

SemEval-2025 Task 2: Entity-Aware Machine Translation

Simone Conia

Sapienza University of Rome
conia@diag.uniroma1.it

Min Li

Apple
min_li6@apple.com

Roberto Navigli

Sapienza University of Rome
navigli@diag.uniroma1.it

Saloni Potdar

Apple
s_potdar@apple.com

Abstract

Translating text that contains complex or challenging named entities—e.g., culture-specific book and movie titles, location names, proper nouns, food names, etc.—remains a difficult task for modern machine translation systems, including the latest large language models. To systematically study and advance progress in this area, we organized the first edition of **Entity-Aware Machine Translation**, or *EA-MT*, a shared task that evaluates how well systems handle entity translation across 10 language pairs. With EA-MT, we introduce XC-Translate, a novel gold benchmark comprising over 50K *manually*-translated sentences with entity names that can deviate significantly from word-to-word translations in their target languages. This paper describes the creation process of XC-Translate, provides an overview of the approaches explored by our participants, presents the main evaluation findings, and points toward open research directions, such as contextual retrieval methods for low-resource entities and more robust evaluation metrics for entity correctness. We hope that our shared task will inspire further research in entity-aware machine translation and foster the development of more culturally-accurate translation systems.

Resources and Links



Website for EA-MT

sapienzanlp.github.io/ea-mt/



Benchmark on HF Datasets

huggingface.co/datasets/sapienzanlp/ea-mt-benchmark



Leaderboard on HF Spaces

huggingface.co/spaces/sapienzanlp/ea-mt-leaderboard



Official Scorer on GitHub

github.com/SapienzaNLP/ea-mt-eval

1 Introduction

Background. The emergence of multilingual large language models (LLMs) and the wide availability of massive multilingual datasets have significantly advanced the field of Machine Translation (MT) (Fan et al., 2021; Tang et al., 2021; Costa-jussà et al., 2022; Kudugunta et al., 2023, *inter alia*). These developments have led to MT systems that not only perform exceptionally well in high-resource languages but also support a growing number of low-resource languages (Fan et al., 2021; Tang et al., 2021; Costa-jussà et al., 2022; Kudugunta et al., 2023, *inter alia*). Nevertheless, the research community still faces several unresolved challenges in MT. Among these, the translation of text that contains entities is still a hard task, especially with particular categories of entities, e.g., movies, books, food, locations, and sometimes even people, to name a few. Indeed, *word-for-word*, or *literal*, translations of their names may not be suitable due to culture-specific references, which can vary depending on social, geographical, and historical contexts, among other factors (Hershcovich et al., 2022).

Motivation. In this context, the challenge lies in accurately identifying when and how to translate entities whose names are significantly different across languages, not because of differences in the script (e.g., English and Chinese) but because of differences in the cultural context. This step is crucial, as relying on literal translations may not convey the intended meaning, risking the effectiveness of the entire translation process (Gaballo, 2012; Díaz-Millón and Olvera-Lobo, 2023; Conia et al., 2024). For example, if we were to translate word-for-word “Qual è la trama de *Il Giovane Holden*?” from Italian to English, we could obtain “What is the plot of *The Young Holden*?”, which is grammatically correct but semantically incorrect. The correct translation “What is the plot of *The Catcher in the Rye*?”

necessitates not only fluency in both the source and target languages but also knowledge of the cultural contexts involved. Current systems often struggle with this task; however, the research community lacks i) a comprehensive benchmark specifically designed to evaluate the performance of MT systems in translating text containing entities, and ii) evaluation metrics that can accurately measure the quality of the translations produced by these systems, as current metrics (e.g., BLEU and COMET) are not designed to capture the quality of entity translations.

Summary of the task. To address this challenge, we organized the first edition of **Entity-Aware Machine Translation (EA-MT)**, a new shared task whose goal is to track the progress and encourage the development of MT systems that can better handle the translation of text containing entities with names that are significantly different across languages. Given a sentence s in English containing an entity e , the task is to translate s into a target language while adapting the name of e to the target language to preserve the original meaning of the sentence. The first edition of EA-MT focuses on:

- text containing entities from various categories, such as movies, books, food, locations, and people, among others;
- entities whose names are significantly different across languages, e.g., *Il Giovane Holden* and *The Catcher in the Rye*;
- translating simple sentences, as the challenge shall lie in the translation of the entity names rather than in the complexity of the sentence structure.

In this task, we provide participants with a dataset containing English sentences with entities and their translations into 10 other languages: *Arabic, Chinese, French, German, Italian, Japanese, Korean, Spanish, Thai, and Turkish*.

Contributions. The first edition of EA-MT attracted around 50 teams, who submitted around 300 runs. Among these, 25 teams submitted their final results for the official leaderboard and 18 of them illustrated their approaches and results in system description papers. In summary, the main contributions of the first edition of EA-MT are:

- **XC-Translate**, a novel benchmark for evaluating the performance of MT systems in trans-

lating text containing entities with names that are significantly different across languages;

- **M-ETA**, a new evaluation metric that can accurately measure the quality of the translations produced by MT systems, focusing on the translation of entity names;
- an analysis of the approaches introduced by the participants and their results, highlighting the strengths and weaknesses of current MT systems in handling entity translation.

We release the benchmark and the evaluation metric to the research community, with the hope that they will encourage further research in this area.

2 Entity-Aware Machine Translation

Task definition. We introduce Entity-Aware Machine Translation (EA-MT), a new shared task that evaluates how well systems handle entity translation across 10 language pairs. More formally, EA-MT is defined as follows: given a source sentence s_e in English that contains an entity mention e , the goal is to produce a translation t in a target language l_{target} that correctly adapts the entity to its culturally-appropriate equivalent e' in that language. For each sentence s_e in English, we have:

$$f(s_e, l_{\text{target}}) \rightarrow t \quad (1)$$

where f represents the translation function, l_{target} is one of the 10 target languages in our benchmark, and t contains the culturally-appropriate equivalent e' of the entity e .

Key challenges. The main challenge of EA-MT lies in the fact that we selected a set of entities whose names are significantly different across languages, e.g., *The Catcher in the Rye* (English), *Il Giovane Holden* (Italian), *El guardián entre el ceniteno* (Spanish), and *L'atrape-cœurs* (French).¹ To address this challenge, an MT system must ensure that the entity name is adapted to its culturally-appropriate equivalent in the target language, which may require a transcreation step instead of a literal translation. To stress the importance of translating the entity names correctly, we also introduce M-ETA, a new evaluation metric that focuses on the quality of the translations of entity names, as described in Section 4. With M-ETA, systems that

¹We provide more details about how we selected “challenging” entities in Section 4.

produce fluent translations but fail to adapt the entity names correctly are strongly penalized, revealing the limitations of current MT systems—and evaluation metrics—in handling entity translation.

Differences with previous MT tasks. EA-MT differs from previous MT tasks in that it focuses on the translation of text containing entities with names that are significantly different across languages, which is a challenge that has not been systematically studied before with a dedicated shared task, across multiple languages and at a significant scale. Previous benchmarks and shared tasks on MT have focused on other aspects, such as low-resource languages (Pal et al., 2023; Sánchez-Martínez et al., 2024), multimodal translation (Specia et al., 2016; Barrault et al., 2018), code-switched data (Chen et al., 2022), and general translation.

