

Tiny Titans at WMT26: Task-Calibrated Quantization and Activation-Aware Low-Rank Compression of Gemma 3 12B

Vania Raya-Rios* Aleksandr Klimchuk* Nicolas M. Cuadrado†

Ivo Gollini Navarrete Samuel Horvath Martin Takac

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

Abu Dhabi, United Arab Emirates

*Equal contribution

†Correspondence: nicolas.avila@mbzuai.ac.ae

Abstract

We present the TINY TITANS submissions for the WMT26 Model Compression shared task, which focuses on Czech-to-German translation. Starting with the instruction-tuned Gemma 3 12B model, we compare two approaches: 4-bit GPTQ, which uses task-specific calibration examples, and a more aggressive low-rank factorization method that replaces attention and feed-forward projections with truncated Singular Value Decomposition (SVD) factors. On a 300-sentence development set, the calibrated GPTQ retains 99.7% of the dense model’s COMET score while reducing GPU memory usage from 26.6 GiB to 17.6 GiB. The recovered SVD model reduces the number of parameters by 46.6% and reaches 0.7804 COMET in a separate evaluation using 1,997 sentence pairs. In an external evaluation using 457 WMT25 paragraphs, the GPTQ, SVD, and SVD-plus-quantization submissions received COMET scores of 0.5751, 0.4263, and 0.3354, respectively. These findings suggest that while post-training quantization effectively reduces memory usage, combining low-rank truncation with low-bit quantization may increase approximation errors without clear efficiency gains.

1 Introduction

Large multilingual language models are capable translation systems, but their memory footprint and inference cost make deployment difficult (Wang et al., 2024a). Compression is especially challenging for machine translation: small perturbations can change lexical choice, agreement, or adequacy even when the generated text remains fluent (Aji and Heafield, 2020). The WMT26 Model Compression shared task (Gowda et al., 2026) evaluates this trade-off directly through translation quality, model size, GPU memory, and decoding speed. In the constrained track, all teams use Gemma 3 12B (Team et al., 2025) and rely solely on a predefined data pool.

We participated as TINY TITANS, investigating whether two compression paradigms can be combined effectively. Weight-only quantization reduces numerical precision while preserving the model architecture, whereas low-rank factorization reduces the dimensionality of its linear maps while preserving the numerical format. The latter is attractive because Transformer projections contain substantial algebraic redundancy (Michel et al., 2019), but naive truncation can degrade task-relevant directions.

Our experiments compare how numerical and structural compression behave under the same translation task. In particular, we examine whether parameter reduction yields practical gains in memory or runtime, and whether low-rank truncation can be safely combined with low-bit quantization. Our comparison of the three packaged submissions under a common external protocol provides evidence about their relative performance independently of the dense baseline.

2 Related Work

Recent work on LLM compression spans quantization, pruning, distillation, and compact architectures (Wang et al., 2024a). GPTQ applies approximate second-order information to one-shot weight quantization (Frantar et al., 2023). AWQ instead identifies salient weight channels through activation statistics and protects them by an equivalent scaling transformation (Lin et al., 2024). These methods preserve the original architecture and can therefore benefit from existing low-bit inference kernels. Our primary system follows this direction, but calibrates GPTQ using Czech–German translation examples.

Low-rank compression instead changes the model’s structure by replacing dense matrices with smaller factors. SVD-LLM (Wang et al., 2024b) and ASVD (Yuan et al., 2025) replace

