TranslateGemma for ES-EN Financial Reports: Exploring Adaptability to Variable-Sized Contexts

Yanco Amor Torterolo Orta, Melina Chatzi, Antonio Moreno-Sandoval
UAM, UNED
{yanco.torterolo, melina.chatzi}@estudiante.uam.es, antonio.msandoval@uam.es

Abstract

This paper explores bidirectional financial Machine Translation (MT) between Spanish and English, focusing on the specialized domain of annual reports from IBEX 35 companies. Fine-tuned models are compared against zero-shot scenarios through a series of experiments, testing factors such as prompting strategies and model size. On the one hand, this work studies a combination of existing fine-tuning strategies aimed at improving the adaptability of MT models to variable-sized contexts, and, on the other hand, it analyzes the limitations detected in current evaluation metrics. Results are mixed: fine-tuned models show an improvement in both short and long-context scenarios in traditional metrics, while zero-shot predictions are clearly favored by neural metrics. In fact, reference-free assessment of the source and the human reference received worse scores than the off-the-shelf prediction models. Consequently, fine-tuning on the human-made dataset hardly improves the neural metrics against zero-shot generations. This suggests that neural metrics tend to favor the fluency of MT generations and literalness over creativity, among other technical limitations regarding long-context adaptability. From a practical standpoint, the low Translation Edit Rate (TER) scores suggest that specialized fine-tuning remains the most viable path for companies to implement efficient Machine Translation Post-Editing (MTPE) workflows, given the stylistic alignment.

Keywords: financial machine translation, translategemma, long-context evaluation, annual reports

1. Introduction

Companies all over the world use translation as a means of reaching a broader audience beyond their borders. As highlighted by Herrero Rodes and Román Mínguez (2015), these companies draft their annual reports and have them translated into other languages. This process is vital for their business strategy since it allows them to address potential shareholders. Spain is no different, with IBEX 35 companies publishing their annual reports on their websites (Quesada and Espada, 2024).

Natural Language Processing (NLP) tasks have evolved dramatically in the past few years, and MT is no exception. Generative LLMs offer increasingly larger context windows, which is beneficial for nearly every kind of NLP task. This is especially true of financial MT, even reaching document-level (Wang et al., 2023), as the context usually contains elements key to the target text.

This paper explores how the context window affects Machine Translation (MT) and how different metrics respond to different context sizes. Inspired by several existing works and methodologies (Tiedemann and Scherrer, 2017; Johnson et al., 2017; Ding et al., 2021), a Data Augmentation (DA) strategy consisting of duplicating the dataset with a version made of concatenated Translation Units (TU), combined with a bidirectional, variable-sized exposure to the dataset during fine-tuning is tested. It is aimed at improving adaptability to long contexts while assessing the viability of current metrics. To this end, a compact, local model was used for bidirectional financial ES-EN MT. This choice stems from data privacy and environmental footprint concerns, which are key considerations for companies. The democratization of LLMs on consumer-grade hardware is also promoted.

This research focuses on the annual reports of IBEX 35 companies. A parallel corpus consisting of aligned ES-EN TUs was used. It was compiled from 34 pairs of annual reports, amounting to 41,951 TUs of varying sizes—ranging from titles and lists to paragraphs. The word count totals 1,257,458 (ES) and 1,098,426 (EN). A selection of this dataset is available in the following repository: Moreno-Sandoval et al. (2025).

Several research questions will be addressed: (1) Is it worth fine-tuning models on the annual reports of IBEX 35 companies, or similar contexts, given the zero-shot capabilities of current LLMs for MT? (2) From a corporate perspective, does local MT provide an acceptable draft as a starting point for human translators? (3) Does this variable-sized context strategy effectively improve scores compared to fine-tuning on a default dataset? (4) Are there limitations in current metrics?

The rest of the paper is structured as follows: Section 2 reviews related work; Section 3 lays the groundwork for the experiments; and Section 4 provides a detailed analysis of the results. Finally, in Section 5, conclusions are drawn based on the findings.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 87–97

2. Related work

The most similar work found examines Arabic (AR-EN) financial MT (Alghamdi et al., 2023), where it is noted that off-the-shelf neural MT models exhibit an inability to translate domain-specific texts. However, LLMs have improved significantly since 2023.

In this regard, generative LLMs are being increasingly used, as recent iterations of the WMT suggest (Kocmi et al., 2023, 2024, 2025). This trend applies to both translation and evaluation, which is highly relevant to the present work, since several LLMs are used for evaluation and TranslateGemma—a recently released state-of-the-art (SOTA) model—is employed. The subtitles of these findings reports illustrate the rapid evolution of LLMs: *2023: LLMs Are Here But Not Quite There Yet*, *2024: The LLM Era is Here but MT is Not Solved Yet*, and *2025: Time to Stop Evaluating on Easy Test Sets*.

Regarding long-context settings, several studies address document-level MT, such as (Herold and Ney, 2023; Wang et al., 2023). Furthermore, this work provides a document-level financial test set for MT (Nakhlé et al., 2025). A notable study analyzing MT metrics can be found in (Di Natale et al., 2025).