3 XC-Translate: A Novel Gold Benchmark for Entity-Aware MT

In this section, we introduce XC-Translate, a novel gold benchmark for evaluating the performance of MT systems in translating text containing entities with names that are significantly different across a set of 10 diverse languages. Creating XC-Translate represents a core contribution of this SemEval task, as it was specifically designed to address the challenges of entity-aware machine translation. While we encourage readers interested in the full technical details to refer to our dedicated publication (Conia et al., 2024), we provide an overview of the benchmark creation process in this section. To create XC-Translate we employ a four-step process:

1. **Entity selection:** We first select the entities of interest for the task. These entities were chosen to be significantly different across languages, e.g., *Il Giovane Holden* and *The Catcher in the Rye*.
2. **Sentence generation:** We generate sentences containing the selected entities. These sentences are simple, as the challenge should lie in the translation of the entities rather than in the complexity of the sentence.
3. **Multi-reference translation:** Each sentence is translated into 10 target languages by at least three native translators.
4. **Translation validation:** Each translation is reviewed by one native speaker of the target

Language Pair	Sample	Dev	Test	Total
EN → Arabic	70	722	4,546	5,338
EN → Chinese	73	722	5,181	5,976
EN → French	75	724	5,464	6,263
EN → German	70	731	5,875	6,676
EN → Italian	73	730	5,097	5,900
EN → Japanese	73	723	5,107	5,903
EN → Korean	73	745	5,081	5,899
EN → Spanish	72	739	5,337	6,148
EN → Thai	73	710	3,446	4,229
EN → Turkish	75	732	4,472	5,279
EN → XX	727	7,278	49,606	57,611

Table 1: Statistics of the EA-MT benchmark provided to participants, divided by language and split. The values indicate the number of instances in each split. The last row shows the total number of instances in the benchmark, where XX indicates all the languages combined.

language, who checks both the overall quality of the translations and the translations of the entity names.

In the following, we describe each step in detail, and provide an overview of the resulting benchmark. We also provide a summary of the statistics of the benchmark in Table 1.

3.1 Selecting “Challenging” Entities

Since the focus of EA-MT is on the translation of text containing entities, we do not randomly select sentences to translate; previous MT tasks that have relied on random sentence selection have been shown not to i) include enough entities to evaluate the translation of entities, and ii) contain entities whose names are significantly different across languages (Zeng et al., 2023). Instead, we will first select the entities of interest for the task, and then generate sentences containing these entities. Although Wikidata (Vrandečić and Krötzsch, 2014) has been shown to be incomplete in terms of entity name coverage (Conia et al., 2023), we use it as a starting point to select entities for EA-MT, and then manually verify the selected entities.

Criteria for entity selection. To avoid entities whose names are similar across languages, we select a random sample of entities that satisfy the following criteria: an entity is valid if and only if its English name and its word-for-word translations have less than a 50% character overlap with the corresponding names in French, German, Italian, and Spanish. Our assumption is that, if the entity has a name in these five languages, then it is a rel-

atively well-known entity, i.e., not belonging to a niche domain or to the tail of the popularity distribution. Moreover, if it satisfies these criteria across these five languages, which mostly share the script (unlike, for example, English and Chinese), then there is a high chance that the translation of such entity requires more than a word-for-word translation. We refer to this set of entities as *challenging* entities, i.e., entities whose names are significantly different across languages and are likely to require a transcreation step instead of a literal translation.

3.2 From Entities to Sentences

Having selected the entities, we generate sentences containing these entities using an LLM, namely, GPT-4. More specifically, given an entity name and its Wikidata description that provides some context about the entity, we prompt the model to generate a simple question pertaining to the entity in English.

Generating a simple question rather than a statement allows us to i) keep the sentence structure simple and short (less than 25 words), ii) ensure that the entity is the most important part of the sentence to translate, iii) provide just enough context to disambiguate the entity, and – most importantly – iv) avoid issues related to the factuality of the generated text. Keeping the sentence structure simple and short is important in our case, as the most challenging part of the task should be the translation of the entity names rather than the complexity of the sentence structure. Moreover, generating a question like “Is *The Catcher in the Rye* a book?” is less likely to generate factuality-related issues than a statement like “*The Catcher in the Rye* is a book”, which may be factually incorrect if the entity is not a book.

Calibrating the complexity of the task. Although factuality is not the main focus of this task, we want to avoid generating sentences that are factually incorrect or misleading, as this would not only affect the quality of the translations but also make it difficult to evaluate the performance of the systems. Future editions of EA-MT or future work could explore the use of more complex sentence structures, such as longer sentences or paragraphs, to evaluate the performance of MT systems. Given the current complexity of the task for traditional MT systems and modern LLMs as shown in the results of the first edition of EA-MT (see Section 6), we believe that the current task is already challenging enough without introducing additional complexity.

Furthermore, increasing the length of the sentences may introduce additional challenges from the perspective of the evaluation metrics, as longer text may include more entities and require coreference resolution and disambiguation of the entities by the evaluation metrics.

3.3 Creating High-Quality Translations

Finally, we translate our set of simple questions in English into the 10 target languages using native translators. To ensure the quality of our translations, we employ a 4-step translation process: first, we check the validity of each generated question; second, each sentence is translated by at least three native translators; third, each translation is reviewed by a native speaker of the target language; and finally, we ask the translators to provide valid aliases for each entity name in the target language.

Multi-reference translation. The entire translation process is guided by a set of instructions and guidelines that we provide to the annotators. Moreover, we require the annotators to be fluent in English, native speakers of the target language, and resident in a country where the target language is spoken.² Before starting the translation process, we also require the annotators to pass an entrance test to further verify their language proficiency and their comprehension of the instructions and guidelines; otherwise, they are not allowed to participate in the annotation task. Finally, the annotators are periodically evaluated on a set of test questions: if they fail on them, they are excluded from the pool of annotators. Since each English question is formulated from a given entity, we aid the translators by providing the entity name(s) from Wikidata in the target language as a hint, the English and target language descriptions of the entity from Wikidata, and the English and target language Wikipedia pages of the entity, which are fundamental resources to grasp the context and background of each entity.

Translation validation. To ensure the quality of our translations, we ask a pool of native speakers to review the translations.³ Similar to the translation step, we provide the reviewers with a set of instructions and guidelines to follow, asking them to check both the overall quality of the translations (e.g., fluency, adequacy, and correctness) and the translations of the entity names (e.g., whether the

²We are not able to verify the residency of the annotators; residency is self-reported.

³The pool of reviewers and translators may overlap.

entity names are translated correctly, whether the most common translation is used, whether the entity name has been adapted to the context of the sentence, etc.). Each translation is reviewed by at least one annotator: if the reviewer finds any issue with the translation, the translation is discarded and the sentence is re-inserted into the pool of sentences to be translated.

Focusing on entity names and aliases. Since our focus is on the translation of entity names, we include an additional step in the translation process: we ask the annotators to provide valid names for each entity in the target language, i.e., other names that can be used to refer to the same entity in the target language. These names must be valid translations of the entity name that can be used interchangeably with the main translation. Importantly, we require the annotators to provide at least one valid name for each entity *in the given context*, i.e., the name must be valid in and adapted to the context of the translated sentence. Sometimes the annotators may deem that there is only one valid name for the entity—i.e., the one used in the translation—and they are allowed to do so: in this case, they must provide the name used in the translation as the only valid name for the entity, and the list of valid names for the entity will contain only one name. Having a list of valid names and aliases for each entity is important for our evaluation, as described in Section 4. It also allows annotators to indicate borderline cases, increasing the robustness of the evaluation process and agreement among annotators.