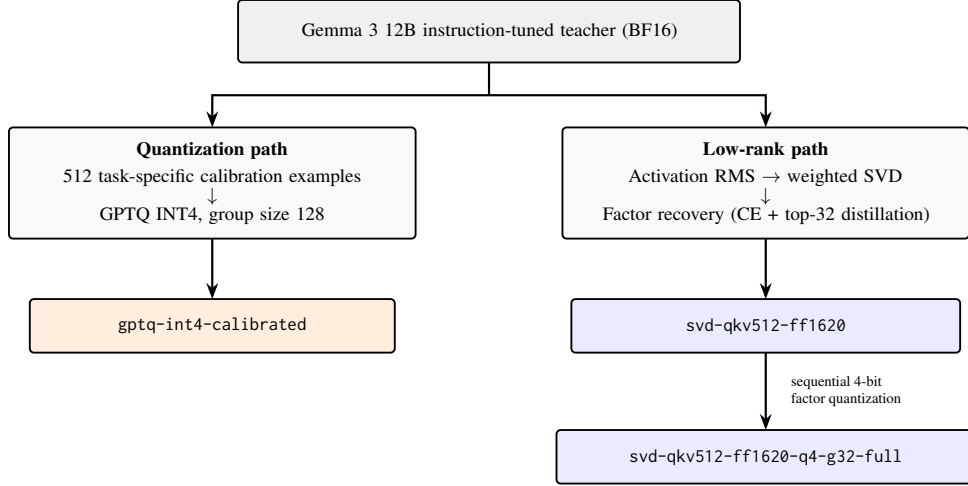


Figure 1: The two TINY TITANS compression paths and three submitted systems. The primary submission, gptq-int4-calibrated, is highlighted in orange.

pretrained weights with truncated factors, thereby reducing the deployed parameter count. ASVD’s use of activation statistics is particularly relevant in multilingual settings, where language-dependent activation structure has been observed (Liu et al., 2024).

Beyond structural parameter reduction, knowledge distillation offers a route to recover performance lost during compression (Hinton et al., 2015). Our recovery stage distills only the teacher’s sparse top-32 output distribution at the target positions, along with the supervised translation loss.

3 Task and System Overview

The WMT26 constrained track fixes Gemma 3 12B as the source model and includes Czech-to-German among its translation directions (Organizers, 2026). Systems are assessed for translation quality using COMET, MetricX, and an LLM judge, as well as disk size, peak GPU memory, and inference speed, in a standardized H100 80 GB environment. This multi-objective setup rewards Pareto-efficient systems rather than any single metric.

Gemma 3 is natively multimodal and multilingual (Team et al., 2025). Since the task input is text, our quantized checkpoint retains the causal language-model component and omits the vision tower. All systems use the instruction-tuned google/gemma-3-12b-it checkpoint and deterministic greedy decoding. To translate, we used the prompt:

Translate the following text from Czech to German.

Return only the translation.
{source}

Figure 1 summarizes the two compression paths, and Table 1 lists the identifiers of the three systems submitted to the organizers.

4 Compression Methods

4.1 Activation-Aware Low-Rank Factorization

For a linear layer with weight $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, ordinary truncated SVD minimizes unweighted reconstruction error. This treats every input channel equally, despite large differences in how often and how strongly channels activate. Following the motivation of activation-aware SVD (Yuan et al., 2025) and related truncation-aware factorization (Wang et al., 2024b), we estimate a per-input-channel RMS statistic

$$s_j = \sqrt{\frac{1}{N} \sum_{n=1}^N x_{n,j}^2} \quad (1)$$

from calibration activations. With $D = \text{diag}(\max(s, \epsilon)^\alpha)$, we compute a rank- r approximation

$$WD \approx U_r \Sigma_r V_r^\top. \quad (2)$$

The deployed factors are

$$A = U_r \Sigma_r^{1/2}, \quad B = \Sigma_r^{1/2} V_r^\top D^{-1}, \quad (3)$$

so that $W \approx AB$ and a dense operation becomes $x \mapsto A(Bx)$.

Submission ID	Role	Compression	Weights
gptq-int4-calibrated	Primary	Standard GPTQ, 4-bit, symmetric, group size 128	7.09 GiB
svd-qkv512-ff1620	Contrastive — SVD	Activation-aware SVD and factor recovery in BF16	11.71 GiB
svd-qkv512-ff1620-q4-g32-full	Contrastive SVD+INT4	— Recovered SVD model with symmetric 4-bit factors, group size 32	4.95 GiB

Table 1: Tiny Titans submissions. “Weights” is the sum of public checkpoint weight shards; it is not the organizers’ official disk-size measurement.

