

Findings of the WMT 2026 Shared Task on Model Compression: No Free Lunch at Extreme Compression

Thamme Gowda

Microsoft

thammegowda@microsoft.com

Roman Grundkiewicz

Microsoft

rogrundk@microsoft.com

Marco Gaido

FBK

mgaido@fbk.eu

Matteo Negri

FBK

negri@fbk.eu

Abstract

We present the second edition of the WMT 2026 Model Compression (WMT26MC) shared task, which studies how large language models can be made efficient for machine translation under model-size and latency constraints. The task offers a constrained track based on Gemma 3 12B and an unconstrained track for compressing models below 20B. Both tracks cover Czech→German, English→Chinese (Simplified), and English→Arabic (Egyptian). Participants submit complete runnable systems, which we run in a standardized environment on a single H100 GPU. Evaluation jointly considers translation quality, model size on disk, and decoding speed, analyzing the resulting quality–size and quality–speed trade-offs. We received 41 submissions from 13 teams. The proposed systems span weight and activation quantization, structured and expert pruning, low-rank factorization, knowledge distillation, vocabulary and vision-component removal, mixed precision, and inference-time reranking. Evaluation with quality-estimation metrics on the blind test sets reveals that post-training quantization is the reliable lever: INT4 and FP8 systems shrink the constrained base to roughly one-quarter and one-half of its original disk size, respectively, and deliver up to 20× the steady-state throughput while staying within quality-estimation noise of the uncompressed model, whereas aggressive low-rank and structural pruning collapse quality.

1 Introduction

Large language models (LLMs) have become strong general-purpose systems for machine translation (MT), but their broad capabilities come with substantial storage, memory, and computational requirements (Brown et al., 2020; Ding et al., 2024; Patil and Gudivada, 2024). These costs complicate deployment on commodity accelerators, lo-

cal servers, mobile hardware, and other resource-constrained platforms. Compression methods such as quantization, pruning, low-rank factorization, and knowledge distillation can reduce these requirements, although aggressive compression may degrade translation quality and some compression techniques fail to yield a corresponding reduction in end-to-end latency (Zhu et al., 2024). A useful comparison therefore has to measure the resulting systems as deployable artifacts rather than considering parameter count or nominal precision alone.

The WMT 2025 Model Compression shared task was introduced with the objective to provide such a comparison for text translation (Gaido et al., 2025). It continued a line of WMT shared tasks on efficient translation (Birch et al., 2018; Hayashi et al., 2019; Heafield et al., 2020, 2021, 2022), complementing a parallel initiative in the speech translation field at IWSLT (Abdulmumin et al., 2025; Adelani et al., 2026). The second edition retains the central objective—reducing the footprint and cost of a general-purpose LLM while preserving machine-translation quality—but updates the benchmark to reflect both newer models and more realistic system deployment.

The WMT 2026 Model Compression (WMT26MC) shared task¹ replaces Aya Expand 8B (Dang et al., 2024) with Gemma 3 12B (Team et al., 2025) in the constrained track, and the unconstrained track permits any source model whose original size is below 20B parameters. Both tracks use the same three language directions: Czech→German, English→Chinese (Simplified), and English→Arabic (Egyptian). As in the previous year, participants submit the complete runnable artifact rather than only model outputs, allowing organizers to measure translation quality, disk footprint, and decoding speed under the same

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

hardware and operating-system conditions.

The evaluation is designed to address three main questions:

- How far can a general-purpose LLM be compressed while retaining its translation quality on the targeted language directions?
- Which combinations of compression and inference techniques provide the strongest quality–size and quality–speed trade-offs?
- How do methods developed under a fixed-model, fixed-data constraint compare with unconstrained model selection and training?

This paper describes the submitted systems, the evaluation pipeline, and reports results offering the following key takeaways:

- Quantization (at INT4) enables reduction to roughly a quarter of the original size with limited quality degradation, while compression at half of the original size (FP8) is essentially lossless.
- Solutions based on aggressive structural pruning fail to retain original translation capabilities even after recovery finetuning.
- Inference speed depends primarily on the serving runtime rather than on model size: reduction in numerical precision leads to faster inference only if the runtime can effectively exploit the reduced precision.

2 Task Description

The task asks participants to reduce the deployable footprint of a general-purpose LLM while maintaining strong translation performance. Systems may target any subset of the three language directions, and rankings are computed separately for each track and language direction.

2.1 Tracks

The **constrained track** establishes a common starting point by requiring all systems to be directly derived from `google/gemma-3-12b-it`.² Gemma 3 12B was selected for its permissive licensing terms and its balance of size and translation capability across the task languages. Any compression technique is permitted. However, the final model must remain a compressed derivative of the original model: for example, a distilled student is eligible only if its architecture is itself obtained from Gemma 3 12B rather than chosen in-

dependently. In addition, constrained-track participants can use only the public data released for the WMT26 campaign and test sets from previous WMT General MT shared task editions for calibration or fine-tuning.

The **unconstrained track** uses the same language directions and evaluation protocol but allows participants to choose any original model below 20B parameters. It also permits additional calibration and training data. By holding the evaluation environment fixed while relaxing model and data choice, this track tests whether model selection and task specialization offer better deployment trade-offs than compressing a common base model.

2.2 Evaluation

The official evaluation uses held-out WMT26 General MT data for the selected directions. Evaluation inputs are exposed to submitted systems as UTF-8 text with one source record per line. Each system must return exactly one translation per input line. This line-oriented interface keeps data handling identical across model frameworks and prevents system-specific preprocessing from changing the number or order of examples.

Submissions are evaluated separately for each track and language along three dimensions:

1. **Translation quality**, assessed with reference-free metrics: COMET-Kiwi-XXL (Rei et al., 2023) and MetricX-XXL-QE (Juraska et al., 2024).
2. **Model footprint**, measured as the on-disk model artifact size. Peak GPU-memory use could not be measured reliably across the heterogeneous runtimes and is not reported.
3. **Decoding speed**, measured from the wall-clock time required to translate the evaluation set under controlled batch settings.

All systems run on an organizer-managed x86_64 machine with Ubuntu 24.04 and a single NVIDIA H100 GPU with 80 GB of VRAM. The same submission interface, input order, decoding workload, and hardware allocation are used for every system.