We provide more details about the benchmark creation process in Conia et al. (2024), including the guidelines provided to the annotators.

3.4 Benchmark Overview

As shown in Table 1, the resulting benchmark, XC-Translate, contains over 50K sentences in English with their translations into the following 10 languages: *Arabic, Chinese (Traditional), French, German, Italian, Japanese, Korean, Spanish, Thai, and Turkish*. Since multiple valid translations are possible for each sentence, we provide multiple references for each sentence in the benchmark, resulting in a total of over 100K manually-created and manually-verified translations. We split XC-Translate into:

- **Sample:** a small sample of sentences for each

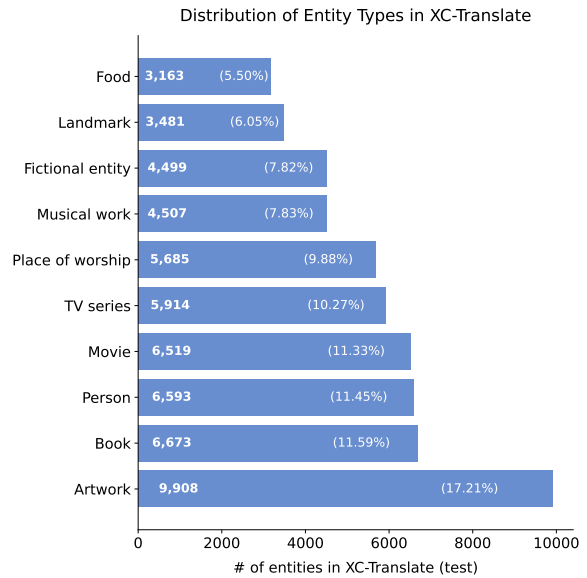


Figure 1: Distribution of the entities in XC-Translate by entity type (top-10 entity types by frequency).

language pair, which participants were encouraged to use for rapid prototyping;

- **Development:** a small development set for each language pair, which participants were encouraged to use for tuning and testing their systems, as the gold references were available;
- **Test:** the official test set for each language pair, which participants were not allowed to use for tuning their systems, as the gold references were released only after the end of the evaluation period.

The split is performed randomly, ensuring that the same entity is not present in two different splits. XC-Translate also indicates a coarse-grained type for each entity, which can be used to group the entities by category, as shown in Figure 1.

4 Evaluation Metrics

To evaluate translation quality, we employ a combination of two metrics: COMET and M-ETA.

Dealing with multiple references. XC-Translate contains multiple references for each sentence, which is often considered a best practice in MT tasks to account for the fact that there are multiple valid ways to convey the same meaning in a target language. When evaluating the translations produced by the systems, we use the best reference approach, which selects the best reference translation for each system output based on the highest

score of the evaluation metric, instead of using the average score across all references.

COMET. COMET (Rei et al., 2020) is a neural metric that evaluates overall translation quality by comparing system outputs against human references. Unlike traditional lexical overlap metrics such as BLEU, COMET leverages contextualized embeddings to capture semantic similarity between translations. We use COMET-22 (Rei et al., 2022), which has shown strong correlation with human judgments across multiple languages. However, while COMET measures general translation quality, it may not specifically capture entity translation accuracy.

M-ETA. To address this limitation, we introduce Manual Entity Translation Accuracy (M-ETA), a simple specialized metric that focuses exclusively on entity translation correctness. Given a set of gold entity translations and predicted translations, M-ETA computes the proportion of correctly translated entities. Importantly, M-ETA accounts for valid aliases of entity names, recognizing that entities often have multiple acceptable translations in a target language. This is crucial for culturally-specific entities where literal translations would be inaccurate. Formally, we define M-ETA as follows:

$$\text{M-ETA} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(e_i \in \mathcal{E}_i) \quad (2)$$

where N is the number of entities in the test set, e_i is the predicted translation of the i -th entity, and \mathcal{E}_i is the set of valid aliases for the i -th entity.

Overall score. The final evaluation score is the harmonic mean of COMET and M-ETA:

$$\text{Overall Score} = \frac{2 \times \text{COMET} \times \text{M-ETA}}{\text{COMET} + \text{M-ETA}} \quad (3)$$

This combined score ensures that systems must perform well on both general translation quality and entity translation accuracy, preventing systems from achieving high rankings by excelling in only one dimension. The harmonic mean particularly penalizes systems that perform poorly in either metric, emphasizing the importance of balanced results.

5 Overview of Participating Systems

In total, 54 participants registered for the task on our CodaBench competition and submitted 322 runs on

the test set, each run containing the predictions of a single system on at least one language pair. Among these, 25 teams submitted the final results of 53 systems for the official leaderboard and 18 teams submitted system description papers. We provide an overview of the systems at EA-MT in Table 3.

As shown in Figure 2, we can observe a clear trend toward the use of large language models (LLMs) for entity-aware machine translation, as 86.8% of the systems are based on LLMs and only 13.2% of the systems are based on traditional NMT models. Moreover, retrieval-augmented generation (RAG) is one of the most common techniques used by the participants, as 26.4% of the systems use this technique to improve their performance. There is still a significant number of participants (47.2%) who opted to fine-tune their models on a training dataset, which is a considerable proportion since only open-source models can be directly fine-tuned. Not surprisingly, the most used LLMs for this task align with the most popular and best-performing LLMs in the general NLP community, i.e., GPT-4o, Qwen-2.5, and LLaMA-3.

6 Results, Analysis, and Discussion

The number of submissions to the official leaderboard allows to provide a birds-eye view of the performance of different systems on the task, and analyze a few interesting trends, which are discussed in depth in Appendix B.

General results. We report the results of the official leaderboard in Table 2, which reports the scores of the systems obtained during the main evaluation phase.⁴ We distinguish between two main categories of systems: systems that use “gold” information during the translation process (i.e., systems that take in input the manually-identified Wikidata ID of the entity appearing in the sentence to be translated) and “end-to-end” systems that do not use this information. In other words, the first category reflects scenarios where the entity to be translated is known in advance, while the second category reflects a more general scenario where it is not known in advance if the sentence to be translated contains an entity and which entity it is. Among the systems using “gold” information, Qwen2.5-72B-LoRA by Pingan Team achieves the best score, with a COMET of 94.7 and an M-ETA of 89.1. Instead, among the

⁴Participants were also allowed to submit additional results during the post-evaluation phase, but these results are not included in the official leaderboard.

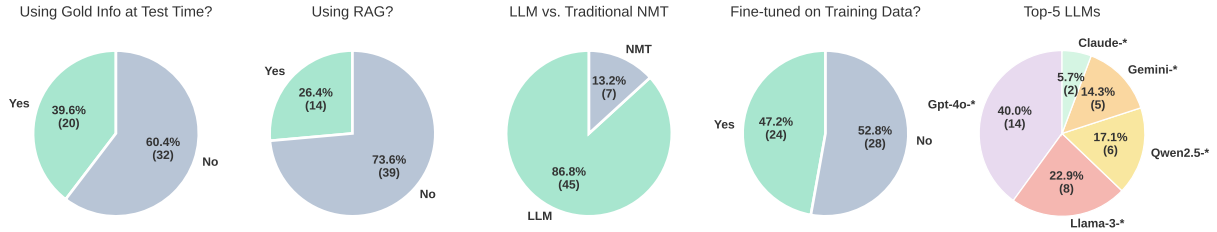


Figure 2: Overview of the systems submitted to EA-MT. Here, we distinguish between systems that: 1) use “gold” information during the translation process; 2) employ RAG; 3) are based on traditional NMT or modern LLMs; 4) are fine-tuned on a training dataset. We also report the top-5 LLMs used among the participants.

