

Layer-Aware Native Quantization for Compact Machine Translation: A WMT26 English-to-Chinese Submission

Alonso Palomino
mail@alonsopg.com

Abstract

This paper describes a constrained English-to-Simplified-Chinese submission to the WMT26 Model Compression shared task. The submitted system is derived from google/gemma-3-12b-it and applies native 4-bit NormalFloat quantization only to the text-transformer MLP projections: gate_proj, up_proj, and down_proj. Attention projections, embeddings, normalization layers, the language-model head, and vision modules remain in BF16. This work continues the selected-layer compression line introduced in prior speech-translation work, but shifts it from storage-oriented codec compression to deployment-oriented native quantization for text MT. The central finding is that selective native quantization is a quality-preserving compromise: it recovers significant quality lost by global q4 while avoiding the BF16-like memory expansion of the earlier q3+Zstd codec. On a local 332-segment WMT25 English-Chinese reference set, the system obtains 28.30 chrF and 31.37 BLEU, compared with 27.67 chrF and 30.80 BLEU for the organizer global-q4 baseline. The submitted artifact is 10.99 GiB and reaches 14.64 GiB peak VRAM locally, placing it between the smaller global-q4 baseline and the higher-quality BF16 model.

1 Introduction

Large language models have become strong general-purpose machine translation (MT) systems, but their storage, memory, and decoding costs remain a barrier to deployment. Compression can reduce these costs, yet translation quality often responds unevenly to the amount and location of compression (Mohammadshahi et al., 2022). This creates a practical shared-task problem: a useful compressed MT system must preserve quality while also reducing disk size, accelerator memory, and inference time.

This paper describes a submission to the WMT26 Model Compression shared task for English-to-Simplified-Chinese MT. The constrained track fixes google/gemma-3-12b-it as the source model, and participants submit complete runnable systems that are evaluated for translation quality, model size, VRAM, and speed on organizer hardware.¹ The submission contains a single primary system, layeraware-native-mlp-q4, for English-Chinese.

The approach is motivated by a simple layer-aware hypothesis. In a previous IWSLT 2026 speech-translation submission, applying lossy codec compression only to transformer MLP projections preserved quality much better than applying one lossy codec globally (Palomino, 2026). This paper continues that line of work, but changes the setting and the compression representation: from speech translation to text-only MT, and from storage-oriented q3+Zstd coding to native low-bit modules that remain compressed during inference.

Using the WMT26 constrained English-Chinese setting, the following research question is addressed: *To what extent can selected-layer MLP compression, implemented as native low-bit quantization, preserve MT quality and reduce deployment memory for Gemma 3 12B compared with BF16, global q4, and storage-codec baselines?* (**RQ**).

To answer **RQ**, the selected-layer policy is retained but the execution representation is changed. The submitted system quantizes only the text transformer’s gate_proj, up_proj, and down_proj modules to native BitsAndBytes NF4 with double quantization and BF16 compute (Dettmers et al., 2023). All other modules remain BF16. This yields a mixed checkpoint in which the selected weights are serialized and loaded as low-bit modules rather than decoded back into ordinary BF16 tensors.

¹<https://www2.statmt.org/wmt26/model-compression.html>

This is therefore a deployment-oriented continuation of the previous selected-layer compression study rather than a new quantization algorithm. The paper reports: (1) a runnable mixed-precision Gemma 3 12B submission; (2) local controlled comparisons against BF16, the organizer global-q4 baseline, and a transferred selected-layer q3+Zstd codec; and (3) an analysis that separates quality, serialized size, peak VRAM, and local generation speed. The main local result is that selective native q4 significantly improves over the smaller organizer global-q4 baseline while retaining much of the BF16 quality at about half the artifact size.

2 Related Work

Model compression is increasingly central to deployable MT. In multilingual MT, post-training pruning and quantization can preserve average quality while still affecting languages and conditions unevenly (Mohammadshahi et al., 2022). The WMT 2025 Model Compression shared task showed a broad interest in quantization, pruning, distillation, and hybrid methods for LLM-based MT, while also emphasizing the tension between compactness and translation quality (Gaido et al., 2025).

Low-bit quantization is a common practical compression route because it is supported by modern inference stacks. QLoRA introduced NF4 and double quantization as effective tools for memory-efficient adaptation of large models (Dettmers et al., 2023). In this work, those mechanisms are used for post-hoc compression rather than fine-tuning: the model is not adapted after quantization, and no calibration set is used.