The evaluation pipeline is released with the organizers' repository.³ Before scoring, a sanity check installs each submission, invokes it on a small input, verifies that the output is non-empty, and en-

²<https://huggingface.co/google/gemma-3-12b-it>

³<https://github.com/thammegowda/wmt26-model-compression>

forces exact input/output line-count match. The main evaluator invokes the same entry point for all accepted systems and records the command, exit status, wall-clock time, CPU times, maximum host resident-set size, and timestamps for each run. Invalid or incomplete outputs are excluded from scoring. The throughput protocol profiles nominal batch sizes of 1, 16, 64, and 256, capped by each system’s declared maximum, with repeated full-set and one-sentence warmup runs per condition. We report net throughput as the number of source characters divided by difference between the median full-set wall time and the median one-sentence warmup wall time. The warmup is therefore a startup proxy that includes initialization and one-sentence decoding, rather than a pure model-loading measurement.

2.3 Submission

A valid submission consists of a form entry and a complete system that the organizers can execute without reconstructing the compression procedure. Each submission package contains source code, environment-installation instructions and dependencies, an inference entry point, and the compressed model artifact (or a stable link to it). Under the submission specification, packages provide `setup.sh`, `run.sh`, `requirements.txt`, and `README.md`. The evaluator calls `run.sh` with a language direction, batch size, input path, and output path; diagnostic output must not be mixed with translations.

Teams may submit multiple systems for a track and language. In that case they designate one system as PRIMARY, with the rest treated as CONTRASTIVE. Participants also declare the language directions and maximum supported batch size for each submission.

3 Participants

We received submissions from 13 teams, of which 10 teams participated in the constrained track and 4 teams targeted the unconstrained track (one team entered both tracks). Many participants submitted multiple entries.

Table 1 reports the team names, number of submissions, track codes, and a brief description of explored optimization methods. Quantization is most common, but it is frequently combined with structural specialization: teams remove vision modules unused by text-only MT, prune vo-

cabularies or transformer components, factorize dense matrices, or distill pruned students. The unconstrained entries broaden model choice to Aya-Expanse, Gemma 4, GPT-OSS, and TildeOpen models. Several entries also optimize the inference system itself through custom kernels, continuous batching, and quality-estimation-based selection. The descriptions below consolidate the short paragraph responses in the submission form. They summarize systems as declared by participants. Claims about quality, speed, or memory use are deferred to the evaluation in Section 4.

3.1 alonso (Palomino, 2026)

The alonso entry is a calibration-free English→Chinese (Simplified) system that selectively applies BitsAndBytes NF4 with double quantization to the gate, up, and down projections of all 48 text-transformer MLP blocks. Attention projections, embeddings, normalization layers, the output head, and vision modules remain in BF16. This selective design trades a larger artifact for lower quantization exposure than uniform 4-bit compression.

3.2 arc/ilsp (Athena Research Center) (Sofianopoulos and Prokopicidis, 2026)

The arc/ilsp systems combine three training-free operations: removal of the vision tower, script-aware vocabulary pruning from roughly 262k to 196k entries, and GPTQ W4A16 quantization. The team selects full- or reduced- vocabulary INT4 primaries by language direction and includes an FP8 contrastive. For English→Arabic (Egyptian), a decoding contrastive generates multiple candidates and applies self-MBR selection using chrF consensus without changing model weights.

3.3 CometCut (Bajwa, 2026)

CometCut compares uniform GPTQ and AWQ with a quality-aware mixed-precision scheme. For the latter, the team quantizes one decoder layer at a time, estimates its sensitivity using COMET on held-out data, and solves a knapsack problem that assigns layers to 16-, 8-, or 4-bit precision under a size budget. All variants use a shared multilingual calibration pool and the LLM-Compressor/vLLM stack.

3.4 Dragon1006 (IIIT) (Balaga et al., 2026)

Dragon1006’s unconstrained English→Chinese (Simplified) system starts from Aya Expanse 8B,

ID	Team	Track	Directions	# Var.	Main techniques	Details
Al	alonso	C	en-zh	1	Selective MLP-only NF4 quantization	§3.1
Ar	arc/ilsp	C	all	4	Vision/vocabulary pruning, GPTQ/FP8, self-MBR	§3.2
C	CometCut	C	all	3	GPTQ, AWQ, COMET-guided mixed precision	§3.3
D	Dragon1006	U	en-zh	1	Layer pruning, LoRA recovery, runtime INT8	§3.4
E	ESTS	U	en-zh, en-ar	6	MoE expert pruning, MXFP4, supervised recovery	§3.5
F	FBK	C	cs-de	2	SmoothQuant, GPTQ, W8A8 activation quantization	§3.6
P	Pare4Bit	C/U	all	3	Vision/depth pruning, GPTQ INT4, model selection	§3.7
S	SLICERS	C	en-zh	1	Layer/FFN pruning, NF4, QLoRA	§3.8
Ta	TahomaMT	C	all	4	Vision/vocabulary pruning, FP8/INT4, QE reranking	§3.9
Ti	TildeOpen-wmt	U	cs-de	5	Distillation, NVFP4/FP8 quantization, GRPO	§3.10
Tn	Tiny Titans	C	cs-de	3	GPTQ, activation-aware SVD, distillation	§3.11
TM	TMU-onono	C	cs-de	2	DiBA factorization and diagonal-only recovery	§3.12
V	VIC	C	all	6	Structured pruning, vocabulary pruning, AWQ	§3.13

Table 1: Registered WMT26 participants and their 41 submitted participant variants across all 13 teams. The ID column defines the participant prefix used throughout the tables and figures; system-level labels append a compact variant subscript to this prefix. C and U denote constrained and unconstrained tracks; “all” denotes cs-de, en-zh, and en-ar. Variant counts (# Var.) describe submitted artifacts

reduces its depth from 32 to 28 layers, and restores translation behavior with QLoRA on a 36k-example subset of WMT19. The adapter is merged into the model to remove PEFT routing overhead, and weights are quantized to 8-bit at inference time with BitsAndBytes. Retrieved after the initial inventory snapshot, the int8 artifact is evaluated here on English→Chinese (Simplified) (Table 2); note that its inference wrapper caps generation at 150 new tokens, which truncates the longer paragraphs of the WMT26 set and depresses its measured quality.