EA-MT – OFFICIAL LEADERBOARD			Average across all languages			Rank	
Team	System	Category	M-ETA	Comet	Overall	ALL	AVG
Pingan Team	Qwen2.5-72B-LoRA	🟡 🧠 📖	89.1	94.7	91.8	1	3.7
Pingan Team	Qwen2.5-72B-LoRA + zhconv	🟡 🧠 📖	89.0	94.8	91.7	2	4.2
DeerLu	Qwen2.5-Max-Wiki	🟡 🔍 🧠 📖	89.0	94.8	91.7	3	4.8
RAGthoven	GPT-4o + WikiData + RAG	🟡 🔍 🧠 📖	88.5	95.0	91.6	4	5.0
Pingan Team	Phi4-FullFT	🟡 🧠 📖	88.9	94.4	91.5	5	5.2
UAlberta	WikiEnsemble	🟡 🔍 🧠 📖	88.3	95.0	91.5	6	6.6
CHILL	GPT4o-RAG-Refine	🟡 🔍 🧠 📖	88.5	94.7	91.5	7	6.3
UAlberta	WikiGPT4o	🟡 🔍 🧠 📖	88.1	95.0	91.4	8	7.8
RAGthoven	GPT-4o + Wikidata	🟡 🔍 🧠 📖	87.6	94.8	91.0	9	7.5
Lunar	LLaMA-RAFT-Plus-Gold	🟡 🔍 🧠 📖	87.3	94.7	90.7	10	5.9
YNU-HPCC	LLaMA + MT	🟡 🧠 📖	85.9	94.5	89.9	11	11.6
arancini	WikiGemmaMT	🟡 🧠 📖	85.3	93.6	88.8	12	10.6
Lunar	LLaMA-RAFT-Gold	🟡 🔍 🧠 📖	82.2	92.6	86.8	13	14.4
SALT 🏠	Salt-Full-Pipeline + Gold	🟡 🔍 🧠 📖	80.0	93.3	85.8	14	14.5
Howard University-AI4PC	DoubleGPT	🟡 🔍 🧠 📖	77.9	93.6	84.8	15	15.3
SALT 🏠	Salt-Full-Pipeline	🔍 🧠 📖	77.1	91.8	83.6	1	1.6
SALT 🏠	Salt-MT-Pipeline	🔍 🧠 📖	71.7	92.5	80.4	2	2.7
FII-UAIC-SAI	Qwen2.5-Wiki-MT	🔍 🧠 📖	68.2	91.6	78.2	3	3.6
Lunar	LLaMA-RAFT-Plus	🔍 🧠 📖	62.9	91.8	74.3	4	5.3
YNU-HPCC	Qwen2.5 + M2M	🧠 📖	62.0	91.8	73.9	5	5.7
FII the Best	mBERT-WikiEuRal	🔍 🧠 📖	60.6	89.5	71.4	6	5.6
Lunar	LLaMA-RAFT	🔍 🧠 📖	56.5	90.4	68.8	7	7.3
UAlberta	PromptGPT	🧠 📖	46.7	91.9	61.5	8	9.3
The 5 Forbidden Entities	MBart-KnowledgeAware	🧠 📖	48.3	84.3	60.5	9	9.3
RAGthoven	GPT-4o + RAG	🔍 🧠 📖	45.3	91.7	60.5	10	10.0
The 5 Forbidden Entities	Embedded Entities	🧠 📖	44.3	83.5	56.8	11	10.9
Zero	FineTuned-MT	🧠 📖	33.7	90.3	47.8	12	13.2
HausaNLP	Gemini-0shot	🧠 📖	33.6	89.3	47.7	13	13.1
Muhandro_HSE	NER-LLM	🧠 📖	28.1	88.2	41.3	14	15.7
Silp_NLP	GPT-4o	🧠 📖	13.5	77.6	20.7	15	16.7
SheffieldGATE	Llama-Wiki-DeepSeek	🧠 📖	89.8*	93.3*	91.5*	–	–
Team ACK	Gemini-Pro	🧠 📖	48.3*	90.9*	63.1*	–	–
Sakura	Rakuten7b-P010	🧠 📖	29.5*	90.7*	44.5*	–	–
VerbaNexAI Lab	TransNER-SpEn	🟡 🧠 📖	24.6*	87.1*	38.4*	–	–
GinGer	LoRA-nllb-200-distilled-600M	🧠 📖	22.0*	88.2*	35.1*	–	–
JNLP	MultiTask-mT5	🧠 📖	12.3*	76.7*	21.2*	–	–

Table 2: Official leaderboard for the EA-MT shared task, showing the top-15 systems divided into two sections. 🟡: System uses gold information at test time. 🔍: System uses retrieval-augmented generation. 🧠: System uses a large language model. 📖: System is fine-tuned on training data. *: Scores averaged over a subset of the 10 languages.

“end-to-end” systems, Salt-Full-Pipeline by SALT achieves the best score, with a COMET of 91.8 and an M-ETA of 77.1, –2.9 and –12.0 points lower than the best system using “gold” information

in terms of COMET and M-ETA, respectively.

Gold information is not enough. An interesting finding is that state-of-the-art LLMs are not perfectly able to translate entities, even when provided with “gold” information, e.g., using the manually-identified Wikidata ID of the entity appearing in the sentence to be translated to retrieve the correct entity name in the target language from Wikidata. The M-ETA scores for the top-10 systems using “gold” information range between 87.3 and 89.1, showing that there is a hard core of entities whose names are difficult to adapt to the context of the translated sentence. Notably, there are two orthogonal aspects to consider: i) sometimes systems disregard the provided “gold” entity name in the target language because the transcreation process resulted in a completely different name, leading the MT systems to prefer a word-for-word translation; ii) sometimes systems fail to adapt the provided “gold” entity name to the context of the translated sentence, e.g., its morphology and syntax.

State-of-the-art LLMs lack cross-lingual and cross-cultural knowledge. Recent LLMs have shown impressive performance on a wide range of tasks, including machine translation. Although LLM-based translations are often fluent, coherent and grammatically-correct, XC-Translate demonstrates that they still struggle with entity translation. Indeed, if we based the evaluation of the systems on COMET only, we would conclude that LLMs are able to translate entities correctly. However, the M-ETA score shows that this is far from the truth, especially for current LLMs, e.g., not fine-tuned on a training dataset or using retrieval-augmented generation. There are two main reasons for this: i) XC-Translate contains a large number of entities that are not well-known, and ii) XC-Translate contains a large number of entities whose names are difficult to adapt to the context of the translated sentence. Therefore, we believe that XC-Translate will be valuable in future research not only for MT but also for benchmarking cross-lingual and cross-cultural knowledge in LLMs.

