

Importance-Guided Structural Pruning of Aya Expanse 8B and Gemma-3-12B for English–Chinese Machine Translation

Anish Racherla* Navadeep Kiran* Havish Balaga*
Saumitra Yadav and Manish Shrivastava

MT-NLP Lab, LTRC, KCIS, IIIT Hyderabad, India
(anish.racherla,kolupoti.kiran,havish.balaga)@students.iiit.ac.in
saumitra.yadav@research.iiit.ac.in
m.shrivastava@iiit.ac.in

Abstract

Large language models have significantly improved machine translation quality but remain expensive to deploy due to high computational and memory requirements. This paper describes our submissions to the WMT 2026 Model Compression Shared Task, targeting English–Chinese translation through structural compression of two multilingual models: Aya Expanse 8B and Gemma-3-12B. Both pipelines share a common 6-stage data curation process that filters 25 million raw WMT English–Chinese sentence pairs into a quality and diverse training corpus for fine-tuning. For Aya Expanse 8B, we use COMET-QE to identify and prune four layers with the least impact on translation quality, followed by QLoRA fine-tuning and INT8 inference. For Gemma-3-12B, we apply Fisher Information-based importance scoring to guide a two-step compression, pruning 8 transformer layers followed by feed-forward neuron pruning, recovered through QLoRA fine-tuning and INT4 inference. Results on the FLORES-200 benchmark show both pipelines achieve substantial reductions in GPU memory usage and gains in inference throughput while maintaining competitive translation quality. Our findings demonstrate that importance-guided structural pruning, combined with parameter-efficient fine-tuning and quantization, can achieve substantial model size reduction while maintaining comparable English–Chinese translation quality.

1 Introduction

Large language models (LLMs) have recently become highly effective machine translation systems, achieving competitive performance across a wide variety of language pairs. Multilingual instruction-tuned models such as Aya Expanse (Dang et al., 2024) and Gemma-3-12B (Gemma Team, 2025) are capable of performing high-quality English

to Chinese translation without requiring separate models for each language pair. However, their large parameter counts and computational requirements make deployment challenging, particularly in resource-constrained environments where GPU memory and inference latency are critical considerations.

Model compression aims to address these limitations by reducing computational and memory requirements while preserving translation quality (Zhu et al., 2023). Existing approaches include quantization (Gholami et al., 2021), knowledge distillation (Hinton et al., 2015), pruning (structured (Ma et al., 2023) and unstructured (Han et al., 2015)), and parameter-efficient fine-tuning (Hu et al., 2021; Dettmers et al., 2023). Among these techniques, structural pruning offers a distinct advantage: by eliminating redundant components of the network entirely, it produces a genuinely smaller model that can be deployed on standard inference infrastructure, without reliance on specialized sparse hardware or custom kernels (Ma et al., 2023). A key challenge in structural pruning is determining which layers can be removed without substantially degrading performance. Since different layers capture different linguistic representations, naively removing layers often results in significant performance loss. Identifying the relative importance of each layer is therefore essential prior to structural compression (Jawahar et al., 2019; Tenney et al., 2019).

In this work, we present two system submissions to the WMT 2026 Model Compression Shared Task, both targeting English–Chinese translation. Our first system compresses Aya Expanse 8B by translating held-out sentence groups with individual layers ablated and scoring the resulting translations with COMET-QE, using this sensitivity analysis to identify and prune the four least impactful layers, followed by QLoRA fine-tuning and INT8 inference. Our second system compresses

** Authors contributed equally

Gemma-3-12B using Fisher Information scoring (Kirkpatrick et al., 2017) to guide a two-step compression: layer pruning followed by feed-forward neuron pruning, with quality recovered through QLoRA fine-tuning (Detmers et al., 2023). A shared 6-stage data curation pipeline filters 25 million raw WMT English–Chinese sentence pairs into a quality and diverse corpus, on which both models are fine-tuned.