We estimate the activation statistics from 512 Czech-German examples and use $\alpha = 1$. All four attention projections (Q, K, V, and O) are factorized at rank 512, while the feed-forward projections use an effective rank of 1,664 after hardware-alignment rounding. Despite the legacy qkv512 checkpoint identifier, the output projection is also factorized at rank 512. Across the 48 text layers, this reduces the model from 11.77B to 6.29B parameters, corresponding to a 46.6% reduction. Embeddings and the language-model head remain dense.

4.2 Recovery Training with Sparse Distillation

To recover translation performance after truncation, we train the compressed model on 50k Czech-German sentence pairs from the WMT26 constrained data pool. These pairs provide both the supervised German targets and the inputs for teacher-logit generation, with no overlap with the evaluation sets to avoid train-evaluation leakage. Inputs follow the Gemma chat template, and we update both factors across all 48 layers while keeping other parameters frozen. The factors contain 5.28 billion parameters, making up 84.0% of the compressed model, and should not be viewed as a lightweight adapter method.

The dense, uncompressed Gemma 3 12B model acts as the teacher. To avoid storing full-vocabulary logits, we retain the teacher’s top 32 token indices and log-probabilities at each supervised target position. Teacher and student distributions are renormalized over these teacher-selected tokens, producing the sparse distillation objective

$$\mathcal{L}_{\text{KD}} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} D_{\text{KL}} \left(\tilde{p}_{T,t}^{(32)} \parallel \tilde{p}_{S,t}^{(32)} \right). \quad (4)$$

The complete objective is

$$\mathcal{L} = 0.6 \mathcal{L}_{\text{CE}} + 0.4 T^2 \mathcal{L}_{\text{KD}}, \quad (5)$$

where $\mathcal{L}_{\text{CE}}$ is causal language-model cross-entropy on the German target tokens and $T = 1$.

This follows the standard knowledge distillation principle (Hinton et al., 2015), adapted to sparse stored targets. We recover the factors for three epochs using Adafactor, BF16 training, and a batch size of 128. All non-factorized model parameters remain frozen.

4.3 Task-Calibrated GPTQ

Our primary submission applies GPTQ (Frantar et al., 2023) to the text-only Gemma checkpoint. GPTQ uses approximate second-order information to quantize weights while compensating for induced errors. We use 512 curated Czech-German examples from the official data pool, formatted with the Gemma chat template and containing source-target translation context. Examples shorter than 64 tokens are excluded, and examples exceeding 2,048 tokens are dropped rather than truncated.

Weights are quantized to 4 bits using symmetric groupwise quantization with a group size of 128. We disable activation-order reordering, enable true sequential quantization, and use a damping fraction of 0.01. The resulting 4-bit checkpoint can be executed directly by the shared-task wrapper and vLLM.

4.4 Sequential Quantization of SVD Factors

Our combined submission starts from the recovered factorization $W \approx AB$. Both B (the down-projection) and A (the up-projection) are quantized to 4 bits using symmetric groupwise rounding with group size 32. The factors are processed sequentially: B is quantized first, and its quantized output on the calibration examples is then used to quantize A . Calibration uses up to 1,024 examples of length 512 and retains up to 256 activation samples per module.

This factor quantizer shares GPTQ’s calibration-driven, layerwise workflow but does not implement GPTQ’s Hessian-inverse error propagation. We therefore refer to it as

sequential factor quantization, reserving GPTQ for the standard algorithm in our primary submission. This distinction is material: factor truncation and independent groupwise rounding introduce two sources of error, while the runtime also requires two quantized linear operations in place of one.

5 Experiments

5.1 Data and Evaluation

All task-specific data comes from the resources permitted by the WMT26 constrained track. We use 512 Czech-German examples to collect activation statistics, 50k sentence pairs for recovery training and teacher-logit generation, and 512 packed News Commentary examples for GPTQ calibration. Inputs are formatted with the Gemma chat template.