3.5 ESTS (UCLA) (Martin et al., 2026)

ESTS builds unconstrained translation specialists from the MXFP4-quantized GPT-OSS-20B mixture-of-experts model. It estimates expert importance from routing on task-proxy generations and allocates retained capacity across layers using the Jensen–Shannon divergence between source- and target-side routing distributions. Experts are physically removed, after which the model is recovered with full-parameter supervised training on GPT-generated pseudo-parallel data. Separate English–Chinese and English–Egyptian-Arabic families vary the number of retained experts.

3.6 FBK@ModelCompression (Suman and Bentivogli, 2026)

FBK submits two W8A8 Czech→German systems. Both use GPTQ for 8-bit weights and round-to-nearest quantization for activations, executing the matrix operations directly in 8-bit precision through the compressed-tensors/vLLM stack. One

variant first applies SmoothQuant to move part of the activation quantization difficulty into the weights; the other omits this smoothing step.

3.7 Pare4Bit (Fakhrutdinov, 2026)

For the constrained track, Pare4Bit rebuilds Gemma 3 12B as a text-only model without the vision tower and applies group-wise GPTQ W4A16 while retaining tied embeddings in BF16. A contrastive additionally drops four transformer layers, uses LoRA knowledge-distillation recovery, and then applies GPTQ. Its unconstrained entry treats model selection as part of compression, choosing a Gemma 4 12B quantization-aware-trained INT4 checkpoint and serving it through vLLM.

3.8 SLICERS (IIIT Hyderabad) (Balaga et al., 2026)

SLICERS targets English→Chinese (Simplified) with a combination of structural and precision compression. Fisher-information estimates guide removal of eight transformer layers and 30% of intermediate feed-forward neurons. The structurally pruned model is then quantized to 4-bit NF4 and adapted with QLoRA on translation data.

3.9 TahomaMT (Microsoft) (Gowda, 2026)

TahomaMT removes the unused SigLIP vision tower and multimodal projector and specializes the tokenizer to the Latin, Arabic, and Han scripts used in the task, reducing its vocabulary from 262,208 to 192,347 entries with byte fallback. Two calibration-free variants use either FP8 weights and per-token FP8 activations (W8A8) or group-wise asymmetric INT4 weights with FP8 activa-

tions (W4A8). Contrastive best-of-four systems sample four candidates and select one with a bundled 4-bit Cometoid quality-estimation model. All variants use the team’s C++/CUDA Tahoma runtime.

3.10 TildeOpen-wmt (Tilde) (Kronis et al., 2026)

Tilde’s unconstrained submission starts from a task-adapted TildeOpen 15B Czech–German model, itself distilled from a 30B model and further trained with supervised fine-tuning and GRPO. The submission compares GPTQ with NVFP4 weights against FP8 weights and dynamic FP8 activations. A second compression axis distills the 15B model to an 8B student, and the team also combines this student with each quantization scheme.

3.11 Tiny Titans (MBZUAI) (Raya-Rios et al., 2026)

Tiny Titans submits three Czech→German systems. The first applies group-wise 4-bit GPTQ using a 1,024-example calibration set. The second factorizes selected weight matrices with activation-aware SVD and then performs task-aware distillation from the original Gemma model, using stored top-32 teacher probabilities. The third combines this low-rank/distillation pipeline with 4-bit GPTQ of the resulting factors.

3.12 TMU-onono (Tokyo Metropolitan University) (Ono, 2026)

TMU-onono applies DiBA (Diagonal and Binary Matrix Approximation), replacing dense linear and tied embedding/output weights with interleaved diagonal scalings and binary mixing factors. The binary factors remain fixed while the diagonal factors are adapted on teacher translations of the WMT26 Czech→German training data, a recovery procedure the team calls DiBARD. Its primary system uses a direct Triton backend, while a contrastive system caches unpacked factors at runtime.

3.13 VIC (Vicomtech) (Ponce and Etchegoyhen, 2026)

VIC removes Gemma’s vision component and learns differentiable masks over feed-forward channels and attention groups while keeping the model weights fixed. The masks are materialized as structured pruning at several target ratios; selected models are then recovered by supervised fine-tuning on multilingual parallel data. The submission family

also explores language-specific vocabulary pruning and applies 4-bit AWQ after pruning, with inference served through vLLM.

4 Results

We evaluate all submissions in a single standardized environment and report the three task dimensions separately. Because the WMT26 test-set references are not available at the time of writing, the quality numbers here are *reference-free* quality-estimation (QE) scores. Unless noted otherwise, quality is COMET-Kiwi-XXL (Rei et al., 2023), cross-checked against MetricX-XXL-QE (Juraska et al., 2024). The constrained track compresses a common base (Gemma 3 12B, 22.7 GiB on disk); the unconstrained track uses heterogeneous bases, so its systems are compared against the constrained base only as a reference line, not on equal footing. Throughout, the uncompressed base and the two bitsandbytes organizer baselines (bnb-q4, bnb-q8) anchor the plots. The full per-system results are consolidated in Table 2, the central table of this report; the figures that follow break out each dimension.

4.1 Translation Quality

The translation quality results in Table 2 indicate the following: post-training *quantization* preserves quality, whereas aggressive *structural* compression does not. Weight-only and weight-activation quantization (INT4/W4A16, FP8/W8A8, INT8) land within QE noise of the uncompressed base, while low-rank factorization (Tiny Titans’ SVD, COMET-Kiwi-XXL 24–32 for Czech→German), binary/diagonal approximation (TMU-onono’s DiBA, 27–28), and 50% FFN pruning without sufficient recovery (Vicomtech, down to 49 for en-ar) collapse quality. The failures reflect capacity removal that the submitted recovery (distillation or light fine-tuning) did not restore. Some models using calibrated data improve on COMET-Kiwi-XXL over baselines but worsen MetricX24 scores. The disagreement underscores the importance of evaluating quality through multiple metrics to enforce sound conclusions.