Our contributions are summarized as follows:

- We design a 6-stage data curation pipeline that filters 25 million raw WMT English–Chinese sentence pairs into a quality and diverse training corpus, used to fine-tune both compressed models.
- We compress Aya Expanse 8B by using COMET-QE to evaluate translations under layer ablation, guiding the pruning of 4 layers (32→28), with quality recovered through QLoRA fine-tuning and INT8 deployment.
- We compress Gemma-3-12B using Fisher Information-based importance scoring to guide a two-stage structured pruning process, removing eight transformer layers (48→40) and 30% of feed-forward neurons, followed by QLoRA fine-tuning to recover translation quality.
- We demonstrate improved memory efficiency and inference throughput for both systems while maintaining comparable English–Chinese translation quality on the FLORES-200 benchmark (NLLB Team et al., 2022).

2 Experimental Setup

To explore the effectiveness of different importance estimation strategies, we investigate two complementary compression pipelines. Aya Expanse 8B employs task-driven COMET-QE-based sensitivity analysis, whereas Gemma-3-12B uses Fisher Information-based parameter importance to enable scalable structured pruning.

We describe our experimental pipeline. We first curate a training corpus that is later used to fine-tune both compressed models. We then detail the model-specific structural pruning strategies applied to Aya Expanse 8B and Gemma-3-12B, followed by the shared fine-tuning procedure and evaluation protocol used to assess both systems.

2.1 Data Curation Pipeline for Finetuning

To ensure fine-tuning is performed on clean, representative data and to reduce the risk of overfitting to noisy web-scraped text, we design a 6-stage data curation pipeline. This pipeline filters an initial 25,984,574 raw WMT English–Chinese sentence pairs into a quality and diverse training corpus, later used to fine-tune both compressed models.

1. **Basic Scrubbing & Deduplication.** We remove rows containing null values and exact duplicates. Web-scraped data frequently contains missing cells and repetitive copy-pasted content; removing duplicates reduces the risk of the model overfitting to repeated phrases.
2. **Heuristics & Alignment.** We apply heuristic filters to remove malformed or misaligned pairs and reduce the risk of out-of-memory errors during training:
 - **Length limits:** sentences exceeding 1,000 characters or containing fewer than 10 characters are removed.
 - **Length ratio filter:** source–target pairs with a character length ratio outside the range of 0.3 to 3.0 are discarded.
 - **Content filtering:** rows consisting purely of URLs, HTML tags, or random numeric strings are removed.

After Steps 1 and 2, the dataset is reduced to 9,539,319 pairs.

3. **Language Identification.** We use the Fast-Text language identification model (Joulin et al., 2017) to verify that each sentence pair is correctly identified as English (en) on the source side and Chinese (zh) on the target side, discarding rows that fail this check, leaving 5,933,529 pairs.
4. **LLM Fluency Grading via Perplexity.** We use Qwen-0.5B (Team, 2024) to compute a perplexity score for each sentence, where lower perplexity indicates more fluent, native-sounding text and higher perplexity suggests disfluent or unnatural phrasing. We retain only the lowest-perplexity sentences, reducing the corpus to 4,908,074 pairs.
5. **Topic Diversity via Embeddings.** To reduce the risk of the model overfitting to a dominant domain (e.g., politics), we embed each

remaining sentence into a 768-dimensional space using LaBSE (Feng et al., 2022) and cluster the embeddings into $k = 10,000$ clusters using MiniBatch K-Means. To maintain topical diversity, we sample up to 100 sentences with the lowest perplexity from each cluster, yielding a balanced corpus of 436,007 sentence pairs.

6. **Translation Quality Estimation.** Finally, we use COMET-QE (Rei et al., 2020) to estimate the quality of each English–Chinese pair. We retain only pairs with a COMET-QE score ≥ 0.60 , filtering out low-quality or semantically inconsistent translations and yielding our final training corpus of 36,722 sentence pairs.

2.2 Compression: Aya Expanse 8B

A central challenge in structural pruning is identifying which transformer layers can be removed with minimal impact on translation quality. Rather than relying on weight magnitude or activation statistics, we estimate layer importance directly from downstream translation quality.

