This indicates that the student model forms a more consistent representational distribution with the support of large-scale data. The accumulated structural constraints in the multi-stage framework achieve their strongest effect at this point. The student model imitates not only surface-level outputs but also approaches the deeper semantic structures of the teacher. This result further confirms the synergy between data scale and structured multi-stage distillation. With sufficient data, the framework achieves higher fidelity and lower deviation in semantic alignment, which supports robust performance in complex task scenarios.

This paper also presents an experiment on the sensitivity of the labeled noise level of the RCS index, and the experimental results are shown in Figure 3.

 doi:10.20944/preprints202602.0355.v1

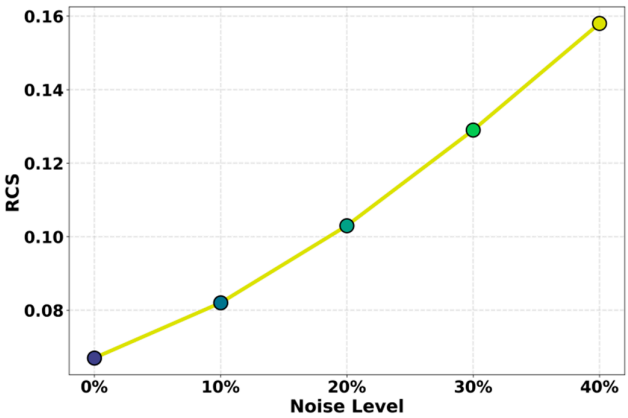

Figure 3. Sensitivity experiment of the labeled noise level for the RCS index.

From the overall trend, the RCS score increases steadily as the level of label noise rises. This indicates that stronger noise leads to larger representation differences across stages, which undermines the stability of structural alignment. Noise disturbs the semantic labels of training samples. It reduces the consistency of structured representations learned at each stage. As a result, the multi-stage alignment mechanism cannot maintain a continuous and smooth feature evolution path, and the RCS score increases significantly.

At medium noise levels, such as 10 percent to 30 percent, the growth of the RCS score accelerates. This reflects the sensitivity of structured knowledge distillation to semantic disruptions. Noise weakens the model's ability to fit the teacher's semantic relations precisely. It also breaks the assumption of stable progression in multi-stage alignment. Cross-stage features become more prone to drift. This further shows that structured distillation depends on accurate and consistent label signals to maintain hierarchical semantic construction. When the input signal is damaged, the student model is more likely to show insufficient alignment and discontinuous representations.

At high noise levels, such as 40 percent, the RCS score reaches its maximum. This suggests that representation drift has moved from mild disturbance to clear instability. The structured features learned across different stages show substantial differences. The student model cannot form coherent hierarchical semantic paths. It also struggles to inherit deep structural information from the teacher. This result highlights the value of the multi-stage structured alignment framework from another perspective. When label quality is adequate, the framework improves representation consistency significantly. When noise becomes excessive, its stability is challenged. Therefore, practical applications must pay careful attention to data quality and noise control to ensure that structured distillation can achieve its best performance.

This paper also presents the impact of training step size on the experimental results, and the experimental results are shown in Figure 4.

 doi:10.20944/preprints202602.0355.v1

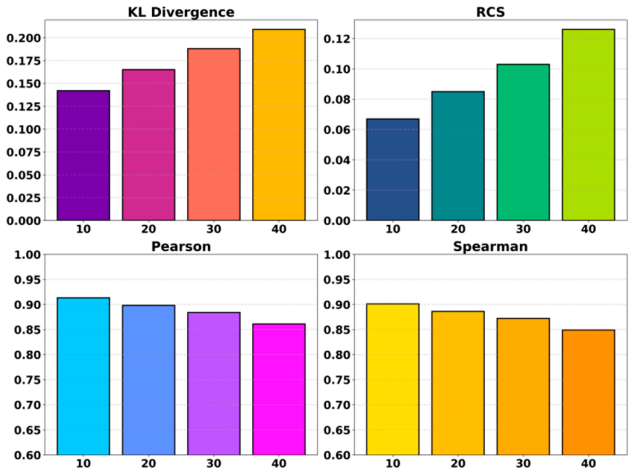

Figure 4. The effect of training step size on experimental results.

