

WMT26 Model Compression Task: Quantising and Distilling TildeOpen

Martins Kronis

Toms Bergmanis

Ingus Jānis Pretkalniņš*

Mārcis Pinnis

Tilde

Vienības gatve 75a, Riga, LV-1004, Latvia

firstname.lastname@tilde.lv

Abstract

This paper describes Tilde’s submission to the unconstrained track of the WMT 2026 Model Compression task in the Czech-German direction. Our baseline is TildeOpen-15B-64k, a model distilled from the TildeOpen-30B-64k European foundation model. We adapt it for CS-DE translation, using supervised fine-tuning followed by GRPO-based reinforcement learning. From this baseline, we explore two axes of compression: post-training quantisation and knowledge distillation. For quantisation, we use LLM Compressor to produce a GPTQ-calibrated model with 4-bit NVFP4 weights (10.1 GB, $\times 3$ smaller than the 30.3 GB footprint of the 15B baseline) and a data-free RTN model with FP8 weights and activations (16.3 GB). In parallel, we prune and distil the base 15B model into an 8B student and fine-tune it with the same supervised fine-tuning recipe. Applying the two quantisation schemes to the 8B model yields our smallest models at 8.9 GB and 5.7 GB. The latter is over $\times 5$ more compact and $\times 1.8$ faster than the 15B baseline. Evaluation on four test sets across a variety of metrics shows that quantisation and distillation have minimal impact on translation quality. The NVFP4-quantised 15B model offers the best quality/size/speed trade-off and is our primary submission for this competition. All six models are publicly released.

1 Introduction

This system description paper details Tilde’s submission to the WMT 2026 Model Compression task¹, which challenges participants to reduce the deployment cost of a large language model—in terms of size and inference speed—while preserving translation quality.

We take part in the unconstrained track and participate only in the Czech-German direction. We

pick this direction because all our models are ultimately based on TildeOpen-30B (Bergmanis et al., 2026), a multilingual European foundation model developed by Tilde with the explicit goal of equitable performance across all European languages. Building the submission on the TildeOpen family lets us leverage first-hand knowledge of the model internals and our prior experience deriving production MT systems² from it, and CS-DE is a language pair where this European focus translates into a strong, competitive baseline.

Our submission explores two complementary compression axes and their combination: (i) post-training quantisation to FP8 and NVFP4 using LLM Compressor³, and (ii) structured pruning followed by knowledge distillation into an 8B student, after which the same quantisation schemes are applied to the distilled model. This results in six models ranging from 30.3GB on disk down to only 5.7GB: a 15B baseline, its two quantised versions, an 8B distilled model, and its two quantised versions.

The 15B baseline is itself a product of distillation: it was obtained by pruning and distilling TildeOpen-30B-64k into TildeOpen-15B-64k. However, since the unconstrained track requires the original model to be below 20B parameters, TildeOpen-30B cannot serve as the starting point of a submission, and the 30B $\rightarrow$ 15B distillation consequently falls outside the scope of the shared task. The resulting TildeOpen-15B-64k, on the other hand, is eligible and thus serves as our starting point. We nevertheless describe the 30B $\rightarrow$ 15B distillation procedure in detail, both as a necessary context for the baseline and because the same pruning and distillation recipe is used to derive the 8B models included in this submission. All large-

*Work carried out while at Tilde

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

²<https://tilde.ai/translate/tilde-machine-translation/>

³<https://github.com/vllm-project/llm-compressor>

scale pruning, distillation and fine-tuning runs were performed on the JUPITER supercomputer⁴ at the Jülich Supercomputing Centre.

All six submission models are publicly available in a dedicated HuggingFace collection⁵, and the submission repo with inference code is available on GitHub⁶.

2 Baseline Model

Our starting point is TildeOpen-15B-64k, distilled from TildeOpen-30B-64k⁷ – the 64k-context version of TildeOpen-30B⁸ (Bergmanis et al., 2026). This section covers the pruning and distillation used to obtain the 15B model, followed by its adaptation to the CS-DE direction via supervised fine-tuning and reinforcement learning. The resulting model is our baseline for the WMT 2026 compression task.

2.1 Model Pruning

We closely follow NVIDIA’s work on model pruning and knowledge distillation (Muralidharan et al., 2024; Sreenivas et al., 2024). We first perform TildeOpen-30B-64k $\rightarrow$ TildeOpen-15B-64k pruning by following the Minitron-8B recipe (Muralidharan et al., 2024) and using a custom HPC fork of the NVIDIA-NeMo framework⁹. We use a small calibration set of 1024 examples, sampled from the context extension phase of TildeOpen 30B-64k. The calibration data uniformly covers all the target European languages and consists of a mixture of TildeOpen-30B base training data (Bergmanis et al., 2026), including code and parallel data, FinePDFs¹⁰ and synthetic data based on the KNOT approach (Tian et al., 2025).