How did the participants address the limitations of current LLMs? As current LLMs still do not encode the cross-lingual and cross-cultural knowledge required to translate entities correctly, retrieval-augmented generation (RAG)—often combined with fine-tuning—is one of the most common techniques used by the participants to improve

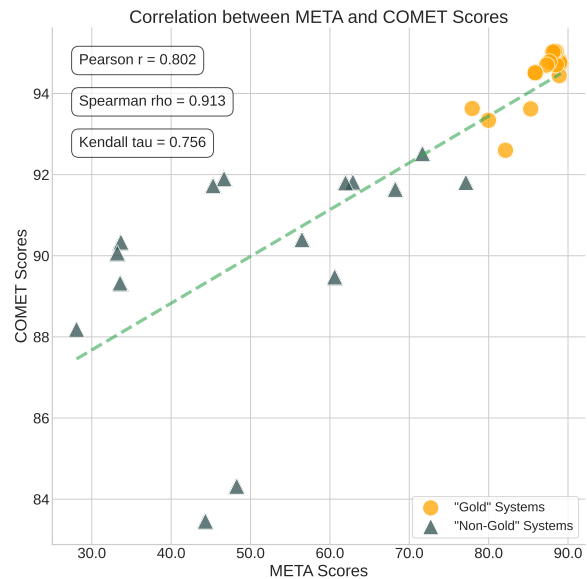


Figure 3: Correlation between M-ETA and COMET using Pearson r, Spearman rho, and Kendall’s tau.

their performance. Different participants used different retrieval strategies to retrieve different types of information (e.g., entity names, descriptions, Wikipedia pages, etc.). For instance, the SALT team used a SQL-based approach to retrieve the entity name in the target language from Wikidata before translating the sentence, whereas the Lunar team took advantage of the function calling capabilities of recent LLMs to retrieve entity-related information. Alternatively, some teams used external tools for entity recognition and linking, e.g., WikiNeural (Tedeschi et al., 2021) and ReLiK (Orlando et al., 2024). Finally, some teams used a combination of these techniques, e.g., the UAlberta team used both retrieval-augmented generation and ensemble learning to improve their performance. We provide more details about the systems submitted to the EA-MT shared task in Appendix A.

COMET is not a good proxy for entity translation. We can observe that the systems that achieve the best overall scores are not necessarily the ones that achieve the best M-ETA scores. Interestingly, the “gold” systems with the best COMET scores rank in 5th, 7th, and 8th place in terms of M-ETA scores, namely GPT-4o + Wikidata + RAG by RAGthoven, WikiEnsemble by UAlberta, and WikiGPT4o by UAlberta. This is even more evident for the “end-to-end” systems: Salt-Full-Pipeline by SALT—the 1st system in terms of M-ETA score—is only separated by 0.1 points in terms of COMET score (91.8 vs 91.7) from GPT-4o

+ RAG by RAGthoven, which ranks 10th in terms of M-ETA score (77.1 vs 45.3). As shown in Figure 3, COMET and M-ETA are correlated; however, even if we remove the outliers, a small shift in the value of COMET can lead to a very large shift in the value of M-ETA. This suggests that COMET is not a good proxy for entity translation, as COMET is often affected more by the fluency and coherence than by the correctness of the entity translation.

Not a universal solution across all languages. Independently of the evaluation metric, different systems do not perform equally well across all languages, as shown by the average rank of the systems across all languages in Table 2, which is computed by averaging the ranks of the systems across all languages. For instance, the best “gold” system, Qwen2.5-72B-LoRA by Pingan Team, only ranks 1st in 2 languages out of 10, achieving only 8th and 9th place in Arabic and Korean, respectively. Vice versa, WikiEnsemble by UAlberta ranks 1st in Arabic but only 6th on average across all languages. Similar to the best “gold” system, the best “end-to-end” system, Salt-Full-Pipeline by SALT, ranks 1st only in 5 languages. Therefore, among the many systems proposed by the participants there is no universal solution for entity-aware machine translation, showing that there is room for improvement in two areas: i) combining the different techniques used among the task participants, and ii) leveraging language-specific features and resources to create better models for each language.

Different research directions and open questions for different settings. The results of the EA-MT shared task show that there are still many open questions and challenges in the field of entity-aware machine translation. Here, we highlight some of the most important ones that emerged from the task, divided into two main categories: i) “gold” systems and ii) “end-to-end” systems.

- For “gold” systems: (1) even with gold entity annotations, *language-specific optimization* seems necessary for now, as no single system uniformly excels across all languages; (2) *knowledge retrieval quality* is uneven across languages, motivating language-specific retrieval strategies rather than a universal approach; and (3) *script adaptation mechanisms* are still an important challenge, as many systems struggle to achieve consistent results across languages with different scripts.

- For “end-to-end” systems: (1) *entity recognition and linking* is a crucial step for improving the performance of these systems, as it allows to identify the entities in the source sentence and link them to their corresponding entity names in the target language; (2) *hybrid systems* are a promising direction for improving the performance while reducing the computational cost; and (3) *cross-lingual and cross-cultural knowledge* is still a challenge for these systems, as they often struggle to adapt the entity name to the context of the translated sentence.

We provide an in-depth analysis of these challenges in Appendix B. In general, we believe that these challenges will be valuable for future research in the field of entity-aware machine translation, and we encourage researchers to explore these directions to improve the performance of their systems.

7 Conclusion and Future Work

In this paper, we presented the Entity-Aware Machine Translation (EA-MT) shared task, which was part of SemEval-2025. The goal of EA-MT is to evaluate the ability of machine translation systems—traditional NMT and modern LLMs—to translate text that contains challenging entities, e.g., entities that are affected by cultural and linguistic differences across languages. To this end, we created XC-Translate, a benchmark for entity-aware machine translation that contains over 50K sentences in English with their translations into 10 languages, with a total of over 100K manually-created and manually-verified translations. We also proposed a new evaluation metric, M-ETA, which focuses exclusively on entity translation correctness. Finally, we analyzed the results of the official leaderboard and discussed the key trends in the systems submitted to the EA-MT shared task. In general, XC-Translate has shown that state-of-the-art LLMs still struggle with entity translation, and that different approaches are needed to bridge the gap between LLMs and human translators. In the future, we plan to extend XC-Translate with additional language-pairs—including pairs where the source language is different from English—and domains, and to introduce new challenges, such as code-switching, low-resource languages, and larger coverage of emerging entities.

Limitations

XC-Translate. XC-Translate is a benchmark for entity-aware machine translation that contains over 50K sentences in English with their translations into 10 languages, with a total of over 100K manually-created and manually-verified translations. However, XC-Translate is limited in several ways. First, XC-Translate only provides translations from English to 10 languages, which limits its applicability to other language pairs. We avoided reversing the translation direction to avoid the risk of introducing noise in the translations and dealing with known issues in back-translation, including the increase of translationese artifacts. Second, XC-Translate only contains translations of entities that are present in Wikidata and feature at least one alias in the source and target languages. This means that XC-Translate may not be representative of entities belonging to domains that are not well covered by Wikidata. Third, while we strived to cover as many languages as possible, we focused on the most widely spoken languages. Increased attention to low-resource languages is needed to ensure that XC-Translate is representative of the diversity of languages spoken around the world. Fourth, XC-Translate only includes questions (see Section 3). This means that XC-Translate may not be representative of other types of text, such as narratives or technical documents.