4.2 Model Size

Figure 1 shows the quality–size trade-offs separately for each language direction under both QE metrics. Both tracks appear in every panel. The headline result is that the base

Constrained track (compress Gemma-3-12B)

ID	System		Data	Size ↓	speed (cs→de)		cs→de		en→zh		en→ar	
					WU (s) ↓	Net c/s ↑	CK ↑	MX ↓	CK ↑	MX ↓	CK ↑	MX ↓
B ₁₆	baseline--uncompressed★	HF	—	20	22.74	13.9	626	56.26	6.31	76.95	2.77	61.51
B ₈	baseline--bnb-q8★	HF+BnB	—	16	12.34	21.5	257	56.38	6.30	76.84	2.80	61.05
B ₄	baseline--bnb-q4★	HF+BnB	—	10	7.78	11.7	838	55.47	6.49	76.65	2.81	60.66
A _{MF4}	● alonso--layeraware-native-mlp-q4	HF+BnB	—	14	10.99	12.2	346	56.38	6.23	76.69	2.78	—
A _{r4}	● arc-ilsp--int4	VLLM	cal	9	7.08	42.1	3,012	56.10	6.52	76.46	2.85	—
A _{r4V}	● arc-ilsp--vocaball-int4	VLLM	cal	8	6.64	41.2	3,046	56.11	6.42	—	—	60.10
A _{r8}	○ arc-ilsp--fp8	VLLM	cal	18	13.81	49.3	3,023	56.30	6.31	76.91	2.75	60.84
A _{IMBR}	○ arc-ilsp--en-ar-mbr	VLLM	cal	8	6.64	42.2	745	56.12	6.40	—	—	61.42
C _M	● cometcut--comet-mixed-precision	VLLM	cal	12	8.53	58.6	10,820	55.98	6.38	76.98	2.75	60.19
C _G	● cometcut--gptq	VLLM	cal	11	7.90	50.1	6,690	56.29	6.45	76.55	2.76	60.58
C _A	○ cometcut--awq	VLLM	cal	11	7.90	53.9	6,526	55.33	6.43	76.02	2.82	60.99
F _G	● fbk--gptq-rtn	VLLM	cal	17	12.72	36.2	7,963	56.97	6.91	—	—	—
F _{5Q}	● fbk--sq-gptq	VLLM	cal	17	12.72	37.5	9,160	56.80	6.91	—	—	—
P ₄	● pare4bit--int4-gptq	VLLM	cal	9	7.07	104.2	4,244	56.75	6.46	76.44	2.78	60.14
P _{P4}	○ pare4bit--pruned4-healed-int4	VLLM	FT	8	6.64	37.7	4,051	55.69	7.41	75.83	2.97	61.99
S _P	● slicers--navadeep-gemma3-12b	HF+BnB	FT	19	15.44	13.7	309	10.54	22.38	56.88	5.62	—
T ₈	● tahomamt--fp8	Tahoma	—	13	10.72	23.1	12,670	56.24	6.42	76.74	2.78	60.99
T _{8B4}	○ tahomamt--fp8-bok4	Tahoma	—	13	10.73	16.0	2,704	56.96	6.37	77.49	2.66	63.12
T _{4B4}	○ tahomamt--int4-bok4	Tahoma	—	7	6.03	22.5	2,500	56.08	6.46	77.06	2.69	64.31
T ₄	○ tahomamt--int4	Tahoma	—	7	6.02	23.2	9,923	54.92	6.52	75.74	2.82	62.11
T _{n5}	● tiny-titans--svd-qkv512-ff1620	HF	FT	15	11.75	10.3	352	31.53	19.78	—	—	—
T _{n4}	● tiny-titans--gptq-int4-calibrated	HF+GPTQ	cal	9	7.12	15.3	20	986	56.27	6.36	—	—
T _{n54}	● tiny-titans--svd-qkv512-ff1620-q4-g32-full	HF	FT	5	4.99	10.6	1,529	24.31	21.68	—	—	—
T _{M7}	● tmu-onono--diba-triton_direct	DIBA/Triton	FT	1	2.33	14.7	413	27.87	18.85	—	—	—
T _{MC}	○ tmu-onono--diba-cached_unpacked	DIBA/PT	FT	1	2.33	16.3	2,483	27.40	18.80	—	—	—
V _{50F}	● vicomtech--r0.50_ffn_sft_ck4802_awq4	VLLM	FT	4	4.27	33.3	20,190	44.33	14.38	58.04	5.70	58.30
V ₂₅	○ vicomtech--r0.25_ffn_awq4	VLLM	FT	6	5.71	33.3	17,250	53.70	10.37	72.48	3.67	70.69
V ₅₀	○ vicomtech--r0.50_ffn_awq4	VLLM	FT	4	4.27	35.2	23,580	26.47	16.76	55.57	5.74	48.85
V _{50Fz}	○ vicomtech--r0.50_ffn_sft_ck4802_awq4_en-zh	VLLM	FT	3	3.74	33.5	20,560	44.28	14.59	58.05	5.77	—
V _{50Fa}	○ vicomtech--r0.50_ffn_sft_ck4802_awq4_en-ar	VLLM	FT	3	3.66	29.8	22,030	44.26	14.23	—	—	58.26
V _{50Fc}	○ vicomtech--r0.50_ffn_sft_ck4802_awq4_cs-de	VLLM	FT	2	3.60	29.9	19,950	44.28	14.23	—	—	—

Unconstrained track (heterogeneous base < 20B params)

ID	System		Data	Size ↓	speed (cs→de)		cs→de		en→zh		en→ar	
					WU (s) ↓	Net c/s ↑	CK ↑	MX ↓	CK ↑	MX ↓	CK ↑	MX ↓
D _{A8}	● dragon1006--aya-expense-8b-int8	HF+BnB	cal	7	7.66	—	—	—	66.02	4.52	—	—
E _{A22}	● ests--gptoss-arz-k22	VLLM	FT	6	6.33	—	—	—	—	—	62.02	5.84
E _{Z26}	● ests--gptoss-zho-k26	VLLM	FT	3	5.15	—	—	—	72.44	3.35	—	—
E _{A24}	○ ests--gptoss-arz-k24	VLLM	FT	5	5.74	—	—	—	—	—	60.81	6.05
E _{A26}	○ ests--gptoss-arz-k26	VLLM	FT	3	5.15	—	—	—	—	—	56.78	7.01
E _{Z27}	○ ests--gptoss-zho-k27	VLLM	FT	2	4.85	—	—	—	71.40	3.45	—	—
E _{Z28}	○ ests--gptoss-zho-k28	VLLM	FT	1	4.55	—	—	—	68.90	3.74	—	—
P _{G4}	● pare4bit--unconstrained-gemma4	VLLM	—	10	9.59	123.1	4,146	58.65	6.08	79.08	2.62	68.22
Ti _{15N}	● tildeopen--15B-GPTQ-nvfp4a16	VLLM	FT	9	9.40	92.3	2,545	59.26	6.22	—	—	—
Ti ₈	○ tildeopen--8B-distill	VLLM	FT	12	15.21	25.2	5,134	58.11	6.31	—	—	—
Ti _{15F}	○ tildeopen--15B-RTN-fp8-dyn	VLLM	FT	11	15.15	98.3	2,840	58.85	6.27	—	—	—
Ti _{8F}	○ tildeopen--8B-distill-RTN-fp8-dyn	VLLM	FT	8	8.30	24.3	3,812	57.94	6.27	—	—	—
Ti _{8N}	○ tildeopen--8B-distill-GPTQ-nvfp4a16	VLLM	FT	4	5.28	86.0	6,548	57.96	6.39	—	—	—

