Quantize, Qualify, Rerank: A Recipe for Compressing LLMs Without Losing Quality

Thamme Gowda
Microsoft
thammegowda@microsoft.com

Abstract

Compressing large language models lowers serving cost, but a single automatic metric may miss quality loss. For the WMT26 Model Compression constrained track, we compress `google/gemma-3-12b-it` (24.4 GB in bf16) for three translation directions on one H100. Paired bootstrap analysis exposes metric disagreement on 4-bit quantization; guided by WMT24 human meta-evaluation, we use MetricX-24-XXL as primary and CometKiwi-XXL as a cross-check. Removing the unused vision stack and non-task vocabulary and using int8 weights or fp8 activations causes no resolved loss, whereas int4 leaves a small residual. QLoRA healing overfits its calibration domain. A Cometoid best-of-4 reranker at T=0.3 improves both the 8-bit and 4-bit systems under WMT25 test set. On post-submission WMT26, both remain competitive with bf16 under reference-free metrics; the small difference between the reranked variants is treated as noise.

1 Introduction

General-purpose large language models (LLMs) spend most of their capacity on broad competence. A machine translation (MT) deployment restricted to three fixed language pairs asks a different question: which parts can be removed, quantized, or replaced without a user noticing? We treat "without noticing" as a measurement problem rather than infer it from a raw metric ordering. We first compare several metrics with paired bootstrap resampling; when they disagree, we use WMT Metrics Shared Task's meta-evaluation as an external anchor. Methods that learn from examples, including calibration, adapter tuning, pruning recovery, and QLoRA-style adaptation (Dettmers et al., 2023), must retain gains across domains.

No MT metric is uniformly reliable across model and output regimes, particularly for close systems. WMT therefore meta-evaluates evolving metric panels against human judgments each year (Callison-Burch et al., 2008; Freitag et al., 2024; Lavie et al., 2025). When our local metrics disagree, the WMT24 human ranking supplies study-specific external evidence. We test data-dependent compression under domain shift.

We study this question in the WMT26 Model Compression constrained track,[1] specializing the prescribed Gemma-3-12B model for three translation directions. We evaluate compression levers alone and in composition, then test whether their apparent gains survive metric qualification, held-out domains, and the post-submission WMT26 blind set.

Contributions and findings.

1. **Metric qualification.** We show that many popular metrics can miss or disagree on quantization losses, and present a paired-bootstrap qualification procedure that measures score-ordering stability and anchors metric choice in human meta-evaluation (Section 4).
2. **Data-free compression.** We combine vision/vocabulary removal, int8 embeddings, and int8/fp8 quantization without calibration, and show that qualified metrics resolve no 8-bit loss, while int4 retains a small residual (Sections 3.1, 5).
3. **Generalization gate.** We show that QLoRA healing can recover in its calibration domain while substantially degrading quality out of domain, motivating held-out-domain validation for learned compression (Section 5.1).

[1] https://www2.statmt.org/wmt26/model-compression.html

4. **Inference-time recovery.** Best-of-k reranking improves both compressed systems under held-out WMT25 metrics and remains competitive with bf16 on WMT26; small differences between reranked variants are treated as noise (Section 6).

2 WMT26 Model Compression Task

The constrained track compresses `google/gemma-3-12b-it` (Gemma Team, Google DeepMind, 2025) for Czech→German, English→Simplified Chinese, and English→Egyptian Arabic on one H100. Systems are ranked on a quality×size×speed Pareto frontier; size includes disk and peak load-time VRAM, and speed is end-to-end tokens/s.[2] We used the 1121-segment WMT25 set for development and scored WMT26 only after submissions were fixed.

We serve the text path with *tahoma*, a C++ runtime with native int4/int8 (Wilkinson, 2024) and fp8 kernels. Its serving architecture follows vLLM (Kwon et al., 2023), including paged attention, continuous batching, and CUDA graph replay.