We decided to prune width only and shrink the embedding dimension and the FFN intermediate size to 2/3 of the original, i.e., $6,144 \rightarrow 4,096$ and $21,504 \rightarrow 14,336$, respectively. We performed no further architecture search. For embedding channels, we apply pruning to MLP, MHA and Layer-Norm layers. For neuron pruning (FFN intermedi-

ate size), only the MLP layer is trimmed. This reduces the model size from 30.678B to 15.167B. In line with the Minitron-8B recipe, we use activation-based importance estimation, calculated only during the forward pass. Pruning was performed on the 4 NVIDIA GH200 GPUs on the Jupiter HPC.

2.2 Knowledge Distillation

After pruning the model, we employ knowledge distillation with the original 30B-64k model as the teacher, as knowledge distillation has been shown to yield much better downstream results than simple retraining, even when accounting for the compute overhead of knowledge distillation in a compute-controlled setting (Muralidharan et al., 2024; Sreenivas et al., 2024). For knowledge distillation, we use training data from the TildeOpen-30B-64k context extension phase. Therefore, we do not need to employ teacher correction. The original Minitron recipe uses 100B tokens to recover an 8B pruned student from a 15B teacher (Muralidharan et al., 2024). Because we work at roughly $\times 2$ the scale—recovering a 15B-pruned student from a 30B teacher—we aim for at least $\times 2$ the recovery data and settle for 300B tokens, since, contrary to Minitron, our models are heavily multilingual. This results in $\times 1.5$ upsampling of the previously mentioned 200B tokens from the context extension phase data.

We train the student model on the teacher’s logits using only the Kullback–Leibler (KL) divergence loss (Hinton et al., 2015) and no hard-label cross entropy. We use learning rate cosine annealing (Loshchilov and Hutter, 2017), with 200-step linear warmup for a total of 18,000 steps. Training was performed on 1024 Nvidia GH200’s from Jupiter HPC.

The obtained model is then compared against the original 30B-64k baseline on English-only versions of ARC (Clark et al., 2018), MMLU (Hendrycks et al., 2020), and HellaSwag (Zellers et al., 2019). Although these datasets are Anglo-centric and therefore not a good proxy for the multilingual capabilities of the TildeOpen model family, we nonetheless chose them to enable a direct comparison with Minitron-8B’s reported performance degradation. The results are shown in Table 1 and were obtained using the Language Model Evaluation Harness¹¹ (Gao et al., 2024).

⁴<https://www.fz-juelich.de/en/jupiter>

⁵<https://huggingface.co/collections/TildeAI/wmt26-compression-task>

⁶<https://github.com/tilde-nlp/wmt26-compression>

⁷<https://huggingface.co/TildeAI/TildeOpen-30b-64k>

⁸<https://huggingface.co/TildeAI/TildeOpen-30b>

⁹<https://github.com/nvidia-nemo>

¹⁰<https://huggingface.co/spaces/HuggingFaceFW/FinePDFsBlog>

¹¹<https://github.com/EleutherAI/lm-evaluation-harness>

	TildeOpen-30B	TildeOpen-15B
ARC-C (25)	0.596	0.525
MMLU (5)	0.592	0.540
HellaSwag (10)	0.829	0.787

Table 1: Performance comparison of TildeOpen-30B and TildeOpen-15B on English-only benchmarks. Reported value is *acc_norm* (*acc* for MMLU), and the number in parentheses indicates the fewshot amount.

The results show a 5-12% degradation across the English benchmarks w.r.t. the original model when distilling a multilingual model supporting 34 languages. This is comparable to the 4-10% degradation reported for the English-only Minitron-8B (Muralidharan et al., 2024) and indicates that the distillation was successful.

To further validate the success of distillation, we perform a downstream evaluation by fine-tuning TildeOpen-30B and TildeOpen-15B for the translation task. We use the WMT24++ (Deutsch et al., 2025) evaluation set to evaluate the results on select European translation pairs, mirroring the evaluation in our previous work (Bergmanis et al., 2026). Both models are trained using the same training recipe on the same closed-source dataset. We report average COMET¹² (Rei et al., 2022) score for each model–source language combination in Table 2, where the target group (XX) consists of the following languages: BG, CS, DA, DE, EN, ES, ET, FI, FR, HR, HU, IS, IT, LT, LV, NL, NO, PL, PT, RO, RU, SK, SL, SR, SV, and TR.

WMT24++	TildeOpen-30B	TildeOpen-15B
EN-XX	0.839	0.836
DE-XX	0.831	0.829
FR-XX	0.825	0.822
PL-XX	0.827	0.824
UK-XX	0.822	0.820
LT-XX	0.817	0.815

Table 2: Results of automatic translation quality evaluation between translation fine-tuned TildeOpen-30B and TildeOpen-15B, as measured by COMET on the WMT24++ dataset.

The results in Table 2 show COMET score differences of 0.002 to 0.003, indicating that the fine-tuned 15B model is on par with the 30B model with respect to the automatic metric and the WMT24++ dataset. These results further validate the distillation process.