Table 2: **Model compression task results along size, speed, quality dimensions.** All systems are categorized by track (constrained=top and unconstrained=bottom), and grouped by participant ID. Primary submissions are shown first (solid circles, bold text), followed by contrastive submissions (hollow circles). The serving runtime is shown in the grey tag; the **Data** column marks whether the method is data-free (—, on-the-fly quantization), calibration-only (cal), or uses fine-tuning/distillation/healing (FT). Metric columns: **Size** (GiB); **WU** (s), the median one-sentence warmup wall time; **Net c/s**, peak reliable Czech→German throughput after subtracting that warmup from full-set wall time; **CK**, COMET-Kiwi-XXL QE on a 0–100 scale (↑); and **MX**, MetricX-XXL error (↓). Warmup and Net c/s come from the same batch, selected to maximize net throughput; estimates for which warmup exceeds 75% of full-set wall time are omitted. Each cell is shaded by the system’s rank *within its own block and column*, from green (best) through yellow to red (worst), with the rank number in the top-left corner; near-identical scores are grouped to share a rank and colour (grouping thresholds: Size 0.1 GiB, warmup 1 s, speed 100 chars/s, CK 0.1, MX 0.01). ↑/↓ in the header gives the “better” direction; the two tracks are ranked independently. Participant IDs follow Table 1. ★: organizer baseline.

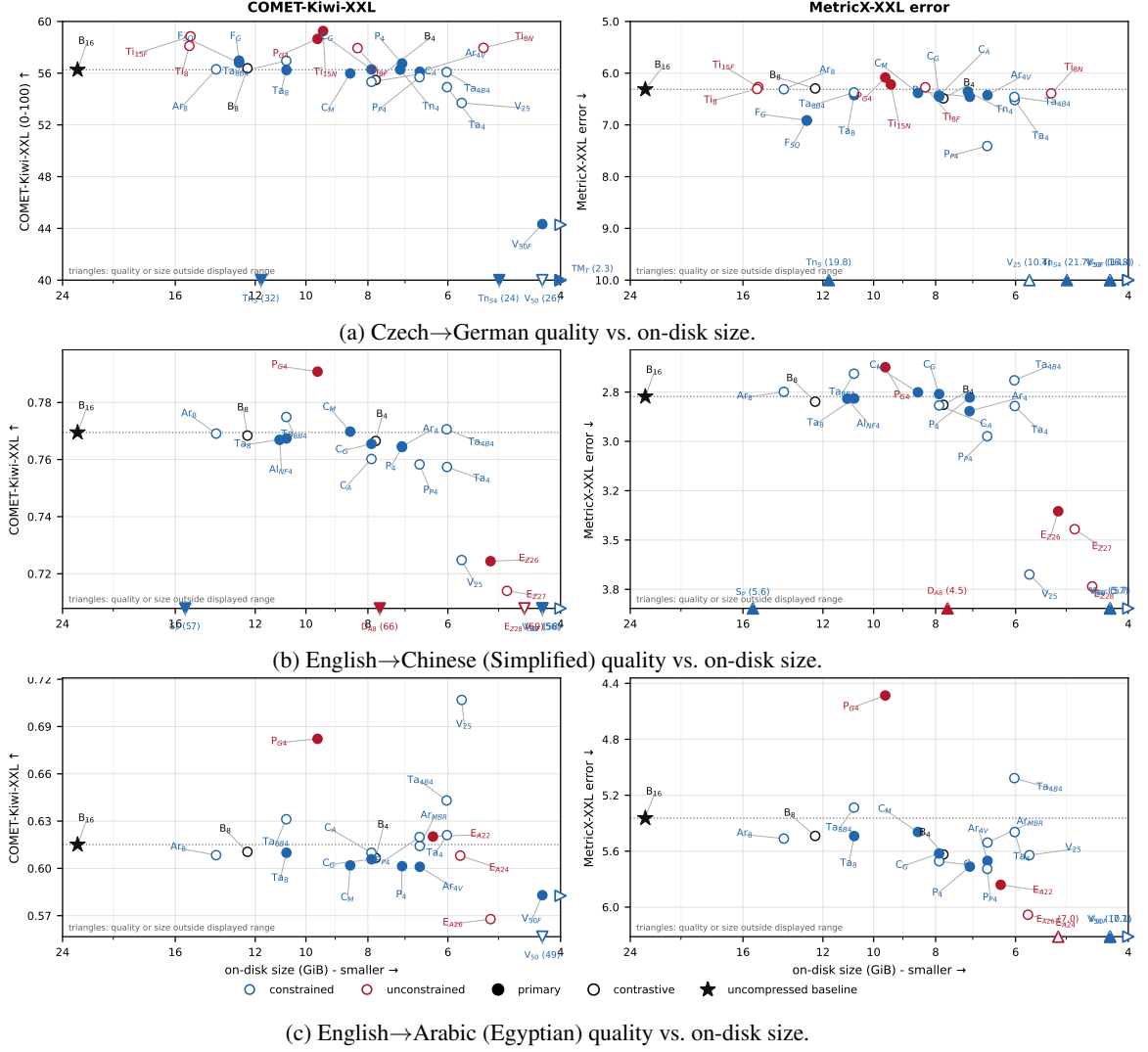


Figure 1: Quality vs. on-disk size for the three translation directions. Blue and red denote constrained and unconstrained systems; filled and hollow markers denote primary and contrastive submissions. Labels are participant IDs from Table 1; the star is the uncompressed baseline.

model can be shrunk to roughly a quarter of its on-disk size at essentially no quality cost: the smallest near-lossless systems are INT4 quantizations at 6.0–6.6 GiB (26–29% of the base)—arc/ilsp/vocaball-int4 for Czech→German, TahomaMT/int4-bok4 for English→Chinese (Simplified), and TahomaMT/int4 for English→Arabic (Egyptian)—all within ± 0.01 COMET-Kiwi of the uncompressed base. FP8 (W8A8) is essentially lossless at about half the size (10.7 GiB), and several INT8 and mixed-precision systems sit just below the base. Vision-tower removal and script-specific vocabulary pruning contribute additional, quality-neutral savings on top of quantization. Beyond quantization, the frontier degrades sharply: the very smallest artifacts (2.3–4.6 GiB) are the structural-compression

systems whose quality has collapsed, so they are Pareto-optimal only in the trivial sense of being small.

