

Iterative Layer-wise Distillation for Efficient Compression of Large Language Models

Grigory Kovalev¹[0009–0008–5356–4527] and Mikhail
Tikhomirov¹[0000–0001–7209–9335]

Lomonosov Moscow State University, Russia
kaengreg@ya.ru tikhomirov.mm@gmail.com

Abstract. This work investigates distillation methods for large language models (LLMs) with the goal of developing compact models that preserve high performance. Several existing approaches are reviewed, with a discussion of their respective strengths and limitations. An improved method based on the ShortGPT approach has been developed, building upon the idea of incorporating iterative evaluation of layer importance. At each step, importance is assessed by measuring performance degradation when individual layers are removed, using a set of representative datasets. This process is combined with further training using a joint loss function based on KL divergence and mean squared error. Experiments on the Qwen2.5-3B model show that the number of layers can be reduced from 36 to 28 (resulting in a 2.47 billion parameter model) with only a 9.7% quality loss, and to 24 layers with an 18% loss. The findings suggest that the middle transformer layers contribute less to inference, underscoring the potential of the proposed method for creating efficient models. The results demonstrate the effectiveness of iterative distillation and fine-tuning, making the approach suitable for deployment in resource-limited settings.

Keywords: Distillation · Knowledge Transfer · Fine-tuning · Neural models

1 Introduction

Recent advances in natural language processing (NLP) have been driven by large language models (LLMs) such as GPT [2] and BERT [7], built on the transformer architecture [26] and trained on massive corpora like Common Crawl¹ and Wikipedia². These models have achieved remarkable performance in tasks ranging from text generation [3], machine translation [6], and question answering [18,4], to information retrieval [1,14,29].

Despite their success, LLMs are computationally expensive, requiring significant memory and computational resources, which poses challenges for deployment on mobile, embedded, or other resource-constrained devices. Their size -

¹ <https://commoncrawl.org/>

² <https://www.wikipedia.org/>

often tens of billions of parameters - limits real-world use where efficiency is crucial.

Model distillation addresses this challenge by transferring the knowledge of a large teacher model to a smaller student model, as illustrated in Figure 1 [20].

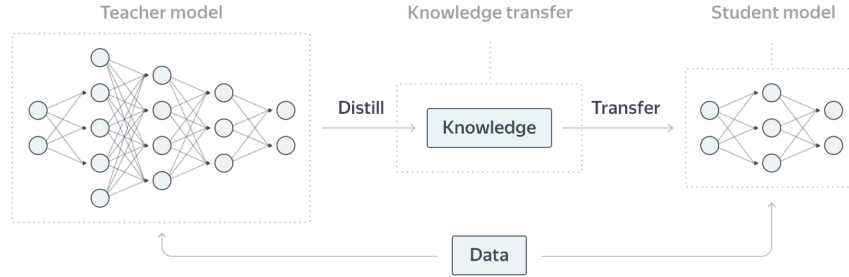


Fig. 1. General scheme of model distillation [8]

In this process, the student learns not only from original data but also from the soft labels (probabilistic outputs) of the teacher, which encode richer information than hard labels. This enables the student to generalize better and be more robust, while reducing the model’s size and inference cost.

Distilled models thus combine strong performance with efficiency, making LLM’s accessible on devices with limited resources, such as smartphones and IoT devices.

In this work, we address the challenge of compressing large language models by developing an improved distillation framework that combines iterative, data-driven evaluation of layer importance with targeted quality restoration techniques. Our method systematically identifies and removes less important transformer layers based on data-driven performance assessment and employs a combined loss function to fine-tune the reduced model on both the logits and hidden states of the original model. Experiments on the Qwen2.5-3B³ model demonstrate that this approach enables substantial reductions in model size - removing up to a third of layers with only a minor drop in quality - thus making high-performing language models more suitable for deployment in resource-constrained environments. These results underscore the effectiveness of iterative distillation and pave the way for further advances in efficient large language model adaptation. To facilitate reproducibility and future research, we publicly release the full implementation at GitHub⁴.

³ <https://huggingface.co/Qwen/Qwen2.5-3B>

⁴ https://github.com/kaengreg/layer-wise_distillation

2 Related Work

2.1 Knowledge Distillation

Knowledge distillation in its modern form was introduced by Hinton et al.[11], proposing that a compact student model can be trained to mimic a large teacher model by using the teacher’s predicted probabilities (soft labels). The distillation objective combines standard cross-entropy with a Kullback-Leibler divergence term, encouraging the student not only to match hard labels but also the probability distribution produced by the teacher:

$$\mathcal{L}_{KD} = \frac{1}{N} \sum_{i=1}^N \left(- \sum_{j=1}^K y_{ij} \log p_{ij} + \lambda D_{KL}(\mathbf{p}_i \| \mathbf{q}_i) \right)$$

where y_{ij} is the hard label, p_{ij} and q_{ij} are the student and teacher class probabilities, respectively. This “dark knowledge” in the teacher’s output captures class relationships, providing richer supervision than hard labels alone.

