MLLM-as-a-Judge for Financial Document Image Machine Translation

LLM juez multimodal para traducción automática de documentos financieros a partir de imágenes

Yanco Amor Torterolo Orta,[1,2] **Alexia Stanescu,**[1]
Antonio Moreno-Sandoval,[1] **Ana García Serrano**[2]
[1]Universidad Autónoma de Madrid
[2]Universidad Nacional de Educación a Distancia
{ytorterolo, agarcia}@lsi.uned.es,
maria.stanescu@estudiante.uam.es, antonio.msandoval@uam.es

Abstract: Document Image Machine Translation (DIMT) enables direct translation from visuals, but traditional Machine Translation (MT) metrics ignore the source image during evaluation. This paper explores an end-to-end (E2E) DIMT approach using Gemma 4 on financial reports from IBEX 35 companies. We propose a Multimodal LLM-as-a-judge (MLLMJ) to provide a grounded and explainable evaluation that considers both visual and textual information. It follows a rubric to assign scores and provides both a justification and an internal thinking process leading to the assigned score. Additionally, we estimate the system's potential upper bound through Oracle-Guided Self-Refinement (OGSR), using the judge's feedback to rectify initial predictions. Our study evaluates various configurations, including zero-shot and fine-tuning scenarios across different model sizes. By focusing on image-aware evaluation, this work attempts to offer a framework that provides reasoning cues for improving E2E DIMT systems and enables future knowledge distillation.
Keywords: DIMT, MLLM-as-a-judge, financial MT, oracle-guided self-refinement, Gemma 4

Resumen: La traducción automática de documentos a partir de imágenes (DIMT) permite traducir directamente desde elementos visuales, pero las métricas tradicionales ignoran la imagen. Este trabajo explora un enfoque DIMT *end-to-end* (E2E) que utiliza Gemma 4 en informes financieros de empresas del IBEX 35. Proponemos un LLM Multimodal como Juez (MLLMJ) para obtener evaluaciones fundamentadas y explicables según la información visual y textual. El juez utiliza una rúbrica para puntuar y ofrece una justificación junto a su correspondiente proceso de pensamiento interno. Además, estimamos el *upper bound* mediante un autorrefinamiento guiado por oráculo (OGSR). El modelo corrige sus predicciones iniciales usando el *feedback* del juez. Analizamos varias configuraciones, como *zero-shot*, *fine-tuning* y distintos tamaños. Al centrarse en una evaluación que tiene en cuenta la imagen, este trabajo ofrece un marco que proporciona elementos de razonamiento para mejorar sistemas DIMT E2E y permitir la futura destilación de conocimiento.
Palabras clave: DIMT, MLLM-as-a-judge, TA financiera, autorrefinamiento guiado por oráculo, Gemma 4

1 Introduction

Machine Translation (MT) used to be limited to textual content. However, recent advances in Multimodal Large Language Models (MLLM) enable, for instance, direct end-to-end (E2E) MT from audio to text and from image to text (Pu et al., 2025). Regarding image-to-text E2E MT, most studies refer to it as Document Image Machine Translation (DIMT) when it involves actual documents, or simply IMT if it involves real-world scenes containing text. While it is also called multimodal translation, this term is less specific, as it can also encompass audio-to-text MT. Regarding DIMT, the multimodal models employed for visual tasks are usually referred to as Vision-Language Models (VLMs). These models can interact with images in several

ways—not only for E2E MT, but also for image description, image captioning, Optical Character Recognition (OCR) (Chen et al., 2024), and more.

Traditionally, monomodal MT has been evaluated using text-based metrics, as images were not part of the process (Kocmi et al., 2023; Kocmi et al., 2024; Kocmi et al., 2025). Although neural MT metrics have emerged in recent years to capture the nuances of outputs from generative models, they do not consider the image due to their inherent textual focus. To the best of our knowledge, no metric has been specifically designed to evaluate DIMT with image awareness. However, incorporating the source image is essential to accurately assess the validity and fidelity of this type of translation, where visual context is key. Neglecting to account for this aspect during evaluation limits explainability and prevents a holistic understanding of the process.

This paper aims to develop a Multimodal LLM-as-a-judge (MLLMJ) approach for the assessment of E2E DIMT predictions (Chen et al., 2024). Furthermore, financial documents often present intricate layouts that hinder OCR, much less DIMT. Therefore, this work focuses on three main aspects: (i) conducting E2E DIMT on IBEX 35 companies' annual reports using a state-of-the-art (SOTA) VLM, Gemma 4; (ii) assessing the performance of these E2E DIMT systems with our MLLMJ approach; and (iii) estimating the performance upper bound through Oracle-Guided Self-Refinement (OGSR), where the reasoning produced by the MLLMJ serves as a critique to rectify initial predictions —a process that could pave the way for future knowledge distillation (Kang, Mun, and Han, 2019; Yoon et al., 2023). Additionally, various experiments involving fine-tuning and zero-shot scenarios are conducted locally, comparing two variants of the model.

This E2E DIMT process is significantly more difficult than performing OCR first and then applying MT to the text (Liang et al., 2025a), as there is potential error accumulation resulting from both the visual interpretation of the image and the subsequent translation. The model is forced to perform both tasks in a single process; consequently, we do not anticipate high performance levels.

Several research questions are posed: (i) Is the E2E DIMT approach mature enough to deliver acceptable translations of financial annual reports? (ii) Is Gemma 4's off-the-shelf zero-shot inference acceptable, given its visual reasoning capabilities? (iii) Does fine-tuning improve results? (iv) Does the MLLMJ approach provide a solid evaluation? (v) Is the generation of synthetic reasoning, derived from the MLLMJ evaluation, viable for future distillation?