In the unconstrained track (Figure 1), *model selection acts as compression*. Pare4Bit’s INT4 Gemma-4-12B exceeds the Gemma-3 base on all three directions (+0.02 to +0.07 COMET-Kiwi) at 9.6 GiB, and MetricX confirms the gain (e.g. English→Arabic (Egyptian) 4.49 vs. 5.36). Tilde’s distilled 8B student, quantized to NVFP4, stays above the Gemma-3 base for Czech→German at only 5.3 GiB, and UCLA’s expert-pruned GPT-OSS-20B (ESTS) reaches 4.6–6.3 GiB by trading measured quality for size. These systems show that a well-chosen, newer sub-20B base—compressed—can dominate a compressed older base.

4.3 Inference Speed

Table 2 reports peak reliable net throughput on the Czech→German speed benchmark, with the corresponding one-sentence warmup time. Compression buys large speedups: the uncompressed base reaches 626 net chars/s, while the near-lossless TahomaMT/fp8 system reaches 12,670 net chars/s—about 20× the base throughput at near-baseline quality.

Crucially, throughput is governed as much by the serving runtime as by the precision. Optimized runtimes (Tahoma’s C++/CUDA engine, vLLM continuous batching) dominate, whereas a naive `bitsandbytes` INT8 baseline is pathologically slow: on the gross Czech→German batch-1 latency probe it takes 11,960 s (≈ 3.3 h) versus 85 s for the fastest system, a $\sim 140\times$ spread. Reducing precision is thus necessary but not sufficient for speed—the deployment stack must exploit it. Figure 2 shows the resulting quality–speed trade-offs on the Czech→German speed benchmark. At near-baseline quality, TahomaMT’s FP8 system reaches the fastest measured steady-state rate, while optimized vLLM systems form the next tier. The figure reports peak reliable net throughput: source characters divided by full-set wall time after subtracting the one-sentence warmup time. We report throughput scaling with batch sizes {1, 16, 64, and 256} only for Czech→German (the direction for which most submissions opted-in) in Table 3.⁴

Detailed Inference Speed Figure 3 shows how net throughput scales with batch size on the Czech→German speed subset, and Table 3 lists per-system net rates at each batch. Throughput grows almost linearly with batch size until a system reaches its declared maximum batch or exhausts memory: the fastest systems are Vi-comtech’s FFN-pruned AWQ variants (~ 23.6 k chars/s at batch 256) and TahomaMT’s FP8/INT4 on its custom engine (~ 10 – 13 k at batch 256), an order of magnitude above the uncompressed and `bitsandbytes` baselines. Rates are warmup-subtracted (steady-state); cells where the one-sentence warmup dominated the full-set wall time are omitted as unreliable.

⁴We suspect some submissions may not have strictly enforced batch size as a strict parallelism limiter due to the use of continuous batching in the underlying serving runtimes such as vLLM.

5 Conclusion and Future Directions

The second edition of WMT Model Compression shared task attracted participants with a variety of methods, which offer useful data points to draw a clear practical map. Post-training quantization—INT4 weight-only and FP8 weight-activation, often combined with vision-tower and vocabulary removal—is the dependable recipe: it compresses the Gemma-3-12B base to about a quarter of its on-disk size and yields up to a 20× steady-state throughput gain while remaining within quality-estimation noise of the uncompressed model. Structural and low-rank methods (SVD, diagonal/binary approximation, heavy FFN pruning) remain risky at this scale: without stronger recovery they trade away quality that light fine-tuning does not restore. In the unconstrained track, treating model selection as a compression decision pays off—a compressed, newer sub-20B base can beat the compressed older base outright. Finally, speed is a property of the serving stack more than the model size. Future editions should freeze peak-VRAM instrumentation, add human/LLM-judge quality, evaluate systems on several hardware configurations, and report speed with model-load time separated from steady-state decoding.

6 Limitations

Several caveats apply to this work. First, quality evaluation relies on reference-free QE metrics with strong meta-evaluation performance. High-quality references and large-scale human evaluation are costly, and strong QE metrics often provide comparable insights to reference-based evaluation (Freitag et al., 2024). Our conclusions remain subject to these metrics’ limitations and biases.

Second, footprint is measured on disk: peak GPU memory could not be measured reliably, and host resident-set sizes are not comparable across runtimes, so we defer a VRAM comparison. Third, the subtracted warmup includes initialization and one-sentence decoding, not model loading alone. We omit estimates above a 75% warmup fraction to limit subtraction instability. Batch-1 latency is available only for Czech→German. Fourth, measurements use a single H100 configuration; rankings may differ on other hardware. Lastly, unconstrained systems use heterogeneous bases; the constrained base is a reference, not an equal-footing comparison.