2.2 Distilling Step-by-Step

The “Distilling Step-by-Step” approach [12] extends classical distillation by leveraging large language models (LLMs) as teachers to provide not only labels, but also natural language rationales - explanations justifying the predictions. Inspired by Chain-of-Thought prompting [27], the teacher generates both pseudo-labels and explanations for each input. The student is then trained on triplets (x^p, r^p, y^p) : input, rationale, and label, which enhances learning, especially in low-resource settings.

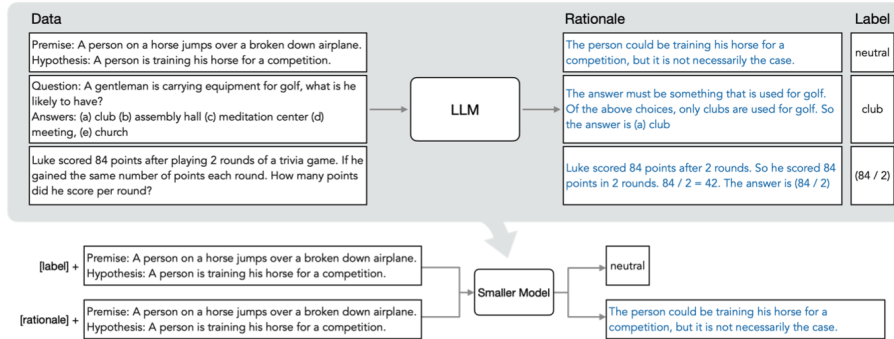


Fig. 2. An example of how teacher labels and explanations are used during student training [12].

The method involves two steps: generating teacher rationales for unlabeled data, and training the student using both text and rationale as input. The loss

combines label and rationale generation:

$$\mathcal{L} = \mathcal{L}_{\text{label}} + \lambda \mathcal{L}_{\text{rationale}},$$

where λ controls the importance of the rationale term. At inference, the student predicts without access to rationales, improving efficiency.

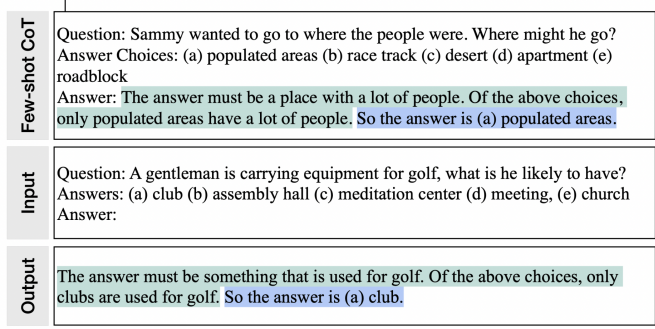


Fig. 3. Illustration for constructing input-rationale-label triplets [12].

2.3 MiniLLM

MiniLLM [9] refines knowledge distillation for generation tasks by focusing on the reverse Kullback-Leibler divergence $KL[q_\theta || p]$ instead of the forward $KL[p || q]$. While the standard approach forces the student to cover all teacher modes (sometimes yielding unlikely generations), minimizing the reverse KL encourages the student to focus on the teacher’s main modes and avoid improbable outputs. The reverse KL is minimized with a policy gradient method [22], and training is stabilized using one-step decomposition, mixed teacher forcing, and length normalization.

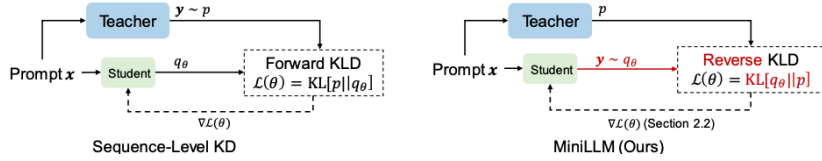


Fig. 4. Scheme of distillation using reverse-KLD [9].

This approach produces student models that are more efficient and accurate for text generation.

2.4 Distillation via Linear Pruning

The authors [19] propose an innovative distillation approach that exploits the linearity inherent in transformer architectures. By identifying and pruning highly linear layers in transformer architectures, substituting them with linear approximations, and employing mean squared error (MSE) layerwise distillation, this method ensures the distilled model retains strong performance despite significant pruning. Evaluations on datasets such as WikiText[17] and ARC-easy[5] demonstrate that this technique effectively reduces performance degradation associated with layer removal, providing a robust strategy for developing efficient transformer models.

2.5 ShortGPT

ShortGPT [16] focuses on reducing model redundancy by identifying less influential transformer layers using the Block Influence (BI) metric:

$$BI_i = 1 - \mathbb{E}_{X,t} \frac{X_{i,t}^T X_{i+1,t}}{\|X_{i,t}\|_2 \|X_{i+1,t}\|_2}$$

where $X_{i,t}$ is the hidden state vector of token t at the i^{th} layer. A low BI score indicates that $X_{i,t}$ and $X_{i+1,t}$ have high cosine similarity, meaning the layer produces only minimal transformations. Such layer is therefore considered redundant and more suitable for pruning. The method demonstrates that removing layers with low BI can significantly shrink models like Llama2-7B [25] and Baichuan2-7B [28] with minimal quality loss.

