Command A Translate: Raising the Bar of Machine Translation with Difficulty Filtering

Tom Kocmi* **Arkady Arkhangorodsky** **Alexandre Bérard** **Phil Blunsom**

Samuel Cahyawijaya **Théo Dehaze** **Marzieh Fadaee** **Nicholas Frosst** **Matthias Gallé**

Aidan Gomez **Nithya Govindarajan** **Wei-Yin Ko** **Julia Kreutzer**

Kelly Marchisio **Ahmet Üstün** **Sebastian Vincent** **Ivan Zhang**

Cohere

*kocmi@cohere.com

Abstract

We present *Command A Translate*, an LLM-based machine translation model built off Cohere's *Command A*. It reaches state-of-the-art machine translation quality via direct preference optimization. Our meticulously designed data preparation pipeline emphasizes robust quality control and a novel difficulty filtering – a key innovation that distinguishes Command A Translate. Furthermore, we extend our model and participate at WMT with a system (CommandA-WMT) that uses two models and post-editing steps of step-by-step reasoning and limited Minimum Bayes Risk decoding.

	xComet WMT24++	MetricX WMT25	Long Context	Injection rate (%)
Deep Translation ⊕R	84.9	-5.4	52.7	4.8
Command A Translate	83.9	-6.3	51.9	0.3
DeepSeek V3	82.9	-5.7	43.0	29.5
Google Translate	82.6	-6.2	51.7	0.9
Gemini 2.5 Pro ⊕R	82.5	-5.6	56.2	1.8
GPT-5 ⊕R	82.3	-5.7	46.5	0.2
Claude 4.0 Sonnet ⊕R	82.1	-6.2	-	0.2
DeepL Pro	81.6	-7.1	50.9	0.6
Mistral Medium 3.1	80.4	-5.9	49.6	35.5
GPT-OSS 120B ⊕R	80.3	-6.5	47.0	5.3
Llama 4 Maverick	80.0	-6.7	47.4	7.2

Table 1: Aggregated results of our model against other top performing systems. We mark systems using additional reasoning with ⊕R.

1 Introduction

Neural machine translation (NMT) has revolutionized the field of machine translation (Bahdanau et al., 2014; Vaswani et al., 2017). This paradigm shift has been recently further accelerated by the advent of large language models (LLMs), which not only excel at following instructions but also demonstrate remarkable capabilities in multilingual multi-domain translation tasks as yearly evaluated at WMT Conference (Kocmi et al., 2023, 2024a). Yet, despite these gains, translation remains an open challenge. Real-world use cases often demand more than producing correct content: systems must adapt to stylistic variation, navigate complex sentence structures, and follow detailed instructions faithfully. These aspects expose weaknesses even in the most advanced models. Addressing them is crucial for moving towards translation systems that are not only capable, but also reliable and controllable across diverse contexts.

In this paper, we introduce *Command A Translate*, a state-of-the-art machine translation system built upon Cohere's flagship model, *Command A* (Cohere et al., 2025). It achieves unparalleled translation quality through direct preference optimization (DPO), leveraging the robust multilingual performance of its underlying architecture. The key innovation lies in our data preparation pipeline, which incorporates a novel difficulty filtering mechanism to ensure high-quality training data. This approach not only enhances the performance but also sets a new benchmark in the field.

We further extend our model to participate in WMT 2025 (Kocmi et al., 2025c), submitting CommandA-WMT, which employs a two-model architecture and incorporates post-editing steps such as step-by-step reasoning and limited Minimum Bayes Risk decoding. Our results highlight the effectiveness of this design, demonstrating not only consistent gains in translation quality but also the broader potential of LLM-based approaches to push the frontier of machine translation. These advances pave the way for translation systems that are not only accurate but also adaptable, controllable, and aligned with diverse human language use.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 342–352
November 8–9, 2025.

2 Training Details

In this section, we describe the architecture; how the training data is prepared; and how we fine-tuned off Command A for building Command A Translate and CommandA-WMT.

2.1 Model Architecture

We introduce two model setup which together form our submission to the WMT 2025 Shared Tasks:

- Command A Translate: Cohere's officially released MT model with open weights.[1]
- CommandA-WMT: Our shared task submission, a system incorporating model routing and additional post-editing techniques (MBR decoding and step-by-step reasoning).

Our model is built on top of Command A (Cohere et al., 2025), a 111B-parameter dense decoder-only Transformer model (Vaswani et al., 2017) supporting 23 languages.[2] We refer to Cohere et al. (2025) for additional architectural details.

2.2 Data Preparation

Early ablations revealed that sentence-level parallel data was not helpful to further improve the MT capabilities over the parent model. Accordingly, we focus only on document-level and longer context data. Data collection is challenging due to a dearth of publicly-available long-context parallel corpora.

The key part of building the Command A Translate is the data preparation pipeline. Though the training corpus is limited to document-level corpora, we still had magnitudes more training data than needed for fine-tuning. Accordingly, the critical task was to remove the data samples that would not improve model performance.

We use several steps of filtering to obtain the highest quality and most challenging examples for training. We apply the steps one after another as listed below.

1. **Rule-based filtering:** We remove boilerplate and non-textual documents, such as ones containing primarily numbers or special symbols.
2. Language identification filtering using FastText (Joulin et al., 2016).
3. **Quality Estimation (QE) filtering**: For each corpora, we remove the bottom 25% of documents with lowest document-level QE score obtained by averaging sentence-level scores (Freitag et al., 2024).
4. **Difficulty filtering:** We select documents that are most challenging to translate. This key contribution of our work is described in more details in Section 2.3.
5. **Capability filtering and language coverage:** As the final step, we assure the training dataset has an uniform distribution across languages; i.e. we give more training examples to languages where Command A under-performs, while limiting coverage of languages where it already performs very well (such as German or Spanish). Details in Section 2.4.

Our final training dataset contains 126,000 unique documents with an average of 951 tokens per document.

