

The ILSP/ARC submission to the WMT26 Model Compression Task

Sokratis Sofianopoulos and Prokopis Prokopidis

Institute for Language and Speech Processing

Athena Research Center, Greece

{s_sofian,prokopis}@athenarc.gr

Abstract

This paper describes our submission to the constrained track of the WMT26 Model Compression task, filed under the team name arc/ilsp. The task requires the compression of google/gemma-3-12b-it for translation in cs→de, en→zh, and en→ar in the Egyptian register. Our systems combine the removal of the vision tower, the pruning of the vocabulary by Unicode script, and GPTQ W4A16 quantization. The two checkpoints we file as primary models occupy 6.64 and 7.08 GiB, i.e. 29% and 31% respectively of the 22.7 GiB disk footprint of the bf16 original. No interval separates the pruned vocabulary from the full-vocabulary INT4 checkpoint on any direction, and at the paragraph granularity on which the task scores, our vocab-INT4 primary costs 0.0048 COMET against bf16 on cs→de and nothing measurable on the other two directions. We also submit two contrastive systems: an FP8 system that is faster than the bf16 original at lower memory, and a self-MBR decoder that adds roughly 0.009 COMET on en→ar without any additional parameters. We do not submit a depth-pruned system; a default generation cap had clipped the distillation targets and taught the student to end a turn mid-paragraph, a defect whose repair recovers a quarter of the deficit against full depth.

1 Introduction

The WMT26 Model Compression task¹ (Gowda et al., 2026) asks participants to compress general-purpose LLMs while preserving translation quality across three directions, cs→de, en→zh and en→ar in the Egyptian register. We describe our submission to the constrained track, which requires the compression of google/gemma-3-12b-it (Gemma Team, 2025). Systems are run by the organizers on a single H100 and ranked on quality–size and quality–speed Pareto frontiers.

¹Task description and rules: <https://www2.statmt.org/wmt26/model-compression.html>

Our starting point was that the task scores translation and nothing else, so that the capacity serving general abilities, languages outside the task, or vision is redundant. Post-training quantization is available to every participant and does the larger part of the work, taking the 22.7 GiB bf16 original to 7.08 GiB. The question we set out to answer is what task specialization adds on top of that shared baseline. Our submitted systems answer it along two axes, the removal of the vision tower and the pruning of the vocabulary by Unicode script, which together reach 6.64 GiB. The vocabulary increment is small enough that it only earns its place if it costs no quality, and no interval on any direction separates the pruned system from the full-vocabulary INT4 primary (Section 5.2). We also examined structured depth pruning with distillation healing, which would have moved the size further. On this data it is dominated by the two simpler alternatives, so Section 6 presents it as a measurement and not as a system.

We submitted four packages as two form entries sharing one codebase, since the form allows a single primary per entry (Table 1, with submission details in Appendix A). Two INT4 checkpoints serve as primary models, one per direction. The evidence on which we made that split did not survive re-certification, as Section 5.2 reports, and we retain the split as filed, which is the conservative choice. The FP8 and self-MBR packages are contrastive. Table 2 summarises what each compression step contributes. Apart from the systems, we report three findings. The first is that a generation cap silently truncated our distillation targets and taught the student to end a turn mid-paragraph (Section 6). The second is that an apparent cost of the vocabulary prune came from comparing against a different build of the same recipe (Section 5.2). The third is that a pre-registered interaction test does not separate the MBR gain from a general decode-time effect (Section 5.3).

Package	Role	Disk	VRAM	Speed
vocab-INT4	PRIMARY cs→de, en→ar	6.64	7.41*	−2–4%
INT4	PRIMARY en→zh	7.08	7.84	−2–4%
FP8	contrastive, all pairs	13.8	12.69	+16–18%
+ self-MBR	contrastive, en→ar only	as vocab-INT4		~1/16×

Table 1: The submitted portfolio. Disk and VRAM at model load in GiB; speed relative to bf16 at batch size 8. Quality is reported in Section 5.1 at sentence level and Section 5.2 at paragraph granularity. *Measured VRAM minus the memory saved from pruned vocabulary rows; the other VRAM figures are measured at load.

2 Related Work

Quantization and structured pruning. GPTQ (Frantar et al., 2023) and AWQ (Lin et al., 2024) established that 4-bit weight quantization can be nearly lossless, and FP8 is natively accelerated on recent hardware. Our results confirm that both are essentially free for translation at the 12B scale, which makes quantization a shared baseline and not a distinguishing factor. Minitron (Muralidharan et al., 2024) couples structured pruning with distillation-based recovery and reports that width pruning outperforms depth pruning at matched size, but only after retraining. A related line of work argues that many transformer layers are redundant, whether measured by Block Influence (Men et al., 2025), by the ineffectiveness of deeper layers (Grosov et al., 2024), or implicitly by early-exit methods (Elhoushi et al., 2024; Schuster et al., 2022). Recent work argues that redundancy is not an intrinsic structural property but is conditioned on the objective under which it is measured (Kim et al., 2026). Our negative result adds evaluation granularity to that picture, because per-layer redundancy measured on sentences did not justify depth removal at paragraph scale.