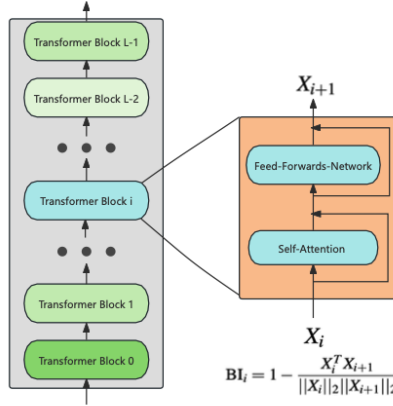


Fig. 5. Block Influence (BI) metric for transformer layers [16].

This metric guides pruning, reducing computational costs and enabling efficient deployment of LLMs.

In summary, these recent advances in distillation - leveraging soft-labels, rationales, modified loss functions, and structural metrics - enable the creation of compact, high-quality models suitable for efficient deployment.

3 Methodology: Iterative Evaluation and Distillation

3.1 Improvements to the ShortGPT Approach

The ShortGPT method, described in [16], involves evaluating the importance of layers only for the original model. This approach can be improved by measuring layer importance iteratively after each layer removal, since such an operation can affect the model’s internal structure and, consequently, the generation process in which the Block Influence metric is calculated.

Fine-tuning can be performed in the following ways:

1. Fine-tuning the student model on the logits of the teacher model.
2. Fine-tuning the student model using the hidden states of the teacher model.
3. Fine-tuning the student model simultaneously on the logits and hidden states of the teacher model.
4. Fine-tuning for a specialized task or domain.

In these and subsequent formulations, the teacher model refers to the original model, and the student model is the model, initially the same model as a teacher, from which k layers have been removed.

In this work, two modifications of the ShortGPT approach are investigated and compared:

1. Iterative measurement of layer importance after each layer removal.
2. Iterative measurement of layer importance after each layer removal, with additional fine-tuning of the model on the outputs of the teacher model to restore quality.

3.2 A New Approach to Layer Importance Evaluation

The evaluation of layer importance in ShortGPT [16] is based on the Block Influence metric, but it does not always correlate well with the actual model quality. Therefore, the following algorithm is proposed to provide a more objective assessment of layer importance:

1. Using the current model with n layers, form n variants by removing one layer at a time.
2. Each model is tested on the selected representative datasets, applying metrics such as F-Score [21], ROUGE [15]; for each layer, the drop in quality is recorded.
3. Layers whose removal results in minimal quality loss are considered the least important.

This approach, as well as the modified ShortGPT method, is analyzed in two variants:

1. Iterative determination of layer importance after their sequential removal.
2. Iterative determination of layer importance with subsequent fine-tuning on the outputs of the teacher model to improve student model quality.

4 Datasets

4.1 Evaluation Datasets

In this work, seven representative datasets were selected to assess layer importance. These datasets cover a range of natural language processing domains, including comprehension, text generation, translation, and general knowledge evaluation. This selection allows for a thorough and objective evaluation of model quality following layer removal. The datasets are introduced below, with each chosen to highlight a different capability of the model:

1. **enMMLU (English Massive Multitask Language Understanding) [10]**: This dataset consists of multiple-choice questions in English, covering various areas of knowledge, including natural sciences, humanities, and professional skills. It is designed to test a model’s ability to understand and utilize knowledge in different contexts, making it useful for assessing the model’s general knowledge.
2. **ruMMLU (Russian Massive Multitask Language Understanding)**: This dataset is similar to enMMLU but adapted for the Russian language and includes multiple-choice questions. It is important for testing models in Russian-language environments.
3. **treewayabstractive**: This dataset is designed for abstractive summarization tasks - the model must process a text in Russian and compose its summary. This allows the assessment of how well the model generates coherent and informative text.
4. **cp_doc_ru**: This dataset is used for document copying tasks in Russian. The model must reproduce the entire document, which helps assess its accuracy when copying text.
5. **cp_para_ru**: Similar to cp_doc_ru, this dataset is for paragraph-level copying tasks in Russian. It assesses the model’s ability to reproduce text at the paragraph level.
6. **flores_en_ru [6]**: This dataset includes sentence pairs for translation from English to Russian. It is intended to test machine translation quality.
7. **flores_ru_en**: The counterpart of flores_en_ru, but for translation from Russian to English. This dataset is important for evaluating bidirectional translation quality and the model’s ability to handle Russian as the source language.

The model’s overall performance is reported as the Aggregate Score, defined as the average of its results across all selected datasets. This provides a single number for comparison, but the detailed per-dataset results for the main considered method are also presented in the Table 3.

4.2 Dataset for Distillation

Selecting a diverse and high-quality dataset for the fine-tuning stage is a critical factor in achieving superior model performance. For this study, we adopted the dataset utilized in [24], which is constructed upon the comprehensive `IlyaGusev/rulm`⁵ collection. This dataset draws from a broad spectrum of sources, including Russian and English Wikipedia articles, literary works, social media content, and reputable news outlets. Such diversity and quality ensure a robust and representative data foundation, facilitating effective knowledge transfer and enhancing the outcomes of the distillation process.

5 Experiments

5.1 Experiments with Loss Functions and Hyperparameter Tuning

To determine the most effective loss functions for the distillation stage following layer pruning, we conducted a series of experiments with various configurations. We evaluated `KLDivLoss`⁶ for aligning logits and probability distributions, as well as `MSELoss`⁷ for matching hidden states and intermediate outputs. Both individual and combined loss functions were systematically assessed to identify which best preserved model quality during distillation.