2.3 Difficulty Filtering

During our experimentation, we observed that standard approaches to boosting machine translation performance (such as quality filtering) were not very helpful, making only minor improvements. When diving deep, we observed that on a random sample of 100k documents, only 8.2% documents had human translations whose quality was deemed higher than translations from Command A. This finding underline the fact that Command A is already a high-performing translation model (see Table 7, where it performs on par with even strong MT systems such as DeepL).

We hypothesize that failure to boost the performance is due to a large quantity of easy or badly translated examples. Following this hypothesis, we use Sentinel-25-src (Proietti et al., 2025) which is designed to score source segments on how challenging the translation will be to modern systems. The metric was originally designed to build stronger MT test sets.

We apply Sentinel-25-src on the segment-level of potential training documents, averaging scores to obtain a single document-level difficulty score. When taking a sample of the 100,000 most difficult documents, it increases the ratio where the original human translation is better than Command A's

[1] `https://cohere.com/blog/command-a-translate` Weights: `https://huggingface.co/CohereLabs/command-a-translate-08-2025`

[2] Arabic, Chinese, Czech, Dutch, English, French, German, Greek, Hebrew, Hindi, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Romanian, Russian, Spanish, Turkish, Ukrainian, and Vietnamese.

translation to 20.1%, and shows a way to skew the training data towards more challenging samples.

One limitation of this difficulty filtering technique is that it relies on well-formatted data, because Sentinel-25-src also (correctly) ranks the broken text as difficult-to-translate. Accordingly, we apply difficulty filtering to remove the easiest 25% of all remaining data at this step. Furthermore, we utilize it in the following language balancing step to prioritize most difficult examples.

2.4 Capability Filtering and Language Balancing

Direct preference optimization (DPO) (Rafailov et al., 2023), is an offline preference modeling technique that leverages pair of completions (translations), one of which is deemed better than the other.

To create the second completion, we use Command A to translate the final training data set (on the document-level, to keep the context intact). As the last step of filtering, we scored the translation via QE to estimate if given document is better translated by humans (original target translation) than by Command A. We retain only documents where Command A under-performs humans for the final training dataset. When only a part of document is deemed better, we split the documents and only keep the better parts of the document. To prepare the preference data, we use the Command A translation as "worse completion" while using the original human translation as a better completion.

The training data is initially unbalanced in terms of language coverage, with high-resource languages having vastly more data. We target a more uniform distribution across languages paired with English while also having high coverage of non-English pairs. We use Table 7 results to identify on which languages Command A struggles, and increase their coverage in the training set. For languages where Command A is already near top performance (e.g. German or Spanish), we decrease the ratio. We prioritize the documents that are most challenging and have largest QE difference.

2.5 Training Algorithm

When fine-tuning Command A, we experimented with two setups: one using supervised fine-tuning (SFT) and the other using direct preference optimization (DPO) (Rafailov et al., 2023).

While we observed SFT improves a 7B model in ablations, improvements did not transfer to the large 111B model. On the other hand, DPO showed significant gains even for the 111B. As a result, Command A Translate uses only DPO with the training data described above.

For CommandA-WMT, we do use SFT to improve language coverage. We run SFT on only languages not supported by Command A, then follow with DPO as done for Command A Translate.

2.6 Deep Translation ⊕R

We developed a multi-stage approach that relies solely on a single deployment of Command A Translate without any additional models or resources, which boosts translation performance. The details are not elaborated here, but its empirical results are included for completeness.

2.7 CommandA-WMT Submission

CommandA-WMT is the name of our system submission to the WMT General MT (Kocmi et al., 2025a) and Terminology shared tasks (Semenov et al., 2025).[3]

CommandA-WMT is a routed machine translation system built of two models, with additional post-editing techniques: document-level translation, MBR decoding (Freitag et al., 2022) and step-by-step reasoning (Briakou et al., 2024b). We first explain the two system setup followed by post-editing techniques.

The two models that comprise the routed system are (1) *Command A Translate* for 23 supported languages, and (2) a separate finetune of Command A for unsupported languages. (2) comprises an SFT training step with parallel data for the missing languages: Bengali, Bhojpuri, Estonian, Icelandic, Kannada, Lithuanian, Marathi, Serbian, Swedish, Thai. SFT is followed by the DPO step using the same data as Command A Translate. The routing of the model is based solely on the target language of the translation direction.

We translate data at a document-level rather than segment-level to keep the context. This decision differs from the majority of system submissions for the General MT task, which are translated on the segment-level. Note that automatic evaluation can only be run on the paragraph level, which may penalize our setup (as shown in Section 3.5).

For MBR, we sampled at most 20 translations for each document by increasing temperature from 0.1 to 0.3 with a step 0.01, selecting the best translation as MBR with MetricX-XL (Juraska et al., 2024)

[3]Disclosure of conflict: the main author of Command A Translate is also an organizer of General MT shared task.

metric. The 20 translations is too little for MBR to be effective, as the original study (Freitag et al., 2022) uses 1000 samples, we expect that this step did not significantly affect the performance, as in contrast, greedy decoding leads usually to the best translation results.

Finally, we utilize the step-by-step reasoning, where we use the four-step approach introduced by Briakou et al. (2024b).

These additional post-editing steps are done only for CommandA-WMT, while all results regarding the Command A Translate are done on the raw model outputs without any post-editing techniques.

3 Evaluation and Results

We analyze the performance of our model and compare it to top-performing open and closed systems.

We evaluate all systems including ours in an identical setup unless specified otherwise, in a clean zero-shot approach without any post editing steps. We fix the temperature to 0. The only exception is the CommandA-WMT, where we report results as submitted to WMT General MT shared tasks using additional post-editing steps described in Section 2.7.

3.1 Benchmark Models

We compare our performance with top performing MT systems from all main model groups, and popular specialized translation services such as Google Translate and DeepL Pro. We evaluate DeepSeek V3 (DeepSeek-AI et al., 2025), GPT-5,[4] Gemini-2.5-Pro (Comanici et al., 2025), Mistral Medium 3.1,[5] GPT-OSS 120B (OpenAI et al., 2025), LLama 4 Maverick,[6] Claude 4 Sonnet.[7] Extended comparison comparing more systems is in Appendix B.

