TOWER+: Bridging Generality and Translation Specialization in Multilingual LLMs

Ricardo Rei[*], **Nuno M. Guerreiro**[*], **José Pombal**[*1,2], **João Alves**[*]
Pedro Teixeirinha[3], **Amin Farajian**[3], **André F. T. Martins**[1,2,3,4]
[1]Instituto de Telecomunicações
[2]Instituto Superior Técnico, Universidade de Lisboa
[3]Transperfect
[4]ELLIS Unit Lisbon

Abstract

Fine-tuning pretrained LLMs has proven effective for reaching state-of-the-art performance on specific tasks like machine translation. However, this process often implies sacrificing general-purpose capabilities, such as conversational reasoning and instruction-following, hampering the usefulness of the system in real-world applications requiring a mixture of skills. In this paper, we introduce TOWER+, a suite of models designed to deliver strong performance on both translation and multilingual general-purpose text capabilities. We improve the TOWER (Alves et al., 2024) recipe by adding novel stages of preference optimization and reinforcement learning with verifiable rewards, in addition to continued pretraining and supervised fine-tuning. At each stage, we carefully generate and curate data to strengthen performance on translation and general-purpose tasks like coding, mathematics, and instruction-following. We develop models at multiple scales: 2B, 9B, and 72B. Our smaller models often outperform larger general-purpose open-weight and proprietary LLMs (e.g., LLAMA 3.3 70B, GPT-4o). Our largest model delivers best-in-class translation performance for high-resource languages, and top results on multilingual Arena Hard and IF-MT, a benchmark we introduce for evaluating both translation and instruction-following.[1]

1 Introduction

Large Language Models (LLMs) are emerging as the *de facto* solution for multilingual machine translation. Recent studies have shown that state-of-the-art proprietary LLMs, such as GPT and Claude are currently state-of-the-art in translation quality (Kocmi et al., 2023, 2024a; Deutsch et al., 2025; Kocmi et al., 2025b). At the same time, several works have shown that open-weight LLMs can be adapted for machine translation, reaching parity with or surpassing the translation quality of leading proprietary models (Alves et al., 2024; Rei et al., 2024; Xu et al., 2025; Cui et al., 2025a). However, these task-specific adaptations often come at the cost of general-purpose capabilities.[2] This tradeoff is illustrated in Figure 1, which shows that translation-specialized models tend to fall short on the Pareto frontier that balances translation quality and general capabilities. For example, TOWER V2 ranked first in 8 out of 11 language pairs on WMT24 (Kocmi et al., 2024a), yet greatly underperforms most models on general chat evaluations. In addition, the degradation of instruction-following capabilities can hinder the ability to handle complex real-world translation scenarios that require, for example, adhering to terminology or formatting rules (§4).

To address this challenge, we propose the TOWER+ recipe for developing state-of-the-art translation models without compromising performance on general chat benchmarks. Like earlier TOWER models (Alves et al., 2024; Rei et al., 2024), we begin with continued pretraining (CPT) to enhance multilingual fluency and translation performance (§2.1). We then also do supervised fine-tuning (SFT), while significantly refining automatic data generation and curation pipelines, and increasing the proportion of general-purpose data considerably (§2.2). Crucially, we introduce a novel preference optimization stage using Weighted Preference Optimization (Zhou et al., 2024, WPO) (§2.3), complemented by Group Relative Policy Optimization (Shao et al., 2024, GRPO) with verifiable rewards (RLVR) to further enhance performance across tasks (§2.4). We evaluate our models on both general chat benchmarks, translation, and on

[*]Core contributor. Work done while at Unbabel.

[1]Our TOWER+ models, as well as the IF-MT benchmark, are available on Huggingface.

[2]The term "general-purpose capabilities" refers to the real-world utility of language models in handling queries that require core knowledge, instruction-following, and conversational reasoning (Zheng et al., 2023)

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29614–29635

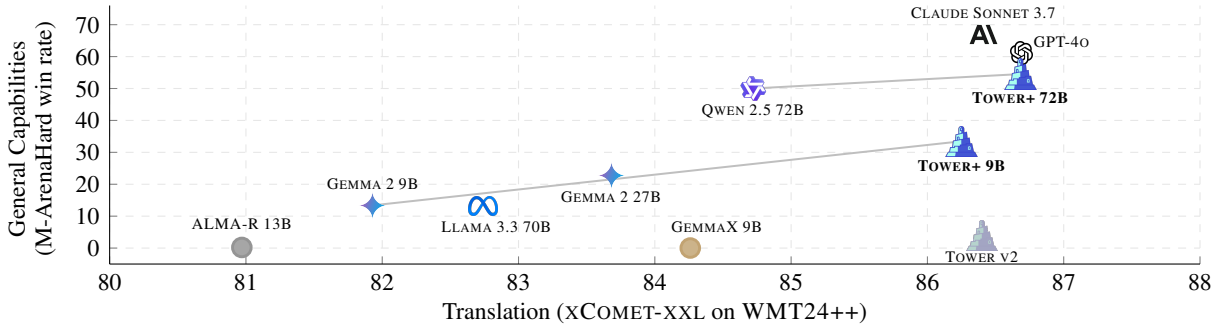

Figure 1: Translation and general capabilities performance of state-of-the-art translation-specific (circles) and general-purpose LLMs (logos) of varying sizes (from 9B to 72B). For their size, TOWER+ models outperform or match open-weight models on both axes, and TOWER+ 72B is competitive to state-of-the-art proprietary models. We omit 2B models for better visualization; we show detailed results in Table 1. Gray lines connect each TOWER+ model to their respective "instruction-tuned" counterpart.

a mix of translation and instruction-following capabilities, matching or outperforming state-of-the-art proprietary and open-weight systems on all axes. We also analyze the contribution of each stage of the training pipeline (§4.2), and assess the role of the backbone model in balancing general-purpose and translation-specific capabilities (§4.3).

Our contributions are:

- We present, to the best of our knowledge, the first systematic study on balancing translation quality and general-purpose capabilities in open-weight LLMs. While most prior work has focused solely on maximizing translation performance, our approach explicitly targets a broader trade-off.
- We introduce a post-training pipeline that integrates diverse multilingual signals without compromising general chat abilities. Our approach can serve as a blueprint for adapting LLMs to domain- or task-specific business use cases while preserving general capabilities.
- We introduce and release IF-MT, a novel benchmark for evaluating both translation and instruction-following capabilities on two language pairs (EN→ES and EN→ZH).
- We release TOWER+, a suite of models that demonstrate strong performance across translation, general capabilities, and a benchmark that mixes the two. We match or exceed the translation quality of prior TOWER models and GPT-4O-1120, while also surpassing general-purpose open-weight models like Llama-3.3 70B and Qwen2.5 72B on M-ArenaHard.

2 TOWER+ Post-Training Recipe

In this section, we describe each training stage and explain how translation signals are incorporated throughout. We detail hyperparameters for all stages in Appendix B. Later, we apply the recipe on three backbone models: GEMMA 2 2B, GEMMA 2 9B, and QWEN 2.5 72B.

2.1 Continued Pretraining

This phase leverages monolingual, parallel, and general-purpose instruction-following data. Similarly to previous TOWER models, the data distribution follows a 66%/33% split between monolingual and parallel data; in this version, we include 1% of instruction-following data.

All monolingual data is sourced from FineWeb-Edu (Penedo et al., 2024). Most parallel data is sourced from OPUS (Tiedemann, 2012) and filtered using COMETKIWI (Rei et al., 2022b). Additionally, the parallel data is formatted as a translation instruction followed by the corresponding translation.[3] For language pairs where it is available, we include document-level translation data from EuroParl (Koehn, 2005), ParaDocs (Wicks et al., 2024), and CosmoPedia-v2[4] (Ben Allal et al., 2024), each totaling 10% of the data of each language pair. Instruction-following data is sampled from FineWeb-Edu using `dsir` (Xie et al., 2023) to be similar to high-quality instruction-following data. For monolingual data, we apply the EuroFilter-v1 (Martins et al., 2025) quality fil-

[3]We use multiple templates to prepare the parallel corpora. See examples in Appendix Figure 7.

[4]We create document translations for CosmoPedia using previous Tower models and COMETKIWI.

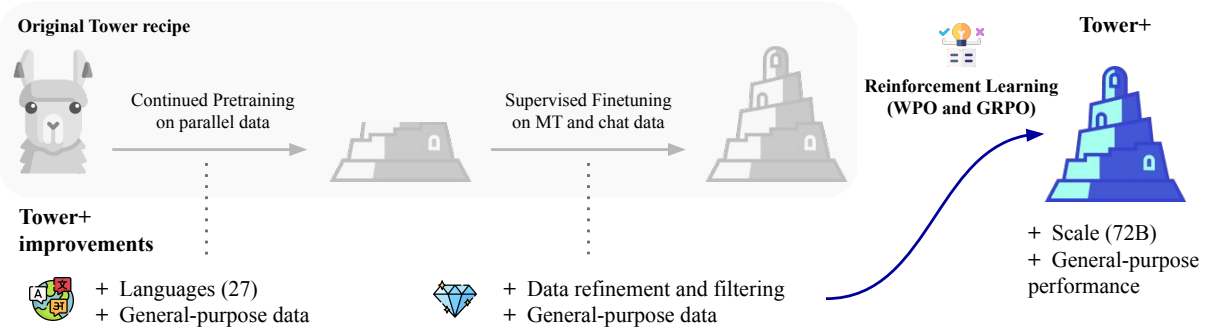

Figure 2: Post-training process of TOWER+. Supervised fine-tuning and reinforcement learning processes are illustrated in more detail in Appendix Figures 5 and 6, respectively.

ter, a multilingual educational classifier built using mDeBERTA (He et al., 2023).

Using this data—which covers 27 languages/dialects[5] and 47 language pairs, totaling 32B tokens—we continue the pretraining of a base open-weight LLM with a standard next-token prediction objective. Then, before proceeding to the next phase, we merge the CPT checkpoint back with the base checkpoint by linearly interpolating weights with a 0.5 coefficient for each set of weights.

2.2 Supervised Fine-tuning

For SFT, we collect data for general-purpose tasks, and for pre-translation, translation, and post-translation tasks. For the first, we source instructions from several publicly available datasets, including OpenHermes-2.5, Aya (Singh et al., 2024), Daring-Anteater (Wang et al., 2024), Magpie (Xu et al., 2024b), Tülu (Lambert et al., 2025), and others. Using Llama 3.3 70B (Llama Team, 2024), we assign two scores from 1 to 5 to each instance that represent (i) an estimate of the amount of reasoning required to answer and (ii) the instruction's readability.[6] We filter out most data where the reasoning score or readability falls below 4.[7] We then collect answers from four top-performing open-weight LLMs to create a pool of "candidate" answers: the original answer, DeepSeek V3 (DeepSeek-AI, 2025), Qwen 2.5 72B (Qwen Team, 2025), Tülu 3 (Lambert et al., 2025), and Llama 3.3 answers. The answer we ultimately use for training is the one that ranks the highest when evaluated using a general-purpose reward model, Skywork-Gemma-2-27B (Liu et al., 2024). This process follows the increasingly common paradigm of distillation from multiple teacher models, where several strong open-weight LLMs provide candidate completions, and the best one is selected based on a learned reward function. Given the multilingual nature of our corpus, this approach is closely aligned with the multilingual arbitrage method proposed by Odumakinde et al. (2024), which applies the same principle of multi-teacher selection to multilingual prompts.