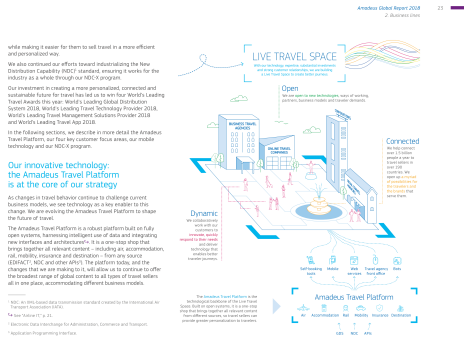
The remainder of this paper is structured as follows: Section 2 presents related work. Section 3 provides details on the dataset, the models, and the experimental configurations. Regarding the specific experiments, Section 4 describes the E2E DIMT approach, Section 5 is devoted to explaining the judge model, and Section 6 details the OGSR process. Section 7 provides the results and analysis; finally, Section 8 presents the conclusions drawn from the findings.

2 Related Work

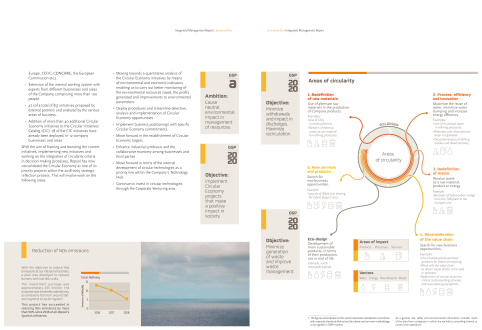
Regarding the DIMT task (Lu et al., 2026; Liang et al., 2024), research of (Liang et al., 2025a) forces the model to perform OCR before MT, which is reported that generate better results. Other studies focus on layout complexity; for example, (Zhang et al., 2025b) proposes reading-order assistance for models. Similarly, (Liang et al., 2025b) focuses on image-to-text alignment to enable lightweight models to perform DIMT. Notably, the emergence of DIMT competitions and benchmarks, such as (Zhang et al., 2025a; Zhuang et al., 2025; Li, Zhu, and Wen, 2025), suggests a strong interest in this field.

Regarding LLM-as-a-judge or MLLMJ approaches, there is a growing body of literature exploring LLMs for textual evaluation across various tasks (Gu et al., 2026; Moreno-Sandoval et al., 2026; Torterolo-Orta, (in press)). (Chen et al., 2024) focuses on multimodal visual tasks, though not specifically on DIMT, whereas (Pu et al., 2025) is not restricted to visual multimodality.

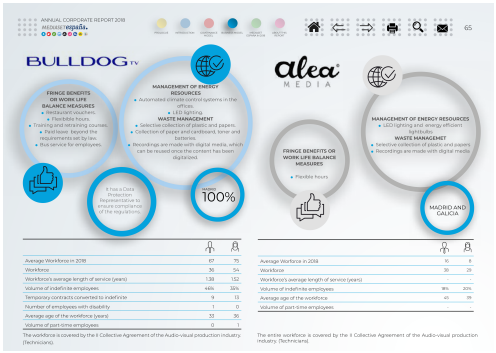
Regarding the use of feedback from a larger model with access to the ground truth—typically for knowledge distillation during fine-tuning—works such as (Kang, Mun, and Han, 2019; Yoon et al., 2023) employ an oracle approach where a teacher model assists a student model. Although our work does not explore distillation, as oracle assistance is applied during inference time, the concept is closely related to our OGSR.



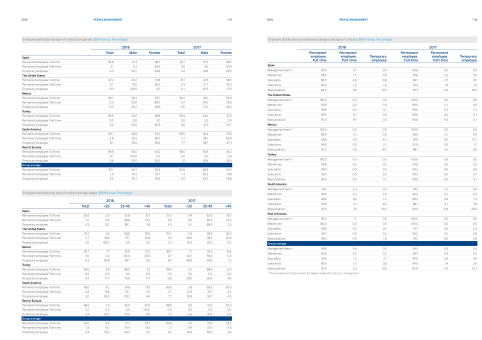
(a) Amadeus sample page



(b) Repsol sample page



(c) Mediaset sample page



(d) BBVA sample page

Figure 1: Representative pages from the IBEX 35 annual reports dataset.

3 *Methodology*

3.1 Dataset

The dataset is made up of Ibex 35 companies' annual reports. These companies make the PDF files available on their respective websites. We used the documents published in 2018, already transcribed. However, this corpus was not originally compiled to be OCR-ready, given that the aim at the time was to obtain clean texts for Natural Language Processing (NLP) tasks such as term extraction. Therefore, structural elements such as page numbers, footnotes, charts, and most tables were omitted by our linguists following the transcription guidelines. For this work, we chose the English PDF versions, which are the official translations of their Spanish counterparts, and their corresponding English transcriptions, which originally consisted of a single TXT file containing all the transcribed narrative content from each individual annual report. Some examples of the dataset can be found in Figure 1.

Given that these financial documents often exhibit an intricate and complex layout that combines odd and even pages in an intertwined manner, PDF pages with a vertical aspect ratio were combined horizontally with the following page, resulting in individual pages. This enables the preservation of the narrative without disrupting the dual-page-dependent reading order. The PDF pages were then converted into PNG files. Afterwards, the corresponding text from each new individual page (whether combined or not) was manually extracted into individual TXT files to match the desired content, thus preparing an OCR-ready dataset.