In general, NLP tasks targeting financial texts have gained significant attention, as shown in works such as (Ke et al., 2025). Additionally, there are benchmarks like WMT24++ (Deutsch et al., 2025) that comprise several domains—literary, news, social, and speech—but lack a specific finance component. A more adequate benchmark would be TransBench (Li et al., 2025), as it includes specialized subdomains such as e-commerce. In this paper, no benchmarks were used due to the task-specific nature of the dataset; however, they remain a valuable consideration for future research.

3. Settings and experiments

3.1. Hardware settings and environment

All experiments described in this paper were conducted on a system equipped with an `AMD Ryzen 7 9800X3D` processor, an `NVIDIA RTX 5080` GPU with 16GB of VRAM, and 32GB of DDR5 RAM. For fine-tuning and inference, the official `Unsloth` Docker image was used to prevent dependency issues. Metrics were computed within a standard Conda environment.

3.2. Models and dataset

The models utilized in this study consist of two variants of the TranslateGemma family: `google/translategemma-4b-it` and `google/translategemma-12b-it` (Finkelstein et al., 2026). Based on the Gemma 3 architecture, these models were specifically fine-tuned by Google to excel in MT across 55 languages. As instruction-tuned variants, they are designed to follow prompts effectively.

Regarding the dataset, the TUs from the corpus were already segmented into paragraph-level structures, although some fragments were short (e.g., titles and lists). Two different configurations were evaluated: (a) the **original version**, which uses TUs not exceeding 697 tokens (including both source and target language references); and (b) the **mixed version**, which combines the original dataset with a reformulated version of itself, where the same TUs are concatenated into larger segments of up to 2,102 tokens. This automated chunking process sought to obtain larger TUs while adhering to rules designed to preserve section and list integrity within each chunk. The original TUs remained intact in the process, as they were not split. Table 1 summarizes the characteristics of each dataset. Each entry includes an ID, the source language (Spanish), and the target language reference (English).

dataset	train	val	test	total
original	39,857	1,047	1,047	41,951
mix	43,738	1,169	1,169	46,076

Table 1: Number of TUs in each dataset split.

It is worth noting that the context window for TranslateGemma is 2,000 tokens. While the model card lists input and output separately, this typically encompasses the source text, prompt, and target output. However, this is a functional limit established during Google's translation-focused fine-tuning. Including the prompt, the mixed dataset contains TUs reaching 2,300 tokens. Since the underlying Gemma 3 architecture supports a context window of up to 128,000 tokens, and the objective of this work is to further fine-tune the models, this should not hinder performance.

3.3. Variable-sized context adaptability

By randomly interleaving individual TUs with concatenated sequences from the mixed dataset, the model was exposed to varying contextual scales. This approach mitigates the "sentence-level bias" (or short-sequence bias) typical of standard MT datasets and encourages discursive consistency across extended financial narratives. Since preceding and subsequent contexts are often crucial for the accurate translation of a segment, leveraging an expanded context window is potentially beneficial. Furthermore, this method serves as a data augmentation technique by offering alternative ways of presenting the training data. This randomized interleaving prevents length-related bias and

promotes flexibility regarding context size. Complementing this, bidirectionality is a core aspect of this strategy. The model is exposed to the dataset for both translation directions (ES-EN and EN-ES). The random sampling of translation directions during training prevents directional bias and results in a more robust system.

3.4. Chat template and prompt

The default inference process in TranslateGemma relies on a complex, hardcoded Jinja2 chat template embedded within the model's tokenizer, which acts as a rigid preprocessing layer. This template, hereafter referred to as **GP**, contains an extensive internal mapping of ISO language codes and enforces a strict conversational structure that includes system role validation and specific token formatting, such as start-of-turn and end-of-turn indicators. While this ensures adherence to Google's official specifications, it introduces a significant computational bottleneck and unnecessary token overhead. In contrast, the official Ollama adaptation[1], referred to as **GPO**, utilizes a more straightforward prompting strategy that bypasses the heavy Jinja2 logic in favor of a direct instructional format. GPO strips away the structural constraints of the hidden template and delivers the core translation task directly to the model. The GPO approach is provided in Figure 1. Effectively, the underlying information and the translation intent remain the same—what varies is the efficiency of the delivery and the reduction in preprocessing latency.

The GP and GPO chat templates are compared alongside a third prompt variant, which consists of a minimal instruction:
`You are a professional translator. Translate the following text from Spanish to English` (or *vice versa*).

> You are a professional {SOURCE_LANG} ({SOURCE_CODE}) to {TARGET_LANG} ({TARGET_CODE}) translator. Your goal is to accurately convey the meaning and nuances of the original {SOURCE_LANG} text while adhering to {TARGET_LANG} grammar, vocabulary, and cultural sensitivities.
>
> Produce only the {TARGET_LANG} translation, without any additional explanations or commentary. Please translate the following {SOURCE_LANG} text into {TARGET_LANG}:
>
> {TEXT}

Figure 1: GPO prompting implementation by Ollama based on Google's implementation (GP).