The closest antecedent is the selected-layer codec approach submitted to IWSLT 2026 (Palomino, 2026). That work applied significant-digit quantization and Zstandard coding to selected MLP projections in Qwen2-Audio. It found that targeted MLP compression preserved speech-translation quality much better than a uniform global codec. The present submission should be read as a continuation of that result. Its novelty is not the NF4 quantizer itself, but the transfer of the selected-layer hypothesis to Gemma-based text MT and its implementation as native low-bit execution rather than as a storage codec that must be decoded before inference.

3 Task and Dataset

Task. To answer **RQ**, the submission uses the WMT26 Model Compression task, which asks participants to reduce the size of LLM-based MT systems while maintaining translation quality. The constrained track fixes Gemma 3 12B as the original model. Participants may submit systems for Czech-German, English-Simplified Chinese, and English-Egyptian Arabic; rankings are reported separately by language direction. This submission covers only English-Simplified Chinese (eng-zho_Hans).

Data. No training, fine-tuning, or calibration data is used. Local analysis uses the reference-backed WMT25 English-Chinese development material prepared by the organizer repository. The usable local set contains 332 source-reference pairs. These examples are used only for evaluation and error-free result collection, not for choosing per-layer bitwidths or fitting any parameters.

Evaluation. Local development measurements cover translation quality, model size on disk, VRAM usage, and decoding speed. Quality is measured with lowercased chrF2 (Popović, 2015) and lowercased BLEU using SacreBLEU 2.5.1 with the Chinese tokenizer (Post, 2018). Paired significance is computed with SacreBLEU paired bootstrap resampling, 10,000 trials, and seed 12345 (Koehn, 2004).

Inference. All systems use the same organizer-style inference wrapper and the same prompt: “Translate the following text from English to Chinese. Return only the translation, with no explanation, labels, or quotes.” The Gemma chat template is applied. Decoding is greedy, with batch size 8 and no sampling. Inputs are length-sorted for batching and restored to original order. Local speed excludes processor and model loading. Peak VRAM is sampled once per second with `nvidia-smi` on one NVIDIA RTX A6000 with 48 GiB.

4 Compression Methods

All compared systems are derived from the same Gemma 3 12B revision and are evaluated under the same local decoding conditions. The systems differ only in how model weights are represented. Table 1 summarizes the policies. Figure 1 shows the submitted compression pipeline. The key step is a fixed architectural decision rule applied to every linear module: if the module is one of the three MLP projections inside a text-transformer block, it

is selected for native q4 conversion; otherwise, it is explicitly preserved in BF16.

4.1 Organizer baselines

The BF16 baseline is the uncompressed google/gemma-3-12b-it model. The organizer global-q4 baseline follows the provided BitsAndBytes preparation path, applying `load_in_4bit=True` to eligible modules and serializing the result. In the local stack, this baseline resolves to FP4 weights, FP32 compute, and no double quantization. This baseline is kept unchanged because it is the reference compression recipe distributed to participants.

4.2 Transferred storage-codec method

To connect this WMT submission to the previous IWSLT approach, the selected-layer q3+Zstd codec is transferred to Gemma 3. The targeted tensors are the same text-transformer MLP projections used by the submitted system. Each selected weight matrix is encoded with `numcodecs.Quantize(digits=3, dtype=f4, astype=f2)` followed by Zstandard level 3 (Alted et al., 2016; Palomino, 2026). Quantization is the lossy step; Zstandard preserves the rounded representation exactly.

This codec variant is included as an ablation, not as a deployment submission. For quality and generation measurements, rounded weights are materialized into the model before inference. Thus the method tests whether q3 rounding of the selected layers preserves quality, and it measures the exact encoded selected-weight payload, but its reported generation time excludes file I/O and Zstandard decompression.

4.3 Layer-aware native MLP q4

Gemma 3 12B contains 48 text-transformer blocks. As shown in Figure 1, the method first inventories the model’s linear modules and then applies one rule: compress a module only if it belongs to the text transformer and is named `gate_proj`, `up_proj`, or `down_proj`. This selects three MLP projections per block, for 144 target modules in total. These matrices contain 8.493 billion weight scalars. All attention projections, embeddings, normalization layers, the language-model head, and vision modules are placed on a skip list and remain BF16.