M-ETA. M-ETA is a specialized metric that focuses exclusively on entity translation correctness. We introduced M-ETA to address the limitations of other metrics, such as COMET and BLEU, which do not specifically capture entity translation accuracy. However, M-ETA also has limitations. First, M-ETA only considers the correctness of entity translations and does not account for other aspects of translation quality, such as fluency and coherence. Therefore, M-ETA should be used in conjunction with other metrics to provide a comprehensive evaluation of translation quality. Second, M-ETA relies on the availability of valid aliases for entity names in the target language, which is a manually-intensive process and may not be feasible for all entities.

Systems. The systems submitted to the EA-MT shared task are based on a variety of approaches, including traditional NMT, retrieval-augmented generation, and large language models. Although the prevalence of LLMs in the submitted systems is a promising trend, it also raises concerns about

the reproducibility and generalizability of the results. Many of the systems are based on proprietary LLMs, which limits their accessibility and reproducibility. Additionally, systems based on closed-source models are difficult to analyze and understand, making it challenging to identify potential biases in the models, which can lead to unfair treatment of certain groups or individuals. Finally, the reliance on large-scale pre-trained models raises questions about the environmental impact of training and deploying these models. Therefore, solutions based on smaller and open-source models may still be competitive and more sustainable in the long run if we consider other factors, such as reproducibility, latency, and energy consumption.

Acknowledgements

We would like to thank the participants of the EA-MT shared task for their valuable contributions and for sharing their systems and results with us. We would also like to thank the annotators and reviewers for their hard work in creating and validating the XC-Translate benchmark. Finally, we would like to thank the organizers of SemEval-2025 for their support, availability, and flexibility, and for providing us with the opportunity to organize this shared task.

Simone Conia gratefully acknowledges the support of the PNRR MUR project PE0000013-FAIR, which fully funds his fellowship. Roberto Navigli also acknowledges the support of the PNRR MUR project PE0000013-FAIR, which partially funds his research.



References

- Abdulhamid Abubakar, Hamidatu Abdulkadir, Rabi'u Abdullahi Ibrahim, Abubakar Khalid Auwal, Ahmad Mustapha Wali, Amina Aminu Umar, Maryam Bala, Sani Abdullahi Sani, Ibrahim Said Ahmad, Shamsuddeen Hassan Muhammad, Idris Abdulmumin, and Vukosi Marivate. 2025. Hausanlp at semeval-2025 task 2: Entity-aware fine-tuning vs. prompt engineering in entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Saurav K. Aryal and Jabez D. Agyemang-Prempeh. 2025. Howard university-ai4pc at semeval-2025 task 2: Improving machine translation with context-aware

- entity-only pre-translations with gpt4o. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraaq Lala, Desmond Elliott, and Stella Frank. 2018. [Findings of the third shared task on multimodal machine translation](#). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.
- Diyang Chen. 2025. pingan-team at semeval-2025 task 2: Lora-augmented qwen2.5 with wikidata-driven entity translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Shuguang Chen, Gustavo Aguilar, Anirudh Srinivasan, Mona T. Diab, and Tamar Solorio. 2022. [CALCS 2021 shared task: Machine translation for code-switched data](#). *CoRR*, abs/2202.09625.
- Simone Conia, Daniel Lee, Min Li, Umar Farooq Minhas, Saloni Potdar, and Yunyao Li. 2024. [Towards cross-cultural machine translation with retrieval-augmented generation from multilingual knowledge graphs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16343–16360, Miami, Florida, USA. Association for Computational Linguistics.
- Simone Conia, Min Li, Daniel Lee, Umar Minhas, Ihab Ilyas, and Yunyao Li. 2023. [Increasing coverage and precision of textual information in multilingual knowledge graphs](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1612–1634, Singapore. Association for Computational Linguistics.
- Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Mar Díaz-Millón and María Dolores Olvera-Lobo. 2023. [Towards a definition of transcreation: a systematic literature review](#). *Perspectives*, 31(2):347–364.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *J. Mach. Learn. Res.*, 22:107:1–107:48.
- Viviana Gaballo. 2012. Exploring the boundaries of transcreation in specialized translation. *ESP Across Cultures*, 9:95–113.
- Delia-Iustina Grigorita, Tudor-Constantin Pricop, Sergio-Alessandro Suteu, Daniela Gifu, and Diana Trandabat. 2025. Fii the best at semeval-2025 task 2: Steering state-of-the-art machine translation models with strategically engineered pipelines for enhanced entity translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Revanth Kumar Gundam, Abhinav Marri, Advait Maladi, and Radhika Mamidi. 2025. Zero at semeval-2025 task 2: Entity-aware machine translation: Fine-tuning nllb for improved named entity translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. [Challenges and strategies in cross-cultural NLP](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.
- Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Thirty-seventh Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*.
- Daniel Lee, Harsh Sharma, Jieun Han, and Sunny Jeong. 2025a. Team ack at semeval-2025 task 2: Beyond word-for-word machine translation for english-korean pairs. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Jaebok Lee, Yonghyun Ryu, Seongmin Park, and Yoonjeong Choi. 2025b. Chill at semeval-2025 task 2: You can’t just throw entities and hope—make your llm to get them right. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Hao Li, Jin Wang, and Xuejie Zhang. 2025. YNU-HPCC at SemEval-2025 Task 2: Local Cache and Online Retrieval-Based method for Entity-Aware Machine Translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Aylin Naebzadeh. 2025. Ginger at semeval-2025 task 2: Challenges in entity-aware machine translation with fine-tuning and zero-shot prompting. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024. [ReLiK: Retrieve](#)

- and LinK, fast and accurate entity linking and relation extraction on an academic budget. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14114–14132, Bangkok, Thailand. Association for Computational Linguistics.
- Santanu Pal, Partha Pakray, Sahinur Rahman Laskar, Lenin Laitonjam, Vanlalmuansangi Khenglawt, Sunita Warjri, Pankaj Kundan Dadure, and Sandeep Kumar Dash. 2023. [Findings of the WMT 2023 shared task on low-resource Indic language translation](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 682–694, Singapore. Association for Computational Linguistics.
- Daniel Arturo Peña Gnecco, Juan Carlos Martinez Santos, and Edwin Puertas. 2025. VerbaNexAI at semeval-2025 task 2: Enhancing entity-aware translation with wikidata-enriched marianmt. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Alberto Poncelas and Ohnmar Htun. 2025. Sakura at semeval-2025 task 2: Enhancing named entity translation with fine-tuning and preference optimization. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. [COMET: A neural framework for MT evaluation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.
- Felipe Sánchez-Martínez, Juan Antonio Perez-Ortiz, Aaron Galiano Jimenez, and Antoni Oliver. 2024. [Findings of the WMT 2024 shared task translation into low-resource languages of Spain: Blending rule-based and neural systems](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 684–698, Miami, Florida, USA. Association for Computational Linguistics.
- Ning Shi, David Basil, Bradley M. Hauer, Noshin Nawal, Jai Riley, Daniela Teodorescu, John Z. Zhang, and Grzegorz Kondrak. 2025. Ualberta at semeval-2025 task 2: Prompting and ensembling for entity-aware translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Sumit Singh, Pankaj Kumar Goyal, and Uma Shanker Tiwary. 2025. [silp_nlp at semeval-2025 task 2: An effect of entity awareness in machine translation using llm](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Demetris Skottis, Gregor Karetka, and Marek Suppa. 2025. Ragthoven at semeval-2025 task 2: Enhancing entity-aware machine translation with large language models, retrieval augmented generation and function calling. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. [A shared task on multimodal machine translation and crosslingual image description](#). In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.
- Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2021. [Multilingual translation from denoising pre-training](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3450–3466, Online. Association for Computational Linguistics.
- Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. [WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Tom Völker, Jan Pfister, and Andreas Hotho. 2025. Salt at semeval-2025 task 2: A sql-based approach for llm-free entity-aware-translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Denny Vrandečić and Markus Krötzsch. 2014. [Wikidata: a free collaborative knowledgebase](#). *Commun. ACM*, 57(10):78–85.
- Lu Xu. 2025. Deerlu at semeval-2025 task 2: Wikidata-driven entity-aware translation—boosting llms with external knowledge. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Xinye Yang, Kalina Bontcheva, and Xingyi Song. 2025. Sheffieldgate at semeval-2025 task 2: Multi-stage reasoning with knowledge fusion for entity translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.
- Zixin Zeng, Rui Wang, Yichong Leng, Junliang Guo, Shufang Xie, Xu Tan, Tao Qin, and Tie-Yan Liu. 2023. [Extract and attend: Improving entity translation in](#)