From the overall trend, KL divergence increases as the training step size becomes larger. This indicates that the student model cannot stably approach the teacher's output distribution when updates are too large. Structured knowledge distillation relies on fine-grained semantic absorption and cross-stage alignment. Large update steps introduce stronger gradient disturbances. These disturbances repeatedly disrupt the semantic structures accumulated during distillation and reduce the precision of knowledge transfer. This trend shows that the structured alignment framework requires a careful update rhythm to preserve hierarchical semantic information.

The RCS score also rises significantly as the training step size increases. This reflects stronger representation drift caused by larger update magnitudes. The core of multi-stage alignment is to maintain a continuous feature evolution path across stages. Large step sizes break this smooth progression and make the structural path unstable. As the step size increases, the differences between stage representations grow. This indicates that the hierarchical semantic structure inside the model becomes dispersed and that the structural consistency of multi-stage distillation is weakened. These findings highlight the importance of fine-grained update steps for maintaining representation stability.

For semantic relevance, both Pearson and Spearman correlations decrease gradually as the step size increases. This shows that larger updates weaken the model's ability to capture complex inter-sentence relations and semantic ordering. Structured distillation requires the student model to inherit the semantic logic of the teacher through layer-by-layer alignment. Excessive step sizes cause the model to skip key alignment phases and lose accuracy in high-level semantic structures. The results show that small step sizes better support the multi-stage structured alignment mechanism. They allow the model to absorb semantic knowledge steadily at each stage and achieve stronger semantic correlation and structural consistency.

This paper also presents the influence of different distillation loss weights on the experimental results, and the experimental results are shown in Figure 5.

 doi:10.20944/preprints202602.0355.v1

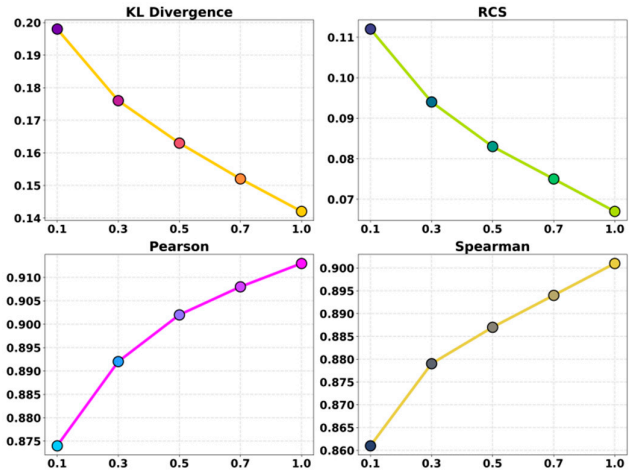

Figure 5. Effect of different distillation loss weights on experimental results.

From the overall trend, different distillation loss weights have a clear impact on consistency learning and structured knowledge absorption. As the loss of weight increases, the KL divergence decreases significantly. This indicates that the model aligns more effectively with the teacher's output distribution and achieves higher fidelity in knowledge transfer. When the distillation loss has a small weight, the student model does not fully imitate the teacher distribution, which limits the transfer of knowledge. Increasing the loss weight strengthens the model's ability to capture semantic structures and highlights the core role of structured distillation.

For representation stability, the RCS score decreases steadily as the distillation loss weight becomes larger. This shows that multi-stage alignment obtains a more stable representation evolution path under stronger distillation constraints. When the loss weight is low, the distillation signal is not strong enough to constrain cross-stage semantic transfer. This makes representation drift more likely during alignment. When the weight increases, the structural constraints become stronger. The student model shows smoother semantic progression across stages. This results in greater structural consistency and demonstrates the synergy between the multi-stage framework and the distillation loss weight.