A growing number of works vary model capacity across depth without agreeing on the profile (Mehta et al., 2021; Ikeda et al., 2025; Bayat et al., 2026; Wu et al., 2026), though always during pretraining, whereas in a frozen checkpoint the allocation is already fixed and has to be recovered by measurement.

MBR decoding. We use consensus MBR (Eikema and Aziz, 2020) with a chrF utility (Popović, 2015) and sampling in the spirit of Freitag et al. (2023). Our contribution is not the decoder itself but a register check specific to Egyptian Arabic and a teacher comparison that isolates whether the gain interacts with compression.

3 Systems

Vision stripping. The released gemma-3-12b-it is multimodal. Its SigLIP tower (Zhai et al., 2023) and projector contribute nothing to translation and complicate quantizer tooling, so we extract a text-only Gemma3ForCausalLM, saving 0.8 GiB, and verify that the extracted model produces identical outputs on a held-out slice. Reductions in size quoted elsewhere in this paper are against the released multimodal checkpoint at 22.7 GiB, which is the artifact participants are given, and not against this text-only extraction.

Vocabulary pruning. Gemma’s 262k-entry vocabulary is dominated by scripts never used in our target languages. We prune the embedding and lm_head rows to a script-based vocabulary subset defined by Unicode script (Latin, Han, and Arabic blocks, universal tokens, and whitespace), retaining roughly 196k rows. We found that validation criteria matter more than the pruning heuristic itself. An earlier candidate passed standard round-trip checks but dropped whitespace tokens introduced by the chat template, producing 21% empty outputs at inference. Consequently, our acceptance criterion requires passing templated prompt tokenization and maintaining end-to-end translation quality.

Quantization. We use GPTQ W4A16 via llmcompressor in the compressed-tensors format (group size 128, static actorder, and lm_head excluded), served using vLLM’s Marlin kernels (Kwon et al., 2023). We calibrate on 196 prompt-formatted en→zh source sentences from an earlier WMT edition disjoint from our evaluation splits. As shown in Section 5.1, calibrating solely on this direction generalizes across all three language pairs. We also evaluate dynamic FP8 quantization as an intermediate quality–throughput baseline. We retain both the pruned and full-vocabulary INT4 checkpoints, assigning primary configurations per direction based on the findings in Section 5.2.

Self-MBR contrastive. For en→ar, we submit an additional contrastive configuration: the vocab-INT4 checkpoint paired with minimum Bayes risk (MBR) consensus decoding, enabled via an environment flag. Decoding generates 16 candidate samples using min- p tail pruning ($p = 0.02$, temperature 1.0, fixed seed) and ranks them using chrF utility. This setup requires no additional parameters but increases generation cost roughly $16\times$. If invoked on any other language pair, the serving wrapper safely falls back to standard greedy decoding with a logged warning.

4 Experimental Setup

Data. We use constrained-track data via the mtda recipes together with past WMT test sets for calibration, and the WMT25 general-MT sets prepared in the task repository for development and validation. One property of that data shaped most of our decisions. Its source lines are paragraphs and not sentences, and the generation script receives them directly, so translation is evaluated a paragraph at a time. Our early development used sentence-split data, which is the origin of the result in Section 6. All figures here are computed on the task’s segmentation unless labelled otherwise. On en→ar the two segmentations coincide, i.e. the same 960 segments serve as both the sentence-level and the paragraph-level set, whereas cs→de and en→zh reduce to 457 and 332 paragraphs respectively. The en→ar direction is evaluated on Egyptian Arabic, for which no constrained training data exists, so we treat it as evaluation-only throughout. Healing on Modern Standard Arabic would pull the model away from the Egyptian register it already has, a risk we confirmed when a depth student healed on cs→de and en→zh collapsed en→ar COMET from roughly 0.69 to 0.54.