[1] https://ollama.com/library/translategemma

3.5. Fine-tuning and inference

Several fine-tuning and inference components were individually analyzed to conduct a comprehensive ablation study. Given the exceptional multilingual capabilities of current LLMs, the inclusion of non-fine-tuned (zero-shot) baselines was deemed essential. Consequently, these models were evaluated in an inference-only setup to establish a comparative benchmark. Furthermore, multiple inference variables were examined, such as the application of **beam search** and the impact of model size. Each translation direction was also analyzed, enabling cross-directional comparisons. The 16GB VRAM constraint necessitated a resource-aware experimental design incorporating QLoRA and careful hyperparameter selection, in which the use of the `Unsloth` library proved instrumental.

3.6. Hyperparameters

Fine-tuning hyperparameters using `Unsloth` remained identical across experiments, with minor adjustments between the 4B and 12B model versions. Both models were loaded in 4-bit precision with a per-device training batch size of 2 and 8 gradient accumulation steps. The evaluation batch size was set to 1, while `eval_accumulation_steps` were set to 4 for the 4B version and 1 for the 12B version to prevent out-of-memory (OOM) errors during evaluation. Training was governed by an early-stopping mechanism (monitoring *eval loss*) with a patience of 3 evaluation calls, occurring every 500 steps. Although 5 epochs were initially scheduled, the models reached early stopping between 2.2 and 4.1 epochs, depending on the dataset, prompt, and model size. The best model (lowest validation loss) was consistently loaded upon completion. Additional parameters included a learning rate of 1×10^{-5}, 400 warmup steps, and the `paged_adamw_8bit` optimizer with a linear scheduler. Gradient checkpointing was enabled (`use_reentrant=False`). Regarding QLoRA configurations, a rank (r) of 16 and a LoRA alpha (α) of 16 were applied to all linear layers.

3.7. Metrics

The metrics chosen for this work are summarized in Table 2. MetricX and XCOMET are established as SOTA metrics, as evidenced by their continued adoption in the most recent WMT25 shared task (Juraska et al., 2025). However, MetricX-24 (Juraska et al., 2024) was utilized instead of the 2025 version, as the latter had not been fully released at the time of writing. According to Juraska (Juraska, 2025), the 2025 iteration "did not outperform other fine-tuned metrics on the WMT25 test set" and "didn't provide a consistent improvement

over MetricX-24." Consequently, they recommend adhering to the MetricX-24 version.

metric	model
MetricX	google/metricx-24-hybrid-large-v2p6-bfloat16
MetricX_QE	google/metricx-24-hybrid-large-v2p6-bfloat16
MetricX_REFQA	google/metricx-24-hybrid-large-v2p6-bfloat16
XCOMET	Unbabel/XCOMET-XL
XCOMET_QE	Unbabel/XCOMET-XL
XCOMET_REFQA	Unbabel/XCOMET-XL
CHREF++	n/a
TER	n/a
BLEU	n/a

Table 2: Metrics and models employed.

MetricX-24 utilizes a hybrid transformer-based architecture—leveraging large-scale encoder-decoder models like mT5—to assess translation quality, consistently demonstrating superior correlation with human judgment (Freitag et al., 2024). The metric operates by encoding the source, reference, and prediction into a shared embedding space, where it performs reference-based regression trained on Multidimensional Quality Metrics (MQM) data. This allows the model to move beyond surface-level lexical overlaps and evaluate semantic fidelity. As a regressive metric trained to predict human-assigned error scores, it operates on an inverse scale where lower values denote higher translation quality.

Complementing this, **XCOMET** provides an error-aware evaluation by integrating a multi-task learning objective into the COMET framework. By utilizing cross-lingual encoders like XLM-RoBERTa and training on both Direct Assessment (DA) and MQM data, it distinguishes stylistic nuances from critical semantic errors to produce a normalized quality score. To overcome the encoder's 512-token constraint in long-form financial documents, a hierarchical dynamic chunking strategy was implemented. This methodology prioritizes line-by-line alignment, followed by sentence-level segmentation and a length-based fallback to ensure comprehensive coverage. Although averaging segment scores provides a global quality estimate, this fragmentation introduces specific risks: the potential loss of cross-chunk cohesion, the risk of alignment drift during fallback splitting, and the dilution of localized critical errors within the aggregate mean.

As a fundamental component of the evaluation framework, an analysis of the intrinsic quality of the gold standard references was performed to determine whether the human-provided translations effectively represent a definitive upper bound for model predictions. This comparative analysis was facilitated by the dual-mode architecture of MetricX and XCOMET, which support both reference-based (source, reference, and prediction) and reference-free evaluation. Specifically, two dimensions were assessed: (a) an evaluation of the source and the reference, hereafter referred to as Reference Quality Assurance (**RefQA**), and (b) an evaluation of the source and the model's prediction, referred to as Quality Estimation (**QE**). Comparing these two dimensions enables the identification of whether a specific score reflects suboptimal model performance or stems from underlying inconsistencies within the reference material itself (Freitag et al., 2023).