¹²<https://huggingface.co/Unbabel/wmt22-comet-da>

2.3 Supervised Fine-tuning

Having successfully distilled the TildeOpen-15B foundational model, we adapt it for CS-to-DE translation.

For training data, we use a mixture of publicly available and proprietary in-house parallel corpora. To bolster the number of unique examples and potentially leverage language similarity in transfer learning, we complement the CS-DE language pair with other Germanic and Slavic languages. More specifically, we include Czech-XX and XX-German translation pairs, where XX denotes DA, NL, EN, DE, SV, BG, HR, CS, PL, SK, SL, and NO. The frequency distribution across all language pairs is close to uniform. In addition to the general translation task data, we add data covering more complementary tasks, such as translation with terminology glossaries, translation memories, and formatting tags. We add the extra tasks to increase the number of unique data points. All data types and languages include sentence and document-level data. Our final supervised fine-tuning (SFT) dataset consists of 200M unique tokens, which we upsample to a maximum of 8 times, yielding a total of 1.46B tokens. A more detailed dataset breakdown is available in Table 3.

We use a custom version of GPT-Neox¹³ (Andonian et al., 2023) on Jupiter HPC to train the model for 370 iterations with a batch size of 4M tokens and 1-sqrt learning rate cool-down from the very start of training.¹⁴

2.4 Reinforcement Learning

We apply reinforcement learning (RL) on the fine-tuned checkpoint to further improve translation quality. We use Group Relative Policy Optimisation (GRPO) (Shao et al., 2024) and follow TranslateGemma (Finkelstein et al., 2026), and Omnilingual MT (Omnilingual MT Team et al., 2026) in choosing MetricX (Juraska et al., 2024) as one of the rewards. COMET is selected as another, making the final reward a 50/50 combination of normalised MetricX and COMET scores.

Following the suggestion that there should be no overlap between SFT and RL training samples (Su et al., 2026), we use a 50k small, mainly sentence-level held-out subset of high-quality CS-DE general translation data from our supervised

¹³<https://github.com/tilde-nlp/llm-gpt-neox>

¹⁴This setup falls into the lower range of update steps due to time limitations; in our experience, training for thousands of updates on smaller batches yields better results.

Task	Samples	Tokens	Trainable tokens	Share
General	5.9M	888M	399M	62.7%
Glossary	61K	161M	65M	10.2%
Translation memory	115K	358M	145M	22.8%
Tags	116K	56M	27M	4.3%
Total	6.2M	1.46B	636M	100.0%

Table 3: Breakdown of the final SFT mixture by translation task. *General* is standard sentence- and document-level translation; *Glossary*, *Translation memory*, and *Tags* add conditioning on a supplied glossary, translation-memory matches, and inline markup. *Share* denotes the percentage of trainable tokens.

fine-tuning mix. For data samples spanning several paragraphs, we compute per-paragraph average MetricX/COMET scores and treat missing or hallucinated paragraphs as 0s in the calculation. We train the model for 1000 updates, with a batch size of 48, a rollout of 8, a sampling temperature of 0.4, and an AdamW optimiser (Loshchilov and Hutter, 2019). To limit divergence from the initial model, we include a low-variance KL regularisation term against the frozen reference policy with a coefficient of 0.003. Training is performed locally, using the VERL¹⁵ framework, on 4 NVIDIA H200 GPUs. Compared to the SFT-only checkpoint, we observe modest improvements across a variety of test sets. The results are shown in Table 6 and are further discussed in Section 5.

The resulting SFT-RL model serves as our Baseline for the WMT 2026 compression task in the CS-DE direction. It has 15B parameters, weighs ~ 30 GB, and achieves 110 tokens/s on a single NVIDIA H100 GPU using vLLM. The model is available on Tilde’s HuggingFace¹⁶ and should be prompted as "Translate the following text from Czech to German: {source}".

3 15B Model Family

From our Baseline model, we derive two 15B quantised versions, one in FP8 and one in FP4. Throughout this submission, we rely on LLM Compressor for quantisation due to its seamless integration with vLLM, our chosen inference engine. Resulting models’ performance metrics and associated evaluation method are covered in Section 5 and Table 7.

¹⁵<https://github.com/verl-project/verl>

¹⁶<https://huggingface.co/TildeAI/TildeOpen15B-64k-wmt26-compression-task-cs-de>

3.1 15B-RTN-fp8-dyn

We compress the Baseline model using data-free round-to-nearest (RTN) W8A8 FP8 post-training quantisation with LLM Compressor, by applying the FP8_DYNAMIC scheme to all linear layers except the final layer. Targeted weights are quantised to FP8 using static per-channel scales, while activations are dynamically quantised to FP8 at inference time using per-token scales. Consequently, the quantisation procedure requires neither a calibration dataset nor calibration forward passes. We chose dynamic activation scaling over static calibration because it adapts to activation ranges observed during inference and is therefore less prone to overfitting in the calibration domain. Additionally, vLLMs’ dedicated kernels make the activation-quantisation runtime overhead small in practice.