Logit-based Fine-tuning with `KLDivLoss` In logit-based fine-tuning, the student model is trained to reproduce the probability distribution produced by the teacher. Logits are the raw outputs of the final model layer prior to activation, encoding the predicted probabilities for each class. The Kullback-Leibler divergence (`KLDivLoss`) minimizes the divergence between the student’s and teacher’s output distributions.

This approach enables the student not only to replicate the teacher’s predictions but also to capture the structural and semantic information present in the teacher’s distribution. The scheme is illustrated in Figure 6.

In several experiments, log-probabilities (logprobs) - obtained from applying softmax to logits - were used instead of raw logits. Logprobs, being the logarithms of class probabilities, are particularly suitable for `KLDivLoss`, which expects probability distributions as input. Using logprobs offers:

1. Simplified computations, as softmax is already applied
2. Reduced influence of extreme logit values, enhancing training stability

Additionally, we experimented with both forward and reverse KL divergence - following the recommendation of MiniLLM [9] - and used log-probabilities as targets to examine how different probability representations influence fine-tuning quality.

⁵ <https://huggingface.co/datasets/IlyaGusev/rulm>

⁶ <https://docs.pytorch.org/docs/stable/generated/torch.nn.KLDivLoss.html>

⁷ <https://pytorch.org/docs/stable/generated/torch.nn.MSELoss.html>

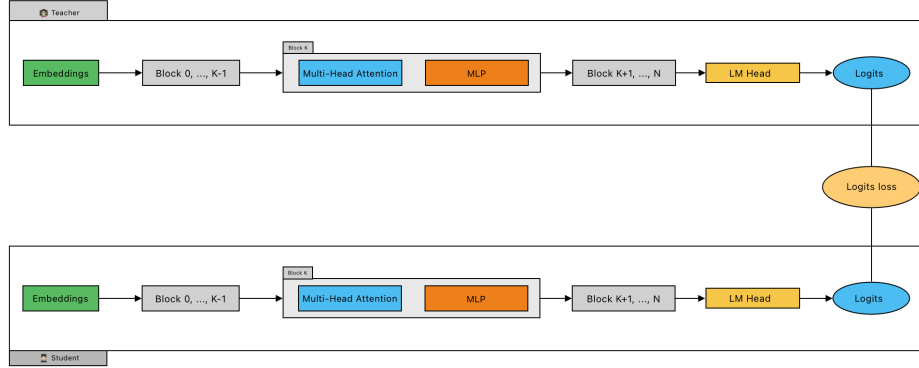


Fig. 6. Logit-based fine-tuning

Fine-tuning on Hidden States with MSELoss For knowledge transfer via intermediate representations, we used **MSELoss** (mean squared error) as a distance measure to align hidden states between teacher and student (see Fig. 7). MSE penalizes large deviations between activation vectors, providing a simple and stable objective for training. Hidden states, the activations at each layer, contain valuable information about the model’s internal structures and semantic relationships. Aligning them encourages the student to reproduce these representations, facilitating deeper knowledge transfer.

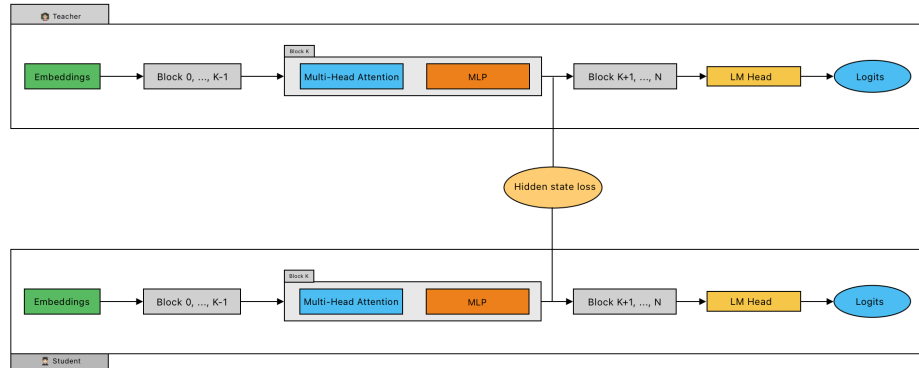


Fig. 7. Fine-tuning on hidden states

We explored the following MSELoss variants:

1. Between outputs of trainable layers: minimizes the discrepancy between corresponding activations in student and teacher, helping the student retain local processing features.

2. Between outputs of the last layers: aligns the final representations before logits formation.
3. Between outputs of the last trainable layers.
4. Combined: sums `MSELoss` across both last trainable and last layers, unifying local and global representation optimization.

Each loss function targets a different aspect of knowledge transfer:

- `KLDivLoss` (logits/logprobs): emphasizes matching output probability distributions, important for tasks where semantic closeness of predictions matters.
- `MSELoss` (hidden states): focuses on aligning internal representations, supporting transfer of deep structural and semantic features, especially valuable in multi-layer architectures.
- Combined `KLDivLoss` and `MSELoss` (Fig. 8): simultaneously optimizes both the output probability distributions and internal representations. To balance their contributions, both normalized and unnormalized variants were evaluated.