3 Compression Levers

Compression choices should follow parameter mass: MLPs account for 72% of the text model, attention for 19%, and the tied embedding/LM-head table for 9%. We consider *structural* removal and *representational* precision changes, evaluating data-free levers before calibration-dependent methods.

3.1 Structural: Removing Deadweight

The vision stack. The base checkpoint is multimodal, but three translation directions never see an image. Its vision stack, a 417M-parameter SigLIP encoder plus multimodal projector occupying 0.84 GB in bf16, is therefore unused. We drop it outright, data-free, before any other lever; the 23.5 GB baseline used throughout is this text-only model. The tokenizer contains further removable capacity.

Vocabulary. The Gemma tokenizer carries 262,208 pieces covering every language and code; three translation directions need a handful of scripts. We categorize every piece by the Unicode *script* of its letters and keep only system/special tokens, the 256 byte-fallback tokens, and pieces whose letters lie in the task scripts (Latin, Arabic, Han); digits, punctuation, and symbols are script-neutral and allowed. As an additional rule, long symbol composites ($\geq$ 3 symbol characters, e.g. code/markup fragments) are dropped while short ones are kept.

This drops 66,831 pieces (25.5%) from non-task scripts. It is safe because any dropped piece still round-trips through byte fallback. The embedding table shrinks from 2.01 GB to 1.50 GB (bf16), and composes with int8 embeddings for a further reduction (Table 1).

Stage	Pieces	Embedding (bf16)
Full tokenizer	262,208	2.01 GB
Script keep-set	195,377	1.50 GB
+ int8 embeddings	195,377	∼0.75 GB

Table 1: Vocabulary specialization: 25.5% fewer pieces with byte-fallback coverage.

System	W/A	Role
bf16	bf16/bf16	native baseline
fp32	fp32/fp32	near-null probe
i8a16	int8/bf16	weight-only, 2×
f8a8	fp8/fp8	speed lever
i4a16	int4/bf16	weight-only, 3.5×
i4a8	int4/fp8	smallest+fastest 4-bit
NF4	FP4/bf16	codebook, cross-engine

Table 2: Precision variants compared throughout the paper (weights/activations format). Section 4 qualifies which metrics resolve differences among them; Section 5 reports their measured scores.

3.2 Representational: Precision Variants

Beyond removing deadweight, we compare the baseline bf16 checkpoint against six alternative numeric representations of the same weights (Table 2). All variants except NF4 run on *tahoma*; NF4 is bitsandbytes' learned-codebook 4-bit quantization, applied online through vLLM rather than tahoma, making it a cross-engine reference.

The baseline checkpoint is natively bf16, so we use it as the reference; fp32 accumulation is an engineering near-null probe, and fp16 is excluded because it overflows Gemma-3 activations. The i4a16 recipe uses asymmetric blockwise quantization with MSE-optimal clip-

[2] https://github.com/thammegowda/wmt26-model-compression

ping. We combine the same int4 weights with fp8 activations in a W4A8 kernel (i4a8), whose matmuls use Hopper fp8 tensor cores.

Blockwise 4-bit scales trade storage for granularity: a smaller block gives finer scales but stores more of them. The deployable window is set by kernel geometry, not preference: each block must divide every quantized weight's input dimension $K \in \{3840, 4096, 15360\}$, whose GCD is 256, so 256 is the largest admissible block and 384/512/768 are runtime-rejected (e.g. $384 \nmid 4096$); below 64, block 32 is refused by the Machete W4A16 kernel, which supports only group sizes $\geq$64. This leaves $\{64, 128, 256\}$ as the only deployable block sizes; Section 5 reports their measured quality (Table 4).