Metrics and significance. We report COMET (wmt22-comet-da, version 2.2.7; Rei et al., 2022) and chrF, and add MetricX-24 (metricx-24-hybrid-xl; Juraska et al., 2024) error predictions for the comparison in Section 6. Differences in chrF are reported as raw point differences. For COMET, we report paired bootstrap resamples over segment-level differences with 95% confidence intervals (10,000 resamples for newly computed comparisons; 1000 resamples for earlier runs, where the minimum attainable two-sided p -value is reported as $p < 0.002$). Because failing to reject a null hypothesis does not establish equiv-

alence, claims of “near-lossless” performance are backed by the margin excluded by the confidence interval. We treat $\Delta\text{COMET} \leq 0.01$ as practically negligible; evaluating a 95% interval against this threshold is more conservative than a standard two one-sided test (TOST) at $\alpha = 0.05$ (which uses a 90% interval). This margin was defined post-hoc. Finally, COMET relies on XLM-R (Conneau et al., 2020), which truncates inputs exceeding 512 subword tokens. This truncation affects 1.1–1.3% of cs→de paragraphs and one of the 960 en→ar segments (counting any instance where source, reference, or hypothesis overflows). Computing chrF over untruncated outputs provides an essential guardrail: a steep drop in chrF alongside steady COMET scores indicates length-related degradation hidden by prefix evaluation, which revealed the truncation issue in Section 6. All cross-system comparisons are evaluated within a single shared harness to avoid cross-environment discrepancies.

Hardware and checks. Development and healing ran on a $4\times$ RTX PRO 6000 (Blackwell) GPU node and a GB10 Grace-Blackwell system, whereas official speed and VRAM benchmarks were reproduced on an NVIDIA H100 to match the evaluation platform. Because absolute COMET scores can exhibit slight cross-hardware variance, we measured this effect empirically. Re-decoding an INT4 checkpoint across separate machines reproduced system-level COMET within 0.0003 on en→zh and 0.0001 on cs→de; the reproduction in Section 5.3 agreed within 0.0002 on en→ar, an order of magnitude below our threshold for a decisive difference. We nonetheless pair every direct comparison within a single machine and execution environment. Finally, throughput is measured using <end_of_turn> as the stop token; this differs from in this architecture, and confusing the two inflates throughput by roughly an order of magnitude.

5 Results

5.1 Quantization

Table 3 shows that INT4 W4A16 is near-lossless on en→zh and en→ar despite single-direction calibration. The cs→de deficit, at -0.003 , is the only one that separates from zero, and we do not describe that direction as near-lossless. This cross-lingual transfer suggests that the underlying activation distributions are largely model-intrinsic rather than

Step	Disk	VRAM	Speed	Δ COMET
bf16 original	22.7	—	—	—
+ vision removal	21.9	—	—	0
+ GPTQ W4A16	7.08	7.84	−2–4%	−0.0030
+ vocabulary prune	6.64	7.41	−2–4%	+0.0018
+ self-MBR	6.64	7.41	$\sim 1/16\times$	+0.0093
FP8 (alternative)	13.8	12.69	+16–18%	−0.001

Table 2: Incremental effect of each step, each row measured against the row above. Disk and VRAM in GiB, speed relative to bf16 at batch size 8. The quality column is not additive, because each delta comes from the comparison that certifies its own step, sentence-level cs→de for quantization and FP8 (Table 3), the cs→de re-certification of Section 5.2 for the vocabulary prune, and en→ar for self-MBR (Table 5). Vision removal is verified by identical outputs and not by a score.

Dir.	Variant	COMET	Δ
cs→de	bf16	0.847	—
	FP8	0.846	−0.001 (ns)
	INT4	0.844	−0.003 [†]
en→zh	bf16	0.849	—
	FP8	0.848	−0.0005 (ns)
	INT4	0.847	−0.002 (ns)
en→ar	bf16	0.689	—
	FP8	0.688	−0.002 (ns)
	INT4	0.693	+0.003 (ns)

Table 3: Sentence-level quantization quality at full evaluation size (cs→de $n=1996$, en→zh 2065, en→ar 960), with en→zh-only calibration. On en→ar this set is also the task’s own paragraph segmentation (Section 4). “ns” denotes a 95% interval containing zero. Deltas are computed at full precision, and the two numeric columns are rounded independently. [†]Nominally significant, practically negligible at the margin of Section 4, and uncorrected across the family of six quantization comparisons reported here, which a Holm correction would require.

language-specific.

Table 3 shows that INT4 W4A16 is near-lossless on en→zh and en→ar despite single-direction calibration. The cs→de deficit, at −0.003, is the only one that separates from zero. The difference is well inside the margin of Section 4, so we treat it as practically negligible; however, because a nonzero deficit is statistically resolved rather than absent, we do not call that direction near-lossless. Our intuition is that the relevant activation ranges are largely model-intrinsic, so that they vary little across directions. On the H100, FP8 dominates bf16 on both ranked axes, using 44% less memory while delivering 16–18% higher throughput on native FP8 tensor cores, at a quality cost whose intervals include zero. INT4 achieves the best quality–size point with 65%

less memory, at a 2–4% batched-throughput cost from Marlin dequantization. A mixed-precision variant protecting the highest Block Influence layers (BI; Men et al., 2025) is dominated by plain INT4. The FP8 result holds at paragraph granularity as well, where no interval separates it from bf16 on any direction (Table 4).