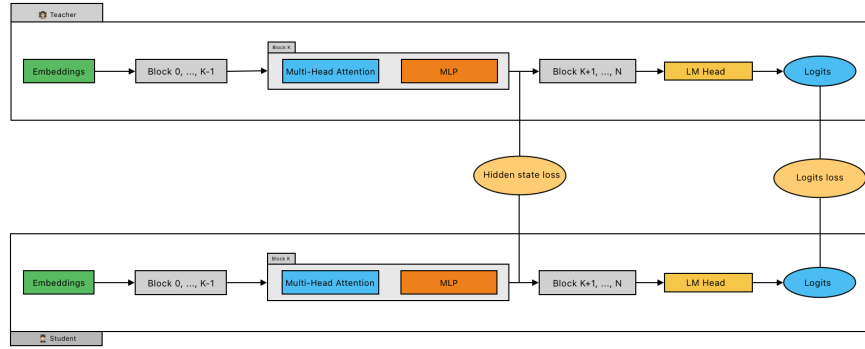


Fig. 8. Fine-tuning with a combined loss function

5.2 Experimental Results for Loss Functions

All experiments used a version of `Qwen2.5-3B` with its two least important layers removed according to the pruning order identified as optimal from evaluations on representative datasets. The results are shown in Table 1.

As shown in Table 1, the highest quality is achieved with a combined loss: `KLDivLoss` on logits (scaled by $1/100$), along with the sum of `MSELoss` over the last trainable and last model layers.

During the experiments, various combinations of hyperparameters were explored. The best balance between efficiency and quality was achieved with a learning rate of $1e-4$ and a maximum sequence length of 2048. These hyperparameters were subsequently used to evaluate the proposed distillation approaches.

Model	Aggregate Score
Original Qwen2.5-3B	0.636
Pruned 2 Layers	0.608
Pruned + HS MSE on All Layers	0.625
Pruned + HS MSE on Last Layers	0.623
Pruned + HS MSE on Last Trainable Layer	0.623
Pruned + HS MSE on Last Trainable + Last Layers	0.624
Pruned + HS MSE on All Trainable Layers	0.624
Pruned + KLDiv + MSE HS $\times 50$	0.623
Pruned + KLDiv Logits / 100 + MSE HS	0.626
Pruned + KLDiv Log-Probs + MSE HS	0.624
Pruned + KLDiv Log-Probs + MSE HS $\times 50$	0.624
Pruned + KLDiv Log-Probs / 100 + MSE HS	0.624
Pruned + KLDiv Log-Probs / 100 + MSE HS $\times 10$	0.623
Pruned + KLDiv / 100 + MSE HS	0.621
Pruned + KLDiv Reverse Teacher Log-Probs + MSE HS	0.624
Pruned + KLDiv Teacher Log-Probs	0.623
Pruned + Reverse KLDiv Student Log-Probs	0.613
Pruned + Reverse KLDiv Teacher Log-Probs	0.624
Pruned + KLDiv	0.608

Table 1. Results of fine-tuning the pruned model with different loss functions. Best loss function is highlighted. MSE HS stands for MSELoss for Hidden States of last trainable and last model’s layers

5.3 Experiments with New Distillation Methods

Experiments were carried out on the **Qwen2.5-3B** model [23], which has 36 layers and 3 billion parameters. The aim was to obtain a reduced model with 28 layers and 2.47 billion parameters, minimizing quality loss.

For comparative analysis of new distillation methods with the original ShortGPT approach [16], a reduced model was constructed using the ShortGPT method.

Five different options for distilling the **Qwen2.5-3B** model are presented:

1. **ShortGPT:** Calculate the importance of all layers in the original model once using the Block Influence metric. Remove the least important layers based on this assessment, without recalculating importance after each removal.
2. **Iterative Block Influence:** Calculate the importance of all layers using the Block Influence metric. Remove the least important layer, then recalculate importance after each removal to determine the next layer to prune.
3. **Iterative Block Influence + Distillation:** Calculate the importance of all layers using Block Influence. After each removal, perform additional fine-

tuning of the model, then recalculate importance considering changes introduced by fine-tuning.

4. **Iterative Layer-wise Pruning:** Calculate the importance of all layers in the original model. After each removal, reassess importance by evaluating on the selected representative datasets, and proceed with the removal of the next least important layer.
5. **Iterative Layer-wise Distillation:** Calculate the importance of all layers in the original model. After each removal, perform fine-tuning, then recalculate layer importance by evaluating on the selected representative datasets before continuing the pruning process.

The models obtained from these approaches are presented in Table 2.

Model Variant	Aggregate Score
Qwen2.5-3B (original)	0.636
Qwen2.5-1.5B (original)	0.601
Pruned - ShortGPT method	0.175
Pruned - Iterative Block Influence	0.142
Pruned - Iterative Layer-wise Pruning	0.179
Pruned - Iterative Block Influence + FT	0.352
Pruned - Iterative Layer-wise Distillation	0.574

Table 2. Comparison of 28-layer models obtained from Qwen2.5-3B and Qwen2.5-1.5B. Pruned models were derived from Qwen2.5-3B using various distillation strategies.