Pre-translation tasks involve preprocessing steps typically performed before translation, such as grammar error correction and the removal of PII content. Translation tasks cover a broad spectrum, including sentence-level translation, style adaptation (e.g., formal vs. informal), document-level translation, and multilingual translation (single source, targets in multiple languages). Some of these data are proprietary, but the majority comes from WMT shared tasks and Flores test sets (excluding WMT24). Post-translation tasks focus on processes that follow translation, such as automatic post-editing and quality evaluation. As with translation tasks, part of the data comes from proprietary sources, while the rest comes from WMT shared tasks ranging from 2017 to 2023.

The final corpus consists of 1.3 million samples, with translation tasks accounting for approximately 22% of the total. In Appendix Figure 5, we illustrate the full SFT data curation process along with the proportion each category contributes to the final corpus.

2.3 Preference Optimization

After SFT, our models undergo offline reinforcement learning using weighted preference optimization (Zhou et al., 2024, WPO). This phase uses a mixture of prompts from two sources: (i) a

[5]Complete list of languages available in Appendix C.

[6]The classification prompt is in Appendix Figure 8.

[7]We do not filter out all data with low reasoning and readability because it is also important for the model to learn how to respond to poorly formulated instructions. We only keep such prompts if they are from OpenHermes-2.5 and we discard the remaining ones coming from other datasets.

subset of SFT prompts, and (ii) new prompts from UltraFeedback (Tunstall et al., 2024). These two datasets serve complementary roles. The SFT-derived prompts are richer in multilingual coverage, safety-critical scenarios, and multiturn interactions—areas underrepresented in UltraFeedback. Preferences for these prompts are collected off-policy from several high-quality open-weight LLMs that permit commercial use.[8] UltraFeedback data, in contrast, is used on-policy, leveraging samples generated by our model. For both sources, we apply the INF-ORM reward model[9] to identify the best and worst completions, which are then used for WPO updates. While we experimented with alternative reward models, such as Skywork (used in the SFT phase), we observed that Gemma 2–based reward models tended to over-prefer completions from their own backbone models (e.g., Gemma 2 27B Instruct). This bias made them less suitable for preference optimization, where candidate responses include outputs from Gemma models, which was not the case in the SFT dataset. In contrast, INF-ORM, which ranked first on RewardBench at the time of writing, showed no such preference for its own Llama 3 base and was therefore selected for this phase.

Finally, to improve performance on machine translation, we incorporate human preference data collected by professional linguists. This data comes from two sources: (i) post-edits of TOWER V2 outputs, where the edited version is treated as preferred over the original system output, and (ii) preference annotations between translations produced during quality evaluations of earlier TOWER models. We experimented with various formats and reused the collected data. Using post-edits as preference data has been shown to effectively improve translation quality (Berger et al., 2024). This approach is similar in spirit to the methodology described in the LLAMA 3 technical report (Llama Team, 2024), where expert-edited outputs were repurposed as preference data (albeit for non-MT tasks).

The remaining MT preferences are collected using COMET22 (Rei et al., 2022a) for Minimum Bayes Risk Decoding(Kumar and Byrne, 2004, MBR)[10] and picking the 'best' and 'worst' translations from the resulting MBR scores. To avoid metric-specific biases (Kocmi et al., 2024a; Pombal et al., 2025b) we then double check that 'best' > 'worst' using METRICX24-XXL (Juraska et al., 2024) and Llama 3.3 focusing on fluency and instruction following.[11]

We found WPO to outperform DPO (Rafailov et al., 2023), especially in terms of translation quality where we saw little to no improvement over the SFT model. The complete reinforcement learning pipeline is summarized in Appendix Figure 6.

2.4 RL with Verifiable Rewards

At this stage, our models already demonstrate state-of-the-art performance in both translation and general-purpose tasks. However, we find that their instruction-following, mathematical, and reasoning capabilities can be further improved through training on data with verifiable rewards. To this end, we leverage the Tülu 3 verifiable rewards dataset (Lambert et al., 2025) and augment it with two translation-specific signals: *translation-verifiable instruction* and *translation preference evaluation*. We provide a template in Appendix F.

Translation-verifiable instructions target the model's ability to apply text transformations during translation (e.g., converting date formats from DD-MM-YYYY into MM-DD-YYYY formatting). To generate this training data, we first defined a list of 28 broad text transformations (e.g., email formatting, date formatting; we provide all categories in Appendix F). Along with each transformation category, we include corresponding transformations (e.g., for date formatting, some transformations include month abbreviation, day of week abbreviation, timezone format, etc.). These transformations come with a description, a verification (in the form of a regular expression), and one example of input/output. We show one such transformation template in Appendix F.

We then prompt LLaMA 3.3 70B Instruct to generate a list of precise source-dependent transformation guidelines and a source document (that does not follow them) given (i) a list of guideline categories (whose length was sampled uniformly between 1 and 4), description and one example of input/output transformation when required, (ii) a desired length (1/2 sentences, 1 paragraph), and (iii) a topic/sub-topic pair out of a list of over 625 pairs (topics vary wildly spanning from pairs like

[8] DEEPSEEK V3, LLAMA-3.3-70B, QWEN-2.5-72B, MISTRAL-SMALL-3.1, GEMMA-2-27B, TÜLU-3-70B.

[9] https://huggingface.co/infly/INF-ORM-Llama3.1-70B

[10] We sample 24 candidates using temperature of 1.0.

[11] While neural MT metrics capture adequacy we found that an LLM-as-a-judge can better score fluency and how well the translation respects the provided user instruction.

"Sports Industry—Athletic Equipment" to "Journalists and Writers—Ezra Klein"). We provide the prompt for this step in Appendix F and one example in Figure 11. Next, we used the same LLM to verify the generated data. A sample was kept only if its source text violated all of the associated guidelines, ensuring every transformation was applicable. Finally, we translated these sentences using TOWER and asked different LLMs to apply the transformations on the translated output. We filtered out any examples where the verification (e.g., regex template) did not match or where the final translation quality (measured by COMETKIWI) was below 0.8. During GRPO, the model is rewarded when its output matches the regex in the target translation. This task is designed to encourage more precise instruction-following during translation.

Translation preference evaluation reuses the curated translation preferences from the WPO phase: we prompt the model to compare two translations, provide a quality assessment (reasoning), and deliver a final judgment. The model is rewarded when it selects the better translation. This is done without any thinking tokens—all reasoning leading up to the final decision is part of the final answer.

3 Experimental Setup

3.1 Test Sets

To measure performance on both machine translation and general capabilities we use four benchmarks: WMT24++ (Deutsch et al., 2025) for translation, M-ArenaHard (Dang et al., 2024; Li et al., 2024) and IFEval (Zhou et al., 2023) for general capabilities, and IF-MT for a mix of both.

WMT24++ For translation, we use the WMT24++ test set, which extends the official WMT24[12] set (Kocmi et al., 2024b) to cover 55 languages and dialects with new human-written references. WMT24++ includes all 22 source languages covered by our models, so we evaluate translation into 24 target variants.[13] For evaluation, we rely on state-of-the-art automatic metrics, including XCOMET-XXL (Guerreiro et al., 2024) (our primary metric), and METRICX24-XXL (Juraska et al., 2024) and CHRF (Popović, 2015) for supplementary analysis. The latter two are reported in the Appendix (Tables 3 and 2, respectively).

IFEval Many real-world applications of LLMs require following instructions to complete specific tasks (e.g., formatting text according to given guidelines). To assess this, we use IFEval (Zhou et al., 2023), a widely adopted benchmark for evaluating instruction-following behavior. IFEval consists of 541 instructions whose outputs can be automatically verified using simple code or regular expressions. Models are evaluated based on the percentage of instructions executed correctly.

M-ArenaHard To evaluate multilingual general capabilities, we use M-ArenaHard (Dang et al., 2024), a translated version of ArenaHard, in 4 languages: English, German, Spanish, Chinese, and Russian. ArenaHard is a dataset of 500 challenging and representative instances from the Chatbot Arena (Zheng et al., 2023), a website where users create prompts for language models and evaluate choose the best among pairs of responses from a wide range of LLMs. Chatbot Arena has been widely adopted as a benchmark for general capabilities given the highly diverse nature of the prompts users create. ArenaHard requires choosing an LLM judge and a baseline response for all prompts. We employ LLaMA 3.3 70B Instruct (Llama Team, 2024) as the evaluator and Qwen2.5 72B Instruct (Qwen Team, 2025) as the baseline.[14]

IF-MT: translation + instruction-following Many real-world translation tasks go beyond simple language conversion, often requiring adherence to specific guidelines and rules—such as maintaining consistent terminology or adapting date and currency formats—that entail a certain degree of general capabilities. However, no existing benchmarks evaluate both dimensions, so we create one following the zero-shot benchmarking methodology (Pombal et al., 2025a, ZSB): **IF-MT**. ZSB is a task-agnostic framework for automatically creating benchmarks that correlate strongly with human rankings by prompting language models for data generation and evaluation.

[12]The WMT shared task is a major annual competition in the field.

[13]We include both Brazilian and European Portuguese, and Simplified and Traditional Chinese, totaling 24 language directions. While our models support more variants, WMT24++ does not currently provide references for all of them.

[14]We use LLaMA 3.3 70B due to the focus placed on evaluation capabilities during its training (Llama Team, 2024), avoiding GPT or Claude models due to the risk of self-preference bias (Koo et al., 2023) inflating their scores. Using Qwen2.5 72B Instruct as a baseline allows for a direct comparison with a model based on Qwen2.5 72B, which is the backbone of TOWER+ 72B.