The selected modules are loaded as BitsAndBytes Linear4bit layers with NF4 quantization,

double quantization, and BF16 computation. After construction, the preparation script verifies that the actual set of Linear4bit modules exactly matches the 144 intended targets. The checkpoint is then saved with SafeTensors, together with a `layer_policy.json` record of compressed and preserved modules. Unlike the codec variant, the submitted checkpoint keeps its low-bit representation when loaded for inference.

5 Results

5.1 Local quality, size, memory, and speed

Table 2 reports the full local comparison on the 332-segment WMT25 English-Chinese reference set. The BF16 model remains the best system in absolute quality and local generation speed. The organizer global-q4 baseline is the smallest and lowest-memory model, but it has the weakest local translation quality among the native inference systems.

The submitted layer-aware native q4 system improves over organizer global q4 by 0.63 chrF and 0.56 BLEU. This quality recovery costs 3.21 GiB in artifact size and 2.60 GiB in peak VRAM, although local generation is slightly faster than global q4. Relative to BF16, the submitted system reduces artifact size by 51.7% and peak VRAM by 44.7%, with quality losses of 0.15 chrF and 0.28 BLEU. It therefore occupies an intermediate operating point between the highest-quality BF16 model and the smallest global-q4 baseline.

5.2 Paired significance

Table 3 compares the submitted system with the main baselines using paired bootstrap resampling. The gains over global q4 are significant for both chrF and BLEU. The difference from the transferred q3+Zstd codec is not significant under either metric. Against BF16, the chrF difference is not significant at the 0.05 level, while the BLEU difference is.

These tests measure whether the local segment-level differences are stable under resampling. The global-q4 comparison is the clearest practical finding: the submitted system recovers a significant amount of lexical-overlap quality relative to the smaller organizer q4 recipe.

5.3 Native q4 versus storage codec

The selected q3+Zstd codec and native selective q4 have nearly identical local quality, but their de-

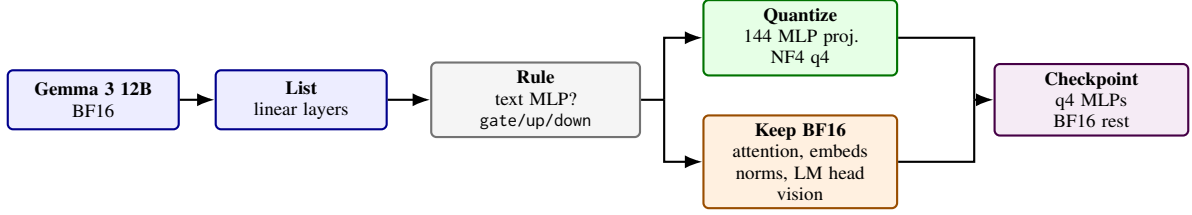


Figure 1: Layer-aware native quantization policy. The rule selects only the three text-MLP projections in each block for native q4 conversion; all other modules stay BF16.

System	Compressed scope	Representation and role
BF16 baseline	None	Original organizer model; quality and performance reference.
Organizer global q4	Eligible linear modules globally	Organizer BitsAndBytes baseline using the default q4 recipe in the provided submission template.
Selected q3+Zstd	144 text-MLP projections	Storage-codec transfer from the IWSLT method; selected weights are rounded and Zstd-compressed, then materialized for inference.
Submitted model: native MLP-q4	144 text-MLP projections	Submitted mixed-precision Gemma 3 system; selected MLP weights are native NF4 modules with double quantization and BF16 compute, while all other modules remain BF16.

Table 1: Compared compression policies. **Submitted model** is the layer-aware native MLP-q4 checkpoint. The global-q4 comparison is operational because target scope and 4-bit configuration both differ.

ployment profiles differ. The codec artifact is estimated at 14.46 GiB and expands to 26.43 GiB peak VRAM once decoded, essentially matching the BF16 model’s 26.46 GiB local memory footprint. Native selective q4 uses a 10.99 GiB artifact and 14.64 GiB peak VRAM. Thus, the main extension over the IWSLT-style codec is not a quality improvement; it is preserving the selected-layer compression in the representation used by the inference engine.

5.4 Submission and reproducibility

The submitted system is the primary constrained English-Chinese entry described in this paper. The public model repository contains SafeTensors shards, tokenizer and processor files, quantization configuration, layer-policy metadata, setup scripts, and the inference wrapper.² The experiment repository records the full preparation pipeline, predictions, metrics, GPU traces, artifact checksums, and the system-source.³

The final model payload is 11,804,446,040 bytes, or 10.99 GiB. The local environment uses Transformers 4.57.6, BitsAndBytes 0.45.3, PyTorch 2.7.0, and SacreBLEU 2.5.1. The submitted artifact passed the organizer sanity check locally, including the required line-count contract.