The next step to obtain an E2E DIMT dataset consisted of translating the dataset with a monomodal MT model that was fine-tuned on the same IBEX 35 annual reports in previous work (Torterolo Orta, Chatzi, and Moreno-Sandoval, 2026). More specifically, the model used was `google/translategemma-12b-it`. We decided not to use the original Spanish versions of the documents, even though we also had them already transcribed. This choice stems from the differences between versions, given that human translators follow a more functional translation approach, with omitted sections, over-explanations, localization decisions, etc. Using MT ensures direct equivalence with the content.

Once the dataset was ready, with individual English PNG files and their corresponding Spanish TXT files, it was converted into the Hugging Face dataset format for convenience. Table 1 shows the dataset split distribution. Since most of the pages were combined, this consequently results in a smaller dataset with high word-density examples.

train	val	test	total
2,104	263	264	2,631

Table 1: Distribution of the dataset sets.

3.2 Explored Models

We explored the Gemma 4 family, a SOTA multimodal model. It has notable visual reasoning combined with a smart and efficient image processor, potentially making it one of the best open-source VLMs. It is worth noting that VLMs usually require low-resolution images, or their processors apply aggressive cropping or tiling to match how they were pretrained. For example, the Gemma 3 family had an 896×896 pixel limit. This is not the case with the Gemma 4 family, which allows higher dynamic resolutions based on a token budget and variable aspect ratios.

Google has released several versions of these models. We utilized **Gemma 4 4B E4B** (where "E" stands for effective) and **Gemma 4 26B A4B** ("A" for active). The former is a compact model designed for edge devices, while the latter is a Mixture of Experts (MoE) model. It is a machine learning approach that divides an AI model into separate sub-networks (or "experts"), each specializing in a subset of the input data, to jointly perform a task. This architecture activates only 4B parameters during inference, even though the entire model must be loaded on the GPU (or partially offloaded). Specifically, for zero-shot experiments, we used `gemma4:e4b-it-q8_0` and `gemma4:26b-a4b-it-q4_K_M` via Ollama. For fine-tuning, we selected `unsloth/gemma-4-E4B-it`. Table 2 summarizes the models and their respective applications in this study.

3.3 Settings

The experiments were carried out locally on consumer-grade hardware. Specifically, the setup features an `AMD Ryzen 7 9800X3D` CPU paired with an `NVIDIA RTX 5080` GPU (16GB of VRAM). Additionally, 32GB of DDR5 RAM was available.

model	role
4B E4B	0-shot inference
4B E4B	fine-tuning and inference
26B A4B	0-shot inference
26B A4B	MLLMJ
26B A4B	OGSR inference

Table 2: Gemma 4 versions and roles.

4 E2E DIMT

The E2E DIMT process followed a structured experimental procedure. Zero-shot inference was conducted via Ollama, while fine-tuning and subsequent inference were executed using Unsloth (Daniel Han and team, 2023). For the sake of replicability, both the system and user prompts are detailed in Figure 2. The system prompt explicitly instructs the model to use the `<|think|>` token to trigger internal reasoning. Given that the dataset lacks specific reasoning annotations explaining how to reach a given translation, this capability was omitted during the fine-tuning experiments. These prompts were tailored to the dataset, explicitly directing the model to omit tables and figures, thus aligning with the formatting conventions of the ground truth.

4.1 DIMT configuration

For all Gemma 4 inference scenarios, we applied the official recommended settings: a temperature of 1, a `top_p` of 0.95, and a `top_k` of 64. To accommodate the high information density of financial reports—where the image requires approximately 1,000 tokens and textual segments total around 5,000 tokens—the context window (`num_ctx`) was set to 16,384. Additionally, the maximum output length (`num_predict`) was capped at 6,144 tokens to provide sufficient budget for verbose reasoning. Regarding Unsloth, the `max_new_tokens` parameter was limited to 4,096, as reasoning was not utilized during the fine-tuning and final inference phases.

The 8-bit Gemma 4 4B E4B version from Ollama was chosen to avoid offloading. Despite being a 4B instruct model, the bf16 size is 16GB. For Unsloth, the full version was loaded, but it was quantized to 4-bit for fine-tuning and to 8-bit for inference. In contrast,

SYSTEM_PROMPT = """<|think|> You are a professional English to Spanish translator and OCR specialist. Translate the English text found in the provided image into Spanish.
Follow these strict formatting rules:
- Produce ONLY the translation. No additional comments or commentary.
- Leave an empty line between each line of generated text (e.g., linea1\n\nlinea2\n\nlinea3).
- Omit irrelevant information such as page numbers, diagrams, schemas, or structural elements.
- Omit tables, even if they include translatable content. Especially if the image is composed of a table, being the main element, even if it includes a title and/or a footnote, omit it entirely (blank output).
- If the image contains no translatable text whatsoever, do not produce any text in the final output (leave it completely blank)."""

USER_PROMPT = "Please translate the text in the provided image."

CHAT_TEMPLATE:

```
{
    "role": "user",
    "content": [
        {"type": "image", "image": sample["img_en"]},
        {"type": "text", "text": f"{SYSTEM_PROMPT}\n\n{USER_PROMPT}"}
    ],
},
{
    "role": "assistant",
    "content": [
        {"type": "text", "text": sample["tra_es"]}
    ]
}
```

Figure 2: System and user prompts used for Ollama inference, and the chat template used for fine-tuning Gemma 4. The `<|think|>` token was not used for the fine-tuning experiments, but the rest of the system prompt remained the same.

the 4-bit Gemma 4 26B A4B was chosen, but offloading could not be avoided due to hardware constraints (16GB of VRAM). The model was 74% loaded on the GPU while the rest was offloaded.