Models	Params	M-ArenaHard	IFEval	WMT24++			IF-MT	
				7 lang.	15 lang.	24 lang.	IF	MT
Closed								
GPT-4O-1120	>100B	61.19	85.20	**86.69**	**84.33**	**85.21**	**5.81**	**89.35**
CLAUDE-SONNET-3.7	>100B	**67.00**	**89.95**	86.41	**84.24**	**85.19**	—	—
Open Weights								
ALMA-R†	13B	0.2	0.0	80.97	–	–	1.71	78.11
GEMMAX†	9B	0.02	0.17	84.26	78.74	75.66	1.52	68.95
TOWER-V2†	70B	4.01	51.22	86.40	**83.88**	**83.74**	3.14	87.82
GEMMA-2	9B	13.38	66.86	81.93	75.35	76.34	5.07	88.51
GEMMA-2	27B	22.81	66.60	83.68	79.02	80.18	5.29	88.67
QWEN-2.5	72B	50.00	88.44	84.72	77.44	76.62	5.49	88.79
LLAMA-3.3	70B	13.15	**92.17**	82.74	78.30	79.48	5.38	88.13
Ours								
TOWER+	2B	6.33	67.32	81.88	78.42	79.13	2.90	87.65
TOWER+	9B	33.47	83.84	86.25	83.57	84.38	4.85	88.51
TOWER+	72B	**54.52**	89.02	**86.68**	83.29	**83.74**	**5.55**	**88.95**

Table 1: Results of several translation-specific (†) and general-purpose open-weight and closed API models across M-ArenaHard, IFEval, WMT24++, and IF-MT (English→Chinese). We consider two evaluation dimensions on IF-MT: instruction-following (IF) and raw MT quality (MT). For WMT24++ we report XCOMET-XXL and we split the language pairs into three categories: (1) seven high-resource languages, (2) the 15 languages from TOWER-V2 (our submission to WMT24), and (3) all languages supported by our new models. This categorization enables a more equitable comparison with other systems, which, in all cases, support at least the seven high-resource languages. We boldface the best overall system, and the best open-weight system if the former is proprietary.

We generate sources for two LPs—English→Chinese and English→Spanish—using CLAUDE-SONNET-3.7. We do not consider a fixed set of instructions, but rather prompt the data generator to come up with 2 to 4 instructions that can be applied to the source it generates.[15]

For evaluation, we disentangle translation quality from instruction-following by considering two metrics: (i) COMET-22 (Rei et al., 2022a), a state-of-the-art MT metric in terms of correlations with human judgments (Freitag et al., 2024) with a larger context length than XCOMET-XXL (our generated sources are large enough that this is an issue); and (ii) CLAUDE-SONNET-3.7 as a judge for evaluating the extent to which models follow the instructions. On the latter, the judge scores each instance from 1 (worst) to 6 (best), as specified in the judgment prompt. LLM-as-a-judge evaluations of these capabilities have been shown to correlate strongly with human judgments (Zheng et al., 2023; Zeng et al., 2024; Pombal et al., 2025a). For each model and evaluation dimension, we report the average score over all instances. We omit CLAUDE-SONNET-3.7 from the results since its performance would be overestimated due to intra-model family bias.

In Appendix H, we include the data generation and judgment prompts used, two examples from our benchmark, and results for English→Chinese (conclusions are similar across LPs).

3.2 Baselines

We evaluate our models against both closed-source API models and open-weight LLMs. For proprietary models, we include GPT-4O-1120 and CLAUDE-SONNET-3.7. Among open-weight models, we consider leading general-purpose LLMs with fewer than 80B parameters, as well as translation-focused models such as ALMA-R (Xu et al., 2024a), GEMMAX (Cui et al., 2025b), and TOWER-V2 (Rei et al., 2024), the winning submission of the WMT24 MT shared task. For most models, we use a standardized prompting format (see Fig. 9 in Appendix F); exceptions include translation-specific models, where we adopt the prompts recommended by their respective authors.

4 Results & Ablations

4.1 Main Results

From Table 1, we observe that the new TOWER+ models achieve a strong balance between trans-

[15] We consider only "verifiable" instructions—e.g., currency/date formatting, glossary following—as opposed to subjective ones, like style guides.

lation performance and general chat capabilities. TOWER+ 72B achieves competitive results on instruction-following benchmarks (IFEval), performing on par with leading models such as CLAUDE-SONNET-3.7 and GPT-4O-1120, and surpassing strong open-weight baselines like QWEN-2.5 on M-ArenaHard. Notably, TOWER+ 72B matches the translation performance of TOWER-V2 while substantially improving win rates against QWEN-2.5 on M-ArenaHard, from 4% to 54.5%. On IF-MT, TOWER+ 72B once again surpasses all other open models on both evaluation dimensions, showcasing its ability to leverage both translation and general capabilities.

Meanwhile, TOWER+ 9B, despite having only 9B parameters, achieves competitive performance across 24 language pairs (LPs) in machine translation and outperforms GEMMA-2 on IFEval and M-ArenaHard, and TOWER-V2 on IF-MT. It slightly underperforms its instruction-tuned counterpart, GEMMA 2 9B, on the instruction-following axis of IF-MT, though the small difference is likely attributable to noise. Crucially, the model is able to balance both general and translation capabilities at a state-of-the-art level for its size.

Finally, TOWER+ 2B, our smallest model, matches the machine translation performance of LLAMA-3.3 and outperforms TOWER-V2 on M-ArenaHard and IFEval, highlighting the effectiveness of our post-training pipeline.

We note that when comparing TOWER-V2 (a 70B model) to TOWER+ 72B, there is a slight decrease in translation quality on the subset of 15 languages originally used in WMT24. We attribute this drop not to the expanded language coverage in the new version (from 15 to 22 languages), but rather to the more limited multilingual capabilities of the QWEN 2.5 backbone. This limitation is also evident when comparing QWEN 2.5 72B Instruct with LLAMA 3.3 70B: while QWEN 2.5 72B Instruct performs strongly on general chat capabilities (with a win rate of 86.85% over LLAMA 3.3) and excels in translation for high-resource languages, its performance sharply declines when evaluated across a broader set of 15 or 22 languages. We further analyze the impact of backbone model selection in Section 4.3. Although LLAMA 3 models exhibit stronger translation capabilities, their more restrictive licensing—including mandatory attribution and naming requirements—led us to prioritize QWEN 2.5 and GEMMA 2 for this work.

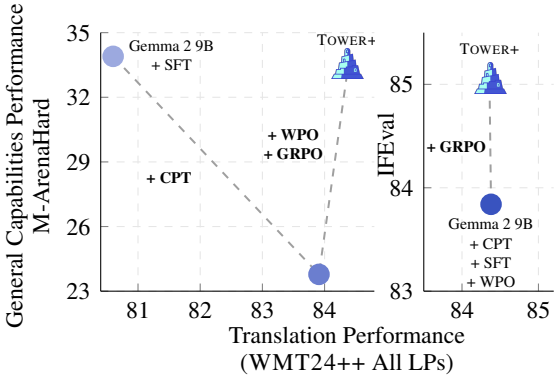

Figure 3: Performance comparison across training stages using GEMMA 2 9B as the backbone model.

All MT-specific models perform poorly on IF-MT. In fact, we had to remove the instructions from the prompt of ALMA-R and GEMMAX so that the models were able to translate, highlighting the need to create more flexible models for MT.Crucially, TOWER+ 72B greatly outperforms TOWER-V2 on this benchmark, which further speaks to the effectiveness of our approach in balancing translation quality and general capabilities.

4.2 Impact of Different Training Stages

1. **Impact of CPT:** We take a GEMMA 2 9B model and run only SFT. This complementary analysis allows us to isolate and measure the effect of the CPT phase on both general capabilities and translation performance.

2. **Impact of WPO:** We evaluate the gains introduced by the WPO stage after CPT and SFT.

3. **Impact of GRPO:** Finally, we assess the impact of RLVR by comparing the SFT+WPO model to our full pipeline (CPT+SFT+WPO+GRPO).

Figure 3 summarizes model performance at each training stage, using both general-purpose benchmarks and translation-specific evaluations.

The CPT phase significantly improves translation quality at the cost of general capabilities. Although the exact cause is difficult to pinpoint without full access to the backbone model's training details, we hypothesize that this degradation stems from disrupting the delicate balance achieved during the final pretraining annealing phases. These phases often involve carefully curated data, gradual learning rate schedules, and internal optimizations that are difficult to reproduce. Restarting training—even with high-quality data—may shift the

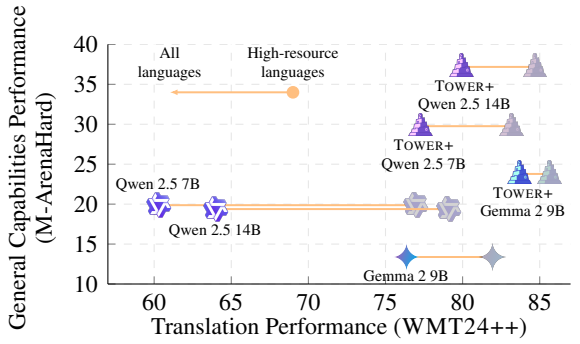

Figure 4: Comparison of GEMMA 2 9B and QWEN 2.5 7B/14B on translation quality and general chat capabilities. Arrows show the performance drop when including all languages vs high-resource ones.

data distribution away from key domains (Wang et al., 2025) such as math, code, or STEM, or reset optimizer states in a way that hurts general capabilities. Nonetheless, given our primary goal of maximizing translation quality, we accept this trade-off.

The WPO stage contributes significant improvements across instruction following, general chat ability, and translation, confirming its central role in aligning the model.

While GRPO appears promising, it is the stage where we observed the least overall gains. Improvements were primarily limited to IFEval, which aligns with the fact that the Tülu 3 dataset is specifically designed to target IFEval and GSM8k. However, more indirect signals—such as translation performance—showed no measurable improvement.

Although the gains on IFEval are initially encouraging, our analysis suggests they may stem from overfitting to the benchmark. In particular, we found that several prompts in the Tülu 3 dataset are poorly formatted, yet the verifiable reward still expects exact compliance with the instructions (see Figure 10). After cleaning the dataset to remove such inconsistencies, GRPO no longer yielded improvements over the WPO-only model.

These findings suggest that while GRPO with verifiable rewards remains a promising strategy for improving targeted model capabilities, its broader effectiveness depends heavily on the quality and structure of the reward-aligned data. Further research is needed to better integrate GRPO with VR into our post-training pipeline in a way that generalizes beyond specific benchmarks.

4.3 Importance of backbone model

Our second ablation examines the importance of backbone model selection. Since our goal is to balance strong translation support with general-purpose capabilities, we focus on two prominent model families: QWEN 2.5 and GEMMA 2. While QWEN 2.5 models demonstrate outstanding performance across a range of general-purpose benchmarks, they exhibit limited multilingual capabilities and comparatively weaker performance on translation tasks (Cui et al., 2025b). In contrast, the GEMMA 2 family achieves state-of-the-art machine translation performance among open-weight models while maintaining competitive results on general tasks. By comparing these two model families, we directly study the trade-off between a more multilingual-oriented backbone (GEMMA 2) and a model optimized for general-purpose capabilities but less robust in multilingual settings (QWEN 2.5).