Within each evaluation protocol, systems use the same prompt, greedy decoding, and output budget; batch sizes may differ across protocols and model formats. We report point estimates for all automatic metrics and do not compute confidence intervals. Consequently, small metric differences are interpreted descriptively and should not be taken as evidence of statistically significant performance differences.

The processed calibration data are publicly available.¹

For quantization development, we evaluate dense and low-bit models on a shared 300-pair Czech-German set, using greedy decoding on a single NVIDIA A100-SXM4-40GB GPU. We report COMET (Rei et al., 2020), MetricX-24 (Juraska et al., 2024), BLEU, chrF, generated tokens per second, and steady GPU memory. Dense BF16 and the INT4 checkpoints are served with vLLM, whereas the BitsAndBytes INT8 baseline uses Transformers. *Load GiB* is the NVML-reported used memory on GPU 0 immediately after model loading, and *Steady GiB* is the corresponding reading immediately after generation completes.

The low-rank study uses 1,997 Czech-German pairs and Transformers inference in BF16. Our controlled comparison varies the attention rank from 256 to 512 and 1,024 while fixing the effective feed-forward rank at 1,664 and using the same three-epoch recovery budget. The dense and compressed models use different batch sizes, so

we compare translation quality but do not claim a controlled improvement in throughput.

Finally, we evaluate the three exact packaged submissions on 457 aligned Czech-German paragraphs from the released WMT25 General MT post-task data.²

All reported scores are point estimates; confidence intervals and significance tests are not reported. Small differences are therefore interpreted descriptively rather than as established performance gains.

5.2 Quantization Development Results

Table 2 compares the primary submission with the dense model and exploratory quantization baselines. Calibrated GPTQ reaches 0.8258 COMET, only 0.0023 below dense BF16, and reduces steady memory by 9.0 GiB (33.8%). Its throughput is 8.4% below dense in this vLLM setup, showing that smaller weights do not automatically improve speed.

BitsAndBytes INT8 essentially preserves dense quality, but is more than 10× slower, making it unattractive for the speed objective. AWQ (Lin et al., 2024) is the fastest, but has the lowest quality among the reported 4-bit systems. Task-calibrated and off-the-shelf GPTQ are nearly tied locally: calibration changes COMET by only +0.0001 and MetricX by +0.0030 (worse). We therefore do not claim a large local calibration gain. We additionally evaluate the three submitted checkpoints using a common protocol on the released WMT25 data, enabling a direct comparison among our systems.

Figure 2 plots the three pairwise trade-offs among quality, steady GPU memory, and throughput for the five systems in Table 2; no values beyond that table are used. Dense BF16 and calibrated GPTQ are both close to the COMET-memory frontier, while AWQ dominates on throughput at a small quality cost. BitsAndBytes INT8 is Pareto-dominated on the memory-speed plane by every 4-bit system. We restrict this comparison to the quantization systems, because the SVD and combined checkpoints in subsection 5.3 were benchmarked on a different evaluation set and inference engine, so their throughput and memory numbers are not comparable on the same axes.

¹[alexsandyr/Calibration on Hugging Face](#)

²[WMT25 General MT post-task data](#)