Gemma 4 models require the image to be provided before any other element, as shown in the chat template in Figure 2. While Gemma 4 supports dynamic resolution with a token budget of up to 1,120 tokens, the resolution must be explicitly specified, especially when using Unsloth, as it defaults to 512x512 pixels (roughly 144 tokens) if left undefined. We set the budget to the maximum 1,120-token limit and the resolution to 1536x1536 pixels for both Ollama and Unsloth experiments (which corresponds to approximately 1,024 tokens). The images were automatically adjusted to the established size while preserving their original aspect ratio with padding, taking full advantage of Gemma 4's compatibility with variable aspect ratio images.

5 MLLM-as-a-Judge and Metrics

The E2E DIMT experiments were evaluated with the MLLMJ, but they were also complemented by MT metrics. Even though these metrics do not account for the image when assigning scores, they are valuable for comparison purposes. Metrics employed are: MetricX (**MX**) (Juraska et al., 2024),

SYSTEM_PROMPT = """<|think|> You are a professional English to Spanish translator and OCR model undergoing an expert supervision phase.
Previously, you generated a faulty translation for the attached document. An expert Judge reviewed your output against the Gold Reference.

Your objective:
1. Examine your old answer.
2. Read the Judge's critique carefully.
3. Observe the Gold Standard Reference.
4. Correct your structural and semantic mistakes internally during your thinking phase to learn how to produce the Gold Standard naturally.
5. Provide the flawless final output obeying the strict formatting rules.

Formatting rules:
- Produce ONLY the translation. No additional comments or commentary outside of your reasoning.
- Leave an empty line between each line of generated text.
- Omit irrelevant information (page numbers, schemas).
- Omit tables entirely (if the image is a table/contains a main table, your output must be completely blank).
- If the Gold Standard Reference says there is no translation expected (BLANK), you must output absolutely nothing."""

user_content = f"""=== YOUR PREVIOUS PREDICTION ===
{display_old_pred}

=== JUDGE EXPERT FEEDBACK ===
Score: {old_score}/5
Feedback snippet: {judge_reasoning}
Judge's internal logic: {judge_thinking}

=== GOLD STANDARD REFERENCE ===
{display_tra_es}

Please correct your answer and provide the perfect final extraction."""

Figure 3: System and user prompts employed for the OGSR inference stage.

xCOMET (**XC**) (Guerreiro et al., 2024), **chrF** (Popović, 2015), Translation Edit Rate (**TER**) (Snover et al., 2006), and **BLEU** (Papineni et al., 2002). Additionally, the dataset was evaluated using the reference-free capabilities of MX and XC, comparing the EN source text with the ES reference translation. We refer to this as Reference Quality Estimation (REFQA), which enables us to establish an upper bound for these two metrics on this specific dataset.

Regarding the MLLMJ evaluation, the model is asked to provide a prediction with a score from 1 to 5, following a Likert scale. A rubric with guidelines on how to assign the scores is given to the model in the prompt (Appendix A). The MLLM has access, in this order, to the source EN image, the prompt containing the rubric, the textual ES reference translation, and the textual ES prediction to be assessed. The model must provide a score and a short explanation of its decision, and the thinking process is extracted to enable next OGSR phase and future distillation. The thinking process in this case is detailed internal reasoning that led the MLLMJ to assign a specific score. It is useful for explainability and traceability.

Task	Gemma	Dataset			Prediction	MLLMJ		
		image (EN)	OCR (EN)	trans. ref. (ES)		score	reasoning	thinking
E2E DIMT	4B 4B ft 26B	V	X	X	output[1]	N/A	N/A	N/A
MLLMJ	26B	V	X	V	V[1]	output[2]	output[3]	output[4]
OGSR	26B	V	X	V	V[1] output	V[2]	V[3]	V[4]

Table 3: Input and output flow for each experimental stage and model role. Superscripts denote the data flow between stages: generated outputs (*output*) from one stage become the inputs (V) for subsequent tasks matching the same numeric index.

5.1 MLLMJ configuration

Given that the model must be able to ingest a higher number of tokens, `num_ctx` was set to 16,384 tokens, and `num_predict` was set to 8,192 tokens. Ollama was used for MLLMJ inference, with the same settings for temperature, `top_p`, and `top_k`. The image preprocessing is also the same.

In order to avoid unnecessary computing, several hard-coded heuristics and execution optimization mechanisms were implemented: (i) if the textual EN source and the textual ES reference are blank, it means that the prediction should be empty, bypassing the MLLMJ and assigning a 5 when this happens; (ii) if both reference elements are blank again, but the prediction contains text, it automatically assigns a 1; and (iii) if both reference elements contain text, but the prediction is blank, it also assigns a 1.

6 Oracle-Guided Self-Refinement

Thanks to the score, the short justification, and the internal thinking process provided by the MLLMJ during the evaluation phase, an OGSR process is possible. This usually involves improving translation quality by having an LLM iteratively critique and revise translations, often leveraging "oracle" (ground truth or high-accuracy) feedback to correct specific errors rather than relying solely on re-training the model. Specifically, our OGSR approach consists of using the MLLMJ's feedback to correct the prior prediction made during E2E DIMT. Figure 3 helps explain the whole pipeline so far, including input and output flow. Table 2 can also help clarify which model version was used during each phase. The system prompt reflecting this correction task can be found in Figure 3. The model is provided with the system prompt followed by a user prompt that incorporates: the source image, the previous prediction, the numeric judge score (1–5), the short judge justification, the judge's internal reasoning (thinking trace), and the gold standard Spanish reference translation.

Since this inference requires the largest number of tokens, `num_ctx` was set to 32,768 and `num_predict` to 8,192. The remaining settings are the same as in previous sections. A heuristic rule has also been implemented: if a prediction received a score of 5 in the previous evaluation, it bypasses the model, as no improvement is required.