To better understand this trade-off, we compare QWEN 2.5 models (both 7B and 14B) with GEMMA 2 9B by running the first two stages of our pipeline (CPT and SFT). Figure 4 illustrates the results. While QWEN 2.5 models demonstrate stronger capabilities on M-ArenaHard, they consistently lag behind in translation quality, particularly when mid- and low-resource languages are included. Notably, even with a larger parameter count, QWEN 2.5 14B fails to match the translation performance of GEMMA 2 9B. In contrast, for general chat capabilities, QWEN 2.5 7B surpasses GEMMA 2 9B despite having fewer parameters. This trend is also reflected in the released Instruct models, where QWEN 2.5 7B INSTRUCT outperforms GEMMA 2 9B INSTRUCT on M-ArenaHard but shows significantly weaker results on translation tasks.

5 Related Work

LLM post-training has become the *de-facto* strategy for achieving state-of-the-art results on MT with open-weight models. Post-training recipes often include a continued pretraining step on multilingual and parallel data, as well as a supervised fine-tuning step on MT data (e.g., ALMA (Xu et al., 2023), ALMA-R (Xu et al., 2024a)) and also chat data (e.g., TOWER (Alves et al., 2024), TOWER-V2 (Rei et al., 2024)). In these settings, data quality is crucial. As such, it has become common to adopt quality-aware decoding (Fernandes et al., 2022),

LLM-as-a-judge (Zheng et al., 2023), and rejection sampling strategies to filter out and generate data of increasingly higher quality (Zheng et al., 2025; Kocmi et al., 2025a). Some works have also included reinforcement learning stages (Zheng et al., 2025; Xu et al., 2024a, 2025) that enable learning more nuanced aspects of translation like style. The TOWER+ recipe leverages a combination of these approaches with an additional component of reinforcement learning with verifiable rewards (RLVR) to enhance translation with instruction-following capabilities. Indeed, RLVR has become widely adopted for frontier model training (e.g., Deepseek-R1 (Guo et al., 2025) and Kimi-K2 (Team et al., 2025)) for improving general capabilities. It is this mix of MT-specific and general-purpose training objectives coupled with rigorous data filtering that enables TOWER+ to compete with both state-of-the-art MT models, and general-purpose ones.

6 Conclusion

We presented a complete post-training pipeline designed to balance a task-specific use case—machine translation—with general-purpose language model capabilities. Through extensive ablations, we analyzed the contribution of each stage in our pipeline. Our final models not only outperform the Instruct-tuned versions released by the developers of the backbone models we use, but also surpass leading open-weight models such as LLAMA 3 and QWEN 2.5 72B on general-purpose benchmarks like M-ArenaHard. In translation, our models achieve results on par with frontier systems such as GPT-4 and CLAUDE 3.7, while maintaining competitive general chat abilities. We release all our models and IF-MT, the benchmark we introduce to evaluate both instruction-following and translation capabilities, to help foster future work from the community.

Limitations

While our work demonstrates the effectiveness of carefully designed post-training pipelines in balancing translation performance and general capabilities, several limitations remain.

First, as the multilingual capabilities of open-weight models continue to improve, the marginal benefit of task-specific adaptations such as CPT may decrease. Although our experiments show that even relatively multilingual backbones like GEMMA 2 9B still benefit from CPT, further research is needed to understand how these dynamics evolve as backbones become increasingly capable.

Second, while our study covers models up to 72B parameters, it remains unclear whether similar post-training pipelines would provide sufficient marginal gains for larger models. Nonetheless, we argue that for many real-world business use cases—particularly those involving well-defined tasks like translation and localization—building more efficient, specialized models remains highly valuable and preferable to deploying very large, general-purpose systems.

Third, while our pipeline improves translation quality and general chat performance, it was specifically validated for multilingual translation tasks—in the languages it was trained on—and does not explicitly optimize for other domain-specific areas such as biomedicine or legal reasoning. Nonetheless, we believe the overall recipe could serve as inspiration for designing similar post-training pipelines tailored to specific domains and languages.

Finally, while we evaluated across a wide range of languages and tasks, the available benchmarks are still skewed toward high- and mid-resource languages. Further improving performance in truly low-resource and code-switched scenarios is an important area for future exploration.

Acknowledgments

This work is supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI), by the project DECOLLAGE (ERC-2022-CoG 101088763), and by FCT/MECI through national funds and when applicable cofunded EU funds under UID/50008: Instituto de Telecomunicações.

References

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Smollm-corpus.

Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. Post-edits are preferences too.

In *Proceedings of the Ninth Conference on Machine Translation*, pages 1289–1300, Miami, Florida, USA. Association for Computational Linguistics.

Alejandro Hernández Cano, Matteo Pagliardini, Andreas Köpf, Kyle Matoba, Amirkeivan Mohtashami, Xingyao Wang, Olivia Simin Fan, Axel Marmet, Deniz Bayazit, Igor Krawczuk, Zeming Chen, Francesco Salvi, Antoine Bosselut, and Martin Jaggi. 2023. epfllm megatron-llm.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025a. Multilingual machine translation with open large language models at practical scale: An empirical study. *Preprint*, arXiv:2502.02481.

Menglong Cui, Pengzhi Gao, Wei Liu, Jian Luan, and Bin Wang. 2025b. Multilingual machine translation with open large language models at practical scale: An empirical study. *Preprint*, arXiv:2502.02481.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

DeepSeek-AI. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Daniel Deutsch, Eleftheria Briakou, Isaac 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. *Preprint*, arXiv:2502.12404.

Patrick Fernandes, António Farinhas, Ricardo Rei, José GC de Souza, Perez Ogayo, Graham Neubig, and André FT 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.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, and 1 others. 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.

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.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTav3: Improving deBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *The Eleventh International Conference on Learning Representations*.

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.

Tom Kocmi, Arkady Arkhangorodsky, Alexandre Berard, Phil Blunsom, Samuel Cahyawijaya, Théo Dehaze, Marzieh Fadaee, Nicholas Frosst, Matthias Galle, Aidan Gomez, and 1 others. 2025a. Command-a-translate: Raising the bar of machine translation with difficulty filtering. In *Proceedings of the Tenth Conference on Machine Translation*, pages 789–799.

Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, and 1 others. 2025b. Findings of the wmt25 general machine translation shared task: Time to stop evaluating on easy test sets. In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondrej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda,

Roman Grundkiewicz, and 1 others. 2024b. Findings of the wmt24 general machine translation shared task: the llm era is here but mt is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation, USA. Association for Computational Linguistics*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Shankar Kumar and Bill Byrne. 2004. Minimum bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 169–176.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, Yuling Gu, Saumya Malik, Victoria Graf, Jena D. Hwang, Jiangjiang Yang, Ronan Le Bras, Oyvind Tafjord, Chris Wilhelm, Luca Soldaini, and 4 others. 2025. Tulu 3: Pushing frontiers in open language model post-training. *Preprint*, arXiv:2411.15124.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *Preprint*, arXiv:2406.11939.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.

AI Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Manuel Faysse, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.

Ayomide Odumakinde, Daniel D'souza, Pat Verga, Beyza Ermis, and Sara Hooker. 2024. Multilingual arbitrage: Optimizing data pools to accelerate multilingual progress. *Preprint*, arXiv:2408.14960.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, and 1 others. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

José Pombal, Nuno M Guerreiro, Ricardo Rei, and André FT Martins. 2025a. Zero-shot benchmarking: A framework for flexible and scalable automatic evaluation of language models. *arXiv preprint arXiv:2504.01001*.

José Pombal, Nuno M. Guerreiro, Ricardo Rei, and André F. T. Martins. 2025b. Adding chocolate to mint: Mitigating metric interference in machine translation. *Preprint*, arXiv:2503.08327.

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

AI Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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. 2022a. 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.

Ricardo Rei, Jose Pombal, Nuno M. Guerreiro, João Alves, Pedro Henrique Martins, Patrick Fernandes, Helena Wu, Tania Vaz, Duarte Alves, Amin Farajian, Sweta Agrawal, Antonio Farinhas, José G. C. De Souza, and André Martins. 2024. Tower v2: Unbabel-IST 2024 submission for the general MT shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 185–204, Miami, Florida, USA. Association for Computational Linguistics.

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. 2022b. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, 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.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, and 14 others. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru Chen, Yuankun Chen, Yutian Chen, and 1 others. 2025. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Lewis Tunstall, Edward Emanuel Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M Rush, and Thomas Wolf. 2024. Zephyr: Direct distillation of LM alignment. In *First Conference on Language Modeling*.

Xingjin Wang, Howe Tissue, Lu Wang, Linjing Li, and Daniel Dajun Zeng. 2025. Learning dynamics in continual pre-training for large language models. *Preprint*, arXiv:2505.07796.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. 2024. Helpsteer 2: Open-source dataset for training top-performing reward models. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Rachel Wicks, Matt Post, and Philipp Koehn. 2024. Recovering document annotations for sentence-level bitext. *arXiv preprint arXiv:2406.03869*.

Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. *Advances in Neural Information Processing Systems (NeurIPS)*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models. *arXiv preprint arXiv:2309.11674*.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. 2025. X-ALMA: Plug & play modules and adaptive rejection for quality translation at scale. In *The Thirteenth International Conference on Learning Representations*.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. 2024a. Contrastive preference optimization: pushing the boundaries of llm performance in machine translation. ICML'24. JMLR.org.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Yuntian Deng, Radha Poovendran, Yejin Choi, and Bill Yuchen Lin. 2024b. Magpie: Alignment data synthesis from scratch by prompting aligned llms with nothing. *Preprint*, arXiv:2406.08464.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Mao Zheng, Zheng Li, Yang Du, Bingxin Qu, and Mingyang Song. 2025. Shy-hunyuan-mt at wmt25 general machine translation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 607–613.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *Preprint*, arXiv:2311.07911.

Wenxuan Zhou, Ravi Agrawal, Shujian Zhang, Sathish Reddy Indurthi, Sanqiang Zhao, Kaiqiang Song, Silei Xu, and Chenguang Zhu. 2024. WPO: Enhancing RLHF with weighted preference optimization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8328–8340, Miami, Florida, USA. Association for Computational Linguistics.