Besides neural metrics, **chrF++** and **TER** (Translation Edit Rate) were included to evaluate different aspects of the translation. While chrF++ measures character-level accuracy, TER estimates the effort a human would need to correct the text. **BLEU** metric was also implemented via the SacreBLEU toolkit as a standard benchmark. Even though BLEU has limitations in capturing full meaning, it remains one of the most widely used metrics in the field, enabling comparability with broader research. Regarding their interpretation, higher chrF++ and BLEU scores, and lower TER values denote superior translation quality. These traditional metrics are more transparent and provide a reliable baseline to check if the main neural metrics suffer from metric bias. This phenomenon is an instance of reward hacking , where the utility function (the reward model) improves but the system's behavior diverges from actual quality or human preferences (Kovacs et al., 2024).

4. Results and discussion

4.1. GP vs GPO

The impact of the chat template on model behavior was initially evaluated, as it informs several subsequent experimental decisions. Table 3 provides a performance and efficiency comparison between the GP and GPO approaches. Both configurations were tested using `google/translategemma-4b-it` in a zero-shot setup. To prioritize inference speed, all 4B models were loaded in 16-bit precision. While the difference in translation quality remains negligible, the efficiency gains are substantial: GPO achieves a reduction in both execution time and energy consumption of over 98% relative to the standard GP template.

It is important to clarify that the GPO version was executed via Ollama, whereas the GP version utilized the `Unsloth` fast inference framework. Given that `Unsloth` is highly optimized for speed, the disproportionate latency observed in

Metric	4b_gp	4b_gpo
Translation Metrics		
MetricX (MX) ↓	**2.8019**	2.8158
XCOMET (XC) ↑	**0.8738**	0.8730
chrF++ ↑	64.46	**64.69**
TER ↓	46.33	**46.25**
BLEU ↑	31.54	**31.76**
Environmental Impact		
Duration (min) ↓	753.79	**10.24**
Emissions (g) ↓	589.88	**6.39**
Total Energy (kWh) ↓	3.39	**0.037**
GPU Energy (kWh) ↓	2.5119	**0.0359**

Table 3: GP and GPO efficiency comparison.

the GP run suggests a substantial bottleneck inherent to the complexity of the original template rather than the inference engine itself. Consequently, this remains a valid comparison, as it highlights how template complexity can negate engine-level optimizations. Furthermore, Ollama lacks native support for resource-intensive Jinja2-based templates, necessitating streamlined adaptations for operational viability. Accordingly, all subsequent experiments were conducted using `Unsloth`, except for `4b_noft_gpo` variants.

To ensure computational efficiency and minimize the environmental footprint, this comparison was restricted to the ES-EN direction, excluding GP regardless of the engine. GPO was used in the remaining experiments, along with the previously mentioned minimal prompt. The absence of "GPO" in a system ID denotes the use of the minimal prompt.

4.2. Prompt, fine-tuning and direction

Regarding prompting, the performance of the GPO configuration is compared against the minimal prompt. As shown in Table 4, the performance gap between `4b_noft_gpo` and `4b_noft` is consistent across both translation directions, particularly in TER. In a zero-shot setting, the prompt serves as the sole mechanism to activate the model's translation capabilities. GPO more effectively triggers the model's internal weights by aligning with the linguistic distribution encountered during its primary training.

As for the fine-tuned models (indicated by the "ft" suffix), distinct patterns emerge depending on the translation direction. On the one hand, Table 4 shows that for ES-EN, performance is generally superior with the minimal prompt. On the other hand, it reveals the opposite trend for EN-ES, where GPO-based systems consistently outperform the minimal approach. In the EN-ES direction, GPO is superior as it aligns the model with its native instruction-tuning patterns (Finkelstein et al., 2026) and Spanish morphological requirements. Furthermore, using original, well drafted, Spanish annual reports as references during evaluation allows GPO to achieve native-level prose, effectively avoiding *translationese* (Zhang and Toral, 2019) and obtaining extremely high fluency scores in MetricX. Conversely, ES-EN finds benefit from using the minimal prompt. As demonstrated by Etxaniz et al. (2024), LLMs often fail to leverage their full potential when prompted in non-English languages, confirming an English-centric bias where less structural guidance in English leads to more natural generation. This is further supported by Richburg and Carpuat (2024), who found that the impact of translation fine-tuning is inherently uneven across language pairs. Finally, neural metrics like MetricX apparently exhibit a fluency bias, potentially over-rewarding the natural phrasing of the generated Spanish text, sometimes at the expense of strict lexical fidelity (Freitag et al., 2024).