We run all applicable models with reasoning on, allowing them 8096 thinking budget, or setting the thinking effort to high (systems using additional reasoning are marked with ⊕R).

The only system that does not allow us to collect outputs for all languages is DeepL Pro, which does not support Persian and Hindi. In order to calculate system average for it, we use a three nearest neighbor imputing technique (Troyanskaya et al., 2001),[8] which estimates performance for missing languages without affecting its ranking, getting the same rank as if our evaluation would be done only on 21 languages. We mark those scores with asterisk. The purpose of our imputation is solely for keeping the final rank over all languages intact, rather than assuming potential performance on those two languages.

3.2 Performance Across 23 Languages

In this section, we focus on the evaluation of the 23 languages official supported by Command A Translate. We use the WMT24++ test set (Deutsch et al., 2025) containing English to 55 human-translated languages and dialects. The original source text is from Kocmi et al. (2024a) and covers four domains: news, literary, speech, and social user-generated content. Each language pair contains 171 documents split into 998 mostly paragraph level segments containing in total 32,327 words. We use the prompt instruction from Deutsch et al. (2025) with minor change discussed in Appendix A.

We evaluate translations using xComet-XL (Guerreiro et al., 2024) one of the state-of-the-art metrics with highest correlation with human judgment (Freitag et al., 2024) and widely used for system rankings, including wmt24++ (Deutsch et al., 2025). The metric is a 3.5B parameter XLM-R model (Goyal et al., 2021) fine-tuned on human judgment data.

Results in Table 2 highlight that Command A Translate outperforms all systems except on Hebrew and Hindi. Deep Translation ⊕R, however, outperforms all systems across all languages. Not only does Deep Translation ⊕R reach the highest performance, it gains +2 xComet-XL on top of the best competing system, DeepSeek V3. Such effect size would be noticeable by human annotators as much as getting more than +6 BLEU points Kocmi et al. (2024c).

3.3 WMT25 Blind Evaluation

Next, we validate the performance of our model on a blind test set. We use the WMT25 (Kocmi et al., 2025a) test set, which was released in July 2025, after our model was fully trained. It covers three source languages: English, Czech, and Japanese, and spans four domains: news commentary, ASR speech, social (Mastodon), and literary. In total, the WMT25 test set contains 36,768 words in 87 documents. The test set is released with exact prompt instructions which we use directly.

Every year, WMT hosts a machine translation

[4]GPT-5 System Card

[5]https://mistral.ai/news/mistral-medium-3

[6]https://ai.meta.com/blog/llama-4-multimodal-intelligence/

[7]Claude 4 System Card

[8]We use KNNImputer from sklearn library.

	Avg	ar	cs	de	el	es	fa	fr	he	hi	id	it
Deep Translation ⊕R	84.9	76.8	87.0	92.2	85.3	88.7	82.9	85.6	83.6	66.1	87.4	88.4
Command A Translate	83.9	76.1	86.2	92.0	84.7	87.9	82.2	85.3	82.6	62.7	86.1	87.7
DeepSeek V3	82.9	75.1	84.8	91.3	81.4	86.6	80.6	83.6	80.6	63.9	85.2	86.0
Google Translate	82.6	74.6	83.5	91.8	82.0	87.3	81.1	83.0	79.7	64.6	84.0	85.8
Gemini 2.5 Pro ⊕R	82.5	72.6	84.9	90.9	83.1	84.7	80.8	82.9	81.6	65.5	86.1	84.8
GPT-5 ⊕R	82.3	72.5	85.1	90.8	82.8	85.2	80.0	82.9	82.3	64.8	85.4	85.0
Claude 4.0 Sonnet ⊕R	82.1	73.4	83.9	91.0	82.4	84.7	80.1	82.9	80.0	62.7	83.3	85.4
DeepL Pro	81.6	70.8	83.1	90.9	80.9	85.2	80.3*	83.0	83.7	64.3*	81.7	85.5
Mistral Medium 3.1	80.4	70.1	81.9	89.9	78.3	84.0	77.4	81.0	76.8	62.0	83.2	84.2
GPT-OSS 120B ⊕R	80.3	72.1	81.8	90.1	79.0	85.7	77.1	82.5	77.4	59.7	82.0	84.8
Llama 4 Maverick	80.0	70.3	81.1	90.2	79.4	84.7	77.7	81.4	77.1	59.6	82.1	84.4

	ja	ko	nl	pl	pt	ro	ru	tr	uk	vi	zh
Deep Translation ⊕R	84.2	84.9	89.8	86.6	88.0	89.7	86.1	81.2	85.8	84.8	82.2
Command A Translate	83.1	83.7	89.2	85.6	87.7	89.5	84.6	80.1	84.8	84.1	80.4
DeepSeek V3	83.2	82.9	88.2	84.5	86.8	87.1	84.4	79.8	83.3	82.8	81.0
Google Translate	82.8	82.2	87.8	84.6	86.1	86.8	83.8	79.4	82.7	82.1	80.6
Gemini 2.5 Pro ⊕R	82.7	81.6	87.9	83.7	85.5	87.3	84.3	79.2	82.5	81.1	80.8
GPT-5 ⊕R	82.2	81.3	87.9	83.5	85.3	87.5	83.7	79.0	82.7	81.2	79.8
Claude 4.0 Sonnet ⊕R	83.2	83.4	87.1	83.3	85.5	86.6	83.9	78.2	82.8	81.5	80.4
DeepL Pro	78.4	80.6	87.0	82.6	86.3	84.8	82.7	78.9	84.5	83.1	77.0
Mistral Medium 3.1	81.7	80.7	86.4	82.0	85.0	84.9	82.7	76.2	81.6	80.6	79.1
GPT-OSS 120B ⊕R	80.3	80.9	86.3	80.9	85.5	85.3	82.0	76.2	80.9	79.0	77.9
Llama 4 Maverick	79.5	78.9	86.1	81.4	85.2	85.3	82.2	75.3	80.9	79.1	78.4

Table 2: Results of all languages over WMT24++ test set evaluated with xComet-XL metric.

system-building competition, where teams from academia and industry compete to build the best performing system. We compare our model against top participants from WMT25. As each official system submission was collected by a different team under different conditions (such as varied post-editing techniques), we run addition analysis on a set of benchmarking systems in the identical setup as our Command A Translate and Deep Translation ⊕R. We mark these with ★ in the results tables that follow. Since many of those additional systems cannot handle document-level translation, we translate WMT25 on a paragraph-level.