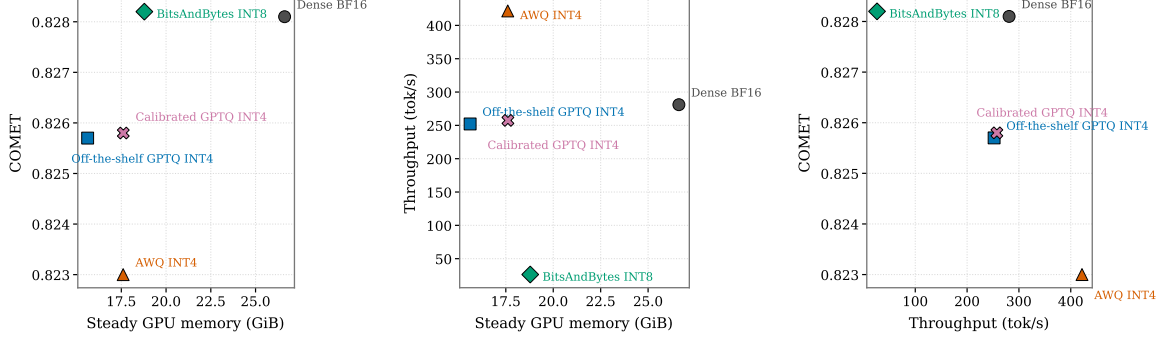


Figure 2: Quality, memory, and throughput trade-offs for the dense and quantized models on the development set.

System	COMET↑	MetricX↓	BLEU↑	chrF↑	tok/s↑	Load GiB↓	Steady GiB↓
Dense BF16	0.8281	7.2734	18.5820	47.4504	281.04	25.427	26.623
BitsAndBytes INT8	0.8282	7.1959	18.4202	47.4239	26.23	14.849	18.787
Off-the-shelf GPTQ INT4	0.8257	7.4496	18.5595	47.3810	252.13	15.478	15.619
AWQ INT4	0.8230	7.5426	17.6066	47.1886	421.70	15.761	17.603
Calibrated GPTQ INT4	0.8258	7.4526	18.4445	47.3299	257.43	17.468	17.611

Table 2: Quantization development results on the shared 300-pair evaluation. Bold marks the best compressed result in each column; dense BF16 is shown as an uncompressed reference.

5.3 Low-Rank Development Results

Table 3 reports the low-rank study. Increasing the attention rank from 256 to 512 and 1,024 raises COMET from 0.7323 to 0.7804 and 0.7986 while reducing parameter savings from 49.4% to 40.8%. Because the compressed models, unlike the dense reference, are recovered on 50k pairs, this primarily supports comparisons among recovered low-rank configurations rather than a controlled dense-versus-compressed comparison. The dense model’s unusually low BLEU (14.82) may also partly reflect differences in output format or instruction following, which we did not analyze qualitatively. The parameter reduction does not produce a proportional speedup. Each factorized projection requires two general matrix multiplications instead of one, and our implementation does not use fused low-rank kernels.

5.4 External WMT25 Evaluation

Because the development experiments use different protocols, we compare the exact submissions separately under a shared external evaluation. Table 4 shows a clear ordering: GPTQ obtains 0.5751 COMET and 44.7 chrF, while SVD loses 0.1488 COMET and 22.7 chrF. Quantizing the recovered factors causes a further decline, resulting in four empty translations.

Table 2 presents the matched dense-versus-GPTQ comparison under our quantization development protocol. Table 4 instead compares the three exact packaged submissions under a common protocol; because it contains no dense reference, it is used only to compare the submitted systems, not to assess the preservation of dense-model quality.

These results support selecting GPTQ as the primary submission. Low-rank factorization first approximates the original weights, and subsequent quantization then perturbs both recovered factors, preventing the model from adapting.

6 Analysis

The similarity between off-the-shelf and task-calibrated GPTQ in Table 2 indicates that Gemma 3’s weights are already fairly robust to 4-bit rounding on this development set, so calibration contributes little additional quality here. In the 300-pair quantization development evaluation, the practical value of our primary submission lies in retaining nearly all of the dense reference’s COMET score while substantially reducing steady GPU memory usage.