The resulting quantised checkpoint is our 15B-RTN-fp8-dyn model for the WMT 2026 compression task in the CS-DE direction. It has 15B parameters, consumes ~ 16 GB of disk space, and achieves 156 tokens/s on a single NVIDIA H100 GPU using vLLM. The model is available on Tilde’s HuggingFace¹⁷ and should be prompted the same as Baseline.

3.2 15B-GPTQ-nvfp4a16

We compress the Baseline model using GPTQ-based (Frantar et al., 2022) NVFP4 weight-only post-training quantisation with LLM Compressor, by applying the NVFP4A16 scheme to all linear layers except lm-head. Targeted weights are quantised to 4-bit NVFP4 with a global per-tensor scale and local scales applied to groups of 16 weights, while activations remain unquantised at 16-bit precision. Unlike RTN, GPTQ uses calibration forward passes to collect layer-input statistics and reduce the re-

¹⁷<https://huggingface.co/TildeAI/TildeOpen15B-64k-wmt26-compression-task-cs-de-fp8-dyn>

construction error.

Our calibration data consists of 512 examples of the general CS-DE task from the SFT dataset. We find that adding data seen during GRPO does not yield any improvements, and calibrating on in-domain, but unseen SFT data leads to degradation.

The resulting quantised checkpoint is our 15B-GPTQ-nvfp4a16 model for the WMT 2026 compression task in the CS-DE direction. It has 15B parameters, weighs ~ 10.1 GB, and achieves 156 tokens/s on a single NVIDIA H100 GPU using vLLM. The model is available on HuggingFace¹⁸ and should be prompted the same as Baseline.

Unfortunately, Hopper-generation GPUs, including the H100 used for WMT 2026 evaluation, do not support native NVFP4 computation. vLLM therefore falls back to a Marlin weight-only kernel, which keeps the weights packed in FP4 but dequantises them on the fly. The model thus retains its smaller disk and GPU memory footprint but cannot benefit from native FP4 arithmetic acceleration and performs similarly to FP8 on the evaluation hardware.

4 8B Model Family

In addition to our 15B model family, we develop 3 more CS-DE models of 8B size. We start by further distilling the previously distilled TildeOpen-15B-64k student model into the TildeOpen-8B-64k base model. This version is then fine-tuned and quantised using approaches similar to those for the 15B models, resulting in the 8B model family. The performance metrics and associated evaluation methods for the resulting models are covered in Section 5.

4.1 Distillation of 8B Base Model

We first prune the TildeOpen-15B-64k base model to ~ 8 B parameters by targeting the model width, as well as shrinking the embedding dimension and the FFN intermediate size to 11/16 of the original, i.e., $4,096 \rightarrow 2,816$ and $14,336 \rightarrow 9,856$, respectively. For everything else, we follow the exact same approach as in Section 2.1, using Jupiter HPC’s 4 NVIDIA GH200 GPUs, thereby reducing the model size from 15.167B to 8.157B.

Similarly, we follow the distillation method of Section 2.2, using the TildeOpen-15B-64k as the teacher. We expect a greater quality drop when

distilling to an 8B student model compared to a 15B student model. To address this, we increase the distillation dataset to 400B tokens, which is 100B more than used in the 30B $\rightarrow$ 15B distillation. We use Jupiter HPC’s 512 Nvidia GH200 GPUs and acquire the distilled 8B student in 47600 steps. We use the same distillation procedure as before.

To validate the distillation, we follow the two-step validation described in Section 2.2. Results for English-only versions of ARC, MMLU, and HellaSwag are shown in Table 4, and results for the downstream translation task are shown in Table 5.

	TildeOpen-15B	TildeOpen-8B
ARC-C (25)	0.525	0.508
MMLU (5)	0.540	0.495
HellaSwag (10)	0.787	0.738

Table 4: Performance comparison of TildeOpen-15B and TildeOpen-8B on English-only benchmarks. Reported value is *acc_norm* (*acc* for MMLU), and the number in parentheses indicates the fewshot amount.

WMT24++	TildeOpen-15B	TildeOpen-8B
EN-XX	0.836	0.833
DE-XX	0.829	0.828
FR-XX	0.822	0.820
PL-XX	0.824	0.822
UK-XX	0.820	0.818
LT-XX	0.815	0.813

Table 5: Results of automatic translation quality evaluation between translation fine-tuned TildeOpen-15B and TildeOpen-8B, as measured by COMET on the WMT24++ dataset.

Table 4 shows performance differences across ARC, MMLU, and HellaSwag that are consistent with those reported in Minitron-8B (Muralidharan et al., 2024) for the same size distillation, and do not exceed 9%. Similarly, results from Table 5 show COMET decreases between 0.001 and 0.003, indicating a successful distillation.