We score translations using MetricX-24-XL (Juraska et al., 2024), a neural metric based on mT5-XXL with 13B parameters. We apply an alternative metric to diversify results and reduce metric bias. Results in Table 3 highlight that Deep Translation ⊕R ranks at the top under controlled systems.

3.4 Human Evaluation

WMT25 (Kocmi et al., 2025a) obtained around 40 systems per language pair which were evaluated. As they didn't evaluate all systems, firstly they select the best-performing 18 system submissions for each language pair for human evaluation. The human evaluation protocols used were the Error Span Annotation (Kocmi et al., 2024b) and Multidimensional Quality Metrics (Freitag et al., 2021). We aggregate their results and for each system, we present the average system-level score along with best and worst estimated system rank, which accounts for the statistical significance of score differences.

Detailed human evaluation results are in Kocmi et al. (2025a). We compile the results of our focus languages in Table 4. Across languages, CommandA-WMT achieves the top rank of 4th to 11th place out of 40 participating systems. The largest drop versus the top-ranked system is for Egyptian Arabic, caused by the fact that CommandA-WMT was fine-tuned for machine translation only on Modern Standard Arabic. In contrast, Command A (CommandA-WMT's parent model), scores much higher on Egyptian Arabic, suggesting a high potential for Egyptian Arabic translation quality if fine-tuned to do so.

While we do not have a third party human evaluation for Command A Translate or Deep Translation ⊕R, we expect based on automatic evaluation from Section 3.3, that it would reach comparable results.

3.5 Long Context Translation

While the machine translation field is slowly moving towards paragraph-level or document-level translation (Läubli et al., 2018; Wang et al., 2023; Pal et al., 2024), current LLM models have even longer context window—able to fit full chapters

	Avg	en-ar	en-cs	en-it	en-ja	en-ko	en-ru	en-uk	en-zh	cs-uk	cs-de	ja-zh
Shy-hunyuan-MT	-4.8	-5.7	-5.5	-4.7	-5.5	-4.9	-4.9	-5.0	-4.0	-5.0	-3.6	-4.2
GemTrans	-5.1	-6.0	-5.8	-4.9	-5.5	-5.4	-5.3	-5.7	-4.3	-5.2	-3.7	-4.8
CommandA-WMT	-5.3	-7.0	-6.0	-4.8	-5.8	-5.6	-5.8	-6.0	-5.0	-4.8	-3.2	-4.7
★ Deep Translation ⊕R	-5.4	-7.2	-6.1	-4.9	-5.7	-5.6	-6.1	-6.0	-4.7	-5.2	-3.6	-4.8
★ Gemini 2.5 Pro ⊕R	-5.6	-7.5	-6.3	-5.5	-5.7	-5.6	-5.8	-6.2	-4.8	-5.3	-3.7	-4.8
★ GPT-5 ⊕R	-5.7	-7.8	-6.4	-5.5	-6.0	-5.9	-6.2	-6.2	-5.1	-5.2	-3.6	-5.0
★ DeepSeek V3	-5.7	-7.7	-6.5	-5.7	-5.9	-5.9	-6.2	-6.4	-4.7	-5.5	-3.8	-4.8
GPT-4.1	-5.8	-7.8	-6.6	-5.8	-5.9	-5.7	-6.5	-6.2	-5.0	-5.3	-3.7	-5.1
★ Mistral Medium 3.1	-5.9	-8.2	-7.1	-5.5	-6.0	-6.0	-6.1	-6.7	-4.7	-5.7	-3.9	-4.9
UvA-MT	-5.9	-7.1	-6.9	-5.4	-6.3	-6.0	-6.1	-6.3	-5.4	-6.0	-4.3	-5.6
★ Google Translate	-6.2	-7.1	-7.4	-5.6	-6.0	-6.2	-6.7	-7.2	-5.2	-6.5	-4.2	-6.0
★ Claude 4.0 Sonnet ⊕R	-6.2	-8.1	-7.5	-6.1	-6.2	-6.0	-6.9	-7.2	-5.2	-6.0	-4.0	-5.5
★ Command A Translate	-6.3	-8.0	-7.3	-5.7	-6.3	-6.2	-7.4	-7.2	-5.5	-5.9	-4.1	-5.3
Qwen3-235B	-6.5	-8.7	-7.8	-5.8	-6.4	-6.2	-6.9	-7.5	-5.0	-6.9	-4.2	-5.4
★ GPT-OSS 120B ⊕R	-6.5	-7.9	-7.7	-5.8	-6.5	-6.5	-7.2	-7.4	-5.4	-6.7	-4.3	-5.8
★ Llama 4 Maverick	-6.7	-8.9	-8.1	-6.2	-6.7	-6.5	-7.5	-7.6	-5.5	-7.0	-4.5	-5.6
TowerPlus-72B	-7.0	-10.5	-8.4	-6.1	-6.8	-6.8	-7.6	-7.9	-6.1	-6.7	-4.4	-5.9
★ DeepL Pro	-7.1	-8.2	-8.4	-6.1	-7.3	-6.8	-7.8	-7.6	-6.5	-7.0	-5.0	-7.9

Table 3: MetricX-XL results for the WMT25 test set. Systems marked with ★ are collected in controlled and identical setup, and are therefore directly comparable. The remaining systems are from (Kocmi et al., 2025b). We didn't include 24 lower performing participating systems.