5.2 Vocabulary pruning

The pruning of the 262k vocabulary entries to roughly 196k using Unicode script removes 0.44GiB from the INT4 checkpoint, reducing it from 7.08 to 6.64GiB. This reduction is commonly assumed to be zero-cost for retained scripts, since the discarded rows are never addressed during generation. We certified this pruning against the full-vocabulary INT4 baseline, decoding both variants on the same machine within a single session and verifying weight checksums against the released model repositories.

On this comparison the assumption holds on all three directions. cs→de gives +0.0018 COMET (95% CI [−0.0028, +0.0060]), en→zh +0.0000 ([−0.0023, +0.0023]), and en→ar −0.0038 ([−0.0093, +0.0020]). None of the three separates from zero, with en→ar exhibiting the widest interval. All three intervals fall inside the margin of 0.01 COMET stated in Section 4, and this is what we mean by near-lossless (though under a stricter 0.005 margin, only en→zh would qualify). The en→ar interval is wide in absolute terms. Its lower bound admits a prune cost on en→ar of 0.0093 COMET, which is the size of the MBR gain we report for the same direction in Section 5.3, so we interpret the en→ar result strictly as a failure to detect a cost rather than proof that none exists.

However, this is not the measurement on which we based our filing. Before the deadline we compared the pruned checkpoint against a different full-vocabulary INT4 build and obtained a small but significant cost on en→zh, of −0.0026 COMET (95% CI [−0.0051, −0.0002], $p=0.026$), which is why our filing assigns primary models per direction. The two builds share a quantization recipe but they are not the same weights. They agree on 3.2% of their outputs on the en→ar set, which is the largest set both builds decoded, and the other build scores 0.0029 COMET higher on en→zh paragraphs, which accounts for the whole of the apparent prune cost. It is not hardware variation, given that the same checkpoint re-decoded on a second machine reproduces to within 0.0001–0.0003

COMET, an order of magnitude below the effect. A comparison of this kind is internally valid and nothing in a per-segment score exposes it, so the check we recommend is to checksum the weights that produced a reported number against the artifact that was released, which is the only check that a plausible directory name cannot satisfy. The consequence for the systems is limited, in that one primary would have served all three directions, so our filing is conservative and we keep the per-direction split as filed.

We also certified the submitted checkpoints against the uncompressed model at the paragraph granularity on which the task scores (Table 4). On $cs \rightarrow de$ (WMT25, 457 documents) the bf16 model reaches 0.8005 COMET and vocab-INT4 reaches 0.7957 at 6.64 GiB, a deficit of -0.0048 (95% CI $[-0.0099, -0.0004]$, $p=0.030$) for a reduction of 71% in disk footprint against a bf16 original of 22.7 GiB. This deficit is smaller than the margin of Section 4 but separates from zero, so we report it as a measured cost and not as a lossless result. No other direction shows a measurable cost. The same checkpoint gives -0.0019 on $en \rightarrow zh$ and -0.0030 on $en \rightarrow ar$, and neither interval excludes zero. The full-vocabulary INT4 checkpoint behaves the same way, namely -0.0045 on $cs \rightarrow de$ and nothing measurable elsewhere, so the cost of this compression stack at paragraph granularity is confined to one direction and the vocabulary prune adds nothing to it. FP8 is not separable from bf16 on any direction.

All three directions exhibit isolated degenerate outputs, in which the model loops a single fragment until exhausting the generation budget. We counted them across the four arms of Table 4 on $en \rightarrow ar$, which is where they concentrate. Five distinct segments are affected, two under bf16, two under FP8 and two under vocab-INT4, with one segment shared between bf16 and FP8 and none under full-vocabulary INT4. The sources inputs are short (15–64 tokens), and each failure runs to the 1024-token ceiling. Because this behaviour appears in the uncompressed model and is absent from full-vocabulary INT4, we do not attribute it to compression. However, it noticeably influences corpus-level scores: the two INT4 arms score 0.5 chrF higher than bf16 on $en \rightarrow ar$ despite scoring at or below it on COMET, a divergence directly driven by loop counts. Each looping instance penalizes the document by 0.08 to 0.26 chrF, even though lexical surface forms do not distinguish an intended turn termination from an exhausted token

budget.

5.3 MBR decoding

Table 5 sets self-MBR against the greedy baseline. Sampling sixteen candidates and returning the chrF-consensus one improves exactly one direction, the weakest of the three.

MBR marginalizes over the model’s posterior, so a gain is expected where the posterior is rough and greedy decoding leaves headroom, which describes $en \rightarrow ar$ at 0.68 while the other two directions are saturated. The gain is not free at inference. It costs sixteen decodes of the source plus the chrF utility computation over the candidate set, and we file it as a contrastive quality option and not as a point we expect to be competitive on the quality–speed frontier. We carried out two pre-registered checks in order to delimit this claim.