neural machine translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1697–1710, Toronto, Canada. Association for Computational Linguistics.

Andrea Zenotto, Vincenzo Catalano, and Ruben Tetamo. 2025. Arancini at semeval-2025 task 2: Multilingual translation enhanced by lightweight llms, ner, and rag for named entities. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*. Association for Computational Linguistics.

A Participating Systems

In this section, we provide an overview of the systems submitted to EA-MT, sorted by team name. For each system, we briefly describe the approach used by the participants to tackle the task and summarize their main findings. For more details, we refer the reader to the corresponding system description papers.

Arancini: “*Multilingual Translation Enhanced by Lightweight LLMs, NER, and RAG for Named Entities*” (Zenotto et al., 2025). This work introduces a multilingual translation pipeline that combines lightweight large language models (LLMs), a dedicated NER module, and a retrieval-augmented generation (RAG) mechanism to improve named-entity handling. The authors benchmark several models (e.g., M2M100 variants, Qwen2.5, Gemma2-9B) and show that integrating entity linking with Wikidata resources, guided by either gold IDs or automatically detected entities, significantly boosts M-ETA (entity-level accuracy) and maintains strong COMET scores (overall fluency). The authors demonstrate that even when NER introduces errors, the retrieval-based approach preserves high semantic fidelity, underscoring the pipeline’s robustness for real-world scenarios in which perfectly labeled data is unavailable.

CHILL: “*You Can’t Just Throw Entities and Hope — Make Your LLM to Get Them Right*” (Lee et al., 2025b). The authors present a system that enhances entity-aware translation by fusing retrieval-augmented generation (RAG) with an iterative self-refinement mechanism. In particular, they retrieve entity labels and descriptions from Wikidata, embedding these details into prompts for GPT-4o to ensure accurate handling of named entities. Crucially, the system self-evaluates each translation on both entity correctness and overall linguistic quality, iterating until it meets a predefined performance threshold or exhausts the allotted refinement steps. This feedback-driven procedure, grounded in large language models, consistently yields improvements in entity accuracy (M-ETA) without sacrificing global translation quality (COMET). A further analysis reveals minimal correlation between label similarity (quantified via Levenshtein distance) and entity-translation precision, underscoring that the critical gains stem from leveraging oracle entity context and iterative revision rather than label similarity alone.

Deerlu: “*Wikidata-Driven Entity-Aware Translation — Boosting LLMs with External Knowledge*” (Xu, 2025). The authors introduce an entity-aware machine translation system that enhances large language models (LLMs) with external knowledge from Wikidata. Their approach involves two strategies: one that uses gold Wikidata IDs for cross-lingual entity retrieval, and a practical alternative that leverages ReLiK to identify and link entities automatically with an external knowledge base. Experiments across multiple language pairs demonstrate significant improvements in named entity translation accuracy with up to a 63-point gain in M-ETA while maintaining strong overall translation quality as measured by COMET. Notably, the system ranks third overall and first among non-finetuned entries on the SemEval-2025 Task 2 leaderboard. Further enhancements tailored to specific linguistic nuances, such as simplified-to-traditional character conversion for Chinese, boost performance and highlight the practical applicability of external-knowledge integration for robust and accurate entity-aware machine translation.


FII the Best: “*Steering State-of-the-art Machine Translation Models with Strategically Engineered Pipelines for Enhanced Entity Translation*” (Grigorita et al., 2025). The authors propose two complementary pipelines for enhancing entity-aware machine translation, with a shared emphasis on integrating structured knowledge into large language models. In the first approach, a multilingual NER module (mBERT trained on WikiNEuRal) identifies entities in the source text, which are then aligned with Wikidata translations and merged back into placeholders to preserve context. Notable refinements include punctuation normalization and replacing general MT with the Gemini API to address grammatical coherence issues. The second approach leverages LLMs (Qwen 2.5 Instruct) guided by carefully engineered prompts to separate named entities, fetch accurate translations from Wikidata, and maintain fluency when reinserting them into the transformed text. Comparative results show that the second strategy consistently yields stronger COMET and M-ETA scores across ten languages, especially for underperforming cases like Chinese. Future work involves substituting Gemini 1.0 with more advanced LLMs and unifying both strategies into a single, robust framework for entity-centric translation.

GinGer: “*Challenges in Entity-Aware Machine Translation with Fine-Tuning and Zero-Shot Prompting*” (Naebzadeh, 2025). The authors tackle named entity translation by experimenting with two distinct strategies: 1) parameter-efficient fine-tuning (PEFT) of multilingual seq2seq models, and 2) zero-shot prompting using open-source LLMs. Their PEFT approach uses LoRA-based low-rank updates to mitigate heavy computational requirements, and applies it to various Transformer-based NMT backbones. They also explore zero-shot translation with models under 10 billion parameters, using carefully constructed prompts that emphasize preserving entity integrity. Empirical results on Arabic, Italian, and Japanese highlight the persistent difficulty of accurately translating named entities, especially under data-scarce conditions or model size constraints. While both PEFT and prompt-based approaches yield improvements over naive baselines, entity translation remains suboptimal, suggesting that more robust integration of external knowledge or domain adaptation is needed. Nonetheless, the work underscores how smaller NMT or LLM models can be practically adapted for entity-aware translation, even with limited computational resources.

HausaNLP: “*Entity-Aware Fine-tuning vs. Prompt Engineering in Entity-Aware Machine Translation*” (Abubakar et al., 2025). The authors explore both fine-tuning a distilled NLLB-200 model and zero-/few-shot prompt-based methods with Gemini for entity-aware machine translation from English into ten target languages. Their fine-tuning strategy includes augmenting training data with named entities (NE) extracted from Wikidata to refine translation performance, while prompt-based approaches either rely on minimal instructions (zero-shot) or incorporate a few examples (few-shot) to promote correct NE usage. By comparing these strategies, they uncover that Gemini consistently achieves higher M-ETA (entity accuracy) than the fine-tuned NLLB-200 model, particularly for European languages. Their findings highlight that the gap between zero- and few-shot prompting is small, suggesting that extensive prompt engineering may not be necessary for robust entity-centric translations.