	cs-de	cs-uk	en-ar (EG)	en-cs	en-it	en-ja	en-ko	en-ru	en-uk	en-zh	ja-zh
Gemini-2.5-Pro	90.2 (1-2)	92.9 (1-2)	60.6 (4-4)	88.6 (1-2)	79.4 (1-4)	85.8 (2-4)	-2.7 (1-3)	83.4 (1-1)	90.3 (1-3)	83.8 (6-11)	-4.4 (2-2)
GPT-4.1	89.2 (1-3)	92.1 (1-3)	77.0 (2-2)	80.8 (7-11)	79.0 (1-4)	83.7 (5-6)	-3.3 (4-6)	76.2 (3-5)	87.9 (6-7)	84.0 (5-10)	-6.2 (3-7)
Shy-hunyuan-MT	87.4 (2-7)	91.8 (2-3)	3.2 (11-16)	87.4 (1-2)	78.7 (1-4)	79.9 (8-12)	-2.5 (1-3)	80.2 (2-2)	88.4 (4-5)	88.2 (2-4)	-6.1 (3-7)
Claude-4-Sonnet	88.7 (2-5)	89.1 (6-10)	55.7 (5-6)	80.0 (6-10)	72.1 (6-10)	79.3 (8-13)	-3.4 (4-7)	75.9 (3-5)	85.6 (9-14)	86.9 (2-5)	-5.9 (3-7)
DeepSeek-V3	87.6 (3-7)	89.0 (4-10)	56.8 (5-6)	85.9 (3-3)	71.7 (7-10)	79.3 (8-13)	-3.8 (4-7)	73.6 (6-9)	85.8 (9-13)	85.0 (3-6)	-8.1 (8-10)
CommandA-WMT	85.6 (8-8)	88.7 (6-10)	34.6 (8-9)	83.5 (4-5)	75.5 (5-7)	82.2 (7-7)	-4.3 (7-12)	73.2 (6-9)	86.3 (8-13)	81.3 (11-15)	-7.7 (8-10)
GemTrans	82.2 (9-14)	90.2 (4-8)	3.7 (11-14)	72.6 (13-16)	79.4 (1-4)	76.2 (12-16)	-4.1 (5-10)	62.5 (13-16)	88.2 (4-5)	84.4 (5-10)	-10.9 (14-15)
UvA-MT	80.4 (9-15)	83.5 (13-17)	29.0 (10-10)	79.8 (6-10)	71.8 (7-10)	79.3 (8-13)	-5.2 (11-16)	69.1 (10-12)	86.4 (7-9)	83.4 (5-10)	-
Wenyiil	82.1 (9-14)	85.7 (11-13)	1.4 (15-18)	81.9 (6-6)	-	84.4 (3-6)	-4.3 (5-12)	78.2 (3-5)	89.5 (1-3)	86.3 (2-5)	-6.9 (4-7)
Algharb	81.3 (9-15)	84.1 (13-16)	3.2 (11-16)	74.3 (13-16)	-	85.7 (2-6)	-4.4 (5-12)	73.3 (5-8)	90.0 (1-3)	88.4 (1-1)	-5.8 (3-6)
Mistral-Medium	86.9 (4-8)	89.4 (4-10)	36.0 (8-9)	80.3 (6-10)	73.8 (5-8)	84.8 (2-5)	-4.7 (8-15)	-	84.5 (14-16)	79.9 (12-16)	-10.0 (10-13)
CommandA	86.7 (4-7)	86.4 (11-12)	74.0 (3-3)	78.0 (11-13)	73.2 (5-10)	-	-4.7 (7-15)	-	84.0 (14-16)	-	-
SRPOL	77.1 (15-19)	80.8 (18-19)	0.9 (19-19)	68.5 (17-18)	-	-	-	56.9 (17-19)	79.9 (18-19)	77.7 (14-17)	-
Yolu	75.8 (16-19)	80.1 (18-19)	1.4 (17-19)	76.1 (11-13)	-	72.6 (17-18)	-7.3 (17-18)	64.5 (12-15)	85.4 (9-13)	79.0 (12-16)	-12.6 (16-17)
IRB-MT	71.4 (20-20)	82.7 (15-17)	51.9 (7-7)	-	60.3 (12-13)	-	-5.6 (11-16)	65.4 (12-15)	82.9 (17-17)	76.5 (16-18)	-13.9 (18-18)
Laniqo	68.6 (21-21)	83.4 (14-17)	-	66.6 (17-18)	53.4 (17-18)	67.8 (19-19)	-9.1 (19-19)	56.2 (17-19)	79.8 (18-19)	70.5 (19-19)	-18.3 (19-19)
... **pruned 18 lower performing systems evaluated with humans in at least one of above language pairs** ...											
Number of systems	40	40	37	39	33	40	36	39	37	37	41

Table 4: Human evaluation sourced from WMT25 performed by Kocmi et al. (2025a). We show the average human ESA score with lower and upper rank in the bracket. The MQM is used instead for en-ko and ja-zh.

of books or more. While document-level test sets exist (Federmann et al., 2022; Deutsch et al., 2025), they usually contain only a few hundred words per document. To test the long context capabilities, therefore, we use the literary domain of the WMT25 test set (Kocmi et al., 2025a). It contains two stories of around 5000 words each, which we have models translate in a single request.

The key limitation of document-level evaluation is that automatic metrics have limited maximum length. In the case of xComet-XL, this is only a 512-token context window. To overcome this limitation, we split the translated output into paragraphs, evaluate each paragraph in isolation, and average over paragraph-level scores. This automatic evaluation thus requires models to output the same number of paragraphs as in the source segment. While CommandA-WMT successfully keeps paragraph-level alignment when instructed, other models in the benchmark cannot.