The first concerns register. A chrF-consensus decoder could raise COMET on Egyptian Arabic simply by drifting output toward Modern Standard Arabic, the higher-mass register that the metric’s training data favors. Measured on the arms of Table 5, on per-segment counts of Egyptian and MSA markers, the mean drift is -0.080 , i.e. the MBR output is in aggregate more Egyptian than the greedy output, and not less. The per-segment picture is less clean, in that the gain co-varies weakly but significantly with movement toward MSA (Spearman $+0.113$, $p=0.0005$), which accounts for a little more than one percent of the variance. We stop short of claiming that the gain is register-neutral, and claim only that the decoder does not move the output toward MSA in aggregate and that the component of the gain which tracks register movement is small. A marker count is a coarse instrument for this question. It is insensitive to dialectal syntax and to lexical choices that no marker list contains. In future work, we plan to repeat the check with a trained dialect classifier and with judgements from speakers of Egyptian Arabic.

The second concerns compression. If the compressed model had a damaged posterior, MBR might be repairing compression damage instead of exploiting general headroom. In a bf16-teacher 2×2 with a pre-registered interaction contrast, decoded in a separate run from Table 5 and at a budget of 1024 tokens, MBR improves the teacher by $+0.0059$ ($p=0.002$) and the compressed model by $+0.0075$ ($p<0.002$). The interaction is $+0.0017$ with a 95% CI of $[-0.0035, +0.0070]$ ($p=0.53$), and we base our reading on that interval, which

Direction	Variant	COMET	chrF	Δ COMET vs bf16 (95% CI)
cs→de ($n=457$)	bf16	0.8005	63.1	—
	FP8	0.8014	63.0	+0.0009 [−0.0010, +0.0028]
	INT4	0.7959	62.5	−0.0045 [−0.0089, −0.0008]
	vocab-INT4	0.7957	62.6	−0.0048 [−0.0099, −0.0004]
en→zh ($n=332$)	bf16	0.8198	28.2	—
	FP8	0.8192	28.1	−0.0005 [−0.0018, +0.0008]
	INT4	0.8204	27.5	+0.0007 [−0.0019, +0.0035]
	vocab-INT4	0.8178	27.5	−0.0019 [−0.0044, +0.0005]
en→ar ($n=960$)	bf16	0.6879	34.7	—
	FP8	0.6859	34.2	−0.0020 [−0.0056, +0.0015]
	INT4	0.6878	35.2	−0.0002 [−0.0056, +0.0053]
	vocab-INT4	0.6849	35.2	−0.0030 [−0.0084, +0.0023]

Table 4: Quality at the paragraph granularity on which the task scores, against the uncompressed model. All arms were decoded through one entry point at greedy decoding with a budget of 1024 tokens, and scored in one pass. Intervals are paired bootstraps over per-segment differences with 10 000 resamples. Only the two cs→de quantized arms separate from zero, and every interval in the table lies inside the margin of Section 4. Absolute scores are not comparable with those of Table 3, which comes from a different harness. This holds for en→ar too, although both tables score the same 960 segments.

Dir.	greedy	MBR	Δ (95% CI)
cs→de	0.8004	0.7990	−0.0014 (ns)
en→zh	0.8149	0.8152	+0.0003 (ns)
en→ar	0.6795	0.6889	+0.0093 [.0057, .0132]

Table 5: Self-MBR ($N = 16$, chrF consensus) against greedy decoding on the same compressed checkpoint. COMET on the development sets, which are also the sets on which we selected the decoder, so these are development-set figures and not estimates of out-of-sample performance. Intervals are paired bootstraps with 10 000 resamples.

is the pre-registered contrast. MBR behaves as a general decode-time technique, with the caveat that $n=960$ bounds a small interaction without excluding it. This is consistent with, though not proof of, INT4 plus vocabulary pruning having preserved the posterior. The run in which both checks were first carried out is not archived. We recomputed the register check on the archived arms of Table 5, which is what produced the weaker statement above, whereas the teacher cells of the 2×2 exist nowhere else, so we report those figures unchanged and note that they have not had the same scrutiny.

One component of the gain is mechanical. The greedy arm degenerates into a repetition loop on two of the 960 segments and the consensus decoder discards the looping candidate in both. Removing them lowers the gain to +0.0085 (95% CI [+0.0048, +0.0123]), so 8.6% of the gain is rescue and the remainder is selection among well-formed

candidates. We recomputed the +0.0093 on separate hardware and obtained the same system scores to within 0.0002. Both come from the development set that selected the method, so the official test set is the first out-of-sample measurement.

6 Depth Pruning

We include this axis because its failure mode is specific to the evaluation granularity that this task uses, and it is therefore easy to miss.