Since Iterative Layer-wise Distillation approach demonstrated the best performance, it was further investigated and extended to a model with 24 layers and 2.161 billion parameters, so that this model could be classified as a “1.5 billion” class model.

Figure 9 illustrates how model quality depends on the number of layers removed. It is evident that models produced using the ShortGPT approach, as well as those employing pruning without a distillation step, experience a substantial quality drop - exceeding 70% after the removal of eight layers. The Iterative Block Influence + Distillation method achieves approximately double the quality of these baseline approaches, yet the decline relative to the original model remains significant. In contrast, Iterative Layer-wise Distillation demonstrates the best performance among all methods considered. Notably, even after removing 12 layers using this approach, the resulting model achieves a score of 0.519 on representative datasets. Thus, eliminating one-third of the layers leads to only an 18% reduction in quality compared to the original model. The model distilled using the Iterative Layer-wise Distillation approach is publicly available on Hugging Face⁸.

⁸ <https://huggingface.co/kaengreg/Qwen2.5-2B-layerwise-distilled>

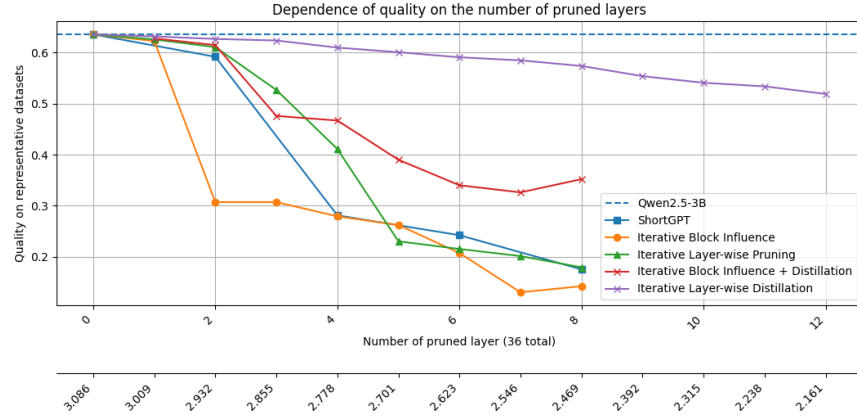


Fig. 9. Results of various approaches on the Qwen2.5-3B model

As a result of all experiments, statistics were collected (Fig. 10) on the layers removed in the approaches considered:

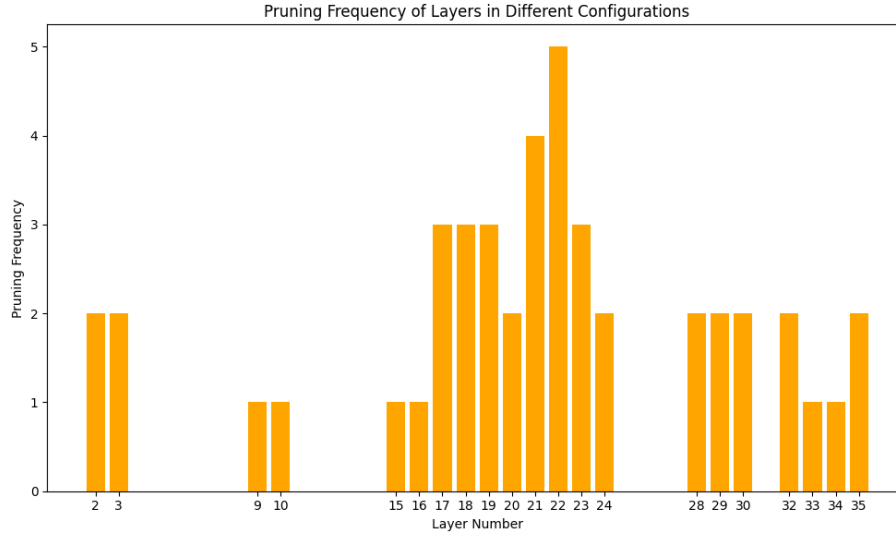


Fig. 10. Statistics on removed layers for the considered approaches

Analysis of the layer removal frequency (Fig. 10) indicates that layers 17–24 were most often identified as the least significant in our experiments. This is explained by the fact that they form intermediate representations, many of which

partially duplicate each other. In transformer architectures, the middle part of the model performs transformations that contribute little to inference and can be removed without significant quality loss. Such layers serve as aggregation blocks, but do not directly participate in either feature extraction or final generation.

In contrast, removing late (final) layers is highly undesirable. These layers play a key role in forming the final logits, i.e., they are responsible for converting internal representations into probability distributions over the vocabulary. Moreover, the uppermost layers concentrate the most semantically rich and contextually relevant features. Their removal severely degrades text generation and destroys the model’s ability to produce meaningful responses.

6 Conclusion

This work explored and systematically evaluated distillation methods for large language models, with a focus on developing compact architectures that preserve high quality while reducing computational requirements. Central to the proposed method is an iterative, data-driven assessment of layer importance using representative datasets, combined with targeted fine-tuning on teacher logits and hidden states via KLDivLoss and MSELoss.