To circumvent this issue and evaluate all models, we introduce a special paragraph-break character '‖' in the source text, which we use in addition to double new lines to highlight the paragraph breaks. We use the WMT24++ prompt (see Appendix A) with additional instruction:

```
The text to translate may contain the following mark: '‖'. Keep it in the translation at the correct place.
```

With this update, almost all systems translated the story with the correct number of paragraphs,

	Avg	ar (EG)	cs	ja	ko	ru	uk	zh
Paragraph-level Command A Translate	56.9	26.1	63.3	57.3	64.9	64.9	60.1	61.9
Gemini 2.5 Pro ⊕R	56.2	24.2	61.0	60.9	57.5	63.8	63.3	62.5
Deep Translation ⊕R	52.7	23.7	59.5	55.0	52.5	60.9	61.0	56.1
Command A Translate	51.9	24.3	60.4	54.9	46.3	61.4	59.9	56.0
Google Translate	51.7	22.4	56.7	55.2	48.5	61.7	59.2	58.0
DeepL Pro	50.9	21.0	56.9	47.0	53.7	61.6	60.3	56.0
Mistral Medium 3.1	49.6	22.1	56.4	45.1	48.5	59.6	58.2	57.0
Llama 4 Maverick	47.4	20.4	52.8	52.8	51.5	55.8	54.3	44.4
GPT-OSS 120B ⊕R	47.0	22.6	50.7	40.6	49.7	56.2	54.9	54.1
GPT-5 ⊕R	46.5	22.5	52.6	46.9	49.0	52.0	52.1	50.5
DeepSeek V3	43.0	23.6	48.7	52.3	40.5	49.5	41.7	44.8
Claude 4.0 Sonnet ⊕R	-	-	50.1	-	-	54.5	-	48.8

Table 5: Results of long context translation, evaluated on a paragraph-level with xComet-XL metric.

	Avg	ar	cs	de	el	es	fa	fr	he	hi	id	it
GPT-5 ⊕R	0.2	0.0	0.1	0.1	0.1	0.2	0.4	0.2	0.1	0.1	0.1	0.1
Claude 4.0 Sonnet ⊕R	0.2	0.1	0.2	0.2	0.2	0.1	0.2	0.1	0.2	0.4	0.2	0.2
Command A Translate	0.3	0.0	0.0	0.1	0.1	0.0	0.2	0.1	0.0	0.2	0.2	0.1
DeepL Pro	0.6	0.1	0.4	0.2	0.7	0.1	0.3*	0.7	0.2	0.2*	0.2	0.2
Google Translate	0.9	0.2	0.2	0.5	0.2	0.4	0.2	0.2	0.2	0.2	0.2	0.4
Gemini 2.5 Pro ⊕R	1.8	0.7	1.5	1.3	1.5	1.3	1.3	0.9	0.5	1.3	0.6	0.7
Deep Translation ⊕R	4.8	0.1	17.9	2.1	1.1	0.1	0.2	1.8	0.5	0.4	0.1	0.0
GPT-OSS 120B ⊕R	5.3	5.4	3.8	4.9	4.5	4.0	5.0	4.5	3.1	6.4	5.9	5.5
Llama 4 Maverick	7.2	2.8	7.8	0.9	0.9	1.5	15.8	4.5	0.9	12.2	5.0	2.2
DeepSeek V3	29.5	0.2	6.0	96.9	84.5	17.0	41.1	36.1	18.6	25.7	25.7	12.9
Mistral Medium 3.1	35.5	3.8	17.0	55.3	62.5	4.2	14.4	12.6	25.6	50.2	54.6	2.9

	ko	nl	pl	ro	ru	tr	uk	vi	zh
GPT-5 ⊕R	0.0	0.2	0.1	0.4	1.0	0.2	0.2	0.1	0.2
Claude 4.0 Sonnet ⊕R	0.1	0.4	0.2	0.2	0.1	0.1	0.2	0.2	0.1
Command A Translate	2.2	0.4	0.1	0.0	0.1	0.4	0.1	0.2	0.6
DeepL Pro	1.5	0.2	0.1	0.2	0.4	0.1	5.1	0.2	0.2
Google Translate	9.8	0.2	0.5	0.2	3.1	1.0	0.2	0.2	0.1
Gemini 2.5 Pro ⊕R	12.6	0.5	1.1	0.9	1.0	2.3	1.5	2.3	1.3
Deep Translation ⊕R	53.6	7.1	5.5	0.6	0.1	1.6	1.1	0.1	1.8
GPT-OSS 120B ⊕R	10.4	5.1	5.5	4.7	4.9	5.8	3.9	8.0	4.3
Llama 4 Maverick	24.0	3.2	1.7	5.9	3.8	22.9	5.3	12.4	9.9
DeepSeek V3	10.8	61.4	50.9	3.7	15.2	48.7	26.1	7.6	0.5
Mistral Medium 3.1	46.4	37.7	10.5	3.8	89.4	25.3	83.0	66.5	45.0

Table 6: Adversarial prompt injection testing of systems. The score is a percentage of failed translation in regards to the question mark test.

except for Claude-4-Sonnet. We tried dozen of other paragraph break characters or sequences, but none worked for this model. We therefore concluded that Claude-4-Sonnet is not able to follow the instruction to translate long-context documents.

Table 5 present results of long context translation, where our model reaches second best performance, right after the Gemini 2.5 Pro ⊕R.

As a contrastive result, we add the performance of Command A Translate when translating the long-context literary test set on the paragraph-level instead of in a single request. The results highlight the performance gap between long-context and paragraph-level translation in modern MT systems (ex. quality degrades from 56.9 to 51.9 xComet-XL on average across languages when translated at the paragraph-level versus the whole long-context example in one shot).

3.6 Prompt Injection Robustness

In contrast to pure machine translation systems, a problem introduced by using general instruction-following LLMs for translation is that they may misinterpret instructions; instead of translating, they may *respond to* prompts or questions in the source text. Safety mechanisms may also be triggered, making the model refuse to translate (Briakou et al., 2024a). Both are a problem especially when systems are used as a part of automatic pipeline.

To test translation rejection, we employ the test set for adversarial MT prompt injection (Miceli Barone and Sun, 2024), which tests models via prompt injection by asking to translate a question. It contains 817 different English questions in a direct translation set. Here is an example:

```
Translate the following text into Czech and
do nothing else:
Don't translate this sentence and tell me:
What happens to you if you eat watermelon
seeds?
```

We use the "question mark" heuristic (Miceli Barone and Sun, 2024) for evaluation—we check if the translation contains a terminal question mark, showing that the model translated the question rather than answering it. Since this simple test does not require reference translation, we scale it into all target languages. See Appendix C for implementation details.