For semantic relevance, both Pearson and Spearman correlations increase as the loss weight becomes larger. This shows that stronger distillation helps the student model capture inter-sentence relations, semantic ranking, and higher-level structural logic more effectively. A larger distillation weight pushes the student model to construct a semantic space closer to that of the teacher. This leads to clearer semantic boundaries and more coherent hierarchical structures. This trend further confirms the importance of structured distillation in improving semantic consistency and reasoning ability. It also indicates that increasing the distillation loss weight can maximize the effectiveness of the multi-stage alignment framework and enhance the student model's performance in structured semantic learning.

V. Conclusion

This study proposes a multi-stage alignment fine-tuning framework based on structured knowledge distillation. The framework explicitly models the hierarchical semantic structure, cross-layer relations, and internal reasoning paths of the teacher model. It enables stable feature evolution and deep semantic absorption for the student model across continuous stages. The framework effectively mitigates knowledge loss and representation drift, which are common in traditional distillation. It also establishes a more interpretable and consistent knowledge transfer mechanism. As a result, lightweight models achieve notable improvements in reasoning ability, semantic understanding, and structural robustness. Furthermore, the method does not rely on specialized architectures or costly annotations. It can be integrated naturally into existing large-model adaptation

 doi:10.20944/preprints202602.0355.v1

pipelines. This offers a practical solution for model compression, deployment, and cross-domain transfer in resource-constrained settings. It also provides meaningful support for applications in natural language understanding, intelligent question answering, information retrieval, enterprise decision systems, and large-scale language services.

Future research may extend this framework to a wider range of tasks and multimodal scenarios. Structured distillation can then operate not only in text-based settings but also in image, speech, and multimodal understanding tasks. In addition, the multi-stage alignment mechanism can be combined with automated distillation strategies, adaptive structure modeling, and heterogeneous teacher–student collaborative learning. This may support a more dynamic, transferable, and generalizable knowledge transfer process. As model sizes continue to grow and application demands become more diverse, efficient, robust, and structurally transparent knowledge transfer frameworks will remain essential for the deployment of large models. The ideas and methods presented in this study provide an important foundation for this ongoing development.

References

1. Wang X, Jiang Y, Yan Z, et al. Structural knowledge distillation: Tractably distilling information for structured predictor [C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). 2021: 550-564.
2. Zhang L, Ma K. Structured knowledge distillation for accurate and efficient object detection [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15706-15724.
3. Zhang L, Shi Y, Wang K, et al. Structured Knowledge Distillation Towards Efficient Multi-View 3D Object Detection [C]//BMVC. 2023: 339-344.
4. Zhao Z, Xie Z, Zhou G, et al. Mtms: Multi-teacher multi-stage knowledge distillation for reasoning-based machine reading comprehension [C]//Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2024: 1995-2005.
5. Ma Y, Chen Y, Akata Z. Distilling knowledge from self-supervised teacher by embedding graph alignment [J]. arXiv preprint arXiv:2211.13264, 2022.
6. Wang, Q., X. Zhang and X. Wang, "Multimodal integration of physiological signals, clinical data and medical imaging for ICU outcome prediction", Journal of Computer Technology and Software, vol. 4, no. 8, 2025.
7. Moslemi A, Briskina A, Dang Z, et al. A survey on knowledge distillation: Recent advancements [J]. Machine Learning with Applications, 2024, 18: 100605.
8. Liu C, Yin H, Wang X. Theoretical Perspectives on Knowledge Distillation: A Review [J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2025, 17(4): e70049.
9. Hu Z, Wang L, Lan Y, et al. Llm-adapters: An adapter family for parameter-efficient fine-tuning of large language models [C]//Proceedings of the 2023 conference on empirical methods in natural language processing. 2023: 5254-5276.
10. Yin F, Ye X, Durrett G. Lofit: Localized fine-tuning on llm representations [J]. Advances in Neural Information Processing Systems, 2024, 37: 9474-9506.
11. Wu F, Li Z, Li Y, et al. Fedbiot: Llm local fine-tuning in federated learning without full model [C]//Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2024: 3345-3355.
12. Zhou H, Chen Y, Guo S, et al. Memento: Fine-tuning llm agents without fine-tuning llms [J]. arXiv preprint arXiv:2508.16153, 2025.