Furthermore, reference quality is a primary factor explaining the discrepancy in MetricX scores across language directions, where EN-ES performance appears significantly superior to ES-EN (1.4895 vs. 2.5841). This may seem counterintuitive, as the source corpus is originally Spanish. As noted in the WMT23 findings: "Metrics might be guilty, but references are not innocent" (Freitag et al., 2023). By comparing the REFQA scores of the references with the QE scores of the model predictions, a clear limitation can be identified. If a metric deems a reference poorly aligned with the source, fine-tuning the model to mimic that reference may propagate those perceived flaws. This is evident when comparing fine-tuned results to the 4b_noft_GPO zero-shot scores. In the ES-EN direction, the MetricX gap between REFQA (2.9171) and QE (2.3320) indicates that the model's independent translations outperform the human references according to the metric. The same trend is observed in EN-ES, though absolute error scores are lower (REFQA 1.9424 vs. QE 1.6859). This proportional difference confirms that the EN-ES direction offers more room for improvement in terms of the metric.

4.3. Size of the model

Table 4 suggests a small difference between the 4b and the 12b variants (`4b_ft_bs_mix` vs. `12b_ft_bs_mix`). These two configurations were almost identical except that the 12b variant was loaded in 4-bit precision for inference and required minor tweaks due to VRAM constraints. The 12b variant is slightly superior, but given the increased resource consumption, the 4b variant appears to be a better option for deployment.

ID	MX ↓	MX_QE ↓	XC ↑	XC_QE ↑	CHRF ↑	TER ↓	BLEU ↑
			ES-EN				
4b_noft_gpo	**2.5841**	**2.3320**	0.9300	**0.9398**	64.69	46.25	31.76
12b_ft_bs_mix	2.6666	2.7468	**0.9342**	0.9392	**70.73**	**38.05**	**40.15**
4b_ft_bs_mix	2.8291	2.8407	0.9253	0.9350	70.19	39.37	39.14
4b_ft_mix	2.8915	2.9337	0.9208	0.9305	69.20	40.51	37.80
4b_noft	2.9153	2.6358	0.9027	0.9139	62.99	223.79	32.16
4b_ft_gpo_bs_mix	2.9734	3.0378	0.9262	0.9362	69.63	39.84	38.51
4b_ft_gpo_mix	3.0343	3.0972	0.9206	0.9286	68.54	41.27	37.22
4b_ft_bs_ori	3.1675	3.4858	0.8885	0.8985	67.21	42.23	36.63
4b_ft_ori	3.2863	3.6783	0.8653	0.8742	65.68	43.72	35.28
4b_ft_gpo_ori	3.3970	3.8776	0.9169	0.9325	65.96	45.44	34.86
4b_ft_gpo_bs_ori	3.8412	4.7432	0.9176	0.9345	64.73	47.91	33.85
REFQA	–	2.9171	–	0.9153	N/A	N/A	N/A
			EN-ES				
12b_ft_gpo_bs_mix	**1.4895**	**1.6859**	**0.9524**	**0.9453**	**72.07**	**36.48**	**43.84**
4b_noft_gpo	1.5761	1.4672	0.9506	0.9544	64.80	47.99	34.24
4b_ft_gpo_bs_mix	1.8024	2.0062	0.9476	0.9429	69.65	40.10	40.99
4b_ft_gpo_mix	1.8058	1.9191	0.9391	0.9345	68.44	41.54	39.35
4b_ft_gpo_ori	1.8823	1.9493	0.9239	0.9231	67.30	42.78	38.35
4b_ft_gpo_bs_ori	1.9665	2.2565	0.9302	0.9305	68.04	41.21	39.79
4b_noft	1.9716	1.8419	0.9122	0.9164	63.24	403.09	34.87
4b_ft_mix	2.0267	2.1913	0.9399	0.9374	67.72	42.41	38.99
4b_ft_bs_mix	2.2443	2.5641	0.9464	0.9432	68.10	41.55	39.78
4b_ft_bs_ori	2.3601	2.9595	0.8261	0.8273	64.28	44.07	36.84
4b_ft_ori	2.5394	3.0886	0.8317	0.8242	62.52	46.51	35.05
REFQA	–	1.9424	–	0.9131	N/A	N/A	N/A

Table 4: Scores sorted by MetricX. Averaged from 1169 samples.

4.4. Variable-sized context results

In order to assess how the variable-sized context strategies (denoted as "mix") compare to regular fine-tuning, the same test set comprising 1,169 samples was split into two groups: one group with contexts (ES + EN pairs) of 1,000 tokens or more, and the other group with contexts of less than 1,000 tokens. This comparison is provided in Tables 5 and 6.

On the one hand, in both language directions, the short-context group performed similarly to those in the default table, with a slight improvement, mainly in MetricX. Since the difference in the number of samples is small (1,121 vs. 1,169), this behavior is expected. On the other hand, the scores of the long-context group (48 samples) are worth discussing. First of all, the biggest difference lies in MetricX, with scores significantly worse than those of the short-context group. In fact, the application of MetricX to extended contexts suggests a significant technical limitation regarding the cumulative nature of its scoring mechanism. As a regression-based model that evaluates a document as a single holistic unit, MetricX tends to aggregate minor stylistic and terminological deviations from the human reference across the entire text. While these discrepancies might be negligible in shorter segments, they apparently accumulate into an artificially inflated error score in long contexts, as the model lacks a mechanism to distinguish between a single catastrophic error and a series of consistent but technically acceptable stylistic variances. Furthermore, the model's underlying calibration is largely derived from human-annotated datasets such as MQM, which are predominantly composed of sentence-level or short-paragraph fragments. Consequently, the metric's regression head is optimized for short-span judgments and may not scale linearly or accurately when forced to process long contexts.