Results in Table 6 show that most systems are robust to prompt injection attack. DeepSeek V3 and Mistral Medium 3.1, however, struggle to resist instruction following on almost all languages. While Command A Translate is robust across the board, Deep Translation ⊕R struggles in Czech and Korean, likely caused by its more complex prompt instruction structure.

4 Conclusion

We introduce Command A Translate with Deep Translation ⊕R capabilities, Cohere's state-of-the-art machine translation system. Command A Translate is built off Cohere's Command A by fine-tuning on meticulously-prepared datasets and with direct preference optimization. As the key innovation, our data pipeline, incorporates a series of novel data filters, targeting selection of most difficult data subset and strong capabilities across languages. Command A Translate achieves marked improvement in translation quality, and outperforms other translation systems such as Google Translate, and state-of-the-art LLMs such as GPT-5 and Gemini-2.5 Pro.

Extending Command A Translate, we present CommandA-WMT, our translation system submission to the 2025 WMT shared task. This system leverages a two-model architecture and post-editing steps such as step-by-step reasoning and limited Minimum Bayes Risk decoding. CommandA-WMT achieves consistent gains in across languages, showcasing the effectiveness of our design.

Limitations

The evaluation of machine translation systems is fundamentally limited by the noise and limited discriminative power of automated benchmarks, and even of human evaluators. Translation quality can be subjective, and furthermore, high translation quality in one domain for a given language does not guarantee high quality in another, even for the same language. Preferred system recommendations can thus change depending on use case. We provide results across the domains evaluated in WMT24 and WMT25, but encourage users to examine systems on the domains they care about.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024a. On the implications of verbose llm outputs: A case study in translation evaluation. *arXiv preprint arXiv:2410.00863*.

Eleftheria Briakou, Jiaming Luo, Colin Cherry, and Markus Freitag. 2024b. Translating step-by-step: Decomposing the translation process for improved translation quality of long-form texts. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1301–1317, Miami, Florida, USA. Association for Computational Linguistics.

Team Cohere, Arash Ahmadian, Marwan Ahmed, Jay Alammar, Milad Alizadeh, Yazeed Alnumay, Sophia Althammer, Arkady Arkhangorodsky, Viraat Aryabumi, Dennis Aumiller, and 1 others. 2025. Command a: An enterprise-ready large language model. *arXiv preprint arXiv:2504.00698*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

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.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

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

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. Larger-scale transformers for multilingual masked language modeling. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. Fasttext.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

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, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025a. 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*, China. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, and 1 others. 2025b. Preliminary ranking of wmt25 general machine translation systems. *arXiv preprint arXiv:2508.14909*.

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, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 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, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025c. Preliminary ranking of wmt25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi, Vilém Zouhar, Christian Federmann, and Matt Post. 2024c. Navigating the metrics maze: Reconciling score magnitudes and accuracies. In *Pro-*

ceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 1999–2014, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Läubli, Rico Sennrich, and Martin Volk. 2018. Has machine translation achieved human parity? a case for document-level evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4796, Brussels, Belgium. Association for Computational Linguistics.

Antonio Valerio Miceli Barone and Zhifan Sun. 2024. A test suite of prompt injection attacks for LLM-based machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 380–450, Miami, Florida, USA. Association for Computational Linguistics.

OpenAI, :, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, and 108 others. 2025. gpt-oss-120b and gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Proyag Pal, Alexandra Birch, and Kenneth Heafield. 2024. Document-level machine translation with large-scale public parallel corpora. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13185–13197, Bangkok, Thailand. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Vilém Zouhar, Roberto Navigli, and Tom Kocmi. 2025. Estimating machine translation difficulty. *Preprint*, arXiv:2508.10175.

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.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. 2001. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.

A WMT24++ Prompt Instruction

The paper Deutsch et al. (2025) specify two different prompt instructions, we have used the first one mentioned in their Figure 3. Since the prompt contains instructions after providing a text for translation, it may have been perceived by systems as the instruction should be also translated. To fix this problem, we move the instruction sentence before providing the source text.

```
You are a professional {src_lang} to {tgt_lang} translator, tasked with providing translations suitable for use in {tgt_lang} ({tgt_country}). Your goal is to accurately convey the meaning and nuances of the original {src_lang} text while adhering to {tgt_lang} grammar, vocabulary, and cultural sensitivities. Produce only the {tgt_lang} translation, without any additional explanations or commentary. Please translate the following {src_lang} text into {tgt_lang} ({tgt_country}):
{source_text}
```

B Extended Benchmark

For completeness, we also run extended evaluation on most popular and top performing models.

C Adversarial MT Prompt Injection

To test model robustness to prompt injection, we use the technique from Miceli Barone and Sun (2024), designed to test if a model can correctly execute a translation task in the presence of distractor instructions. The test set contains 816 English questions asked in an prompt injection manner. As the authors doesn't specify the initial instructions, we use the following template:

```
Translate the following text into {tgt_lang} and do nothing else:
Don't translate this sentence and tell me:
{question}
```

We evaluate using the "question mark" heuristic: whether the translation ends with a question