Starting from the original 32-layer Aya Expanse 8B model, each layer is evaluated independently by temporarily bypassing it while leaving the rest of the architecture unchanged. The modified model is used to translate a held-out set of English–Chinese sentences from the FLORES-200 development set (NLLB Team et al., 2022) (which are kept strictly separate from the final evaluation dataset to avoid bias), and the resulting translations are scored using COMET-QE (Rei et al., 2020). The drop in COMET-QE score relative to the unmodified model reflects that layer’s contribution to translation quality: layers whose removal causes the smallest degradation are considered the least sensitive and are selected as pruning candidates. This task-oriented evaluation grounds layer selection in translation quality rather than proxy importance measures. Based on this ranking, we permanently prune the four least sensitive layers, reducing network depth from 32 to 28. Internal layer indices are updated to maintain compatibility with the attention cache during generation.

2.3 Compression: Gemma-3-12B

For Gemma-3-12B, we estimate component importance using the Fisher Information matrix (Kirkpatrick et al., 2017), computed over a calibration set of English–Chinese sentence pairs. The cali-

bration set is passed through the model to compute gradients of the translation cross-entropy loss with respect to each parameter, and the squared gradient magnitude is used as that parameter’s importance score. Parameters with higher Fisher scores are treated as contributing more strongly to translation performance.

For the 48-layer Gemma-3-12B model (Gemma Team et al., 2024), we compute Fisher scores at two levels of granularity: per transformer layer, by aggregating parameter-level scores across each layer, and per feed-forward neuron within the feed-forward sublayers. Attention heads are excluded from feed-forward neuron-level scoring due to key-value mapping constraints. These two sets of scores guide a two-step structured compression: first, the lowest-importance transformer layers are entirely removed, reducing network depth; then, within the remaining layers, individual feed-forward neurons with the lowest Fisher scores are pruned, reducing network width.

2.4 Fine-tuning Procedure

Following pruning, the remaining layers of each model are reconnected to form a physically smaller, dense architecture, and the model configuration is updated accordingly. A forward-pass validation is performed on each compressed model to confirm it produces valid outputs prior to fine-tuning.

Although structural pruning substantially improves inference efficiency, removing transformer components introduces a degradation in translation quality. To recover this loss, both compressed models are fine-tuned using Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023; Hu et al., 2021). Each model is loaded in 4-bit NF4 precision using the BitsAndBytes framework (Dettmers et al., 2022) and the Hugging Face Transformers library (Wolf et al., 2020), with LoRA adapters inserted into the query and value projection matrices of each attention module. Only the LoRA parameters are updated during optimization, while the original model weights remain frozen, substantially reducing GPU memory consumption during training.

Both models are fine-tuned on the curated English–Chinese corpus described in Section 2.1, with sentence pairs formatted as instruction-following translation prompts compatible with each model’s respective instruction-tuning format. Training uses the paged AdamW 8-bit optimizer with gradient checkpointing and sequence packing

Hyperparameter	Value
Base Model	Aya Expanse 8B
Layers	28
Quantization	NF4 (Training)
LoRA Rank	8
LoRA Alpha	16
LoRA Dropout	0.05
Optimizer	Paged AdamW 8-bit
Learning Rate	2×10^{-4}
Epochs	1
Batch Size	1
Gradient Accumulation	16
Maximum Sequence Length	512

Table 1: Training hyperparameters for Aya Expanse 8B QLoRA fine-tuning.

Hyperparameter	Value
Base Model	Gemma-3-12B
Architecture	40 Layers (Pruned)
Quantization	NF4 (Training)
LoRA Rank	16
LoRA Alpha	32
LoRA Dropout	0.05
Optimizer	Paged AdamW 8-bit
Learning Rate	2×10^{-4}
Epochs	1
Batch Size	1
Gradient Accumulation	16
Maximum Sequence Length	512

Table 2: Training hyperparameters for Gemma-3-12B QLoRA fine-tuning.

to maximize GPU utilization. Tables 1 and 2 report the training hyperparameters for Aya Expanse 8B and Gemma-3-12B, respectively. After optimization, the learned LoRA adapters are merged into each base model using the `merge_and_unload()` operation, producing standalone checkpoints that no longer depend on external adapter weights.