4 Metric Qualification and Bootstrap Analysis

We compare ChrF2 (Popović, 2015; Post, 2018) under both mean-segment and corpus aggregation. We compute COMET22 (Rei et al., 2020), BLEURT20 (Sellam et al., 2020), CometKiwi22 (Rei et al., 2022), and CometKiwi-XXL (Kiwi_{XXL}) (Rei et al., 2023), and Cometoid (Gowda et al., 2023) with PyMarian (Gowda et al., 2024). We additionally score XCOMET-XL (Guerreiro et al., 2024) and evaluate MetricX-24-XXL (Mx24_{XXL})[3] (Juraska et al., 2024) in reference-based ($\text{Mx24}_{\text{XXL}}^{\text{ref}}$) and QE ($\text{Mx24}_{\text{XXL}}^{\text{QE}}$) modes on the variants of Section 3.2.

We run paired bootstrap resampling separately for each metric and system pair (Koehn, 2004). For rows with additive segment scores, let x_i be the signed score difference on segment i: bf16 minus the variant for higher-is-better metrics, and the variant minus bf16 for lower-is-better Mx24_{XXL}. Thus $x_i > 0$ always favors bf16. Each replicate draws N=1121 values with replacement from the observed paired differences $x_1, \ldots, x_N$. If $x_j^{(b)}$ is the jth value drawn in replicate b, we compute

$$\begin{aligned} \bar{x}^{(b)} &= \frac{1}{N} \sum_{j=1}^{N} x_j^{(b)} \\ S_M &= \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\left[\bar{x}^{(b)} > 0\right] \end{aligned} \tag{1}$$

with B=4000 replicates. Here $\bar{x}^{(b)}$ is the mean difference in the bth virtual test set, and $\mathbf{1}[\cdot]$ is one when its condition holds. Therefore S_M=97% means that bf16 obtains the better mean metric score in 97% of the virtual test sets; it is not the probability of a human-quality loss. Values near 50% leave the metric ordering unresolved; it does not sufficiently prove quality equivalence.

The *ChrF2 (mean seg.)* row applies this protocol to sentence-level ChrF2 scores. For *ChrF2 (corpus)*, each replicate instead uses the same paired indices to compute SacreBLEU's corpus-level ChrF2 difference.

We estimate the required sample size as

$$N^{\star} = \left(\frac{1.96\, s}{\bar{x}}\right)^2 \tag{2}$$

where $\bar{x}$ and s are the observed mean and standard deviation of the signed differences for that comparison. It estimates how many segments would be needed for a two-sided 95% interval to exclude zero if the observed effect and variance persist; we show ∞ when $\bar{x} \leq 0$ because the observed direction does not favor bf16. This additive-segment extrapolation is undefined for the nonlinear corpus-ChrF2 aggregation. We treat fp32/int8/fp8 as near-null probes and int4 as the stronger perturbation.

The local study is ambiguous. Both Mx24_{XXL} modes strongly favor bf16 over the two Tahoma int4 variants. Kiwi_{XXL} favors bf16 over i4a16, while i4a8 remains unresolved; several other learned metrics do not resolve a consistent 4-bit effect. ChrF2 also favors bf16.

External anchor. WMT24 ranks Mx24_{XXL} in the top tier against professional MQM/ESA judgments and Kiwi_{XXL} above older baselines (Freitag et al., 2024); ChrF2 ranks much lower. We therefore use Mx24_{XXL} as primary and Kiwi_{XXL} as a reference-free cross-check, describing differences as metric residuals rather than demonstrated human-significant losses.

5 Effective Compression Levers

Precision, engines, and blind-set transfer. Table 7 combines WMT25 development results with the WMT26 test set, kept blind during development. On both sets, i8a16 and f8a8 remain near bf16, whereas i4a16, i4a8, and vLLM NF4 leave larger metric residuals; these

[3] MetricX implementation: https://github.com/thammegowda/metricx.