4.2 8B CS-DE Translation Models

Using the distilled 8B model, we first create our 8B-distil submission model by applying the same post-training recipe as for the baseline model, except that we skip reinforcement learning due to lack of time. The resulting SFT model has 8B parameters, consumes ~ 16 GB of disk space, and achieves 164 tokens/s on a single NVIDIA H100 GPU using vLLM. The model is available

¹⁸<https://huggingface.co/TildeAI/TildeOpen15B-64k-wmt26-compression-task-cs-de-fp4a16>

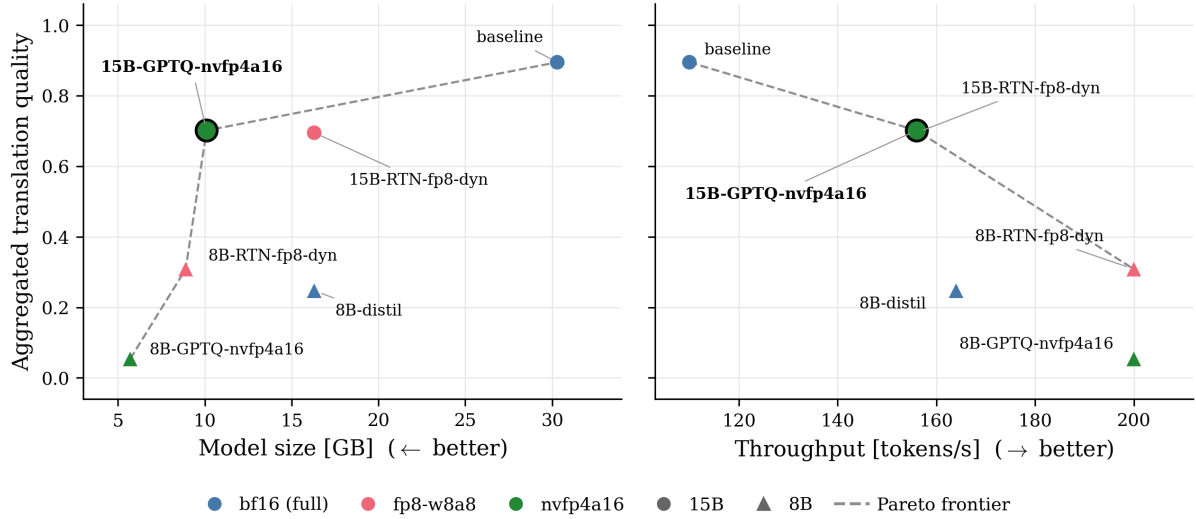


Figure 1: Aggregate translation quality versus model size (left) and decoding speed (right). Dashed lines mark the Pareto frontier; 15B-GPTQ-nvfp4a16 is our primary submission.

Test set	COMET↑	MetricX↓	ChrF++↑	BLEU↑
WMT24++	+0.0004	−0.029	−0.60	−0.41
WMT25	+0.0021	−0.131	−0.18	+0.34
FLORES-200 Sent	−0.0014	+0.015	−0.50	−0.33
FLORES-200 Doc	−0.0018	+0.114	−0.06	+0.12
NTREX Sent	+0.0018	−0.059	+0.04	+0.26
NTREX Doc	+0.0013	+0.054	−0.06	+0.30

Table 6: Effect of RL post-training on the 15B model for CS-DE across six test settings. Entries report score differences between the SFT+RL and SFT-only checkpoints, $\Delta = \text{SFT+RL} - \text{SFT}$. Positive differences indicate improvement for COMET, ChrF++, and BLEU; negative differences indicate improvement for MetricX.

on Tilde’s HuggingFace¹⁹ and should be prompted the same as Baseline.

We then apply the same quantisation recipes as in Sections 3.1 and 3.2 to produce two compressed versions of 8B-distil: 8B-RTN-fp8-dyn and 8B-GPTQ-nvfp4a16. The former consumes ~ 9 GB of disk space, the latter – 5.7 GB; both reach 200 tokens/s on an H100 GPU. Both models are available on HuggingFace^{20,21}.

5 Model Evaluation

To select our primary submission, we evaluate all 6 models using WMT 2026 recommended automatic

translation quality evaluation metrics and the suggested evaluation sets. Namely, we use: WMT25²², WMT24++, FLORES-200 (NLLB Team, 2024), and NTREX (Federmann et al., 2022; Barrault et al., 2019). For FLORES-200 and NTREX, we conduct both sentence-level and document-level evaluations. As for the metrics, we use COMET and MetricX. Additionally, we opt for two non-neural metrics: ChrF++ (Popović, 2017) and BLEU (Papineni et al., 2002). For document-level tests, we report sentence average COMET and MetricX. We report effect of RL training from Subsection 2.4 in Table 6. Table 7 summarises translation quality evaluation results, while Table 9 reports 95% confidence intervals showing the uncertainty in each model’s score differences from Baseline. Finally, Table 8 reports models’ size on disk and throughput, measured using vLLM on a single NVIDIA H100 GPU with no prefix caching and batch size 1.