7 Results and Analysis

Table 4 contains the metrics obtained by the different systems. First of all, it is worth noting that the textual REFQA for MX and XC are acceptable but suboptimal, establishing a functional ceiling for these neural metrics and probably suggesting overall discrepancies with the translations. The textual MT metrics are rather poor in general terms, but this was expected, since the translation is performed from the EN image, not from the source EN text. This reflects the challenges of performing E2E DIMT.

Regarding MLLMJ scores, they are well-aligned with the textual MT metrics, with no apparent outliers. `26B 0-shot R` was the best-performing system for the E2E DIMT task (excluding the OGSR run, which should not be compared with the rest). The larger model unsurprisingly outperforms the smaller one. In contrast, the effect of fine-tuning is worth discussing. `4B ft no R` experienced a significant downgrade compared to its zero-

ID	MLLMJ ↑	MX ↓	XC ↑	CHRF ↑	TER ↓	BLEU ↑
OGSR (not comparable - unfair comparison)						
26B 0-shot R OGSR	**4.4886**	**5.9954**	**0.7735**	**91.09**	**26.10**	**83.88**
E2E DIMT Systems						
26B 0-shot R	**2.6616**	**9.3562**	**0.4055**	**63.98**	**72.96**	**41.91**
4B 0-shot R	2.4015	11.4653	0.2789	55.91	87.81	30.85
4B ft no R	2.0720	12.3217	0.2898	51.05	110.12	24.57
Dataset REFQA (EN source vs ES reference)						
REFQA	N/A	4.9738	0.9009	N/A	N/A	N/A

Table 4: Performance comparison of E2E DIMT systems, sorted by MLLMJ score (1–5). Results represent the mean over 264 samples. "R" stands for reasoning, and "ft" for fine-tuned.

System	0	1	2	3	4	5	Total
26B 0-shot R OGSR	2	22	1	6	22	211	264
26B 0-shot R	1	115	25	31	18	74	264
4B 0-shot R	3	72	91	46	18	34	264
4B ft no R	2	87	109	38	12	16	264
Global	**8**	**296**	**226**	**121**	**70**	**335**	**1056**

Table 5: Score distribution across systems (0—5).

shot counterpart. This clearly stems from the inability to fine-tune it with its reasoning capabilities, as the dataset does not contain reasoning traces leading to the correct answer. So, reasoning traces seems a crucial skill for models to achieve better predictions.

Table 5 shows the distribution of scores across systems. The scores of zero stem from parsing errors in the evaluation JSON and the fallback mechanism. Evidently, despite this being a SOTA model, it can still struggle to generate valid JSON outputs. For E2E DIMT systems, surprisingly, `26B 0-shot R` received more scores of 1 than the 4B versions, but in turn, it received fewer scores of 2 and more scores of 5. Regarding `26B 0-shot R OGSR`, it achieved the best results, with 211 scores of 5 out of 264 predictions.

7.1 Qualitative Analysis

The MLLMJ's evaluations of the `26B 0-shot R` system have undergone an in-depth qualitative analysis to assess both the performance of the best E2E DIMT system and the evaluation quality. Upon inspection, the scores and rationales are accurate and correctly reflect the rubric criteria. This implies that the underperformance of the E2E DIMT systems does not stem from the metric's inability to capture visual and textual errors.

There are several phenomena worth discussing. First of all, there is a significant number of predictions that received a score of 1 because the model ran out of tokens even before generating an actual answer. The culprit is the prior internal reasoning, which consumes significantly more tokens than anticipated, reaching the output limit abruptly. When the output limit was reached during the generation of the actual answer, the score varies from 2 to 3, depending on how much was left to translate and how it captured the core idea. An example is provided in Figure 4. This perfectly aligns with the OGSR not experiencing the same issue, given the massive context and generation limits it was allocated. Consequently, if the model had been granted a larger token limit, the scores could potentially have been notably better.

Other issues related to the quality of the generated translations have been detected, most notably hallucinations. These include errors in proper nouns and figures, which are considered severe mistakes in financial translation. Examples include generating `Cristina Zamora` instead of `Cristina Zamorano`, `AECFA` instead of `AEGFA`, or `Prince of Water` instead of `Prince of Peace`. According to the MLLMJ, these "are not minor typos but significant errors that distort the factual content of the source text." This example received a score of 2, following the

Score: 2
Reason: The student's prediction is significantly incomplete. It completely omits the entirety of page 71 and the "Activities" and "2018 MILESTONES" sections from page 72. Furthermore, the text on page 72 is truncated mid-sentence. While the portion of text that is present is transcribed/reconstructed accurately, the loss of a large, crucial portion of the document's content is a major failure.
Thinking excerpt: Final check: Is there any hallucination? No. Is the text present accurate? Yes, very accurate. Is the text complete? No, very much not complete. Is the missing text "crucial"? Yes, nearly half the image content is gone. er/footer) is not hallucinated, it's in the image.

Figure 4: Example with a score of 2.

rubric (Appendix A). However, the MLLMJ tolerates certain minor hallucinations, especially if the translation is otherwise accurate. In this regard, the following is an example of such tolerance, which received a score of 4: "There are only a few minor issues: a small typo (`innovula` instead of `innovadora`) and a slight omission in one paragraph (the part about the 'review of the end-to-end logistics process' is missing). However, the core semantic meaning remains completely faithful to the original text and the reference."