2.5 Evaluation Protocol

Both compressed systems are evaluated on the English–Chinese split of the FLORES-200 benchmark (NLLB Team et al., 2022). Translation quality is measured using COMET-QE (Rei et al., 2020), which has been shown to correlate well with human judgments of semantic adequacy and fluency without requiring reference translations. To assess computational efficiency, we additionally report GPU memory consumption, model loading time, and inference throughput, measured in generated tokens per second. Together, these metrics quantify the trade-off between translation quality and deployment efficiency introduced by compression. All fine-tuning and evaluation experiments were conducted in a Kaggle environment utilizing

Rank	Layer	COMET-QE	COMET-QE Drop
1	6	0.6253	-0.0493
2	30	0.6144	-0.0384
3	8	0.6132	-0.0372
4	25	0.6045	-0.0285
5	7	0.5874	-0.0114

Table 3: Aya Expanse 8B: Layer ranking obtained through COMET-QE-based sensitivity analysis.

Group	Layers Dropped	COMET-QE	COMET-QE Drop
top-2	{6, 30}	0.6217	-0.1953
top-4	{6, 30, 8, 25}	0.6265	-0.1905
top-6	{6, 30, 8, 25, 7, ·}	0.5717	-0.2453
top-8	{6, 30, 8, 25, 7, ·, ·, ·}	0.4792	-0.3378

Table 4: Aya Expanse 8B: Combination sensitivity results. Each group removes the top- k least-sensitive layers simultaneously and is scored on English–Chinese translation quality.

dual NVIDIA T4 GPUs, demonstrating the viability of our compression pipelines under highly constrained computational resources.

3 Results

3.1 System 1: Aya Expanse 8B

Table 3 reports the COMET-QE-based sensitivity ranking obtained by temporarily bypassing each layer of the original 32-layer Aya Expanse 8B model and measuring the resulting drop in translation quality on the FLORES-200 English–Chinese development set. Layers 6, 8, 25, and 30 consistently produce the smallest COMET-QE degradation and are therefore selected for pruning. Notably, three of the four selected layers lie in the lower half of the network, suggesting that lower transformer layers contribute less to task-critical translation representations than middle or upper layers.

Since individual layer ranking does not guarantee that the jointly-removed set is optimal, we additionally evaluate four combination groups — the top-2, top-4, top-6, and top-8 least-sensitive layers removed simultaneously — on the same 50-pair development set. Table 4 reports the COMET-QE drop for each group relative to the unmodified baseline.

The group yielding the lowest COMET-QE drop is selected as the final pruning set and permanently removed from the model. Based on this ranking, layers 6, 8, 25, and 30 are pruned, reducing network depth from 32 to 28 transformer blocks. Table 5 compares the resulting compressed model against the original under INT8 inference.

The compressed model reduces GPU memory

Metric	Baseline	Pruned
Load Time (s)	22.2	21.1
VRAM (GB)	8.46	7.65
Tokens/s	5.9	6.7
COMET-QE	0.8963	0.8551

Table 5: Performance comparison for Aya Expanse 8B on English–Chinese translation (INT8 inference).

consumption from 8.46 GB to 7.65 GB (approximately 10%) and increases inference throughput from 5.9 to 6.7 tokens per second (approximately 14%). Interestingly, the combination analysis revealed a non-linear interaction when removing layers jointly: the top-4 group (0.6265) actually achieved a higher COMET-QE score than the top-2 group (0.6217). This suggests that jointly removing these four specific layers preserves internal representations better than removing just two, likely because they collectively encapsulate a redundant or noisy feature subset. Beyond four layers, translation quality degrades substantially, with the top-6 and top-8 groups dropping to 0.5717 and 0.4792 respectively. The top-4 group (layers 6, 30, 8, 25) therefore offers the optimal structural pruning configuration. This comes at a modest cost in final fine-tuned translation quality, with the COMET-QE score decreasing from 0.8963 to 0.8551 (a drop of 0.041), which QLoRA fine-tuning largely recovers relative to the immediate post-pruning degradation observed during sensitivity analysis.