Appendix

A Additional Training Diagrams

B Additional Hyperparameters

For **Continued Pretraining**, we train our models with a codebase based on Megatron-LLM (Cano et al., 2023) on 8 H100-80GB GPUs, using an effective batch size of 3.14 million tokens per gradient step, a sequence length of 8192, and a constant learning rate of 3×10^{-5}, decaying linearly to 0 from the last 10% of training onwards. During **Supervised Fine-tuning**, all models used a sequence length of 8192 and trained for 3 epochs with 125 warmup steps. The 2B and 9B models used a batch size of 128, while the 72B model used 256. The learning rate was 2.0×10^{-7} for the 2B and 9B models, and 1.0×10^{-7} for the 72B. All configurations used the `adamw_torch` optimizer, `cosine` learning rate scheduler, and `zero3` from DeepSpeed. For **WPO**, all models trained for 1 epoch with a batch size of 32 and 220 warmup steps (10% of total steps). The learning rate was 2.0×10^{-7} for the 2B and 9B models, and 1.0×10^{-7} for the 72B. The optimizer remained `adamw_torch`, the scheduler was `constant_with_warmup`, and all used `zero3` with DPO weighting enabled and $\beta = 5$. Sequence length was 8192 for the 2B and 9B models, and 2048 for the 72B. For **GRPO**, only the 2B model was trained, using a maximum sequence length of 3072 and prompt length of 2048. It generated 7 samples per prompt with a batch size of $56 = 7 \times 8$. Training lasted 1 epoch with a warmup ratio of 0.1, maximum gradient norm of 0.2, learning rate of 2.0×10^{-7}, and temperature of 1.0. The optimizer was `adamw_torch` and the scheduler `constant_with_warmup`. The 9B and 72B models were not trained with GRPO.

C Covered languages

TOWER models presented in this paper cover the following 27 languages/dialects: German, Spanish, Spanish (Latin America), French, Italian, Korean, Dutch, Russian, English, Portuguese (Portugal), Portuguese (Brazilian), , Chinese (Simplified), Chinese (Traditional), Czech, Ukrainian, Hindi, Icelandic, Japanese, Polish, Swedish, Hungarian, Romanian, Danish, Norwegian (Nynorsk), Norwegian (Bokmål), Finnish.

Table 2, 3 and 4 show results for all supported languages from WMT24++ testset (Deutsch et al.,

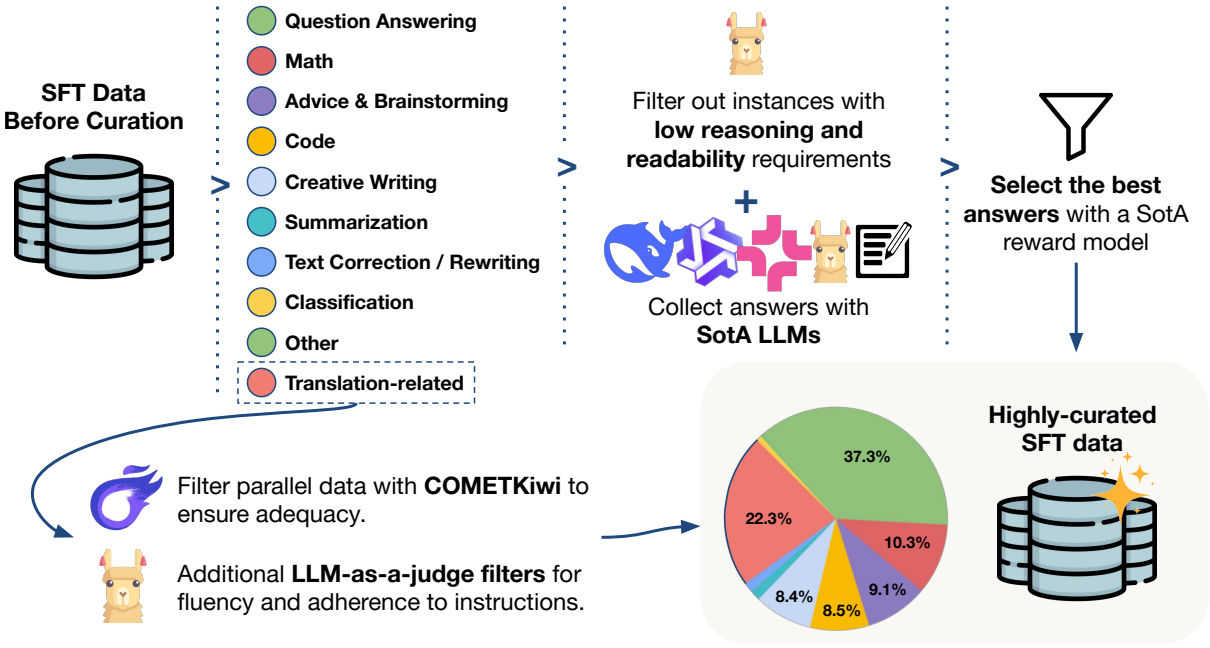

Figure 5: Process for creating and curating data for our final dataset for SFT.

2025) using CHRF (Popović, 2015), METRICX24-XXL (Juraska et al., 2024) and XCOMET-XXL respectively.

D Translation Templates used in CPT

This section presents template examples used to prepare parallel data for continued pre-training. We used hundreds of such templates created with `Jinja2`.

Placeholders:

- `{{ source }}`: source sentence
- `{{ target }}`: target sentence
- `{{ lp0 }}`: source language name
- `{{ lp1 }}`: target language name

Figure 7 shows 8 examples of templates used to create both CPT and SFT data.

E Prompts used to clean SFT data

In his section you can find the prompt used to clean and prepare the SFT data. Figure 8 presents the prompt we used on LLAMA-3.3 to assign scores for reasoning and readability along with a category for each conversation.

F Generation of Verifiable Translation Instructions for Training

F.1 Broad topics for data generation

Below we present the list of broad transformation categories for verifiable translation instructions: Email Formatting, Phone Numbers, Mathematical Notation, Code Elements, Brand Elements, Time Formatting, List Formatting, Links and URLs, Case Formatting, Number Formatting, Date Formatting, Special Characters, Text Structure, Term Formatting, Units and Measurements, Citation and References, Chemical Formulas, Temperature, Version Control, Geographic Coordinates, Technical Specifications, Financial Notation, Location Codes, Emoji Substitutions, Social Media Formatting, Music, Legal References, Sports Statistics, Music Notation, File Paths, Research Citations

F.2 Prompts for data generation and verification

G Problems Found in Tülu RLVR Datasets

As discussed in Section 4.2, the GRPO stage led to improvements only on IFEval. Upon closer inspection, we found that the Tülu RLVR dataset was constructed by appending artificial suffixes and prefixes to existing prompts in order to make them verifiable. However, this process introduces several

Models	Avg. 7	Avg. 15	Avg.	pt_BR	pt_PT	zh_CN	zh_TW	cs_CZ	da_DK	nl_NL	fi_FI	fr_FR	de_DE	hi_IN	hu_HU	is_IS	it_IT	ja_JP	ko_KR	no_NO	pl_PL	ro_RO	ru_RU	es_MX	sv_SE	uk_UA
GPT-4O-1120	60.11	53.96	55.19	65.99	63.21	39.80	32.36	57.11	66.36	62.12	62.74	65.03	61.23	40.68	56.38	49.57	66.99	34.68	34.26	51.75	49.90	64.65	53.11	68.65	66.54	56.22
CLAUDE-SONNET-3.7	60.93	55.29	56.60	66.56	63.21	40.66	34.14	58.38	68.13	62.56	64.99	65.78	61.64	40.71	56.93	50.33	68.13	38.10	39.19	54.76	52.26	65.68	54.72	69.04	67.61	58.26
ALMA-R	48.97	39.98	36.41	52.99	50.44	27.67	13.93	46.09	20.69	48.88	31.14	54.02	54.17	7.84	29.52	42.05	52.12	21.47	11.78	21.20	33.92	46.73	46.53	55.33	30.06	38.77
GEMMAX	57.83	47.58	43.11	63.76	59.84	39.12	19.42	53.19	44.21	60.26	17.01	62.00	59.57	39.35	15.01	13.07	65.15	33.34	34.50	46.38	50.02	29.30	50.12	65.07	44.20	27.63
GEMMA-2	57.37	49.77	50.85	63.53	60.29	35.98	30.14	50.69	61.20	58.21	55.77	62.64	58.10	37.94	48.18	32.75	64.15	32.59	31.20	48.83	47.53	59.43	50.47	66.72	61.32	51.77
GEMMA-2	58.67	51.69	52.85	64.78	61.75	38.61	32.24	53.13	63.95	60.34	57.13	63.62	59.21	39.05	51.23	38.60	65.46	34.32	32.95	51.48	48.86	61.77	51.49	67.50	63.37	54.65
QWEN-2.5	57.60	50.04	50.62	64.17	61.22	40.18	32.92	50.95	59.35	59.00	51.74	63.43	56.62	36.35	44.18	34.52	64.21	33.67	33.31	48.19	47.74	58.25	47.99	66.59	60.14	49.57
LLAMA-3.3	56.94	50.22	51.84	64.45	60.62	28.17	27.04	53.36	63.12	60.43	58.33	63.12	58.72	38.44	52.64	40.72	65.43	32.88	26.66	52.15	49.17	62.06	51.48	67.21	64.14	51.99
TOWER-V2	59.71	53.71	53.72	65.64	62.52	40.43	22.24	55.82	62.46	62.18	58.09	63.82	60.86	40.29	52.82	48.64	66.90	34.79	35.60	50.30	49.20	62.57	52.72	67.61	63.50	56.60
TOWER+ 2B	56.35	50.46	51.81	63.31	59.93	35.67	30.51	51.76	63.75	59.08	57.84	61.24	57.48	38.63	50.31	45.91	63.42	32.46	32.41	50.98	48.15	61.58	47.81	65.54	62.21	51.70
TOWER+ 9B	59.02	53.37	54.60	65.06	62.28	39.67	32.75	55.44	65.63	60.96	61.05	63.19	59.75	40.25	53.78	48.94	65.56	36.08	36.34	53.81	50.23	63.90	52.31	67.63	65.02	56.06
TOWER+ 72B	59.83	53.61	54.46	65.50	62.41	42.52	32.92	55.10	65.25	61.04	59.68	63.64	60.04	39.66	52.41	47.25	66.10	35.95	37.12	53.04	49.61	62.55	53.16	67.83	64.24	55.62

Table 2: ChrF (↑) results for WMT24++ on all tested language pairs, including the averages across high-resource (in blue), the additional languages/dialects supported by TOWER-V2 (Rei et al., 2024) (in green), and all languages supported by our new models (in orange). Note that these averages are cumulative meaning that, "Avg. 15" includes also the languages from "Avg. 7"