It is worth noting that the MLLMJ is able to recognize when the prediction is better than the textual reference by leveraging the visual source, highlighting the importance of multimodal evaluation. Consider, for instance, this justification for a score of 5: "Although there are minor typos (e.g., `Reallia` instead of `Realia`), the translation is much more faithful to the English source text in the image than the provided Spanish reference (which uses `alaos` instead of `halcones peregrinos`). The quality is top-tier." The judge tolerated a minor error detected through the visual input and deemed the prediction superior to the reference. Interestingly, it also detected a mistake in the textual reference. Therefore, its evaluation is not limited to seeking a perfect match to the reference when the latter exhibits mistakes compared to the image.

7.2 OGSR and Future Distillation

The model successfully utilized the feedback instead of simply copying the textual ES reference translation; otherwise, it would have achieved perfect scores. Therefore, given the strong visual, reasoning and instruction-following capabilities of Gemma 4 using such a large context window, it is clear that a future distillation process could yield competitive results.

Regarding distillation, it was beyond the scope of this paper, but it shows promise. Both the reasoning from the MLLMJ and the corrections applied by the OGSR system are clearly suitable for distillation. One of the issues encountered when fine-tuning Gemma 4B was the performance degradation resulting from the omission of reasoning. However, performance is expected to increase with a fine-tuning process that incorporates reasoning, as suggested by the improvements observed in OGSR.

7.3 Emissions and Energy

Figure 5 provides an analysis of the energy consumption and carbon emissions associated with the systems. The most notable finding is that the 26B models are the most efficient systems in terms of inference. For instance, although `26B 0-shot R OGSR` required a massive context window and is significantly larger than its 4B counterpart, its efficiency stems from its MoE architecture. It only activates 4B active parameters during inference, which is beneficial for its environmental footprint. Therefore, MoE models appear to be the optimal choice, provided they can fit within the hardware constraints. In contrast, the most inefficient model was TranslateGemma, which can be explained by its 12B effective size. Regarding the difference between `4B ft no R` and `4B 0-shot R`, the results are expected because the former lacks the internal reasoning process, saving computational resources and time. Reasoning is usually verbose and tends to be longer than the actual prediction.

8 Conclusions

This paper presents several E2E DIMT experiments, describes an MLLMJ system, and introduces a promising OGSR approach to leverage the judge's feedback. Addressing the research questions: (i) the E2E DIMT approach faces more challenges than sequential

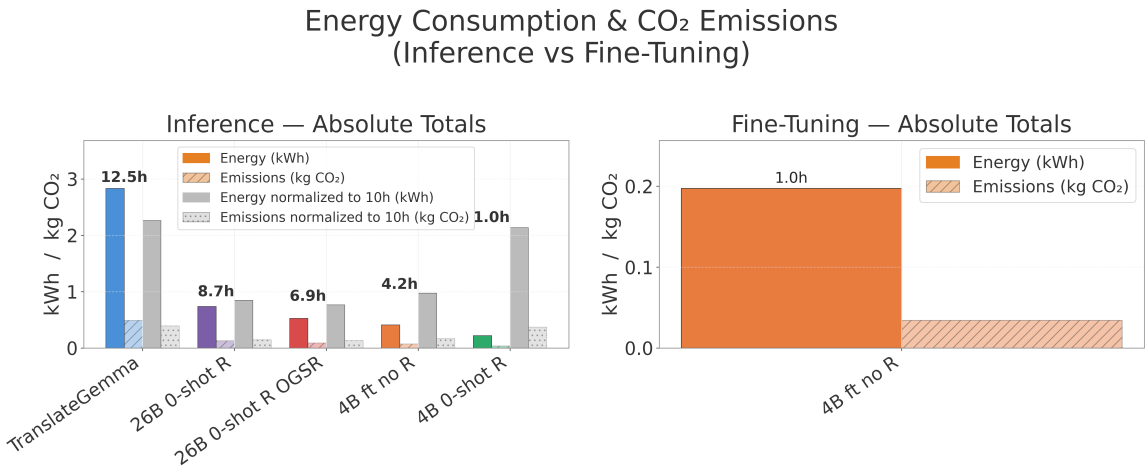


Figure 5: Comparison of energy consumption, carbon emissions, and execution time during inference (left) and fine-tuning (right).

OCR and MT due to potential error accumulation. However, it is promising and can be further optimized, as suggested by OGSR's upper bound, since Gemma 4's visual and reasoning abilities are notable. (ii) Gemma 4's off-the-shelf zero-shot performance is remarkable thanks to its visual reasoning and internal thinking process, which lead the model to better predictions. In contrast, (iii) fine-tuning without reasoning degrades the scores. (iv) MLLMJ has proven to be a step forward in E2E DIMT evaluation, as it incorporates the image into the process and does not rely solely on the textual reference. It also improves explainability during evaluation. Finally, (v) the synthetically generated reasoning obtained from the MLLMJ and the OGSR is a valuable resource resulting from these experiments, as it can potentially enable the distillation of smaller models to mimic the reasoning necessary to reach correct predictions.

Limitations

Due to time and length constraints, the MLLMJ analysis currently lacks quantified correlation with human judgments via Pearson correlation coefficients or similar metrics. Consequently, human evaluation should be conducted to provide a definitive baseline. Additionally, while the OGSR experiment is useful for establishing an upper bound, it does not directly demonstrate the effectiveness of knowledge distillation in smaller models using a reasoning-based dataset powered by the judge's feedback. These aspects will be explored in future work.

Replicability and Open Science

The fine-tuned adapters of the monomodal TranslateGemma model used to prepare the multimodal dataset, described in Section 3, are available at (Torterolo Orta, 2026a). The multimodal Gemma 4 E4B fine-tuned adapters are provided in (Torterolo Orta and Stanescu, 2026). The model outputs, including their respective scores, evaluations, and MLLMJ reasoning, are accessible via (Torterolo Orta, 2026b). These files also contain the textual ground truth references.