Parameter count does not translate directly into latency. Although the submitted SVD checkpoint removes 46.6% of parameters, its implementation requires two matrix multiplications

Checkpoint	Attn. rank	FF rank	Params. (B)	Reduction	Steps	COMET↑	BLEU↑	chrF↑
Dense BF16	–	–	11.766	0.0%	–	0.7755	14.8215	49.0346
FF only	–	2,304	9.643	18.0%	600	0.8281	21.9403	52.0478
Mixed	256	1,664	5.948	49.4%	1,173	0.7323	15.1434	45.3507
Mixed (submitted)	512	1,664	6.288	46.6%	1,173	0.7804	18.0011	48.1896
Mixed	1,024	1,664	6.967	40.8%	1,173	0.7986	19.2156	49.2833
Mixed	512	2,048	7.349	37.5%	600	0.7997	19.1012	49.3338
Mixed	512	2,176	7.703	34.5%	600	0.8060	19.5429	49.7226

Table 3: Low-rank development results on the 1,997-pair evaluation set. The three attention-rank sweep configurations (256/512/1,024) share an effective feed-forward rank of 1,664 and the same recovery budget; additional rows explore feed-forward rank separately. The dense reference does not receive recovery training.

Submission	COMET↑	chrF↑	Empty↓
gptq-int4-calibrated	0.5751	44.7	0
svd-qkv512-ff1620	0.4263	22.0	0
svd-qkv512-ff1620-q4-g32-full	0.3354	13.7	4

Table 4: External evaluation of the packaged submissions on 457 Czech-German paragraphs from the released WMT25 General MT post-task data.

per projection, introducing an intermediate tensor and additional kernel launch; without a controlled dense-versus-factorized throughput comparison, we do not attribute a runtime speedup to the parameter reduction. This is consistent with prior evidence that structural reductions must be paired with kernels designed to exploit them to yield wall-clock speedups (Behnke and Heafield, 2021; Wang et al., 2024a).

Among the factorization choices, the attention rank is sensitive to the attention rank. Within the training-controlled attention-rank sweep, attention compression strongly affects quality. At fixed feed-forward rank and recovery budget, raising the attention rank from 256 to 512 adds 0.34B parameters for a COMET gain of 0.0480. In contrast, a further increase to 1,024 adds 0.68B parameters for a smaller gain of 0.0182 (Table 3). The feed-forward-only configuration has the highest observed COMET among the tested checkpoints, but only an 18.0% parameter reduction. The submitted rank-512 configuration, therefore, favors parameter reduction over the highest score observed in this study.

Composing low-rank truncation with factor quantization compounds these errors rather than simply adding them. If the recovered weight is $W = AB + E_{\text{svd}}$ and quantization perturbs the factors to $A + \Delta_A$ and $B + \Delta_B$, the deployed weight acquires the additional first-order terms $A\Delta_B + \Delta_A B$ together with the second-order term $\Delta_A \Delta_B$. Because recovery training optimizes A and B before these perturbations are introduced,

the model cannot adapt to them. The practical effect is visible in both matched evaluations: chrF falls from 47.4 for the SVD-only checkpoint to 43.5 for the combined model on the public sample, and from 22.0 to 13.7 on the external WMT25 set.

Sparse top-32 distillation was used alongside supervised cross-entropy to recover the factorized model. Because we do not include a cross-entropy-only ablation, we cannot isolate its individual contribution. It does not, however, address the runtime overhead introduced by factorization: distillation shapes what the model predicts, not how efficiently it computes those predictions. Closing this gap would require either recovering the model after quantization with fused low-rank kernels or restricting factorization to the specific projections whose reduced shapes yield a measurable speedup.

7 Limitations

The quantization and low-rank development studies use different datasets and inference engines, so their scores and runtime values are not directly comparable. The low-rank evaluation also lacks MetricX and a controlled dense-versus-factorized throughput measurement. Finally, the hidden WMT26 results were unavailable at the time of writing; therefore, our direct submission comparison relies on the released WMT25 data, which may differ from the current task in domain and difficulty. The dense reference in the low-rank study does not undergo the 50k-pair recovery

training used for the compressed models. Although we verified that the recovery data do not overlap with the evaluation sets, dense-versus-recovered comparisons conflate compression with supervised adaptation. We also did not conduct a qualitative analysis to determine whether differences in instruction-following or output format contribute to the dense model’s low BLEU score. Finally, we report point estimates without bootstrap confidence intervals, which limits conclusions about small differences between systems.