Sentence level. A probe using the Block Influence metric of Men et al. (2025), and not either of the pruning methods proposed there, identifies a contiguous band of low-importance layers, layers 15–26. The removal of these twelve layers without healing collapses the model to COMET 0.20, but sequence-level knowledge distillation (Kim and Rush, 2016) on roughly 300k teacher-generated pairs recovers it to within −0.002 and −0.008 COMET of the baseline on en→zh and cs→de sentences respectively, and a contiguous cut outperforms a distributed one at the same budget. None of these conclusions holds after the move to evaluation at the paragraph level, where the healed student plateaus about 0.03 COMET and 6 chrF below full depth and terminates paragraphs early. Appendix B records the configuration of the pruning and of the healing runs.

Clipped targets teach the student to stop early.

An audit of the distillation corpora, pre-registering what each outcome would mean, found that a default of 256 tokens in the generation script had trun-

cated the teacher outputs that were used as training targets. A teacher output that is cut at the cap is still wrapped by the chat template, which appends a supervised end-of-turn token, so the cap does not withhold supervision at the tail but supplies incorrect supervision. Generating with 1024 tokens available, the student stops at exactly 256 tokens on 58 of the 457 task paragraphs, and several of those outputs end in the middle of a word, for instance after *Wahlwer* or *Hochsch*.

The damage tracks the length support of the supervision and not the proportion of the targets that are clipped. A corpus with 2.8% of its targets clipped produced the same limit of 256 tokens as one with 62%, which is explained by the fact that neither contained any target beyond roughly 260 tokens, against a 99th percentile of 502 in the reference. These lengths, and the cap itself, are counted in the model’s own tokenizer, and not in the tokenizer of the evaluation metric discussed in Section 4. An assertion on length at load time does not detect this, because a clipped target is short, carries a valid end-of-turn token, and is indistinguishable from a genuinely short one. Only the finish reason recorded at generation time separates the two cases, and we now record it at every stage that produces text.

Re-healing on verified supervision. We regenerated the paragraph corpus with verified length support and re-healed the same un-healed student under an otherwise identical recipe, so that the corpus was the only change to the supervision. The learned stop at 256 tokens disappears entirely, and quality improves significantly over the heal on the truncated corpus (+0.0072 COMET, 95% CI [+0.0029, +0.0115], $p=0.002$; +1.93 chrF). The earlier heal also left the embeddings untied, so we repeated the comparison against a tied export of the same student and obtained +0.0059 COMET (95% CI [+0.0017, +0.0100]), the tying itself not separable from zero. The largest part of the deficit against full depth nevertheless remains, at -0.0215 COMET (95% CI $[-0.0286, -0.0145]$, $p<0.002$) and -4.25 chrF, so against a total gap of 0.0287 COMET and 6.18 chrF the re-heal recovers 25% by COMET and 31% by chrF. MetricX-24 reproduces the decomposition on the same data at a recovered share of 23%. Part of the remaining deficit is not capacity. The re-healed student over-generates on five of 457 paragraphs where the full-depth control does not, and our audit attributes roughly 0.004

COMET of the deficit to that behaviour. The training loss plateaued during the re-heal, which weakens an explanation based on the budget of the heal, while a ceiling imposed by the distillation recipe itself remains untested. We therefore describe the remaining deficit as consistent with the architecture and not as confirmed capacity loss, and this axis stays out of the portfolio. In future work, we plan to examine MLP width at matched capacity with all layers retained.

7 Conclusion

In this paper we described the ILSP/ARC submission to the WMT26 Model Compression task. Vision-tower removal, script-based vocabulary pruning and GPTQ W4A16 together produce models 69–71% smaller than the original. No interval separates the vocabulary prune, which contributes the last 0.44 GiB, from the full-vocabulary INT4 primary on any direction, and at paragraph granularity the submitted checkpoint costs 0.0048 COMET against bf16 on cs→de, which is the only direction on which any interval separates from zero. An FP8 contrastive system covers the quality–speed point, and a self-MBR contrastive system adds a quality gain on the weakest direction at no parameter cost.

Two of our findings transfer beyond this submission. The first concerns the supervision of a distillation. Any recipe that generates its own targets under a token cap and wraps them with a chat template will teach the student to end a turn at the cap, and the injury scales with the longest target the corpus contains rather than with the fraction that is clipped, so a corpus with 2.8% of its targets affected fares as badly as one with 62%. The finish reason recorded at generation time is the one signal that identifies this, and we recommend recording it at every stage of a pipeline that produces text. The second is that a compression result is a statement about two specific artifacts, and their identity is what our own records captured least well. The apparent prune cost on one direction, which is why we filed two primary models instead of one, came from comparing against a build that shared the recipe but not the weights. In future work, we plan to compare width profiles at matched capacity.