Due to copyright concerns, we cannot share the dataset directly; however, the original PDFs are publicly available on each IBEX 35 company's website. Appendix B lists the hyperparameters used during fine-tuning for replicability purposes.

Acknowledgments

This work is part of the coordinated Spanish national project GRESEL: UAM (PID2023-151280OB-C21) and UNED (PID2023-151280OB-C22). 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 the UAM-IIC Chair in Computational Linguistics, which funded Alexia Stanescu's contract.

AI Use Statement

Gemini was utilized to enhance grammatical accuracy and refine the sentence structure of the original English versions. Figure 5 was also enhanced using AI-assisted tools.

References

Chen, D., R. Chen, S. Zhang, Y. Liu, Y. Wang, H. Zhou, Q. Zhang, Y. Wan, P. Zhou, and L. Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark.

Daniel Han, M. H. and U. team. 2023. Unsloth.

Gu, J., X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Z. Lin, B. Zhang, L. Ni, W. Gao, Y. Wang, and J. Guo. 2026. A survey on llm-as-a-judge. *The Innovation*, page 101253.

Guerreiro, N. M., R. Rei, D. v. Stigt, L. Coheur, P. Colombo, and A. F. T. Martins. 2024. xCOMET: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Juraska, J., D. Deutsch, M. Finkelstein, and M. Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA, November. Association for Computational Linguistics.

Kang, M., J. Mun, and B. Han. 2019. Towards oracle knowledge distillation with neural architecture search.

Kocmi, T., E. Artemova, E. Avramidis, R. Bawden, O. Bojar, K. Dranch, A. Dvorkovich, S. Dukanov, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, H. Lakougna, J. Lundin, C. Monz, K. Murray, M. Nagata, S. Perrella, L. Proietti, M. Popel, M. Popović, P. Riley, M. Shmatova, S. Steingrímsson, L. Yankovskaya, and V. Zouhar. 2025. Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China, November. Association for Computational Linguistics.

Kocmi, T., E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, M. Karpinska, P. Koehn, B. Marie, C. Monz, K. Murray, M. Nagata, M. Popel, M. Popović, M. Shmatova, S. Steingrímsson, and V. Zouhar. 2024. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In B. Haddow, T. Kocmi, P. Koehn, and C. Monz, editors, *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46, Miami, Florida, USA, November. Association for Computational Linguistics.

Kocmi, T., E. Avramidis, R. Bawden, O. Bojar, A. Dvorkovich, C. Federmann, M. Fishel, M. Freitag, T. Gowda, R. Grundkiewicz, B. Haddow, P. Koehn, B. Marie, C. Monz, M. Morishita, K. Murray, M. Nagata, T. Nakazawa, M. Popel, M. Popović, M. Shmatova, and J. Suzuki. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In P. Koehn, B. Haddow, T. Kocmi, and C. Monz, editors, *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore, December. Association for Computational Linguistics.

Li, B., S. Zhu, and L. Wen. 2025. MIT-10M: A large scale parallel corpus of multilingual image translation. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, and S. Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5154–5167, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Liang, Y., Y. Zhang, C. Ma, Z. Zhang, Y. Zhao, L. Xiang, C. Zong, and Y. Zhou. 2024. Document image machine translation with dynamic multi-pre-trained models assembling. In K. Duh, H. Gomez, and S. Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7084–7095, Mexico City, Mexico, June. Association for Computational Linguistics.

Liang, Y., Y. Zhang, Z. Zhang, Z. Chen, Y. Zhao, L. Xiang, C. Zong, and Y. Zhou. 2025a. Improving MLLM's document image machine translation via synchronously self-reviewing its OCR proficiency. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23659–23678, Vienna, Austria, July. Association for Computational Linguistics.

Liang, Y., Y. Zhang, Z. Zhang, Y. Zhao, L. Xiang, C. Zong, and Y. Zhou. 2025b. Single-to-mix modality alignment with multimodal large language model for document image machine translation. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12391–12408, Vienna, Austria, July. Association for Computational Linguistics.

Lu, J., T. Song, Z. Wu, P. Li, X. Liang, H. Yang, K. Chen, N. Xie, Y. Lu, J. Zhao, S. Sun, and D. Wei. 2026. Global-local dual perception for mllms in high-resolution text-rich image translation.

Moreno-Sandoval, A., J. Porta, Y. Torterolo, A. Stanescu, M. Chatzi, and S. Roseti. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026). In A. Moreno Sandoval and P. Martínez, editors, *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, pages 114–124, Palma de Mallorca, Spain, May. ELRA.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, and D. Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Popović, M. 2015. chrF: character n-gram F-score for automatic MT evaluation. In O. Bojar, R. Chatterjee, C. Federmann, B. Haddow, C. Hokamp, M. Huck, V. Logacheva, and P. Pecina, editors, *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Pu, S., Y. Wang, D. Chen, Y. Chen, G. Wang, Q. Qin, Z. Zhang, Z. Zhang, Z. Zhou, S. Gong, Y. Gui, Y. Wan, and P. S. Yu. 2025. Judge anything: Mllm as a judge across any modality.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla, and J. Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231, Cambridge, Massachusetts, USA, August 8-12. Association for Machine Translation in the Americas.

Torterolo-Orta, Y. A. (in press). AI meets literary corpora: LLMs for advanced synthetic data generation and automated evaluation in Luces de Bohemia. In A. Gil-Casadomet and M. Tordesillas, editors, *Research in Language Sciences: International and Interdisciplinary Perspectives on Languages and Communication*. Peter Lang.