Bootstrap support S_M and one-sided $N^\star$

$S_M \le 5\%$: variant wins in $\ge 95\%$ of resamples | $5\% < S_M < 95\%$: neither wins in $\ge 95\%$

$95\% \le S_M < 99\%$: bf16 wins in 95–99% of resamples | $S_M \ge 99\%$: bf16 wins in $\ge 99\%$

		fp32		i8a16		f8a8		i4a16		i4a8		NF4	
Metric	**Size**	S_M	$N^\star$	S_M	$N^\star$	S_M	$N^\star$	S_M	$N^\star$	S_M	$N^\star$	S_M	$N^\star$
CometKiwi22	565M	45.3%	∞	22.7%	∞	23.1%	∞	73.0%	12.1k	10.7%	∞	81.7%	5.5k
Cometoid	569M	41.0%	∞	74.7%	12.6k	30.0%	∞	41.4%	∞	6.6%	∞	44.8%	∞
BLEURT20	576M	2.8%	∞	12.6%	∞	49.0%	∞	60.2%	68.1k	39.8%	∞	4.9%	∞
COMET22	581M	93.3%	2.0k	60.9%	87.1k	58.6%	103k	58.7%	85.1k	89.0%	2.8k	48.1%	∞
XCOMET-XL	3.49B	6.8%	∞	13.2%	∞	4.2%	∞	35.3%	∞	53.2%	570k	10.1%	∞
Kiwi_{XXL}	10.7B	65.4%	24.2k	55.4%	445k	21.0%	∞	98.2%	989	56.0%	334k	99.4%	711
$\text{Mx24}_{\text{XXL}}{}^{\text{ref}}$	12.9B	70.1%	14.1k	70.8%	16.0k	67.8%	23.0k	100.0%	**172**	100.0%	337	100.0%	450
$\text{Mx24}_{\text{XXL}}{}^{\text{QE}}$	12.9B	79.0%	5.8k	36.7%	∞	65.7%	28.9k	100.0%	**184**	100.0%	**325**	99.9%	500
ChrF2(mean seg.)	–	62.2%	44.0k	67.4%	22.6k	98.8%	892	100.0%	258	100.0%	196	100.0%	129
ChrF2(corpus)	–	57.1%		81.8%		98.2%		100.0%		100.0%		100.0%	

Table 3: Paired-bootstrap support S_M for bf16 over each variant. $N^\star$ estimates the sample size needed to resolve a bf16-favoring effect at 95%; ∞ denotes the opposite observed direction. Corpus ChrF2 has no additive $N^\star$.

are not claims of human-perceptible loss. Both best-of-4 systems remain competitive with bf16 on WMT26. Their differences are small and not significance-tested, so we do not interpret i4a8-bok4's numerical lead as evidence that int4 is better.

The engine controls separate compression from implementation effects. tahoma's bf16 quality is close to hf-bf16 across all four score columns[4], with small numerical differences across engines. vLLM bf16/fp8 changes are mixed, so we do not interpret their apparent metric gains. vLLM quantizes the bf16 snapshot on load; its FP8/NF4 disk values come from equivalent static exports. At steady state, tahoma f8a8 matches vLLM FP8 (6094 vs. 5934 tok/s), while i4a8 exceeds vLLM NF4 (5131 vs. 3190 tok/s). With the same local checkpoint, tahoma cold start is 24–34 s, versus 55–168 s for vLLM FP8 and 92 s for NF4.

4-bit block size. Of the deployable sizes in Section 3.2, block 128 gives the best quality (Table 4); block 256 is slightly smaller and faster but worse, so submissions use 128.

8-bit embeddings are nearly lossless. Per-row int8 quantization cuts the tied embedding/LM-head table from 2.01 to ∼1.0 GB with changes of at most 0.18 Kiwi_{XXL} and 0.04 $\text{Mx24}_{\text{XXL}}{}^{\text{ref}}$ (Table 5); all submissions use it.