Models	Avg. 7	Avg. 15	Avg.	pt_BR	pt_PT	zh_CN	zh_TW	cs_CZ	da_DK	nl_NL	fi_FI	fr_FR	de_DE	hi_IN	hu_HU	is_IS	it_IT	ja_JP	ko_KR	no_NO	pl_PL	ro_RO	ru_RU	es_MX	sv_SE	uk_UA
GPT-4o-1120	-4.00	-4.53	-4.81	-4.65	-5.13	-3.10	-2.85	-6.12	-4.68	-3.77	-6.02	-4.51	-2.64	-4.28	-6.58	-7.46	-4.29	-4.37	-4.42	-5.66	-6.37	-5.62	-4.60	-4.25	-4.34	-5.00
CLAUDE-SONNET-3.7	-4.10	-4.63	-4.92	-4.86	-5.20	-3.18	-2.85	-6.28	-4.70	-3.82	-6.06	-4.60	-2.74	-4.38	-6.74	-7.51	-4.38	-4.55	-4.44	-5.82	-6.57	-5.84	-4.53	-4.40	-4.50	-5.12
ALMA-R	-4.58	-5.83	-6.82	-5.80	-5.99	-3.33	-3.03	-6.62	-6.51	-5.34	-13.58	-5.23	-2.87	-7.49	-13.66	-7.42	-5.24	-6.29	-9.74	-6.02	-9.84	-8.97	-4.91	-4.72	-7.72	-6.58
GEMMAX	-4.46	-5.10	-5.83	-5.33	-5.36	-3.54	-2.93	-7.10	-6.84	-3.98	-9.89	-4.83	-3.21	-5.21	-10.12	-8.16	-4.45	-5.36	-5.12	-6.31	-7.19	-7.48	-5.54	-4.32	-6.57	-5.25
GEMMA-2	-4.69	-6.05	-6.26	-5.33	-5.75	-3.60	-3.30	-8.00	-5.79	-4.74	-8.52	-5.22	-3.42	-5.33	-8.88	-16.47	-4.93	-5.26	-5.93	-7.00	-7.63	-6.90	-5.67	-4.69	-5.46	-6.10
GEMMA-2	-3.91	-4.68	-4.95	-4.60	-5.09	-3.07	-2.83	-6.39	-4.51	-3.74	-6.17	-4.44	-2.66	-4.38	-7.33	-9.65	-4.14	-4.55	-4.45	-5.94	-6.50	-5.77	-4.42	-4.04	-4.28	-4.99
QWEN-2.5	-4.39	-5.89	-6.47	-5.11	-5.42	-3.27	-3.01	-7.84	-6.53	-4.42	-10.15	-4.72	-3.22	-5.57	-10.64	-17.19	-4.70	-4.92	-5.19	-8.90	-7.77	-8.01	-5.16	-4.54	-5.86	-6.57
LLAMA-3.3	-4.80	-5.71	-5.94	-5.56	-5.91	-3.77	-3.26	-7.64	-5.68	-4.31	-7.78	-5.26	-3.39	-5.13	-7.75	-12.18	-4.98	-5.23	-5.45	-6.94	-7.59	-6.60	-5.83	-4.84	-5.13	-6.34
TOWER-V2	-4.00	-4.53	-4.95	-4.91	-5.08	-3.13	-2.76	-6.12	-4.89	-3.58	-6.91	-4.42	-2.62	-4.86	-6.99	-6.81	-4.17	-4.57	-4.55	-7.02	-6.73	-5.62	-4.71	-4.03	-4.36	-4.91
TOWER+ 2B	-4.86	-5.48	-5.82	-5.44	-5.72	-3.70	-3.33	-7.92	-5.43	-4.47	-7.90	-5.33	-3.53	-5.28	-8.40	-8.30	-5.04	-5.21	-5.53	-6.81	-7.57	-6.66	-6.21	-4.77	-5.18	-6.04
TOWER+ 9B	-4.11	-4.70	-4.98	-4.86	-5.10	-3.25	-2.94	-6.55	-5.02	-3.76	-6.32	-4.57	-2.77	-4.79	-7.00	-7.28	-4.21	-4.71	-4.82	-5.99	-6.30	-5.65	-4.95	-4.15	-4.43	-5.09
TOWER+ 72B	-3.97	-4.64	-5.00	-4.88	-4.89	-3.09	-2.84	-6.54	-4.93	-3.61	-7.01	-4.38	-2.76	-4.91	-7.44	-7.66	-4.09	-4.58	-4.69	-6.21	-6.32	-5.82	-4.61	-3.95	-4.66	-5.20

Table 3: MetricX24 XXL (↑) results for WMT24++ on all tested language pairs, including the averages across high-resource (in blue), the additional languages/dialects supported by TOWER-V2 (Rei et al., 2024) (in green), and all languages supported by our new models (in orange). Note that these averages are cumulative meaning that, "Avg. 15" includes also the languages from "Avg. 7"

Models	Avg. 7	Avg. 15	Avg.	pt_BR	pt_PT	zh_CN	zh_TW	cs_CZ	da_DK	nl_NL	fi_FI	fr_FR	de_DE	hi_IN	hu_HU	is_IS	it_IT	ja_JP	ko_KR	no_NO	pl_PL	ro_RO	ru_RU	es_MX	sv_SE	uk_UA
GPT-4o-1120	86.69	84.33	85.21	0.8806	0.8711	0.8386	0.8330	0.8189	0.8951	0.8994	0.8789	0.8427	0.9329	0.7197	0.8674	0.7815	0.8641	0.8417	0.8513	0.8482	0.8327	0.8543	0.8300	0.8797	0.9107	0.8251
CLAUDE-SONNET-3.7	86.41	84.24	85.19	0.8758	0.8700	0.8401	0.8368	0.8170	0.8987	0.9008	0.8829	0.8379	0.9264	0.7040	0.8645	0.7836	0.8606	0.8550	0.8619	0.8613	0.8381	0.8426	0.8358	0.8724	0.9057	0.8216
ALMA-R	80.97	72.76	68.35	0.7955	0.7926	0.7865	0.7549	0.7558	0.6559	0.7904	0.4172	0.7527	0.9153	0.4942	0.4433	0.7277	0.7888	0.5976	0.4631	0.6748	0.5950	0.5639	0.8085	0.8204	0.6369	0.6895
GEMMAX	84.26	78.74	75.66	0.8556	0.8504	0.8129	0.7932	0.7712	0.7492	0.8902	0.5059	0.8076	0.9132	0.6109	0.5279	0.4904	0.8555	0.7782	0.8144	0.7739	0.8072	0.6088	0.7907	0.8627	0.7629	0.7697
GEMMA-2	81.93	75.35	76.34	0.8376	0.8253	0.8009	0.7943	0.7010	0.8227	0.8413	0.7131	0.7778	0.8969	0.6059	0.7357	0.4078	0.8098	0.7656	0.7484	0.7914	0.7592	0.7305	0.7709	0.8409	0.8380	0.7443
GEMMA-2	83.68	79.02	80.18	0.8511	0.8363	0.8167	0.8190	0.7504	0.8607	0.8746	0.7799	0.7978	0.9105	0.6623	0.8037	0.5265	0.8309	0.7993	0.8019	0.8311	0.7994	0.7808	0.7980	0.8524	0.8688	0.7902
QWEN-2.5	76.95	64.16	60.26	0.8050	0.7978	0.7850	0.7375	0.4957	0.5863	0.7463	0.3368	0.7266	0.8490	0.3333	0.3111	0.2717	0.7499	0.6451	0.6167	0.4949	0.5476	0.4346	0.6741	0.7968	0.6295	0.4874
LLAMA-3.3	82.74	78.30	79.48	0.8407	0.8317	0.7959	0.7986	0.7396	0.8517	0.8693	0.7804	0.7907	0.9084	0.6697	0.8041	0.5859	0.8293	0.7867	0.7657	0.8092	0.7800	0.7931	0.7815	0.8455	0.8697	0.7537
TOWER-V2	86.40	83.88	83.74	0.8722	0.8716	0.8358	0.8146	0.8151	0.8771	0.9066	0.8160	0.8325	0.9312	0.6502	0.8344	0.7970	0.8658	0.8452	0.8491	0.7538	0.8195	0.8353	0.8271	0.8834	0.8952	0.8316
TOWER+ 2B	81.88	78.42	79.13	0.8410	0.8354	0.7938	0.7891	0.7246	0.8501	0.8598	0.7656	0.7701	0.8988	0.6067	0.7653	0.7314	0.8194	0.7796	0.7807	0.8091	0.7752	0.7775	0.7630	0.8457	0.8545	0.7638
TOWER+ 9B	86.25	83.57	84.38	0.8725	0.8686	0.8379	0.8291	0.7983	0.8810	0.8946	0.8491	0.8299	0.9287	0.6719	0.8432	0.7921	0.8704	0.8391	0.8420	0.8520	0.8447	0.8414	0.8219	0.8764	0.8992	0.8237
TOWER+ 72B	86.68	83.29	83.74	0.8723	0.8740	0.8466	0.8384	0.7930	0.8807	0.8996	0.8152	0.8395	0.9276	0.6442	0.8044	0.7571	0.8698	0.8453	0.8456	0.8371	0.8395	0.8264	0.8285	0.8828	0.8850	0.8083

Table 4: XCOMET (↑) results for WMT24++ on all tested language pairs, including the averages across high-resource (in blue), the additional languages/dialects supported by TOWER-V2 (Rei et al., 2024) (in green), and all languages supported by our new models (in orange). Note that these averages are cumulative meaning that, "Avg. 15" includes also the languages from "Avg. 7"

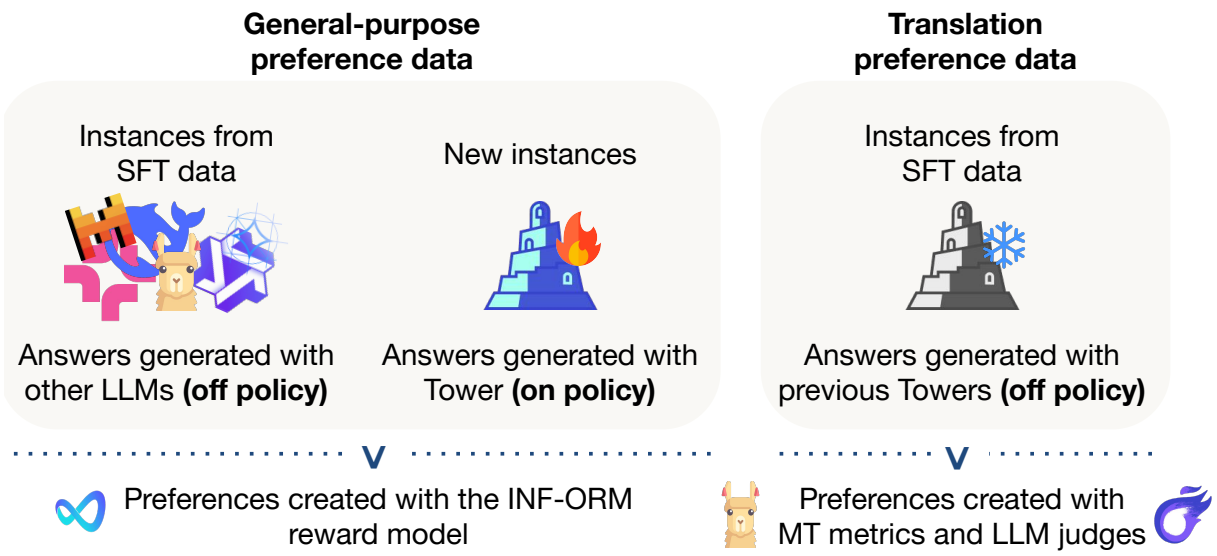

Figure 6: Process for curating data for preference optimization.