Table 6 shows that RL further improves the 15B model’s COMET and MetricX scores on both WMT benchmarks and NTREX Sent, as compared to the SFT checkpoint. The largest neural-metric gains occur on WMT25, where COMET increases by 0.0021 and MetricX decreases by 0.131, accompanied by a 0.34-point BLEU improvement. Notably, BLEU improves in four of the six test settings despite not being included in the RL reward, showing that the observed benefits are not

¹⁹<https://huggingface.co/TildeAI/TildeOpen8B-64k-wmt26-compression-task-CS-DE>

²⁰<https://huggingface.co/TildeAI/TildeOpen8B-64k-wmt26-compression-task-cs-de-fp8-dyn>

²¹<https://huggingface.co/TildeAI/TildeOpen8B-64k-wmt26-compression-task-cs-de-fp4a16>

²²<https://github.com/wmt-conference/wmt25-general-mt>

		Baseline	15B-RTN -fp8-dyn	15B-GPTQ -nvfp4a16	8B-distil	8B-RTN -fp8-dyn	8B-GPTQ -nvfp4a16
WMT24++	COMET	0.817	-0.003	-0.001	-0.003	-0.003	-0.004
	MetricX	2.56	0.04	0.01	0.07	0.05	0.10
	ChrF++	51.17	-0.42	-0.09	-0.65	-0.42	-0.59
	BLEU	24.65	-0.58	0.08	-0.89	-0.54	-0.91
WMT25	COMET	0.784	0.001	-0.003	-0.001	-0.003	-0.003
	MetricX	5.03	-0.03	0.07	0.09	0.07	0.13
	ChrF++	58.68	-0.23	-0.27	-1.14	-1.03	-1.25
	BLEU	37.51	-0.47	-0.45	-1.33	-1.24	-1.63
FLORES-200 Sent	COMET	0.875	0.001	0.000	-0.001	-0.001	-0.001
	MetricX	1.41	0.01	0.00	0.07	0.05	0.04
	ChrF++	55.01	-0.10	-0.05	-0.42	-0.39	-0.38
	BLEU	29.36	-0.19	0.05	-0.66	-0.60	-0.67
FLORES-200 Doc	COMET	0.869	0.000	0.000	-0.002	-0.001	-0.003
	MetricX	2.24	0.03	-0.04	-0.07	-0.02	-0.10
	ChrF++	58.78	-0.15	-0.02	-0.89	-0.77	-0.81
	BLEU	31.27	-0.25	0.09	-1.16	-0.85	-1.06
NTREX Sent	COMET	0.856	0.000	0.001	-0.001	-0.002	-0.003
	MetricX	1.82	0.03	0.02	0.05	0.05	0.04
	ChrF++	52.60	-0.13	-0.22	-0.76	-0.57	-0.83
	BLEU	26.44	-0.26	-0.33	-1.04	-0.79	-1.08
NTREX Doc	COMET	0.858	0.002	-0.005	0.003	0.000	-0.005
	MetricX	4.74	-0.18	0.27	0.01	-0.05	0.15
	ChrF++	60.79	-0.02	-0.08	0.13	0.13	-0.11
	BLEU	29.60	-0.05	-0.06	-0.10	-0.13	-0.34

Table 7: Automatic evaluation of all six submission models on CS-DE across six test sets. The Baseline column reports absolute scores, while the remaining columns report differences relative to Baseline. Higher is better for COMET, ChrF++, and BLEU; lower is better for MetricX.

	Baseline	15B-RTN -fp8-dyn	15B-GPTQ -nvfp4a16	8B-distil	8B-RTN -fp8-dyn	8B-GPTQ -nvfp4a16
Size on disk [GB]	30.3	16.3	10.1	16.3	8.9	5.7
Throughput [tokens/s]	110	156	156	164	200	200

Table 8: Comparison of submission models’ size on disk and throughput.

confined to the optimised metrics. The gains are not universal: both FLORES-200 settings show lower neural-metric performance, NTREX Doc has mixed neural-metric results, and ChrF++ decreases in five settings. Overall, the results support RL as a useful refinement of the SFT baseline, with the clearest benefits on WMT25 and NTREX Sent. However, further human-eval would be necessary to confirm this conclusion.