[4]tahoma's Gemma-3 originally closely reimplemented HuggingFace `transformers`, but inadvertently omitted linear RoPE scaling, slightly degrading quality. The omission was identified by cross-checking vLLM's reference implementation and fixed; the post-fix scores are reported here, with hf and vLLM controls unchanged.

Block	$\text{Kiwi}_{\text{XXL}}\uparrow$	$\text{Mx24}_{\text{XXL}}{}^{\text{ref}}\downarrow$	**Size**	**tok/s**
64	64.61	5.89	7.31	4706
128	**65.22**	**5.81**	6.96	4723
256	64.28	6.11	**6.78**	**4758**

Table 4: Deployable 4-bit block sizes; size is on-disk GB.

	$\text{Kiwi}_{\text{XXL}}\uparrow$		$\text{Mx24}_{\text{XXL}}{}^{\text{ref}}\downarrow$	
Variant	**bf16**	**int8**	**bf16**	**int8**
i8a16	65.10	65.16	5.66	5.68
i4a16	64.58	64.62	5.87	5.89
i4a8	64.96	64.93	5.87	5.85
f8a8	65.24	65.42	5.63	5.67

Table 5: Per-row int8 embedding ablation.

5.1 Recovery via Calibration: Domain Sensitivity

We QLoRA-heal i4a8 on GlobalVoices for CES-DEU and ENG-ARA_EG (Dettmers et al., 2023). Late recovery helps the matched domain but degrades both WMT25 OOD metrics (Table 6), so we reject it.

6 Inference Time Recovery

Although i4a8 has a small $\text{Mx24}_{\text{XXL}}{}^{\text{ref}}$ residual, compressed systems still beat bf16 on some

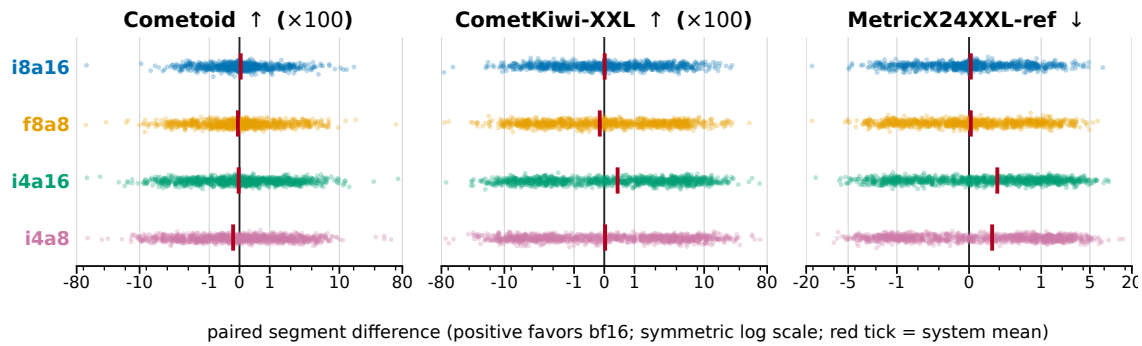

Figure 1: Per-segment differences for i8a16, f8a8, i4a16, and i4a8 against bf16 on WMT25 under Cometoid, Kiwi_{XXL}, and $\text{Mx24}_{\text{XXL}}^{\text{ref}}$. Positive values favor bf16; negative values identify segments where the compressed output scores better. Red ticks mark means; axes use symmetric log scales.

	GlobalVoices		WMT25 OOD	
Stage	**KXXL↑**	**M24XXL↓**	**KXXL↑**	**M24XXL↓**
i4a8	84.21	4.282	65.05	5.837
step=50	81.58	4.637	63.06	6.387
step=100	81.09	4.625	62.82	6.483
step=400	84.61	4.123	59.94	7.234

Table 6: QLoRA trajectory for i4a8: 600 held-out GlobalVoices segments versus the 1121-segment WMT25 OOD set. KXXL=CometKiwi-XXL; M24XXL=MetricX-24-XXL-ref.