Experimental results on the Qwen2.5-3B model demonstrate that distillation strategies incorporating iterative, layer-wise fine-tuning after each pruning step consistently outperform approaches without fine-tuning. Notably, the Iterative Layer-wise Distillation method reduced the model from 36 to 28 layers (2.47B parameters) with less than a 10

Analysis of the pruning process revealed that some intermediate layers (17–24) contribute little to inference, whereas the final layers are essential for preserving model performance. Selective fine-tuning of layers adjacent to those removed was found to be particularly effective for recovering lost dependencies. This approach holds significant promise for scaling down larger models, where such low-contribution layers are more prevalent, thereby enabling their efficient deployment in resource-limited environments.

In summary, combining iterative importance evaluation with targeted fine-tuning enables the construction of efficient, high-performing language models. Future research should further investigate the scalability and adaptability of these methods to a broader range of architectures and tasks.

Acknowledgements

The research was supported by the Russian Science Foundation, project № 25-11-00191, <https://rscf.ru/project/25-11-00191/>.

The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

7 Additional Results

Model Variant	Aggregate Score	enMMLU	ruMMLU	treeway_abstractive	cp_doc_ru	cp_para_ru	flores_en_ru	flores_ru_en
Qwen2.5-3B (original)	0.636	0.68	0.565	0.241	1	0.99	0.445	0.534
Qwen2.5-1.5B (original)	0.601	0.622	0.465	0.223	1	1	0.39	0.51
Pruned - ShortGPT method	0.175	0.281	0.293	0.145	0.02	0.04	0.164	0.283
Pruned - Iterative Block Influence	0.142	0.285	0.256	0.037	0.03	0.01	0.135	0.244
Pruned - Iterative Layer-wise Pruning	0.179	0.594	0.467	0.004	0	0	0.01	0.178
Pruned - Iterative Block Influence + FT	0.352	0.405	0.355	0.169	0.16	0.5	0.387	0.485
Pruned - Iterative Layer-wise Distillation	0.574	0.562	0.458	0.204	0.95	0.98	0.381	0.481

Table 3. Full comparison of 28-layer models obtained from **Qwen2.5-3B** and **Qwen2.5-1.5B**. Pruned models were derived from **Qwen2.5-3B** using various distillation strategies.

References

1. Bajaj, P., Campos, D., Craswell, N., Deng, L., Gao, J., Liu, X., Majumder, R., McNamara, A., Mitra, B., Nguyen, T., et al.: Ms marco: A human generated machine reading comprehension dataset. arXiv preprint arXiv:1611.09268 (2016)
2. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
3. Chiang, W.L., Zheng, L., Sheng, Y., Angelopoulos, A.N., Li, T., Li, D., Zhu, B., Zhang, H., Jordan, M., Gonzalez, J.E., et al.: Chatbot arena: An open platform for evaluating llms by human preference. In: Forty-first International Conference on Machine Learning (2024)
4. Clark, J.H., Choi, E., Collins, M., Garrette, D., Kwiatkowski, T., Nikolaev, V., Palomaki, J.: Tydi qa: A benchmark for information-seeking question answering

- in typologically diverse languages. *Transactions of the Association for Computational Linguistics* **8**, 454–470 (2020)
5. Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O.: Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv:1803.05457v1 (2018)
 6. Costa-jussà, M.R., Cross, J., Çelebi, O., Elbayad, M., Heafield, K., Heffernan, K., Kalbassi, E., Lam, J., Licht, D., Maillard, J., et al.: No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672 (2022)
 7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Burstein, J., Doran, C., Solorio, T. (eds.) *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019). <https://doi.org/10.18653/v1/N19-1423>, <https://aclanthology.org/N19-1423/>
 8. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. *International Journal of Computer Vision* **129**(6), 1789–1819 (2021)
 9. Gu, Y., Dong, L., Wei, F., Huang, M.: Minillm: Knowledge distillation of large language models. In: *The Twelfth International Conference on Learning Representations* (2023)
 10. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300 (2020)
 11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
 12. Hsieh, C.Y., Li, C.L., Yeh, C.K., Nakhost, H., Fujii, Y., Ratner, A., Krishna, R., Lee, C.Y., Pfister, T.: Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. arXiv preprint arXiv:2305.02301 (2023)
 13. Jiao, X., Yin, Y., Shang, L., Jiang, X., Chen, X., Li, L., Wang, F., Liu, Q.: Tinybert: Distilling bert for natural language understanding. arXiv preprint arXiv:1909.10351 (2019)
 14. Kwiatkowski, T., Palomaki, J., Redfield, O., Collins, M., Parikh, A., Alberti, C., Epstein, D., Polosukhin, I., Devlin, J., Lee, K., et al.: Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics* **7**, 453–466 (2019)
 15. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: *Text summarization branches out*. pp. 74–81 (2004)
 16. Men, X., Xu, M., Zhang, Q., Wang, B., Lin, H., Lu, Y., Han, X., Chen, W.: Shortgpt: Layers in large language models are more redundant than you expect. *CoRR* (2024)
 17. Merity, S., Xiong, C., Bradbury, J., Socher, R.: Pointer sentinel mixture models (2016)
 18. Rajpurkar, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
 19. Razzhigaev, A., Mikhalechuk, M., Goncharova, E., Gerasimenko, N., Oseledets, I.V., Dimitrov, D., Kuznetsov, A.: Your transformer is secretly linear. *CoRR* (2024)
 20. Shiyonov, V.: 7.2 knowledge distillation. *Yandex ML-Handbook* (2024)
 21. Sokolova, M., Japkowicz, N., Szpakowicz, S.: Beyond accuracy, f-score and roc: a family of discriminant measures for performance evaluation. In: *Australasian joint conference on artificial intelligence*. pp. 1015–1021. Springer (2006)