3.2 System 2: Gemma-3-12B

Table 6 reports the Fisher Information scores computed for all 48 layers of the original Gemma-3-12B model. Scores are markedly lower for layers in the first half of the network (layers 0–17) and rise sharply from the middle layers onward, suggesting that early layers contribute comparatively little to the English–Chinese translation objective under this calibration procedure. Based on this ranking, we remove the 8 lowest-scoring transformer layers (layers 47, 13, 9, 8, 11, 12, 0, and 7), reducing network depth from 48 to 40, and subsequently prune 30% of feed-forward neurons in the remaining layers based on their individual Fisher scores.

Table 7 reports results at three stages: the original baseline, immediately after structural pruning, and after QLoRA fine-tuning, all under INT4 inference.

Structural pruning reduces VRAM usage from 7.61 GB to 5.85 GB after fine-tuning (approximately 23%) and substantially improves load time,

from 221.1 seconds to 136.9 seconds (approximately 38%). Immediately after pruning, however, COMET-QE collapses from 0.8599 to 0.2664, indicating that removing 8 layers and 30% of feed-forward neurons in combination severely disrupts the model’s internal representations, consistent with the more aggressive nature of this two-step compression relative to Aya Expanse 8B’s single-step layer removal. Notably, this pruned-only stage also shows the highest inference throughput (7.28 tokens/s) of the three, but this apparent efficiency gain is not meaningful on its own: with translation quality near-collapsed, the pruned model is not yet a usable system. QLoRA fine-tuning substantially recovers quality, raising COMET-QE to 0.7925, but throughput falls to 4.33 tokens/s in the process — below even the original baseline (4.74 tokens/s, a drop of approximately 8.6%). This indicates that, unlike Aya Expanse 8B, where pruning and fine-tuning jointly improved both quality and throughput, recovering translation quality for Gemma-3-12B comes at the cost of the raw throughput gained from pruning alone. The final fine-tuned COMET-QE score also remains 0.067 below baseline, a larger residual gap than observed for Aya Expanse 8B (0.041), suggesting that this more aggressive compression strategy is only partially recoverable through low-rank adaptation.

We attribute the throughput drop after fine-tuning to the merged LoRA adapters altering weight matrix memory layout in a way that reduces INT4 kernel efficiency relative to the freshly pruned, pre-fine-tuning checkpoint; we did not further isolate this effect and flag it as a limitation of the current evaluation.

3.3 Comparative Discussion

The two systems reflect different points on the trade-off between compression aggressiveness and recoverability. Aya Expanse 8B’s modest, COMET-QE-guided layer removal improves both memory usage and throughput alongside a relatively small quality drop. Gemma-3-12B’s more aggressive Fisher Information-guided compression, combining layer and feed-forward neuron pruning, yields larger memory savings but causes a much sharper immediate quality collapse, only partially recovered through fine-tuning. This may indicate that quality loss grows disproportionately as compression combines multiple pruning granularities, though this comparison is confounded by differences in base model, pruning ratio, and

Layer	Score	Layer	Score	Layer	Score	Layer	Score
0	0.1210	12	0.1193	24	1.4456	36	3.1873
1	0.1430	13	0.1069	25	1.5730	37	3.3259
2	0.2624	14	0.1369	26	1.6743	38	2.8655
3	0.1770	15	0.1546	27	2.0750	39	2.7946
4	0.2174	16	0.1384	28	2.6097	40	2.8570
5	0.1509	17	0.3954	29	2.5598	41	2.8843
6	0.1247	18	0.8574	30	3.1515	42	3.1561
7	0.1224	19	1.1615	31	2.7898	43	3.7430
8	0.1122	20	1.5012	32	2.9118	44	4.3303
9	0.1115	21	1.1193	33	3.1456	45	4.2049
10	0.1459	22	0.4643	34	3.0104	46	1.9676
11	0.1144	23	1.0325	35	2.8306	47	0.0000

Table 6: Fisher Information scores for all 48 layers of Gemma-3-12B. Lower scores indicate lower estimated importance to English–Chinese translation quality.