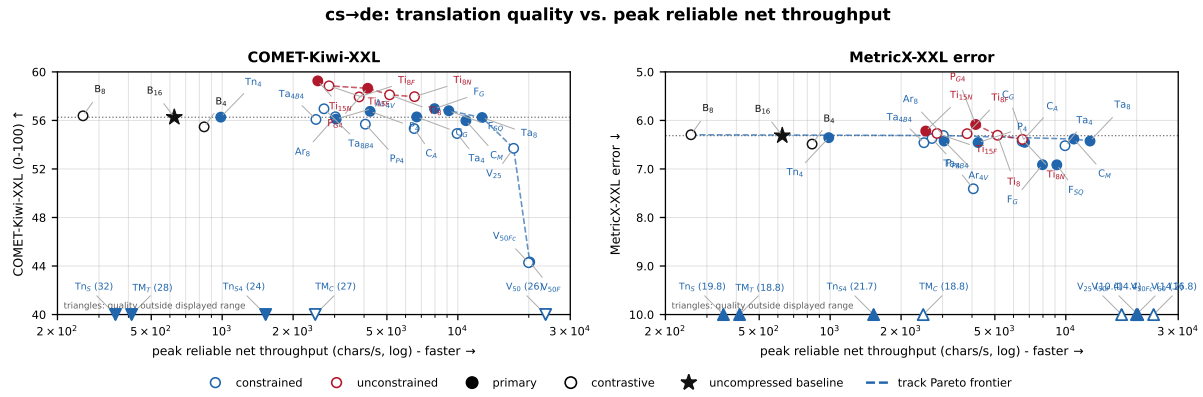


Figure 2: Quality versus peak reliable net throughput on the Czech→German speed benchmark. Throughput is source characters per second after subtracting the median one-sentence warmup time from the median full-set wall time; estimates with warmup exceeding 75% of the full-set wall time are excluded. Blue and red denote constrained and unconstrained systems; filled and hollow markers denote primary and contrastive submissions. Dashed lines show per-track Pareto frontiers. Triangles mark scores outside the displayed quality range.

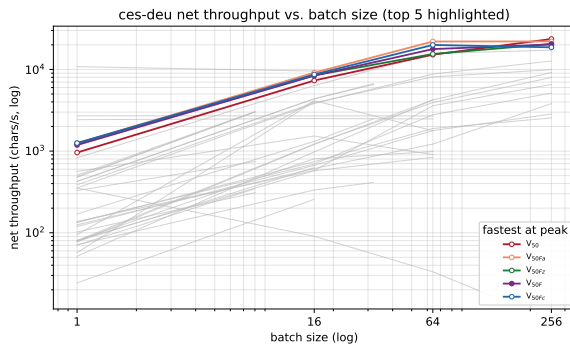


Figure 3: Net throughput vs. batch size (Czech→German, log-log). The five systems with the highest peak rate are highlighted; the rest are drawn in grey.

Acknowledgments

The work of Marco Gaido and Matteo Negri has been funded from the European Union’s Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BETWEEN People).

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, and 33 others. 2025. [Findings of the IWSLT 2025 evaluation campaign](#). In *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pages 412–481, Vi-

enna, Austria (in-person and online). Association for Computational Linguistics.

David Ifeoluwa Adelani, Victor Agostinelli, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Sébastien Bratières, Marine Carpuat, Fabrício Carraro, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, Marcello Federico, Marco Gaido, Mahendra Gupta, HyoJung Han, Ali Hatami, Lewis C. Howe, Dávid Javorský, Yejin Jeon, and 36 others. 2026. [Speech translation and metrics in 2026: Findings of the IWSLT campaign](#). In *Proceedings of the 23rd International Conference on Spoken Language Translation (IWSLT 2026)*, pages 336–422, San Diego, USA (in-person and online). Association for Computational Linguistics.

Angad Ripudaman Singh Bajwa. 2026. Comet sensitivity-guided mixed precision quantization for the wmt26 model compression task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Havish Balaga, Anish Racherla, Kolupoti Navadeep Kiran, Saumitra Yadav, and Manish Shrivastava. 2026. Importance-guided structural pruning of aya expand 8b and gemma-3-12b for english–chinese machine translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

Alexandra Birch, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Yusuke Oda. 2018. [Findings of the second workshop on neural machine translation and generation](#). In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 1–10, Melbourne, Australia. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, J. Kaplan, Prafulla Dhariwal, Arvind Nee-lakantan, Pranav Shyam, G. Sastry, Amanda Askell,

ID	System	GiB	WU (s)	Net chars/s at batch				peak	@ b
				1	16	64	256		
V ₅₀	vicomtech-r0.50_ffn_awq4	4.27	35.2	959	7 318	15 150	23 580	23 580	256
V _{50Fa}	vicomtech-r0.50_ffn_sft_ck4802_awq4_en-ar	3.66	29.8	1 247	9 134	21 980	22 030	22 030	256
V _{50Fz}	vicomtech-r0.50_ffn_sft_ck4802_awq4_en-zh	3.74	33.5	1 242	8 465	15 450	20 560	20 560	256
V _{50F}	vicomtech-r0.50_ffn_sft_ck4802_awq4	4.27	33.3	1 186	8 399	17 750	20 190	20 190	256
V _{50Fc}	vicomtech-r0.50_ffn_sft_ck4802_awq4_cs-de	3.60	29.9	1 253	8 681	19 950	18 650	19 950	64
V ₂₅	vicomtech-r0.25_ffn_awq4	5.71	33.3	830	6 343	15 620	17 250	17 250	256
Ta ₈	tahomamt-fp8	10.72	23.1	354	3 822	8 785	12 670	12 670	256
C _M	cometcut-comet-mixed-precision	8.53	58.6	10 820	9 550	—	—	10 820	1
Ta ₄	tahomamt-int4	6.02	23.2	387	3 911	8 110	9 923	9 923	256
F _{SQ}	fbk-sq-gptq	12.72	37.5	77	1 206	4 239	9 160	9 160	256
F _G	fbk-gptq-rtm	12.72	36.2	80	1 224	3 927	7 963	7 963	256
C _G	cometcut-gptq	7.90	50.1	425	4 364	—	—	6 690	32
Ti _{8N}	tildeopen-8B-distill-GPTQ-nvfp4a16	5.28	86.0	137	580	3 592	6 548	6 548	256
C _A	cometcut-awq	7.90	53.9	424	4 367	—	—	6 526	32
Ti ₈	tildeopen-8B-distill	15.21	25.2	135	696	2 784	5 134	5 134	256
P ₄	pare4bit-int4-gptq	7.07	104.2	52	—	4 244	—	4 244	64
P _{G4}	pare4bit-unconstrained-gemma4	9.59	123.1	95	4 146	1 778	—	4 146	16
P _{P4}	pare4bit-pruned4-healed-int4	6.64	37.7	57	4 051	—	—	4 051	16
Ti _{8F}	tildeopen-8B-distill-RTN-fp8-dyn	8.30	24.3	125	611	1 218	3 812	3 812	256
Ar _{4V}	arc-ilsp-vocaball-int4	6.64	41.2	507	—	—	—	3 046	8
Ar ₈	arc-ilsp-fp8	13.81	49.3	477	—	—	—	3 023	8
Ar ₄	arc-ilsp-int4	7.08	42.1	486	—	—	—	3 012	8
Ti _{15F}	tildeopen-15B-RTN-fp8-dyn	15.15	98.3	119	669	1 780	2 840	2 840	256
Ta _{8B4}	tahomamt-fp8-bok4	10.73	16.0	2 704	2 700	2 694	—	2 704	1
Ti _{15N}	tildeopen-15B-GPTQ-nvfp4a16	9.40	92.3	133	742	1 891	2 545	2 545	256
Ta _{4B4}	tahomamt-int4-bok4	6.03	22.5	2 424	2 500	2 478	—	2 500	16
TM _C	tmu-onono-diba-cached_unpacked	2.33	16.3	170	1 315	—	—	2 483	32
Tn _{S4}	tiny-titans-svd-qkv512-ff1620-q4-g32-full	4.99	10.6	564	1 529	907	—	1 529	16
Tn ₄	tiny-titans-gptq-int4-calibrated	7.12	15.3	80	805	986	—	986	64
B ₄	baseline-bnb-q4	7.78	11.7	81	574	838	—	838	64
Ar _{MBR}	arc-ilsp-en-ar-mbr	6.64	42.2	328	—	—	—	745	8
B ₁₆	baseline-uncompressed	22.74	13.9	78	626	—	—	626	16
TM _T	tmu-onono-diba-triton_direct	2.33	14.7	71	334	—	—	413	32
Tn _S	tiny-titans-svd-qkv512-ff1620	11.75	10.3	352	91	33	—	352	1
Al _{NF4}	alonso-layeraware-native-mlp-q4	10.99	12.2	70	—	—	—	346	8
SP	slicers-navadeep-gemma3-12b	15.44	13.7	102	—	—	—	309	8
B ₈	baseline-bnb-q8	12.34	21.5	24	257	—	—	257	16