²<https://huggingface.co/alonsopg/wmt26-layeraware-native-mlp-q4>

³<https://github.com/alonsopg/wmt26-layeraware-compression>

6 Conclusion

This paper presented a WMT26 constrained English-Chinese model-compression submission based on layer-aware native quantization. The method selectively quantizes only the text-transformer MLP gate, up, and down projections of Gemma 3 12B, while preserving attention, embeddings, normalization layers, the language-model head, and vision modules in BF16. The main conclusion is that selected-layer native quantization is a strong quality-preserving operating point for MT compression: it is not the smallest model, but it recovers significant chrF and BLEU over the organizer global-q4 baseline while retaining much of the BF16 model’s quality.

The comparison with the transferred selected-layer q3+Zstd codec clarifies why the representation matters. Both selected-layer approaches preserve similar local quality, but the codec expands to BF16-like memory after decoding. Native q4 keeps the selected weights compressed during inference, reducing artifact size by about half relative to BF16 and lowering peak local VRAM by 44.7%. These findings support the selected-layer hypothesis for text MT and show that using native low-bit modules turns the idea from a storage codec into a deployment-relevant compression strategy.

Limitations

The results in this paper are local development measurements. They use one RTX A6000, so speed

System	chrF	BLEU	Gen. s/line	Peak VRAM (GiB)	Artifact (GiB)
BF16 baseline	28.45	31.65	1.354	26.46	22.74
Organizer global q4	27.67	30.80	2.271	12.04	7.78
Selected q3+Zstd	28.26	31.39	1.358 [†]	26.43	14.46
Submitted model: native MLP-q4	28.30	31.37	2.175	14.64	10.99

Table 2: Local results on 332 WMT25 English-Chinese segments, batch size 8, one RTX A6000. **Submitted model** is the mixed Gemma 3 checkpoint: 144 native-q4 text-MLP projections and all other modules BF16. Generation excludes loading. [†]The codec run starts after rounded weights are materialized and excludes Zstandard decompression.

Baseline	Δ chrF	p	Δ BLEU	p
Global q4	+0.63	.0003	+0.56	.0061
Selected codec	+0.05	.2561	−0.02	.3576
BF16	−0.15	.0897	−0.28	.0339

Table 3: Layer-aware native q4 minus each baseline locally. Significance uses 10,000 paired-bootstrap trials.

and memory behavior may change under different kernels, drivers, and hardware.

The local evaluation covers only English-Chinese and only 332 WMT25 reference-backed examples. The same policy may behave differently for Czech-German, English-Egyptian Arabic, other domains, or longer inputs. Local quality is also limited to chrF and BLEU, while shared-task ranking may include additional automatic metrics such as COMET, MetricX, and LLM-based assessment.

The comparison with organizer global q4 is an operational baseline comparison, not a complete ablation. Global q4 and layer-aware native q4 differ in both layer scope and low-bit configuration. A full factorial analysis would compare global and selected policies under matched FP4 and NF4 settings.

References

- Francesc Alted, Prakhar Goel, Jerome Kelleher, John Kirkham, Alistair Miles, Jeff Reback, Trevor Manz, Grzegorz Bokota, Josh Moore, Martin Durant, and Paul Branson. 2016. [Numcodecs](#). Python package.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. [QLoRA: Efficient finetuning of quantized LLMs](#). In *Advances in Neural Information Processing Systems*, volume 36.
- Marco Gaido, Roman Grundkiewicz, Thamme Gowda, and Matteo Negri. 2025. [Findings of the WMT 2025 shared task on model compression: Early insights on compressing LLMs for machine translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain. Association for Computational Linguistics.
- Alireza Mohammadshahi, Vassilina Nikoulina, Alexandre Berard, Caroline Brun, James Henderson, and Laurent Besacier. 2022. [What do compressed multilingual machine translation models forget?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2022*, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Alonso Palomino. 2026. [Selected-layer codec compression for compact speech translation models: An IWSLT 2026 English-to-Chinese submission](#). In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, San Diego, USA (in-person and online). Association for Computational Linguistics.
- Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, Lisbon, Portugal. Association for Computational Linguistics.
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, Brussels, Belgium. Association for Computational Linguistics.