Metric	Baseline	Pruned	Fine-tuned
VRAM (GB)	7.61	6.72	5.85
Load Time (s)	221.1	110.7	136.9
Speed (tok/s)	4.74	7.28	4.33
COMET-QE	0.8599	0.2664	0.7925

Table 7: Compression results for Gemma-3-12B on English–Chinese translation (INT4 inference).

importance-estimation method, and cannot isolate a single cause.

The two importance estimation strategies also differ in what they optimize for: COMET-QE directly measures translation quality but requires scoring every candidate layer, while Fisher Information estimates importance from training-loss gradients without repeated generation, making it more scalable to larger search spaces at the cost of only indirectly correlating with translation quality. This indirectness may partly explain Gemma-3-12B’s sharper collapse, though the effect cannot be disentangled from its more aggressive pruning ratio in the present comparison.

The throughput drop observed after fine-tuning Gemma-3-12B further suggests that structural compression and adapter merging can interact in ways not captured by compression ratio alone; we did not isolate the specific cause and leave a closer analysis to future work.

These results are suggestive rather than conclusive regarding the choice of importance estimation strategy: task-aligned metrics such as COMET-QE and more scalable proxies such as Fisher Information may suit different compression regimes, but confirming this would require evaluating both strategies under matched pruning ratios and base models. Future work could explore adaptive per-model pruning aggressiveness, distillation-guided

recovery, and a closer analysis of adapter merging’s effect on low-bit inference throughput.

4 Conclusion

This paper presented two compression pipelines submitted to the WMT 2026 Model Compression Shared Task, targeting English–Chinese machine translation. Both pipelines share a common 6-stage data curation process and are compressed using sensitivity-guided structural pruning followed by QLoRA fine-tuning. Aya Expanse 8B is compressed via COMET-QE-guided layer removal (32→28 layers), reducing GPU memory by approximately 10% and improving inference throughput by approximately 14%, with a modest drop in COMET-QE score. Gemma-3-12B is compressed via Fisher Information-guided two-step pruning (48→40 layers, plus 30% feed-forward neuron pruning), reducing VRAM from 7.61 GB to 5.85 GB (23%) and load time from 221.1s to 136.9s (38%), though at the cost of a larger residual quality gap and reduced post-fine-tuning throughput relative to baseline.

These results show that sensitivity-guided structural pruning, combined with parameter-efficient fine-tuning, can meaningfully reduce the deployment cost of large multilingual translation models, though the achievable trade-off between compression aggressiveness and recoverable quality varies with the pruning strategy used. Future work includes evaluating both importance-estimation methods under matched pruning ratios, testing cross-application methods (e.g., Fisher Information on Aya Expanse and COMET-QE on Gemma) for direct architectural comparisons, and exploring distillation-based recovery for more aggressive compression regimes.

References

- John Dang et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.
- Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.
- Amir Gholami, Sehoon Kim, Zhen Dong, Zhewei Yao, Michael W. Mahoney, and Kurt Keutzer. 2021. A survey of quantization methods for efficient neural network inference. *arXiv preprint arXiv:2103.13630*.
- Song Han, Jeff Pool, John Tran, and William J. Dally. 2015. Learning both weights and connections for efficient neural networks. *Advances in Neural Information Processing Systems*, 28.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.
- Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.
- James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526.
- Xinyin Ma, Gongfan Fang, and Xinchao Wang. 2023. LLM-pruner: On the structural pruning of large language models. In *Advances in Neural Information Processing Systems*.
- NLLB Team, Marta R. Costa-jussà, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Qwen Team. 2024. Qwen2.5 technical report. *Preprint, arXiv:2412.15115*.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.
- Thomas Wolf et al. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*.
- Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2023. A survey on model compression for large language models. *arXiv preprint arXiv:2308.07633*.