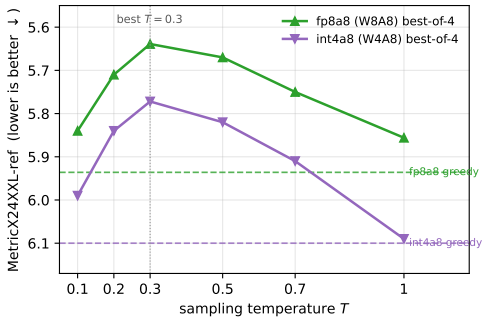

Figure 2: Cometoid best-of-4 temperature sweep on WMT25. Solid lines show held-out $\text{Mx24}_{\text{XXL}}^{\text{ref}}$ (axis inverted); dashed lines show greedy decoding.

segments (Figure 1). We exploit this variation by sampling four candidates and selecting the highest-scoring one with the 569M-parameter, reference-free Cometoid QE model (Gowda et al., 2023). We evaluate with held-out Kiwi_{XXL} and $\text{Mx24}_{\text{XXL}}^{\text{ref}}$.

What temperature to use for sampling? A single T=1.0 sample is worse than greedy, and even best-of-4 barely helps. The WMT25 sweep instead peaks at T=0.3 for f8a8 and i4a8 (Figure 2), improving $\text{Mx24}_{\text{XXL}}^{\text{ref}}$ by 0.17–0.27 (Table 7). We submit these settings as `f8a8-bok4` and `i4a8-bok4`. Their bundled Cometoid selector uses 4-bit weights and int8 embeddings, occupies ∼420 MB, and is included in the reported submission disk size; decoding costs approximately 4× greedy.

7 Related Work

Quantization and adaptation. Post-training LLM compression includes outlier-aware int8 computation (Dettmers et al., 2022), weight-only GPTQ and AWQ (Frantar et al., 2023; Lin et al., 2024), and activation quantization such as SmoothQuant (Xiao et al., 2023). We instead evaluate disk size, peak VRAM, H100 throughput, and generated-text quality jointly: at our serving point, int4 weights primarily reduce size, fp8 activations improve speed, and W4A8 combines both. Calibration-aware quantizers and LoRA/QLoRA depend on representative data (Hu et al., 2022; Dettmers et al., 2023); our GlobalVoices–WMT25 split shows that in-domain healing can reverse out-of-domain, motivating our OOD gate.

Metric qualification. Learned metrics improve over surface-form matching (Rei et al., 2020; Sellam et al., 2020; Rei et al., 2022; Juraska et al., 2024), but meta-evaluations show that score magnitudes and system rankings need not match human preferences (Kocmi et al., 2021, 2024; Freitag et al., 2024; Lavie et al., 2025). Building on paired bootstrap testing (Koehn, 2004), we use controlled precision changes to probe metric resolution, quantify

System	Disk GB	VRAM GB	tok/s	WMT25		WMT26 blind	
				$\text{Kiwi}_{\text{XXL}}\uparrow$	$\text{Mx24}_{\text{XXL}}{}^{\text{ref}}\downarrow$	$\text{Kiwi}_{\text{XXL}}\uparrow$	$\text{Mx24}_{\text{XXL}}{}^{\text{QE}}\downarrow$
bf16	23.5	23.2	4486	65.33	5.493	68.69	4.108
i8a16	11.8	14.4	4891	65.27	5.522	68.77	4.128
i4a16	**6.7**	9.7	5103	64.64	5.715	68.43	4.196
f8a8$^{\checkmark,\star}$	11.5	14.2	**6094**	65.21	5.491	68.78	4.136
i4a8$^{\checkmark}$	**6.7**	9.7	5131	65.40	5.570	68.78	4.128
f8a8-bok4$^{\checkmark}$	12.0	12.9	1750	65.95	5.462	69.83	4.009
i4a8-bok4$^{\checkmark}$	6.9	**7.5**	1456	**66.32**	5.467	**70.02**	**3.986**
hf-bf16	23.5	–	–	65.28	5.529	68.65	4.156
vllm-bf16	23.5	26.1	4917	65.46	5.407	68.68	4.085
vllm-fp8	15.3	15.8	5934	65.40	**5.378**	68.68	4.096
vllm-nf4	8.4	11.8	3190	64.47	5.636	67.99	4.228