Limitations

All of our development sets also informed the selection of methods, so the official results reported in the findings of the task (Gowda et al., 2026) are

our first measurement out of sample. This holds for the $+0.0093$ attached to the $\text{en} \rightarrow \text{ar}$ MBR contrastive system as well, since the same 960 segments both selected the method and produced the figure. Our equivalence statements are read from 95% intervals against a margin fixed after the fact, and the families of related tests we report are uncorrected. The assignment of primary models per direction was filed on a single uncorrected test ($\text{en} \rightarrow \text{zh}$, $p=0.026$) that a re-certification against the submitted checkpoint did not reproduce (Section 5.2); the corrected measurement resolves no difference on any direction, so the split we filed is conservative and the systems remain as submitted. COMET is the least reliable on Egyptian Arabic of the three directions, and our register check counts marker words, which is weaker evidence than a trained dialect classifier would give, and the weak correlation between segment-level gains and MSA markers remains associated with the metric’s intrinsic register bias. The task also evaluates with an LLM as a judge, for which we have no local proxy, so a failure that COMET and chrF tolerate may be penalised there. Our claims across hardware cover Blackwell and Hopper only.

Acknowledgments

This work has been supported by Pharos, the Greek AI Factory for accelerating AI innovation, funded by the European High Performance Computing Joint Undertaking (EuroHPC JU) and participating states under Horizon Europe, grant agreement No. 101234269.

References

- Reza Bayat, Ali Behrouz, and Aaron Courville. 2026. [Tapered language models](#). *Preprint*, arXiv:2606.23670.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- 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 (COLING)*.
- Mostafa Elhoushi, Akshat Shrivastava, Diana Liskovich, Basil Hosmer, Bram Wasti, Liangzhen Lai, Anas Mahmoud, Bilge Acun, Saurabh Agarwal, Ahmed Roman, Ahmed Aly, Beidi Chen, and Carole-Jean Wu. 2024. [LayerSkip: Enabling early exit inference and self-speculative decoding](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoeftler, and Dan Alistarh. 2023. [GPTQ: Accurate post-training quantization for generative pre-trained transformers](#). In *The Eleventh International Conference on Learning Representations (ICLR)*.
- Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. 2023. [Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation](#). *Findings of the Association for Computational Linguistics: EMNLP*.
- Gemma Team. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- Thamme Gowda, Marco Gaido, Roman Grundkiewicz, and Matteo Negri. 2026. Findings of the WMT 2026 shared task on model compression: No free lunch at extreme compression. In *Proceedings of the Eleventh Conference on Machine Translation (WMT 2026)*, Budapest, Hungary.
- Andrey Gromov, Kushal Tirumala, Hassan Shapourian, Paolo Glorioso, and Daniel A. Roberts. 2024. [The unreasonable ineffectiveness of the deeper layers](#). *arXiv preprint arXiv:2403.17887*.
- Wataru Ikeda, Kazuki Yano, Ryosuke Takahashi, Jaesung Lee, Keigo Shibata, and Jun Suzuki. 2025. [Layerwise importance analysis of feed-forward networks in transformer-based language models](#). In *Proceedings of the Conference on Language Modeling (COLM)*.
- 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.
- Minkyu Kim, Vincent-Daniel Yun, Youngra Kim, Suin Cho, Woosang Lim, and Sunwoo Lee. 2026. [Re-thinking layer redundancy: Calibration matters more than search in LLM depth pruning](#). *arXiv preprint arXiv:2604.24938*.
- Yoon Kim and Alexander M. Rush. 2016. [Sequence-level knowledge distillation](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- 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)*, pages 611–626.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. [AWQ: Activation-aware weight quantization for LLM compression and acceleration](#). In *Proceedings of Machine Learning and Systems (MLSys)*.

Sachin Mehta, Marjan Ghazvininejad, Srinivasan Iyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2021. [DeLighT: Deep and light-weight transformer](#). In *International Conference on Learning Representations (ICLR)*.

Xin Men, Mingyu Xu, Qingyu Zhang, Qianhao Yuan, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. 2025. [ShortGPT: Layers in large language models are more redundant than you expect](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20192–20204, Vienna, Austria. Association for Computational Linguistics.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. [Compact language models via pruning and knowledge distillation](#). *Advances in Neural Information Processing Systems (NeurIPS)*.

Maja Popović. 2015. [chrF: character n-gram F-score for automatic MT evaluation](#). In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*.

Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C. Farinha, Taisiya Glushkova, Alon Lavie, Luísa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*.

Tal Schuster, Adam Fisch, Jai Gupta, Mostafa Dehghani, Dara Bahri, Vinh Q. Tran, Yi Tay, and Donald Metzler. 2022. [Confident adaptive language modeling](#). *Advances in Neural Information Processing Systems (NeurIPS)*.