Table 3: Detailed Czech→German inference speed. **ID** is the compact system identifier and **System** gives the corresponding package name; **GiB** is model size on disk. **WU** is the median one-sentence warmup wall time at the batch yielding the listed peak. Net throughput (characters per second) subtracts that warmup from the full-set wall time. “—” marks batches a system did not run or where warmup dominated the wall time (> 75%), making the net estimate unreliable. Primary submissions are bold. Ordered by peak rate.

- S. Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. Henighan, R. Child, A. Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, and 12 others. 2020. [Language models are few-shot learners](#). *ArXiv*, abs/2005.14165.
- John Dang, Shivalika Singh, Daniel D’souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. [Aya expanse: Combining research breakthroughs for a new multilingual frontier](#). *Preprint*, arXiv:2412.04261.
- Tianyu Ding, Tianyi Chen, Haidong Zhu, Jiachen Jiang, Yiqi Zhong, Jinxin Zhou, Guangzhi Wang, Zhihui Zhu, Ilya Zharkov, and Luming Liang. 2024. [The efficiency spectrum of large language models: An algorithmic survey](#). *Preprint*, arXiv:2312.00678.
- Nail Fakhrutdinov. 2026. Pare4bit: Quantization, depth pruning, and model selection for the wmt26 model compression task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chiklu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchichio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs breaking MT metrics? results of the WMT24 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.
- 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*, pages 484–494, Suzhou, China. Association for Computational Linguistics.
- Thamme Gowda. 2026. Quantize, qualify, rerank: A recipe for compressing llms without losing quality. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.

- Hiroaki Hayashi, Yusuke Oda, Alexandra Birch, Ioannis Konstas, Andrew Finch, Minh-Thang Luong, Graham Neubig, and Katsuhito Sudoh. 2019. [Findings of the third workshop on neural generation and translation](#). In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, pages 1–14, Hong Kong. Association for Computational Linguistics.
- Kenneth Heafield, Hiroaki Hayashi, Yusuke Oda, Ioannis Konstas, Andrew Finch, Graham Neubig, Xian Li, and Alexandra Birch. 2020. [Findings of the fourth workshop on neural generation and translation](#). In *Proceedings of the Fourth Workshop on Neural Generation and Translation*, pages 1–9, Online. Association for Computational Linguistics.
- Kenneth Heafield, Biao Zhang, Graeme Nail, Jelmer Van Der Linde, and Nikolay Bogoychev. 2022. [Findings of the WMT 2022 shared task on efficient translation](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 100–108, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Kenneth Heafield, Qianqian Zhu, and Roman Grundkiewicz. 2021. [Findings of the WMT 2021 shared task on efficient translation](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 639–651, Online. Association for Computational Linguistics.
- 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.
- Martins Kronis, Toms Bergmanis, Ingus Jānis Pretkalniņš, and Marcis Pinnis. 2026. Wmt26 model compression task: Quantising and distilling tildeopen. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Liu O. Martin, Lucas Bandarkar, and Nanyun Peng. 2026. Ests at wmt26: Routing-informed expert pruning for model compression. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Nobutaka Ono. 2026. Tmu-onono at wmt26: Diba-based model compression for czech-to-german machine translation. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Alonso Palomino. 2026. Layer-aware native quantization for compact machine translation: A wmt26 english-to-chinese submission. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- R. Patil and V. Gudivada. 2024. [A review of current trends, techniques, and challenges in large language models \(llms\)](#). *Applied Sciences*.
- David Ponce and Thierry Etchegoyhen. 2026. Vi-comtech@wmt 2026: Mask-based pruning for model compression. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Vania Raya-Rios, Aleksandr Klimchuk, Ivo Gollini Navarrete, Nicolas M. Cuadrado Avila, Martin Takác, and Samuel Horvath. 2026. Tiny titans at wmt26: Task-calibrated quantization and activation-aware low-rank compression of gemma 3 12b. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, Josã© Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André Martins. 2023. [Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 841–848, Singapore. Association for Computational Linguistics.
- Sokratis Sofianopoulos and Prokopis Prokopidis. 2026. The ilsp/arc submission to the wmt26 model compression task. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. Association for Computational Linguistics.
- Dhairya Suman and Luisa Bentivogli. 2026. Fbk’s submission to wmt26 model compression task: Improving inference speed for quantized models. In *Proceedings of the Eleventh Conference on Machine Translation*, Budapest, Hungary. 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](#). *Preprint*, arXiv:2503.19786.
- Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. 2024. [A survey on model compression for large language models](#). *Transactions of the Association for Computational Linguistics*, 12:1556–1577.