Tables 7 and 8 guide the selection of the primary submission. To compare models, we construct an aggregate quality score. For each metric–test-setting combination, we rescale the six models’ scores so that the worst receives 0 and the best receives 1, with intermediate scores placed proportionally and lower MetricX scores treated as better. We then average the rescaled scores over

all 24 combinations of four metrics and six test settings.. Figure 1 plots this score against model size and throughput. The dashed line marks the Pareto frontier, which contains the models that no alternative beats in both quality and deployment cost. Both quantised 15B models clearly outperform the 8B variants in quality. At the same time, 15B-GPTQ-nvfp4a16 matches the quality of 15B-RTN-fp8-dyn at a $\times 1.6$ smaller footprint. It also has identical speed on H100 due to the lack of native FP4 support. In practice, we expect the FP4 models to be faster than their FP8 counterparts on Blackwell architecture GPUs. The confidence intervals in Table 9 provide further support for selecting the 15B NVFP4 model. For both quantised 8B models, the BLEU differences from Baseline remain below zero across the full

	15B-RTN -fp8-dyn	15B-GPTQ -nvfp4a16	8B-distil	8B-RTN -fp8-dyn	8B-GPTQ -nvfp4a16
WMT24++					
COMET	-0.0018 / +0.0017	-0.0020 / +0.0020	-0.0030 / +0.0030	-0.0031 / +0.0031	-0.0031 / +0.0031
MetricX	-0.034 / +0.034	-0.040 / +0.039	-0.056 / +0.055	-0.059 / +0.057	-0.062 / +0.062
ChrF++	-0.30 / +0.27	-0.32 / +0.30	-0.52 / +0.48	-0.43 / +0.44	-0.42 / +0.42
BLEU	-0.34 / +0.33	-0.34 / +0.35	-0.97 / +0.49	-0.52 / +0.52	-0.53 / +0.54
WMT25					
COMET	-0.0040 / +0.0047	-0.0039 / +0.0037	-0.0058 / +0.0059	-0.0060 / +0.0058	-0.0063 / +0.0062
MetricX	-0.117 / +0.117	-0.131 / +0.131	-0.185 / +0.180	-0.183 / +0.183	-0.201 / +0.196
ChrF++	-0.43 / +0.48	-0.37 / +0.37	-0.66 / +0.64	-0.55 / +0.55	-0.71 / +0.70
BLEU	-0.52 / +0.59	-0.48 / +0.60	-0.76 / +0.97	-0.84 / +1.01	-0.95 / +1.11
FLORES-200 Sent					
COMET	-0.0007 / +0.0007	-0.0009 / +0.0009	-0.0014 / +0.0014	-0.0014 / +0.0014	-0.0013 / +0.0013
MetricX	-0.017 / +0.017	-0.021 / +0.021	-0.033 / +0.034	-0.033 / +0.033	-0.031 / +0.031
ChrF++	-0.20 / +0.20	-0.24 / +0.24	-0.35 / +0.35	-0.35 / +0.35	-0.35 / +0.35
BLEU	-0.31 / +0.31	-0.34 / +0.34	-0.52 / +0.52	-0.52 / +0.52	-0.52 / +0.52
FLORES-200 Doc					
COMET	-0.0013 / +0.0014	-0.0015 / +0.0016	-0.0022 / +0.0022	-0.0021 / +0.0022	-0.0023 / +0.0023
MetricX	-0.051 / +0.062	-0.042 / +0.039	-0.153 / +0.143	-0.138 / +0.129	-0.162 / +0.148
ChrF++	-0.20 / +0.19	-0.20 / +0.20	-0.36 / +0.36	-0.34 / +0.34	-0.34 / +0.34
BLEU	-0.31 / +0.31	-0.32 / +0.32	-0.52 / +0.52	-0.51 / +0.51	-0.51 / +0.51
NTREX Sent					
COMET	-0.0010 / +0.0010	-0.0011 / +0.0011	-0.0017 / +0.0017	-0.0018 / +0.0018	-0.0018 / +0.0018
MetricX	-0.020 / +0.021	-0.025 / +0.025	-0.037 / +0.038	-0.040 / +0.040	-0.039 / +0.039
ChrF++	-0.18 / +0.18	-0.21 / +0.21	-0.32 / +0.32	-0.32 / +0.32	-0.32 / +0.33
BLEU	-0.27 / +0.27	-0.30 / +0.30	-0.47 / +0.47	-0.46 / +0.46	-0.46 / +0.47
NTREX Doc					
COMET	-0.0053 / +0.0066	-0.0072 / +0.0058	-0.0097 / +0.0101	-0.0106 / +0.0108	-0.0126 / +0.0122
MetricX	-0.352 / +0.310	-0.432 / +0.486	-0.343 / +0.344	-0.388 / +0.380	-0.440 / +0.444
ChrF++	-0.16 / +0.15	-0.21 / +0.20	-0.69 / +1.16	-0.69 / +1.17	-0.72 / +1.16
BLEU	-0.26 / +0.26	-0.31 / +0.31	-0.64 / +0.82	-0.67 / +0.84	-0.69 / +0.83

Table 9: Asymmetric 95% paired-bootstrap uncertainty offsets for the candidate-minus-Baseline differences in Table 7, based on 50,000 resamples. Each entry is $-\ell / +u$, giving the distances from the unrounded difference $\hat{\Delta}$ to its lower and upper percentile limits: $[\hat{\Delta} - \ell, \hat{\Delta} + u]$. Offsets are calculated before rounding, so reconstructing limits from the printed values is approximate. The document runs resample document entries.