22. Sutton, R.S., McAllester, D., Singh, S., Mansour, Y.: Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems* **12** (1999)
23. Team, Q.: Qwen2.5: A party of foundation models (09 2024), <https://qwenlm.github.io/blog/qwen2.5/>
24. Tikhomirov, M., Chernyshov, D.: Facilitating large language model russian adaptation with learned embedding propagation. *Journal of Language and Education* **10**(4), 130–145 (2024)
25. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* (2023)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. *Advances in neural information processing systems* **30** (2017)
27. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35**, 24824–24837 (2022)
28. Yang, A., Xiao, B., Wang, B., Zhang, B., Bian, C., Yin, C., Lv, C., Pan, D., Wang, D., Yan, D., et al.: Baichuan 2: Open large-scale language models. *CoRR* (2023)
29. Zhang, X., Thakur, N., Ogundepo, O., Kamalloo, E., Alfonso-Hermelo, D., Li, X., Liu, Q., Rezagholizadeh, M., Lin, J.: Miracl: A multilingual retrieval dataset covering 18 diverse languages. *Transactions of the Association for Computational Linguistics* **11**, 1114–1131 (2023)

Response to Reviewers

Review 1

1. “Please, describe explicitly what X , i , t , t_2 are in the expression for BI.” — The formula was corrected and a detailed description of X , i , and t was added.
2. “Section 4.1. One benchmark might have some peculiarities, which would lead to biased scores. It is recommended to perform tests on several datasets.” — This was a misunderstanding of the section’s purpose. The section (ex. 4.1, now 5.2) in question described datasets used for distillation, not for evaluation. To avoid confusion, the text was reworked and all datasets used for evaluation were moved into a dedicated **Datasets** section.
3. “As far as the model pruning is iterative, the order in this procedure might also affect the result.” — In this section we investigated models with two pruned layers following a predefined order. This order was determined in preliminary experiments as the most effective. The influence of layer order on results was acknowledged and incorporated into the experimental setup.
4. “Why MSE is utilized to score hidden states. Do we know the residuals in this case are distributed normally?” — MSE was used as a distance measure to align hidden states between teacher and student. It penalizes large deviations between activation vectors, providing a simple and stable training objective. This approach is standard in distillation (e.g., [13]). Corresponding clarifications were added to the text.

5. “Table 1 should explicitly provide the name of the benchmark score used.” — The term **Benchmark score** was renamed to **Aggregate Score** for clarity.
6. “It looks like “Pruned – Iterative Layer-wise Distillation” 3B model is worse than Qwen2.5-1.5B. I believe it should be commented on which cases one might prefer the distilled 3B model over the original 1.5B model.” — While we initially expected to outperform Qwen2.5-1.5B, this did not occur under the presented configuration. Qwen2.5-3B was selected for its moderate size, which enabled a broader series of experiments, while our primary objective was to demonstrate the effectiveness of the proposed method. We note that further fine-tuning may yield competitive or superior results even compared to Qwen2.5-1.5B, and we plan to investigate this in future work, including experiments with larger base models.

Review 2

1. “The paper compares only to ShortGPT, omitting benchmarks against other distillation methods (e.g., Distilling Step-by-Step, MiniLLM, or linear pruning). This weakens claims of superiority.” — Our work builds upon the ShortGPT framework, with our contributions focusing on extending and improving this approach. For consistency, we used ShortGPT as the baseline. We plan to extend comparisons to other distillation methods (e.g., MiniLLM, Distilling Step-by-Step) in future work.
2. “No study on the impact of individual components (iterative reevaluation vs. fine-tuning alone).” — We agree this is an important observation. Due to computational constraints, a detailed ablation study was beyond the scope of this work, but we intend to pursue it in future research.
3. “Layer redundancy analysis (layers 17–24) lacks statistical validation (confidence intervals).” — The observation regarding layers 17–24 is based on empirical evidence: in our experiments these layers were consistently pruned with minimal performance loss. Given the limited number of runs, confidence intervals could not be meaningfully reported. We also clarified in the text that “redundancy” denotes limited contribution to inference performance, rather than a flaw in the original architecture.
4. “Results are limited to Qwen2.5-3B with no experiments on larger models.” — Qwen2.5-3B was selected for its moderate size, which enabled a broader series of experiments. We expect our method to generalize well to larger models, and this is planned for future work.

Review 3

1. “I did not have a systematic presentation of the factors that make up the research. There are many “old” and “new” methods, but is there a clear system for how this zoo is formed?” — Our work was built on the ShortGPT framework, with our contributions representing extensions and improvements to it. Other approaches were included in the related works section, and several ideas from them were integrated into our methods.

2. “Why are datasets classified as methods?” — This has been corrected: datasets used for evaluation are now presented in a separate section.