As for XCOMET scores, given the segmentation approach used to fit the 512-token limit, this metric can potentially fall short of assessing a larger context as a whole. Overall, it provides consistent scores across both groups, with a small decrease in the long-context group. Similarly, traditional metrics suggest good performance in the long-context group, on par with the short-context group, except for the lower half of the ranking, where the original dataset (`ori`) without GPO obtains poor scores, especially in the EN-ES direction.

ID	MX ↓	MX_QE ↓	XC ↑	XC_QE ↑	CHRF ↑	TER ↓	BLEU ↑
Long contexts with ≥ 1,000 tokens (48 samples)							
12b_ft_bs_mix	**8.0723**	5.7770	0.8916	0.9053	**72.35**	**42.36**	**46.93**
4b_noft_gpo	8.1615	**5.6664**	0.9313	0.9263	68.96	46.66	41.09
4b_noft	8.3184	5.8418	0.8820	0.9117	68.70	47.54	40.93
4b_ft_gpo_bs_ori	8.4941	6.2150	0.9013	0.9281	68.11	51.95	41.04
4b_ft_gpo_bs_mix	8.5254	5.9954	0.9131	0.9265	71.63	45.44	45.08
4b_ft_bs_mix	8.5664	6.0687	0.9238	**0.9306**	71.75	46.06	44.90
4b_ft_mix	8.6068	6.1364	0.9020	0.9219	71.37	45.52	44.92
4b_ft_gpo_mix	8.6895	6.2077	0.9032	0.9210	70.65	46.28	44.06
4b_ft_gpo_ori	8.7376	6.4365	0.8951	0.9144	67.29	52.17	40.71
4b_ft_bs_ori	8.7539	7.6302	**0.9425**	0.9100	60.34	57.67	35.90
4b_ft_ori	9.2630	8.7100	0.8743	0.9181	55.17	62.29	32.49
Short contexts with ≤ 1,000 tokens (1121 samples)							
4b_noft_gpo	**2.3453**	**2.1892**	0.9300	0.9404	64.51	46.23	31.36
12b_ft_bs_mix	2.4351	2.6170	**0.9360**	**0.9407**	**70.66**	**37.86**	**39.86**
4b_ft_bs_mix	2.5835	2.7025	0.9254	0.9352	70.13	39.09	38.89
4b_ft_mix	2.6468	2.7965	0.9216	0.9308	69.10	40.30	37.50
4b_noft	2.6839	2.4985	0.9036	0.9140	62.74	231.34	31.79
4b_ft_gpo_bs_mix	2.7356	2.9112	0.9267	0.9366	69.55	39.60	38.23
4b_ft_gpo_mix	2.7921	2.9640	0.9213	0.9289	68.45	41.05	36.93
4b_ft_bs_ori	2.9283	3.3083	0.8862	0.8980	67.51	41.57	36.66
4b_ft_ori	3.0304	3.4628	0.8649	0.8723	66.13	42.93	35.39
4b_ft_gpo_ori	3.1683	3.7681	0.9179	0.9332	65.90	45.16	34.61
4b_ft_gpo_bs_ori	3.6419	4.6802	0.9183	0.9347	64.59	47.74	33.55

Table 5: Ablation analysis of the **ES-EN** pair size impact. Sorted by MetricX.

Regardless of these metric limitations, fine-tuning seems to improve traditional metric scores, with a bigger advantage in the short-context group compared to zero-shot. The `mix` runs using the variable-sized strategies are generally better than their `ori` counterparts in both short and long contexts. However, zero-shot `4b_noft_gpo` offers surprisingly good adaptability to variable-sized contexts off-the-shelf, ranking better than most of the fine-tuned systems in neural metrics, only behind the 12B systems in several metrics, especially the traditional ones.

4.5. Qualitative insights

Regarding `mix` runs, an inspection of the examples with the highest error scores ($MetricX > 10.0$) revealed no evident mistakes made by the models. This further suggests that MetricX has limitations when evaluating long-form contexts.

As for `4b_noft`'s noticeable poor performance in TER, in many instances, this flagging of low-quality segments was attributable to the production of an over-explanation and/or the creation of multiple alternative renderings, causing its output to be heavily penalized by TER's scoring mechanism. This highlights that neural models fail to penalize this kind of typical LLM production, probably exhibiting metric bias or reward hacking. In contrast, traditional metrics, mainly TER, were able to detect the deviation. This configuration with neither fine-tuning nor GPO would also produce its prediction in the source language, consequently yielding extremely poor scores in every metric.