95% confidence interval in five of the six test settings. For 15B-GPTQ-nvfp4a16, this occurs only on NTREX Sent. Under the same criterion, the 8B NVFP4 model also shows COMET losses in three settings, whereas no such loss is detected for the 15B NVFP4 model. Together with the Pareto analysis, these results favour 15B-GPTQ-nvfp4a16 as a compromise that prioritises preserving translation quality over the additional size and speed gains offered by the 8B models. We therefore select it as our primary submission.

6 Conclusion

We presented Tilde’s submission to the WMT 2026 Model Compression task in the Czech-German direction: six models ranging from 30.3 GB to 5.7 GB on disk and from 110 to 200 tokens/s in decoding throughput. All of them are derived from

TildeOpen-15B-64k through post-training quantisation, knowledge distillation, and their combination. Together, the two compression axes trace a spectrum of quality-size-speed trade-offs: quantisation reduces the footprint at minimal quality cost, while distillation to 8B buys further size and speed gains in exchange for a small additional drop in translation quality. Among the six systems, 15B-GPTQ-nvfp4a16 matches the quality of its FP8 counterpart at a $\times 1.6$ smaller footprint. It thus offers the best quality-size-speed trade-off among our models and is our primary submission.

7 Limitations

Our experiments cover a single language pair and a single model family, so the observed quality-size-speed trade-offs may not transfer to other languages, domains, or architectures. The 8B

model received supervised fine-tuning but not the reinforcement-learning stage applied to the 15B baseline, so comparisons between the two families conflate compression effects with post-training differences. Evaluation relies solely on automatic metrics, without human assessment. We report marginal paired-bootstrap confidence intervals, without adjustment for multiple comparisons or direct pairwise testing between compressed models. COMET and MetricX also serve as RL rewards, partially overlapping the optimisation and evaluation criteria. Speed was measured in a single vLLM configuration on one H100 GPU, which lacks native NVFP4 support, so FP4 throughput is understated. Finally, full reproduction of training is limited by proprietary data and supercomputing requirements, though all models and inference code are public.

Acknowledgments

The research leading to these results has received funding from the project “Competence Centre of Information and Communication Technologies for Research” of the Recovery and Resilience Facility, contract No. 5.1.1.2.i.0/2/24/A/CFLA/007 signed between IT Competence Centre and Central Finance and Contracts Agency, Research No. 2.7 “Real-Time Speech Translation and Event Content Processing Tool”.

We gratefully acknowledge the Gauss Centre for Supercomputing e.V. (<https://www.gauss-centre.eu>) for funding this project by providing computing time on the GCS Supercomputer JUPITER at Jülich Supercomputing Centre (JSC) under project TildeLM (Project ID: jureap133).

References

- Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Jason Phang, Shivanshu Purohit, Hailey Schoelkopf, Dashiell Stander, Tri Songz, Curt Tigges, Benjamin Thérien, and 2 others. 2023. [GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch](#).
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Toms Bergmanis, Martins Kronis, Ingus Jānis Pretkalniņš, Dāvis Nicmanis, Jeļizaveta Jelinska, Roberts Rozis, Rinalds Viksna, and Mārcis Pinnis. 2026. [TildeOpen LLM: Leveraging curriculum learning to achieve equitable language representation](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 10634–10652, Palma, Mallorca, Spain. European Language Resources Association (ELRA).
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.
- Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma technical report](#). *Preprint*, *arXiv:2601.09012*.
- Elias Frantar, Saleh Ashkboos, Torsten Hoeffer, and Dan Alistarh. 2022. GPTQ: Accurate post-training compression for generative pretrained transformers. *arXiv preprint arXiv:2210.17323*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. [The language model evaluation harness](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- 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.
- Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *International Conference on Learning Representations (ICLR)*.
- Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.
- 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](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 41076–41102. Curran Associates, Inc.
- NLLB Team. 2024. [Scaling neural machine translation to 200 languages](#). *Nature*, 630:841–846.
- Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Xiang “Tony” Cao, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenhaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, and 13 others. 2026. [Omnilingual MT: Machine translation for 1,600 languages](#). *Preprint*, arXiv:2603.16309.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: A method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Maja Popović. 2017. [chrF++: Words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa 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)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. [Deepseekmath: Pushing the limits of mathematical reasoning in open language models](#). *Preprint*, arXiv:2402.03300.
- Sharath Turuvekere Sreenivas, Saurav Muralidharan, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. 2024. [LLM pruning and distillation in practice: The Minitron approach](#). Technical report, NVIDIA.
- Xiaole Su, Kasey Zhang, and Andy Lyu. 2026. [SFT-GRPO data overlap as a post-training hyperparameter for autoformalization](#). *Preprint*, arXiv:2604.13515.
- Junfeng Tian, Da Zheng, Yang Chen, Rui Wang, Colin Zhang, and Debing Zhang. 2025. [Untie the knots: An efficient data augmentation strategy for long-context pre-training in language models](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1223–1242, Vienna, Austria. Association for Computational Linguistics.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.