Table 7: WMT25 development and WMT26 blind-test operating points. $^{\checkmark}$ marks submissions and * the primary. WMT26 uses reference-free metrics because references were unavailable at the time of writing. Throughput is hot steady state and per retained output; best-of-4 disk includes its 0.42 GB selector. hf-bf16 is a quality-only `transformers` reference.

ordering stability, and anchor downstream metric choice in WMT human meta-evaluation. We therefore report resolved differences as metric residuals, not automatically as human-perceptible losses.

Quality-aware decoding. Prior work questions mode-seeking decoding (Eikema and Aziz, 2020) and uses learned metrics for N-best reranking or minimum Bayes-risk decoding (Fernandes et al., 2022; Freitag et al., 2022). Reference-free QE makes selection deployable without references (Rei et al., 2022, 2023; Gowda et al., 2023). Our cheaper alternative independently scores four low-temperature samples with a small bundled Cometoid model and is evaluated by held-out arbiters. Applying it to f8a8 and i4a8 shows that sampling headroom survives quantization.

8 Conclusion

Our calibration-free stack removes unused vision/vocabulary components, uses int8 embeddings, and selects f8a8 for speed or i4a8 for size. Int8/fp8 show no resolved loss; int4 has a small $\text{Mx24}_{\text{XXL}}{}^{\text{ref}}$ residual, and QLoRA fails the OOD gate. Cometoid best-of-4 recovers WMT25 quality and remains competitive on WMT26; we make no claim that one reranked precision is better than the other.

Limitations

Results cover one model, one H100 runtime, and three translation directions, so speed trade-offs may not transfer to other hardware. QLoRA calibration covers only two directions. We lack direct human comparisons: WMT24 supplies an external metric ranking, not judgments on our outputs.

References

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. 2008. Further meta-evaluation of machine translation. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 70–106, Columbus, Ohio. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Bryan Eikema and Wilker Aziz. 2020. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4506–4520, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1396–1412, Seattle, United States. Association for Computational Linguistics.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *International Conference on Learning Representations (ICLR)*.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, 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.

Gemma Team, Google DeepMind. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Thamme Gowda, Roman Grundkiewicz, Elijah Rippeth, Matt Post, and Marcin Junczys-Dowmunt. 2024. PyMarian: Fast neural machine translation and evaluation in python. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 328–335, Miami, Florida, USA. Association for Computational Linguistics.

Thamme Gowda, Tom Kocmi, and Marcin Junczys-Dowmunt. 2023. Cometoid: Distilling strong reference-based machine translation metrics into even stronger quality estimation metrics. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 751–755, Singapore. Association for Computational Linguistics.

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

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 (WMT)*.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation (WMT)*, pages 478–494. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1999–2014, Bangkok, Thailand. 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*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.

Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Suzhou, China. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2024. AWQ: Activation-aware weight quantization for LLM compression and acceleration. In *Proceedings of Machine Learning and Systems (MLSys)*.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation (WMT)*, pages 186–191.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, Daan van Stigt, Marcos Treviso, Luisa Coheur, José G. C. de Souza, and André F. T. Martins. 2023. Scaling up CometKiwi: Unbabel-IST 2023 submission for the quality estimation shared task. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*, pages 841–848.

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.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Lucas Wilkinson. 2024. Introducing Machete, a mixed-input GEMM kernel optimized for NVIDIA Hopper GPUs. Neural Magic / Red Hat Developer Blog.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. 2023. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*.