Zhaofeng Wu, Oliver Sieberling, Shawn Tan, Rameswar Panda, Yury Polyanskiy, and Yoon Kim. 2026. [Variable-width transformers](#). *CoRR*, abs/2606.18246.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. [Sigmoid loss for language image pre-training](#). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986.

A Submission Details

The submission codebase is at <https://github.com/soksof/wmt26-model-compression> (tag `wmt26-code-submission`). Model weights are on the Hugging Face Hub as `soksof/gemma-3-12b-wmt26-{vocaball-int4,`

`int4, fp8}`. Two form entries share this codebase, entry A with primary `vocaball-int4` on `cs→de` and `en→ar`, and entry B with primary `int4` on `en→zh`. The package named `vocaball-int4` is the system this paper calls `vocab-INT4`, and `int4` is the full-vocabulary `INT4` system. The `MBR` package reuses the `vocaball-int4` weights with consensus decoding ($N=16$) enabled via `submission.env`.

B Depth Pruning and Distillation Configuration

This appendix records the configuration of the runs discussed in Section 6, recovered from the archived launch scripts, training states and model configurations. Fields the archive does not record are marked as unrecorded.

Pruning. The student removes twelve contiguous layers, 15–26, from the 48 layers of the text-only checkpoint, leaving 36. Hidden size (3840), intermediate size (15360), vocabulary (262208 entries) and maximum position (131072) are unchanged from the parent.

Teacher generation. Targets were generated by `scripts/gen_teacher_kd.sh`, which shards the source across seven GPUs and calls `eval/translate.py` on `models/gemma-3-12b-text`, i.e. the vision-stripped `bf16` model of Section 3, at greedy decoding and a memory utilization of 0.9. The script passes no `--max-new-tokens`, so the default of 256 in `eval/translate.py` applied. This is the cap discussed in Section 6.

Corpora. Table 6 reports the corpora we generated. The sentence-level pair holds 150k targets per direction, i.e. the roughly 300k of Section 6, and no corpus was generated for `en→ar`, which we treat as evaluation-only. We count a target as clipped when it falls in the window 254–256 tokens or just beyond it, measured in the model’s own tokenizer. On the sentence-level corpora this reaches 0.01% and 0.03%, which is why the heal that used them was not materially affected. It reaches 2.8% on the `cs→de` half of the paragraph corpus and 62.3% on the `cs→de` half of the eight-sentence corpus, and these are the two corpora compared in Section 6. No corpus contains a target beyond 260 tokens, and prompt and target together never approach the 2048-token maximum length of the

Corpus	Dir.	n	clipped	max
sentence	cs→de	150000	0.01%	256
	en→zh	150000	0.03%	257
paragraph	cs→de	37500	2.8%	258
	en→zh	37499	0.5%	258
8-sentence	cs→de	18750	62.3%	260
	en→zh	18750	13.0%	258
document	cs→de	2874	2.1%	257
doc. shuf.	cs→de	2874	0.6%	257

Table 6: Distillation corpora, with the share of targets clipped at the 256-token generation cap and the longest target in each, both counted in the model’s own tokenizer.

loader, the longest total being 641 tokens. The corpus we regenerated with verified length support, on which the re-heal of Section 6 was trained, is not archived. It was produced after the audit above was run, on the machine that was decommissioned at the end of the submission period, and only the student survived the copy. We can nevertheless measure its effect on the student it produced. That student never stops at exactly 256 tokens on the 457 task paragraphs, where the student healed on the eight-sentence corpus does so 58 times.

Healing. All heals use AdamW ($\beta_1=0.9$, $\beta_2=0.999$, no weight decay), a peak learning rate of 1×10^{-5} under a cosine schedule with a warmup ratio of 0.03, a per-device batch of 1 with 16 gradient accumulation steps, gradient clipping at 1.0, bf16, a maximum length of 2048 tokens, and seed 42. The paragraph heals run for three epochs and the earlier sentence-level heal for one. No evaluation was run during training and no best-checkpoint selection was performed, i.e. the final checkpoint of each run is the one we export and report. For the heal on the document corpus, for which a training state survives, this is 135 optimizer steps over 2874 examples in three epochs, which implies four data-parallel devices and an effective batch of 64, and over which the training loss falls from 0.44 to 0.012. The training arguments of the two students compared in Section 6 are not themselves archived; the settings above are read from sibling runs launched the same day under the same recipe.

Students. The heal on the truncated corpus was exported with untied embeddings in fp32 (10.08B parameters), and both the tied re-export of that student and the re-heal on the verified corpus are tied and bf16 (9.08B parameters), and the differ-

ence of 1.007B parameters is the untied lm_head. The quantized depth student uses the same GPTQ recipe as our submitted systems, namely W4A16 over all linear layers with lm_head excluded, block size 128, dampening 0.01 and static actorder.