It is worth noting the considerable score disparity observed in translations that received very low ratings despite being arguably correct in fine-tuned systems. One plausible explanation lies in the penalization of terminology errors arising from the translation of product names or institutional references. Most banks and financial institutions operate with established in-house terminology and defined translation strategies (for instance, a no-translation policy). Consequently, a fine-tuned system aligned to these stylistic choices is prone to deviate substantially from the metric criteria, since metrics do not consider these institutional style guidelines. Furthermore, professional human translators often adopt a functionalist approach, prioritizing dynamic equivalence Nida (1964, p. 159) and strategic localization over literal mapping. While these human-centric adaptations ensure the text is "fit for purpose" in a corporate context, they are frequently penalized by automated metrics that favor semantic overlap and stylistic uniformity over

ID	MX ↓	MX_QE ↓	XC ↑	XC_QE ↑	CHRF ↑	TER ↓	BLEU ↑
Long contexts with ≥ 1,000 tokens (48 samples)							
12b_ft_gpo_bs_mix	**5.9720**	5.1175	0.9161	0.9252	**73.94**	**40.04**	**50.07**
4b_ft_gpo_bs_mix	6.3643	5.1175	0.8980	0.9232	70.44	46.31	45.48
4b_noft_gpo	6.3792	**4.7002**	**0.9455**	**0.9413**	69.79	47.22	42.54
4b_ft_gpo_mix	6.4082	5.2314	0.8821	0.8958	71.11	46.04	45.49
4b_noft	6.4538	4.7402	0.8954	0.8709	70.22	46.53	43.27
4b_ft_mix	7.1188	5.4089	0.9243	0.9170	65.80	55.37	40.39
4b_ft_bs_mix	7.4414	5.6484	0.9317	0.9308	60.15	58.21	35.67
4b_ft_gpo_ori	8.2959	6.5846	0.9226	0.9326	52.66	65.89	29.83
4b_ft_gpo_bs_ori	8.3613	7.7751	0.9350	0.9148	49.68	67.07	28.74
4b_ft_bs_ori	10.2780	13.6107	0.8307	0.8333	19.86	89.40	7.58
4b_ft_ori	10.9902	14.1523	0.8491	0.7497	18.70	89.53	7.79
Short contexts with ≤ 1,000 tokens (1121 samples)							
12b_ft_gpo_bs_mix	**1.2976**	1.5390	**0.9540**	0.9462	**71.99**	**36.33**	**43.57**
4b_noft_gpo	1.3704	**1.3287**	0.9508	**0.9550**	64.59	48.03	33.89
4b_ft_gpo_bs_mix	1.6071	1.8729	0.9497	0.9437	69.61	39.83	40.80
4b_ft_gpo_ori	1.6077	1.7508	0.9239	0.9227	67.92	41.79	38.72
4b_ft_gpo_mix	1.6087	1.7773	0.9416	0.9362	68.33	41.35	39.09
4b_ft_gpo_bs_ori	1.6927	2.0202	0.9300	0.9312	68.82	40.10	40.26
4b_noft	1.7797	1.7178	0.9129	0.9184	62.94	418.36	34.51
4b_ft_mix	1.8086	2.0535	0.9406	0.9383	67.80	41.86	38.93
4b_ft_bs_ori	2.0211	2.5034	0.8259	0.8270	66.18	42.13	38.09
4b_ft_bs_mix	2.0218	2.4321	0.9470	0.9437	68.44	40.83	39.96
4b_ft_ori	2.1776	2.6149	0.8309	0.8274	64.40	44.67	36.22

Table 6: Ablation analysis of the **EN-ES** pair size impact. Sorted by MetricX.

nuanced, contextual translation. These factors further explain why REFQA scores may underperform relative to QE ones.

4.6. Comparing results to other works

No direct comparisons with prior work were identified. The most closely related research is discussed below. For instance, the WMT24 repository[2] employed `google/metricx-23-xl-v2p0` (Juraska et al., 2023), an earlier version than the one utilized in this study, and `Unbabel/wmt23-cometkiwi-da-xl`, a reference-free metric comparable to XCOMET_QE scores. The winning system of that edition, Unbabel-Tower70B (Rei et al., 2024), achieved a MetricX score of 1.875 and a CometKiwi score of 0.745. At the time of writing, the WMT25 systems repository had not yet been released.

This prior work (Rajaee et al., 2026) reports XCOMET-XL scores for several language pairs, excluding Spanish, with averages ranging from 74.9 to 80.4. Another study (Oncevay et al., 2025) provides chrF++ and COMET scores for various language pairs; for English-source combinations, chrF++ results range from 43.79 to 66.8, while COMET scores range from 73.06 to 90.87. However, these findings are not directly comparable to the experiments in this study.

[2]https://github.com/wmt-conference/wmt24-news-systems

5. Conclusions

Addressing the research questions: (1) Despite TranslateGemma's notable zero-shot performance, fine-tuning on annual reports better aligns predictions with the reference style—as indicated by traditional metrics and qualitative analysis—even when neural metrics exhibit lower alignment with human references. (2) IBEX 35 companies may find value in this approach given the remarkably low TER, which minimizes Machine Translation Post-Editing (MTPE) effort. (3) The variable-sized context strategies for fine-tuning provide slightly better results in both short- and long-context settings. (4) There seem to be limitations in current neural MT metrics, including metric bias toward model predictions, reward hacking effects, and language direction asymmetry. However, a thorough qualitative analysis should be performed on a larger scale in order to empirically affirm whether that is the reason for the REFQA underperformance against zero-shot QE, consequently limiting fine-tuning effectiveness.