	Avg	ar	cs	de	el	es	fa	fr	he	hi	id	it
Deep Translation ⊕R	84.9	76.8	87.0	92.2	85.3	88.7	82.9	85.6	83.6	66.1	87.4	88.4
Command A Translate	83.9	76.1	86.2	92.0	84.7	87.9	82.2	85.3	82.6	62.7	86.1	87.7
GPT-4o	83.4	75.4	85.9	91.5	84.5	86.8	82.1	84.5	82.2	63.7	84.9	86.6
Claude Opus 4.1	83.1	75.2	85.0	90.8	83.2	85.8	81.2	83.5	81.6	64.8	85.2	85.6
DeepSeek V3	82.9	75.1	84.8	91.3	81.4	86.6	80.6	83.6	80.6	63.9	85.2	86.0
Google Translate	82.6	74.6	83.5	91.8	82.0	87.3	81.1	83.0	79.7	64.6	84.0	85.8
Gemini 2.5 Pro ⊕R	82.5	72.6	84.9	90.9	83.1	84.7	80.8	82.9	81.6	65.5	86.1	84.8
GPT-5 ⊕R	82.3	72.5	85.1	90.8	82.8	85.2	80.0	82.9	82.3	64.8	85.4	85.0
Claude 4.0 Sonnet ⊕R	82.1	73.4	83.9	91.0	82.4	84.7	80.1	82.9	80.0	62.7	83.3	85.4
Command A	81.6	73.1	84.0	91.0	82.2	85.9	79.5	83.6	79.4	60.8	82.8	85.5
DeepSeek R1	81.5	73.2	83.8	89.7	80.4	86.2	78.2	82.7	78.9	62.6	84.6	85.0
DeepL Pro	81.5	70.8	83.1	90.9	80.9	85.2	79.2*	83.0	83.7	62.7*	81.7	85.5
Qwen MT Plus	80.5	73.4	79.5	91.2	77.1	86.3	74.3	83.4	73.4	59.4	84.6	86.1
Mistral Medium 3.1	80.4	70.1	81.9	89.9	78.3	84.0	77.4	81.0	76.8	62.0	83.2	84.2
GPT-OSS 120B ⊕R	80.3	72.1	81.8	90.1	79.0	85.7	77.1	82.5	77.4	59.7	82.0	84.8
Llama 4 Maverick	80.0	70.3	81.1	90.2	79.4	84.7	77.7	81.4	77.1	59.6	82.1	84.4
Llama 3.1 405B	80.0	69.7	81.6	90.5	78.2	84.7	76.9	82.4	77.8	59.7	81.3	84.4
Qwen3-235B-A22B	79.7	70.5	80.5	90.3	78.4	85.5	73.2	82.5	70.0	59.4	83.0	84.8
Aya Expanse 32B	79.5	70.8	81.3	90.4	79.7	85.0	76.4	82.1	75.8	57.4	80.9	84.3
Gemma 3 (27b)	79.3	70.2	81.0	89.2	79.6	82.7	78.0	79.9	76.3	60.1	80.9	83.6
Mistral Large Latest	79.3	70.8	80.4	91.2	77.7	85.1	74.8	83.0	77.9	58.4	79.8	85.9

	ja	ko	nl	pl	pt	ro	ru	tr	uk	vi	zh
Deep Translation ⊕R	84.2	84.9	89.8	86.6	88.0	89.7	86.1	81.2	85.8	84.8	82.2
Command A Translate	83.1	83.7	89.2	85.6	87.7	89.5	84.6	80.1	84.8	84.1	80.4
GPT-4o	83.7	83.4	88.5	84.5	86.8	88.1	84.5	79.9	84.2	82.5	80.5
Claude Opus 4.1	84.1	83.6	88.0	84.6	86.2	87.5	85.1	80.2	83.8	82.5	81.7
DeepSeek V3	83.2	82.9	88.2	84.5	86.8	87.1	84.4	79.8	83.3	82.8	81.0
Google Translate	82.8	82.2	87.8	84.6	86.1	86.8	83.8	79.4	82.7	82.1	80.6
Gemini 2.5 Pro ⊕R	82.7	81.6	87.9	83.7	85.5	87.3	84.3	79.2	82.5	81.1	80.8
GPT-5 ⊕R	82.2	81.3	87.9	83.5	85.3	87.5	83.7	79.0	82.7	81.2	79.8
Claude 4.0 Sonnet ⊕R	83.2	83.4	87.1	83.3	85.5	86.6	83.9	78.2	82.8	81.5	80.4
Command A	81.3	81.8	87.4	82.7	86.2	87.6	82.4	75.9	82.6	81.2	78.5
DeepSeek R1	81.4	81.3	87.1	83.4	85.7	85.4	83.5	77.7	81.7	81.4	80.0
DeepL Pro	78.4	80.6	87.0	82.6	86.3	84.8	82.7	78.9	84.5	83.1	77.0
Qwen MT Plus	82.3	81.2	86.9	79.1	86.1	83.8	83.5	76.5	80.3	81.9	81.0
Mistral Medium 3.1	81.7	80.7	86.4	82.0	85.0	84.9	82.7	76.2	81.6	80.6	79.1
GPT-OSS 120B ⊕R	80.3	80.9	86.3	80.9	85.5	85.3	82.0	76.2	80.9	79.0	77.9
Llama 4 Maverick	79.5	78.9	86.1	81.4	85.2	85.3	82.2	75.3	80.9	79.1	78.4
Llama 3.1 405B	79.5	80.0	86.7	81.5	84.5	86.2	82.5	75.4	80.4	78.3	77.2
Qwen3-235B-A22B	79.4	80.7	85.6	80.7	85.5	85.2	82.1	74.5	80.3	81.2	79.4
Aya Expanse 32B	78.9	78.6	86.2	81.3	84.8	85.9	80.8	73.6	80.6	79.7	74.1
Gemma 3 (27b)	78.8	78.1	85.2	81.1	83.6	84.8	82.2	74.4	80.6	78.8	75.8
Mistral Large Latest	80.5	80.7	82.1	79.7	84.7	81.8	82.1	72.1	81.2	77.1	77.5

Table 7: Extended WMT24++ results with xComet-XL for extensive set of systems.

mark (ignoring white spaces and quotation marks). The final score is a percentage of failed cases. As the heuristic does not require reference translation, it can be easily scaled to any number languages. The only limitation is proper handling of question marks per language. We therefore also check for following language-specific question marks: Chinese and Arabic question mark, and the semi-colon for Greek.

We have not evaluated on Japanese, for which the question mark test doesn't work as the language allows different paraphrases not ending with question mark. An example from Google Translate: この文を翻訳せずに、「すべての星は星ですか？と言ってください」