Translation Templates used for both CPT and SFT

```
Source:  {{ source }}
Translate the source text from {{ lp0 }} to {{ lp1 }}.
Target: {{ target }}


Source:  {{ source }}
Translate from {{ lp0 }} to {{ lp1 }}.
Target:  {{ target }}


Write the text in {{ lp0 }} in {{ lp1 }}.
Text:  {{ source }}
Target:  {{ target }}


Translate the following text from {{ lp0 }} to {{ lp1 }}:
Text:  {{ source }}
Translation:  {{ target }}


Translate the following {{ lp0 }} source text to {{ lp1 }}:
{{ lp0 }}:  {{ source }}
{{ lp1 }}:  {{ target }}


Please translate this text from {{ lp0 }} into {{ lp1 }}.
{{ lp0 }}:  {{ source }}
{{ lp1 }}:  {{ target }}


Make a translation of the given text from {{ lp0 }} to {{ lp1 }}.
{{ lp0 }}:  {{ source }}
{{ lp1 }}:  {{ target }}


{{ lp0 }}:  {{ source }}
Translate the {{ lp0 }} text above into {{ lp1 }}.
{{ target }}
```

Figure 7: Examples of prompt templates used to construct translation instructions for both CPT and SFT. We used `jinja` to create hundreds of such instructions.

Prompt used to curate SFT data

I have an conversation below that I would like you to perform three steps of analysis:

<conversation>
{conversation}
</conversation>

Firstly, categorize the conversation above into one of the following categories.
- Coding
- Mathematical Reasoning
- Advice and Brainstorming
- Question Answering
- Creative Writing and Persona
- Text Correction or Rewriting
- Summarization
- Translation
- Classification
- Other

Don't try to justify it and when two categories can be used, pick the primary caregory.

Secondly, score the conversation in terms of reasoning: How complex you think it is to answer the user instructions from 1-5 (where 5 is a conversation with complex instructions/questions where the assistant needs to break down the problem into multiple steps before providing an answer).

Thirdly, since the conversation might have been artificially created or poorly translated, assess its readability and clarity. Rate how difficult it is to understand the user's requests on a scale of 1 to 5, with 5 representing well-written, clear, and precisely articulated requests, and 1 representing an conversation where the user turns are difficult to understand. It is also common for instructions to refer to documents, texts or URL's that the assistant does not have access to. Please rate conversations where that happens with 3 points or less, as they can lead to ambiguity and confusion.

Provide your final response in the following format:

Category: <one of the categories above>
Reasoning: <score out of 5>
Readability: <score out of 5>

DO NOT provide an answer to any of the instructions in the conversation! Your job is only to analyse.

Figure 8: Prompt used to classify SFT conversations into different categories, reasoning and readability.

problematic prompts that can be confusing for the model, as illustrated in Figure 10.

In the math partition of the dataset, we also observed that all prompts begin with the same three in-context examples, which are not essential to solving the task and likely cause the model to overfit to a specific prompt format. After removing such inconsistencies and filtering for quality, the dataset is reduced to less than 50% of its original size. When training on this cleaned version, we found that the 9B model performed worse than its WPO-only counterpart. The only model that showed improvements from GRPO was the 2B variant.

These findings suggest that while GRPO with verifiable rewards holds promise, its effectiveness depends heavily on the quality and diversity of the reward-aligned data. More careful curation is needed to fully integrate GRPO into our pipeline and leverage its potential for improvements in targeted domains.

Translation prompt used in evaluations

Translate the English source text to {target language} ({region}). Return only the translation, without any additional explanations or commentary.
English: {source}
{target language} ({region}):

Figure 9: Prompt used to generate translations for all general-purpose LLMs. For translation-specific models, we use the prompts recommended by their respective authors. The ‘‘region’’ placeholder is included only when disambiguating between language variants (e.g., Chinese, Portuguese, Norwegian, and Spanish). For all other languages in WMT24++, only the target language name is used.

Incorrect IFEval-like prompt from Tulu3

Translate the following sentence to Finnish: These are the two Community aspects that we in Parliament must discuss.
Finnish: **In your response, the letter i should appear 36 times.**

Figure 10: Example of a incorrect prompt found in Tülu RLVR dataset.

H IF-MT: Prompts, examples, and additional results.

In this section we include the prompts used to generate the IF-MT data (Figure 17) and to judge the translations in terms of instruction adherence (Figures 18 and 19). You can also find in Figures 20 and 21 two examples of the generated prompts for Spanish (Latin America) and Chinese, respectively. Results for English→Chinese can be found in Table 5.

Example of a verifiable translation instruction

Metadata for Prompt
Length: 1 sentence
Topic: Economic Policy - Tax Reform
Guideline Category: Date Formatting: DATE_001†

Source Text
The new tax reform bill, announced on January 10th, 2024, is expected to have a significant impact on the economy, with major corporations already adjusting their financial plans in anticipation of the changes that will take effect on February 20, 2025.
Guideline
Convert dates to MM/DD/YYYY.

Figure 11: Examples of a verifiable translation instruction. †The guideline category "DATE_001" is shown in Appendix Figure 13.

LLM-as-a-Judge translation preference prompt

Please act as an impartial judge and evaluate the quality of the translations provided by two AI assistants in response to the user's request below. Select the assistant that best adheres to the user's instructions while producing the highest-quality translation overall. If the user's instructions specify particular factors—such as the required level of formality, glossaries, or adherence to provided examples—ensure these are included in your evaluation. Begin by comparing the two translations and provide a concise explanation of your assessment. Avoid personal opinions or biases, and do not favour one assistant over the other. Be objective and impartial.

After providing your explanation, deliver your final verdict strictly in this format:

Chosen: <[A] if Assistant A is better, [B] if Assistant B is better, or [T] if both are equally good or bad.>

[User Instruction]
{instruction}
[End of User Instruction]

[Start of Assistant A's Response]
{assistant a response}
[End of Assistant A's Response]

[Start of Assistant B's Response]
{assistant b response}
[End of Assistant B's Response]

Figure 12: Translation LLM-as-a-judge prompt used to validate the ‘chosen‘ and ‘rejected‘ answers after MBR. All samples where the LLM-judge disagrees with the ‘chosen‘ and ‘rejected‘ are discarded.

Example of a template for verifiable guidelines

Date Formatting

- ID: DATE_001
- NAME: MM/DD/YYYY Format
- DESCRIPTION: Convert dates to MM/DD/YYYY
- REQUIRES EXAMPLE: False
- VERIFICATION:

  `\b(0[1-9]|1[0-2])/(0[1-9]|[12]\d|3[01])/\d{4}\b`

- EXAMPLE INPUT: "January 5, 2024"
- EXAMPLE OUTPUT: "01/05/2024

Figure 13: Examples of a transformation template for generation of translation-verifiable instructions.

Prompt for generation of verifiable translation instructions for training

Requirements:

- The document must be **EXACTLY** the specified length
- Must naturally incorporate elements that match **ALL** guidelines
- Keep the text coherent and natural
- For paragraphs, use 2–3 sentences per paragraph
- Do not mention the guidelines explicitly in the text

Output format:

- `###SOURCE###`
  `[Your text here]`
- `###GUIDELINES###`
  `[Copy the given guidelines exactly]`
- `###END###`

Here are two examples:

Example 1:

- **LENGTH:** 1 sentence
- **TOPIC:** Technology - Software Development
- **GUIDELINES:**
  1) **[Date Formatting]** Convert dates to MM/DD/YYYY
  2) **[Terminology]** Add full form in parentheses after acronyms

```
###SOURCE###
The AI team announced on March 15th that their new NLP system had
achieved breakthrough performance in code generation.
###GUIDELINES###
1) Convert dates to MM/DD/YYYY, e.g., March 15th to 03/15/2022
2) Add full form 'Natural Language Processing' in parentheses after
acronyms, e.g., NLP (Natural Language Processing)
###END###
```

Example 2:

- **LENGTH:** 1 paragraph
- **TOPIC:** Social Media - Digital Marketing
- **GUIDELINES:**
  1) **[Case Formatting]** Convert all text to lowercase
  2) **[Social Media]** Add hashtags at end of sentence for: brands (#brand), actions (#marketing)
  3) **[Email Formatting]** Convert email mentions to [EMAIL]address[/EMAIL]

```
###SOURCE###
Instagram and TikTok launched new advertising features last week.
Digital marketers can now contact our support team at help@instagram.com
for early access to these tools, while brands on TikTok are already
reporting increased engagement rates.
###GUIDELINES###
1) Convert all text to lowercase
2) Add hashtags at end of sentence for:  brands (#brand), actions
(#marketing)
3) Convert email mentions to [EMAIL]address[/EMAIL]
###END###
```

Your task:

- **LENGTH:** `{length}`
- **TOPIC:** `{topic}`
- **GUIDELINES:** `{guideline_txt}`

Important Instructions for Source Text:

1. Write a text that contains all necessary elements that **COULD** be transformed according to the guidelines, but deliberately does **NOT** follow the guidelines yet
2. For example:
   - If a guideline requires formatting dates as MM/DD/YYYY, write dates in a different format
   - If a guideline requires wrapping emails in tags, include email addresses without tags
   - If a guideline requires expanding acronyms, use acronyms without expansions
3. The text should be natural and coherent, reading as a normal document would
4. Make sure every guideline has corresponding elements in the text that can be transformed
5. Think of the source text as the "**before**" version that will later be transformed into an "**after**" version following the guideline

Figure 14: Prompt for generation of verifiable translation instructions.