6. Acknowledgements

This work is framed under the Spanish National Project GRESEL (PID2023-151280OB-C21). It was also partially funded by grant PTA2023-023812-I (awarded to Yanco Amor Torterolo Orta) through MICIU/AEI/10.13039/501100011033 and the European Social Fund Plus (ESF+); and by an FPI-UAM scholarship awarded to Melina Chatzi.

7. Bibliographical References

Emad A. Alghamdi, Jezia Zakraoui, and Fares A. Abanmy. 2023. Domain adaptation for arabic machine translation: The case of financial texts.

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.

Paolo Di Natale, Elena Chiocchetti, and Egon Waldemar Stemle. 2025. Meta-evaluation of automatic machine translation metrics between Italian and a minor language variety of German. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 371–383, Cagliari, Italy. CEUR Workshop Proceedings.

Liang Ding, Di Wu, and Dacheng Tao. 2021. Improving neural machine translation by bidirectional training.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, Markus Freitag, and David Vilar. 2026. Translategemma technical report.

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.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628, Singapore. Association for Computational Linguistics.

Christian Herold and Hermann Ney. 2023. On search strategies for document-level neural machine translation.

Leticia Herrero Rodes and Verónica Román Mínguez. 2015. English to spanish translation of the economics and finance genres. *InTRAlinea: Online Translation Journal*, (Special Issue: New Insights into Specialised Translation). Revista del Departamento de Traducción e Interpretación de la Universidad de Bolonia.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Juraj Juraska. 2025. Comment on metricx-25 (issue #12). GitHub Issue Comment. Google Research MetricX Repository.

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.

Juraj Juraska, Tobias Domhan, Mara Finkelstein, Tetsuji Nakagawa, Geza Kovacs, Daniel Deutsch, Pidong Wang, and Markus Freitag. 2025. MetricX-25 and GemSpanEval: Google Translate submissions to the WMT25 evaluation shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 957–968,

Suzhou, China. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google Submission to the WMT 2023 Metrics Shared Task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767, Singapore. Association for Computational Linguistics.

Zixuan Ke, Yifei Ming, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Demystifying domain-adaptive post-training for financial LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31033–31059, Suzhou, China. 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 Lundin, Christof Monz, Kenton Murray, Masaaki Nagata, Stefano Perrella, Lorenzo Proietti, Martin Popel, Maja Popović, Parker Riley, Mariya Shmatova, Steinthór Steingrímsson, Lisa Yankovskaya, and Vilém Zouhar. 2025. 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, Suzhou, China. 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, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. 2024. 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, Maja Popović, Mariya Shmatova, and Jun Suzuki. 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.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. 2024. Mitigating metric bias in minimum Bayes risk decoding. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1063–1094, Miami, Florida, USA. Association for Computational Linguistics.

Haijun Li, Tianqi Shi, Zifu Shang, Yuxuan Han, Xueyu Zhao, Hao Wang, Yu Qian, Zhiqiang Qian, Linlong Xu, Minghao Wu, Chenyang Lyu, Longyue Wang, Gongbo Tang, Weihua Luo, Zhao Xu, and Kaifu Zhang. 2025. Transbench: Benchmarking machine translation for industrial-scale applications.

Mariam Nakhlé, Marco Dinarelli, Raheel Qader, Emmanuelle Esperança-Rodier, and Hervé Blanchon. 2025. Dolfin – document-level financial test set for machine translation.

Eugene A. Nida. 1964. *Toward a Science of Translating*. E. J. Brill, Leiden.

Arturo Oncevay, Charese Smiley, and Xiaomo Liu. 2025. The impact of domain-specific terminology on machine translation for finance in European languages. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2758–2775, Albuquerque, New Mexico. Association for Computational Linguistics.

Cristóbal Parra Quesada and Manuela Cañizares Espada. 2024. Relationship of market capitalization of the ibex 35 to corporate social responsibility and transparency. *Corporate Social Responsibility and Environmental Management*, 31(4):3551–3572.

Sara Rajaee, Sebastian Vincent, Alexandre Berard, Marzieh Fadaee, Kelly Marchisio, and Tom Kocmi. 2026. Unlocking reasoning capability on machine translation in large language models.

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.

Aquia Richburg and Marine Carpuat. 2024. How multilingual are large language models fine-tuned for translation?

Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models.

Mike Zhang and Antonio Toral. 2019. The effect of translationese in machine translation test sets. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 73–81, Florence, Italy. Association for Computational Linguistics.

8. Language Resource References

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Sofía Micaela Roseti, Blanca Carbajo-Coronado, and Jordi Porta. 2025. Financial ES-EN parallel corpus from annual reports.