8 Conclusion

Our WMT26 submissions show that task-calibrated GPTQ is the most reliable of the evaluated compression strategies: it preserves nearly all development-set COMET while reducing steady GPU memory by roughly one-third, and it clearly outperforms both low-rank submissions in the matched external evaluation. Activation-aware SVD removes almost half of the parameters, but its quality is sensitive to the attention-rank budget, while its two-factor execution introduces computational overhead that prevents us from equating parameter savings with inference speedups. Quantizing the recovered factors further reduces checkpoint size but compounds truncation and rounding error. Compression methods must therefore be designed together with their execution runtime.

9 Reproducibility

The submission wrappers and entry points are available in the public repository.³ The released checkpoints are:

- `gptq-int4-calibrated`;
- `svd-qkv512-ff1620`; and
- `svd-qkv512-ff1620-q4-g32-full`.

Acknowledgments

We thank the WMT26 Model Compression shared-task organizers for providing the evaluation framework and data resources. We also acknowledge the computing support provided by Mohamed bin Zayed University of Artificial Intelligence.

³Tiny Titans submission repository

References

- Alham Fikri Aji and Kenneth Heafield. 2020. [Compressing Neural Machine Translation Models with 4-bit Precision](#). In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 35–42, Online. Association for Computational Linguistics.
- Maximiliana Behnke and Kenneth Heafield. 2021. [Pruning Neural Machine Translation for Speed Using Group Lasso](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1074–1086, Online. Association for Computational Linguistics.
- Elias Frantar, Saleh Ashkboos, Torsten Hoefer, and Dan Alistarh. 2023. [GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers](#). *arXiv preprint*. ArXiv:2210.17323 [cs.LG].
- Thamme Gowda, Marco Gaido, Roman Grundkiewicz, and Matteo Negri. 2026. Findings of the WMT 2026 shared task on model compression: No free lunch at extreme compression. In *Proceedings of the eleventh conference on machine translation (WMT 2026)*, Budapest, Hungary.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. [Distilling the Knowledge in a Neural Network](#). *arXiv preprint*. ArXiv:1503.02531 [stat.ML].
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google Submission to the WMT 2024 Metrics Shared Task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [AWQ: Activation-aware Weight Quantization for On-Device LLM Compression and Acceleration](#). *Proceedings of Machine Learning and Systems*, 6:87–100.
- Weize Liu, Yinlong Xu, Hongxia Xu, Jintai Chen, Xuming Hu, and Jian Wu. 2024. [Unraveling Babel: Exploring Multilingual Activation Patterns of LLMs and Their Applications](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11855–11881, Miami, Florida, USA. Association for Computational Linguistics.
- Paul Michel, Omer Levy, and Graham Neubig. 2019. [Are Sixteen Heads Really Better than One?](#) In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- WMT26 Model Compression Shared Task Organizers. 2026. [Shared Task: Model Compression](#).
- 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.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. [Gemma 3 Technical Report](#). *arXiv preprint*. ArXiv:2503.19786 [cs.CL].

Wenxiao Wang, Wei Chen, Yicong Luo, Yongliu Long, Zhengkai Lin, Liye Zhang, Binbin Lin, Deng Cai, and Xiaofei He. 2024a. [Model Compression and Efficient Inference for Large Language Models: A Survey](#). *arXiv preprint*. ArXiv:2402.09748 [cs.CL].

Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. 2024b. [SVD-LLM: Truncation-aware Singular Value Decomposition for Large Language Model Compression](#). In *The Thirteenth International Conference on Learning Representations*.

Zhihang Yuan, Yuzhang Shang, Yue Song, Dawei Yang, Qiang Wu, Yan Yan, and Guangyu Sun. 2025. [ASVD: Activation-aware Singular Value Decomposition for Compressing Large Language Models](#). *arXiv preprint*. ArXiv:2312.05821 [cs.CL].