Torterolo Orta, Y. A. 2026a. Bidirectional ES-EN financial translation model: translategemma-12b LoRA adapters using parallel annual reports from IBEX 35 companies. https://doi.org/10.21950/NKX2YN. e-cienciaDatos, V1.

Torterolo Orta, Y. A. 2026b. Results and model outputs from the paper: MLLM-as-a-Judge for Financial Document Image Machine Translation. https://doi.org/10.21950/VBG0HP. e-cienciaDatos, V1.

Torterolo Orta, Y. A., M. Chatzi, and A. Moreno-Sandoval. 2026. TranslateGemma for ES-EN Financial Reports: Exploring Adaptability to Variable-Sized Contexts. In A. Moreno Sandoval and P. Martínez, editors, *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, pages 87–97, Palma de Mallorca, Spain, May. ELRA.

Torterolo Orta, Y. A. and M. A. Stanescu. 2026. EN-ES financial multimodal translation model (DIMT): gemma-4-E4B LoRA adapters (EN image

-> ES text) fine-tuned on annual reports from IBEX 35 companies. https://doi.org/10.21950/HEZRAQ. e-cienciaDatos, V1.

Yoon, J. W., H. Y. Kim, H. Lee, S. Ahn, and N. S. Kim. 2023. Oracle teacher: Leveraging target information for better knowledge distillation of ctc models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2974–2987.

Zhang, Y., Y. Liang, Z. Zhang, Z. Chen, L. Xiang, Y. Zhao, Y. Zhou, and C. Zong, 2025a. *ICDAR 2025 Competition on End-to-End Document Image Machine Translation Towards Complex Layouts*, page 505–522. Springer Nature Switzerland, September.

Zhang, Z., Y. Zhang, Y. Liang, C. Ma, L. Xiang, Y. Zhao, Y. Zhou, and C. Zong. 2025b. Reading when translating: Multimodal document image machine translation with reading flow prediction. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2606–2621.

Zhuang, W., W. Li, Z. Lan, X. Han, P. Li, and J. Su. 2025. Patimt-bench: A multi-scenario benchmark for position-aware text image machine translation in large vision-language models.

A MLLM-as-a-Judge Rubric

```
<|think|> You are an expert evaluator assessing the quality of a multimodal OCR translation.
Your task is to rate how accurately a STUDENT_PREDICTION matches the REFERENCE_ANSWER and visually relates to the provided IMAGE.

RUBRIC (score from 1 to 5):

- 5: Excellent quality: The prediction captures the full meaning exactly as the Reference. It does not contain hallucinated tables, schemas, or extra content. It perfectly reconstructs the textual meaning of the provided Image.
- 4: Good quality: The translation is highly accurate but has very minor formal errors (e.g., punctuation) or added a trivial visual element. The core semantic meaning remains completely faithful to the Reference.
- 3: Medium quality: The prediction captures the core translation but is visibly incomplete, contains obvious OCR mistakes compared to the Image, or includes noticeable structural noise.
- 2: Low quality: Crucial portions of the source text are missing or there are severe semantic translation errors compared to the Reference and Image.
- 1: Very poor quality: The prediction completely failed. It may be hallucinated, left in English, incorrectly structured as a raw diagram/table, or is complete nonsense (e.g., "[ERROR]").

First, briefly explain your reasoning. Then assign a single integer score from 1 to 5.

Return your response as pure JSON with this exact schema (do not use markdown formatting blocks if possible):
{{
"score": <integer 1-5>,
"reasoning": "<short explanation>"
}}

REFERENCE_SPANISH:
{reference}

STUDENT_PREDICTION:
{prediction}
```

B Fine-Tuning Hyperparameters

```
MODEL_NAME = "unsloth/gemma-4-E4B-it"
model, processor = FastVisionModel.from_pretrained(
    MODEL_NAME,
    load_in_4bit = True,
    use_gradient_checkpointing = "unsloth",
    attn_implementation = "sdpa",
)

model = FastVisionModel.get_peft_model(
    model,
    finetune_vision_layers     = True,
    finetune_language_layers   = True,
    finetune_attention_modules = True,
    finetune_mlp_modules       = True,

    r = 32,
    lora_alpha = 32,
    lora_dropout = 0,
    bias = "none",
    random_state = 3407,
)

trainer = SFTTrainer(
    model = model,
    train_dataset = train_dataset,
    eval_dataset = eval_dataset,
    processing_class = processor.tokenizer,
    data_collator = UnslothVisionDataCollator(model, processor),
    args = SFTConfig(
        per_device_train_batch_size = 2,
        gradient_accumulation_steps = 8,
        per_device_eval_batch_size = 1,
        eval_accumulation_steps = 4,
```

```
        warmup_steps = 100,
        num_train_epochs = 5,
        learning_rate = 1e-5,
        logging_steps = 50,

        save_strategy = "steps",
        save_steps = 50,
        eval_strategy = "steps",
        eval_steps = 50,

        optim = "adamw_8bit",
        weight_decay = 0.01,
        lr_scheduler_type = "cosine",
        seed = 3407,
        output_dir = OUTPUT_MODEL_DIR,
        report_to = "none",
        load_best_model_at_end = True,
        save_total_limit = 3,
        metric_for_best_model = "eval_loss",
        greater_is_better = False,

        remove_unused_columns = False,
        dataset_text_field = "",
        dataset_kwargs =
        ↪  {"skip_prepare_dataset": True},
        max_length = MAX_SEQ_LENGTH,
    ),
    callbacks = [
        FullLogger(FULL_LOGS_PATH),
        EarlyStoppingCallback(early_stopping
        ↪  _patience=3)
    ]
)
```