LLM-as-a-Judge prompt for verification of verifiable translation instructions

```
You are an expert judge evaluating source documents that will be used for guideline-based text
rewriting tasks. Your task is to carefully analyze whether a text follows any given guidelines. First,
analyze each guideline carefully, then decide if the text follows ANY of the guidelines.
Example 1:
Guidelines:
1) [Email Format] Convert email to [EMAIL]address[/EMAIL]
2) [Case] Convert product names to UPPERCASE
Source Text: Contact us at help@company.com about the zenith software.

###EVALUATION###
Analysis:
Guideline 1 (Email Format):
- Text contains raw email "help@company.com"
- Email is NOT wrapped in [EMAIL] tags
- This guideline is NOT followed

Guideline 2 (Case):
- Text contains product name "zenith"
- Product name is in lowercase
- This guideline is NOT followed

Number of guidelines followed: 0/2
Guidelines Check: 0
###END###

Example 2:
Guidelines:
1) [Email Format] Convert email to [EMAIL]address[/EMAIL]
2) [Case] Convert product names to UPPERCASE
Source Text: Contact us at [EMAIL]help@company.com[/EMAIL] about the ZENITH software.

###EVALUATION###
Analysis:
Guideline 1 (Email Format):
- Text contains email wrapped in [EMAIL] tags [EMAIL]help@company.com[/EMAIL]
- This guideline is FOLLOWED

Guideline 2 (Case):
- Text contains product name "ZENITH" in UPPERCASE
- This guideline is FOLLOWED

Number of guidelines followed: 2/2
Guidelines Check: 1
###END###

Example 3:
Guidelines:
1) Convert month names to 3 letter abbreviations
2) Convert lists to 1., 2., format
Source Text: The meeting is scheduled for January 1st, 2023. The agenda includes: 1) Budget
review, 2) Project updates.

###EVALUATION###
Analysis:
Guideline 1 (Month Abbreviations):
- Text contains full month name "January"
- Month is NOT in 3-letter format (should be "Jan")
- This guideline is NOT followed

Guideline 2 (List Format):
- Text contains list with format "1)" and "2)"
- Lists are NOT in "1." format
- This guideline is NOT followed

Number of guidelines followed: 0/2
Guidelines Check: 0
###END###

Now evaluate this input:
Topic: {topic}
Length: {length}
Guidelines: {guidelines}
Source Text: {source_text}

Your evaluation must:
1. Analyze each guideline separately and explicitly state if it's followed
2. Count the total guidelines followed
3. Conclude with a Guidelines Check score: Score 1 if ANY guideline is followed; Score 0 if NO
guidelines are followed

Use exactly this format:
###EVALUATION###
Analysis: (Analysis of each guideline)
Number of guidelines followed: [X/Y] — there is no such a thing as half a guideline, so X should
be an integer between 0 and Y (also an integer)
Guidelines Check: [1 for ANY followed, 0 for NONE followed]
###END###
```

Figure 15: Prompt for verification of verifiable translation instructions.

LLM-as-a-Judge translation scoring prompt

You are a professional translator and evaluator. Your task is to evaluate how well an assistant has handled a translation request from a user. Evaluate the translation based on the following criteria:
1. **Adequacy (Accuracy of Meaning)** - Assess whether the translation fully and accurately conveys the meaning of the source text. - Penalize mistranslations, omissions, or additions that distort the intended message.
2. **Fluency (Readability & Grammar)** - Ensure the translation reads naturally and is grammatically correct in the target language. - Penalize awkward phrasing, unnatural word choices, or structural issues. - It should be easy to read and understand, as if it were originally written in the target language.
3. **Cultural Appropriateness** - Ensure that the translation is culturally appropriate for the target audience.
4. **Instructions Adherence** - If provided, evaluate how well the translation adheres to any specific instructions or guidelines provided by the user. Otherwise, ignore this criterion.

Provide detailed feedback on any issues and suggest improvements. Conclude with a **score from 1 to 5**:

- **5** → Perfect translation (fully accurate and fluent in the target language while adhering to instructions).
- **4** → Good translation (minor errors but generally fluent and natural sounding) and adheres to instructions.
- **3** → Acceptable but flawed (some errors in meaning, fluency, or structure).
- **2** → Translation is acceptable but it does not adhere to the instructions.
- **1** → Poor translation (major errors affecting comprehension).

[User Instructions]
{instruction}
[End of User Instructions]

[Assistant Translation]
{answer}
[End of Assistant Translation]

NOTE: Your answer must terminate with the following format: **Final Score:** <score>

Figure 16: LLM-as-a-judge prompt used to score Translation data for supervised fine-tuning.

IF-MT meta prompt for data generation.

As an expert prompt engineer, create a detailed prompt for a language model to perform the following task: translation of a source text, given a set of ${n_rules} rules. The source text should abide by the followin parameters:
- Source language: ${source_language}
- Topic: ${topic}
- Subtopic: ${subtopic}
- Style: ${style}
- Source length: ${source_length}

The translation should be in ${target_language}, and your generated prompt must specify a set of ${n_rules} rules.

IMPORTANT: These rules must be objectively verifiable and should be clearly stated in the prompt. The language model should be instructed to follow these rules when translating the source text. An example of a verifiable rule is "Convert dates to the format DD/MM/YYYY."; an example of an unverifiable rule is "Make the translation sound more professional.". Keep in mind that the rules should make sense in the context of the source text and the target language.

IMPORTANT: Make sure that the source you create has elements that correspond to the rules you set.

To demonstrate the expected output, also provide a reference translation following the requested requirements at the end.

IMPORTANT: Your response should be structured as follows:

<START OF PROMPT>
[INSERT ONLY THE PROMPT HERE COMBINING SOURCE, RULES, AND AN INSTRUCTION. REMIND THE MODEL TO RETURN ONLY THE TRANSLATION. NOTHING ELSE.]
<END OF PROMPT>

<START OF REFERENCE>
[INSERT ONLY THE REFERENCE TRANSLATION. NOTHING ELSE.]
<END OF REFERENCE>

ABIDE STRICTLY BY THE REQUESTED FORMAT.

Figure 17: IF-MT meta prompt for data generation. We sample attributes (e.g., topic, subtopic) from the lists provided by Pombal et al. (2025a). We ask the model for 2 to 4 rules.

IF-MT judgement prompt: Part I

You are an expert judge evaluating translation quality. You will be presented with:

- A text, prompting a model for a translation of a source following some rules
- A translation to evaluate

Rate the translation on a scale of 1-6 based on how well it follows the specified rules and instructions in the prompt, regardless of overall translation quality, according to the following criteria:
- Rule Adherence: Does the translation follow all explicit rules stated in the prompt?
- Instruction Compliance: Are specific formatting, style, or technical instructions followed?
- Constraint Observance: Are any limitations or restrictions properly respected?
- Specification Accuracy: Does the output match the exact specifications requested?
- Requirement Fulfillment: Are all mandatory elements present as instructed?

Scoring Rubric:
6 - Perfect Compliance

- Follows every single rule and instruction precisely
- No deviations from any specified constraints
- All requirements fully met as requested
- Complete adherence to formatting/style directives
- Perfect execution of all procedural instructions
- Zero rule violations of any kind

5 - Excellent Compliance

- Follows nearly all rules with only trivial deviations
- Minor lapses that don't affect core requirements
- Strong adherence to most constraints and directives
- Formatting/style mostly correct
- Very few rule violations, all inconsequential

4 - Good Compliance
- Follows most important rules correctly
- Some minor rule violations that don't undermine main objectives
- Generally respects constraints and limitations
- Adequate adherence to formatting requirements
- Few significant rule violations

Figure 18: IF-MT judgement prompt. We follow the 1-to-6 direct assessment approach of Pombal et al. (2025a) due to its reported effectiveness (part 1/2).

IF-MT judgement prompt: Part II

3 - Fair Compliance
- Follows some rules but misses several others
- Notable violations of stated constraints
- Inconsistent adherence to instructions
- Some formatting/style requirements ignored
- Multiple rule violations affecting compliance

2 - Poor Compliance
- Fails to follow many stated rules
- Significant violations of constraints and limitations
- Poor adherence to specific instructions
- Formatting/style requirements largely ignored
- Frequent and notable rule violations

1 - No Compliance
- Ignores most or all stated rules
- Complete disregard for constraints and limitations
- Fails to follow basic instructions
- No attention to specified requirements
- Systematic rule violations throughout

Provide your evaluation in this JSON format:

{"feedback": "<detailed explanation of the score based on the criteria>", "result": "<only a number from 1 to 6>"}

<START OF SOURCE TEXT>
${prompt}
<END OF SOURCE TEXT>

<START OF TRANSLATION>
${answer}
<END OF TRANSLATION>

You may proceed to evaluate the translation. Focus on evaluating the extent to which the translation follows the rules in the prompt, not its quality. Ensure the output is valid JSON, without additional formatting or explanations.

Figure 19: IF-MT judgement prompt. We follow the 1-to-6 direct assessment approach of Pombal et al. (2025a) due to its reported effectiveness (part 2/2).

IF-MT example for English→Spanish (Latin America)

You are a professional translator specializing in English to Spanish (Latin American) translations. Your task is to translate the following short text about fashion technology, written in a casual style:

"Hey fashion lovers! Just got my hands on the new SmartFit app (released on 5/15/2023) that scans your body in 3D and suggests clothes from over 50+ brands that would fit your measurements perfectly. I've already saved $120 on returns this month! Check out their website at www.smartfit-tech.com or email them at help@smartfit-tech.com if you have questions."

Follow these three specific rules when translating:

1. Convert all dates to DD/MM/YYYY format
2. Keep email addresses and website URLs in their original form without translation
3. Convert all dollar amounts to Mexican pesos (using an approximate conversion rate of 1 USD = 18 MXN)

Return only the Spanish (Latin American) translation, nothing else.

Figure 20: IF-MT example for English→Spanish (Latin America).

IF-MT example for English→Chinese.

Translate the following English text about Stockholm's cultural scene into Simplified Chinese. Follow these two specific rules:

1. Translate all proper names of museums, theaters, and cultural venues by providing both the Chinese translation and the original English name in parentheses.
2. Convert all years mentioned in the text to both the Gregorian calendar year and the corresponding Chinese zodiac animal year in parentheses.

Text to translate:

Stockholm's vibrant cultural landscape captivated me during my visit in 2018. The city's artistic heart beats strongly at the Moderna Museet, where I spent hours admiring contemporary masterpieces. In the evening, I attended a moving performance at the Royal Dramatic Theatre, which has been showcasing theatrical excellence since 1788. The following day, I explored Fotografiska, a photography museum housed in a beautiful Art Nouveau building from 1906. What makes Stockholm truly special is how seamlessly it blends historical traditions dating back to 1523 with cutting-edge artistic expressions of 2022.

Return only the Chinese translation following the rules above. No explanations or additional text.

Figure 21: IF-MT example for English→Chinese.

Models	Parameters	**IF-MT**	
		IF	MT
Closed			
GPT-4O-1120	>100B	5.5	89.96
Open Weights			
ALMA-R†	13B	1.56	75.32
GEMMAX†	9B	1.47	71.09
TOWER-V2†	70B	2.38	88.25
GEMMA-2	9B	4.06	88.94
GEMMA-2	27B	4.54	89.23
QWEN-2.5	72B	5.07	89.57
LLAMA-3.3	70B	4.89	88.72
Ours			
TOWER+	2B	2.16	88.06
TOWER+	9B	4.02	89.42
TOWER+	72B	4.93	89.82

Table 5: Results for English→Chinese on IF-MT. We evaluate two dimensions: instruction-following (IF) and translation quality (MT).