Howard University-AI4PC: “*Improving Machine Translation With Context-Aware Entity-Only Pre-translations with GPT4o*” (Aryal and Agyemang-Prempeh, 2025). The authors propose a three-step pipeline that combines external knowledge from


Wikidata and structured GPT prompts to improve named entity translation. First, they extract target-language labels and descriptions for each entity via Wikidata lookups, ensuring that contextually specific translations are available. Second, they refine these entity translations with a dedicated GPT prompt, guiding the model to produce accurate named entities. Finally, they feed both the original source text and the refined entity translations into a context-aware GPT prompt, generating a translation that preserves semantic integrity while accurately handling named entities. Experiments indicate that this multi-pass strategy yields substantial gains over baseline GPT-only approaches, especially for languages with distinctive tokenization or orthographic conventions (e.g., Arabic, Japanese). Although dependent on Wikidata coverage, the proposed method demonstrates how systematically bridging large language models with external entity information can significantly enhance the quality of cultural- or domain-specific named entity translations.

 **Pingan Team:** Best “gold” system — “*LoRA-Augmented Qwen2.5 with Wikidata-Driven Entity Translation*” (Chen, 2025). The authors present a system for entity-aware translation that leverages a LoRA-based fine-tuning of the 72B-parameter Qwen2.5 model, augmented by a synthetic data generation pipeline. Specifically, they incorporate Wikidata entries to retrieve multilingual entity labels, then synthesize sentence pairs containing these labels to improve named entity translation coverage. LoRA focuses on low-rank updates while maintaining the original model’s generalization ability, enabling domain adaptation without excessive resource overhead from a computational point of view. Experimental results across ten languages show that their approach achieves state-of-the-art performance on the SemEval-2025 Task 2 leaderboard, evidenced by both high COMET scores for global translation quality and substantially improved M-ETA scores for named entity translation accuracy. Notably, the system demonstrates effective handling of rare or culturally specific references, suggesting that combining structured knowledge (Wikidata) with large language models (Qwen2.5) and targeted LoRA fine-tuning can robustly address complex cross-lingual entity mappings.

RAGthoven: “*Enhancing Entity-Aware Machine Translation with Large Language Models, Re-*

trieval Augmented Generation and Function Calling” (Skottis et al., 2025). The authors describe a lightweight, high-impact approach to entity-aware machine translation that combines GPT-4o with Wikidata-based named entity translations, retrieval-augmented generation, and function calling. Implemented in the RAGthoven framework, their system first enriches the source sentence with any existing named entity data from Wikidata. It then retrieves similar (English→Target) sentence pairs from a small parallel corpus to as contextual examples, and finally, uses GPT-4o to generate a final translation that incorporates this information. When the gold entity is not pre-identified, the system invokes a multi-step procedure in which the LLM identifies the named entity, queries the Wikidata API for translations, and re-injects them into the prompt. Empirical results on ten languages show strong gains over a baseline GPT-4o, up to a twenty-point boost when Wikidata entity IDs are provided. The proposed method highlights that carefully orchestrating calls to external knowledge can effectively mitigate typical LLM weaknesses in handling culturally specific or domain-limited entity references.

Sakura: “*Enhancing Named Entity Translation with Fine-Tuning and Preference Optimization*” (Poncelas and Htun, 2025). The authors explore two techniques to incorporate dictionary-based knowledge for named entity translation from English into Japanese: 1) fine-tuning on either individual or batched dictionary entries, and 2) applying preference optimization to rerank the model’s output toward the dictionary references. While fine-tuning with single entries maximizes entity-level accuracy (M-ETA), it can degrade overall quality (COMET, CHRF). Aggregating entries into lists mitigates this trade-off, but still affects fluency and coverage. In contrast, preference optimization yields more balanced improvements, boosting named entity fidelity without significantly harming broader translation performance. Experiments using a curated Wikidata-derived dictionary and a pre-trained RakutenAI-7B model demonstrate that both strategies are effective, with distinct trade-offs in preserving entity translations and maintaining global translation quality.

 **SALT:** Best end-to-end system — “*A SQL-based Approach for LLM-Free Entity-Aware Translation*” (Völker et al., 2025). The authors propose a lightweight two-stage pipeline, SALT, that

bypasses large language models for entity-aware translation and instead uses SQL-based retrieval in combination with constrained neural decoding. The proposed approach identifies source sentence spans via n-gram matching, retrieves corresponding entity translations from a SQL-indexed knowledge base, and then injects these matches into a distilled NLLB-200 model augmented with logit biasing to favor the provided entity translations. Ablation studies reveal that simple string-based retrieval rivals more complex neural methods, that limiting each entity to a single candidate avoids confusion in generation, and that logit biasing effectively improves name-entity accuracy without harming overall translation quality. Despite using far fewer parameters than LLMs, SALT achieves state-of-the-art performance among systems without gold data, narrowing the performance gap with LLM-based methods to less than one percentage point in harmonic mean metrics.

SheffieldGATE: “*Multi-Stage Reasoning with Knowledge Fusion for Entity Translation*” (Yang et al., 2025). The authors introduce a multi-agent entity-aware machine translation system, focusing on precise handling of named entities. Their approach employs a three-stage reasoning pipeline involving entity extraction, knowledge enhancement, and translation decision-making. In the first stage, an LLM identifies named entities and relevant context within the source text. The second stage uses Wikidata-based retrieval, guided by refined LLM-generated queries, to gather candidate entity information with descriptions and alternate names. Finally, in the translation stage, a fine-tuned LLM selectively integrates these candidate entities to produce contextually accurate translations. An additional verification module detects reasoning failures and refines outputs, guarding against omissions or semantic shifts. Experimental results across four language pairs (English–German, English–French, English–Italian, and English–Spanish) confirm significant gains in entity translation accuracy, as measured by M-ETA, while maintaining strong overall translation quality (COMET).

silp_nlp: “*An effect of Entity Awareness in Machine Translation using LLM*” (Singh et al., 2025). The authors propose two strategies for entity-aware translation from English into various target languages: prompting GPT-based models (GPT-4o and GPT-4o-mini) directly, and fine-tuning an NLLB-200 model with LoRA adaptation. An external

Universal NER module identifies named entities in the source text, which are then incorporated into the LLM prompts or appended to the training data for NLLB-200. Automatic results on M-ETA and COMET demonstrate that adding entity annotations boosts overall performance for both approaches, though success is highly dependent on the accuracy of the entity extraction stage. GPT models outperform a fine-tuned NLLB-200, but both approaches benefit from explicit named entity information, reinforcing the value of entity-awareness to resolve the common pitfalls of incorrectly handling rare or ambiguous entities.

Team ACK: “*Beyond Word-for-Word Machine Translation for English-Korean Pairs*” (Lee et al., 2025a). The authors focus on translating English text into Korean by evaluating thirteen models (LLMs and MT systems) on knowledge-intensive question-answer pairs. Their setup combines three automatic metrics, namely, BLEU, COMET, and M-ETA, to measure fluency, general translation quality, and entity-specific accuracy. Notably, they also conduct a comprehensive human annotation of 650 samples to identify error types and construct an interesting error taxonomy. Empirical results highlight that LLMs generally outperform traditional MT approaches but still often fail to preserve cultural nuances when adapting entity references between English and Korean. The authors classify the most frequent error types (e.g. incorrect responses, misaligned or phonetic entity translations), and further note how entity popularity and type can influence outcomes. They conclude that current automatic metrics often overlook finer cultural nuances, underscoring the continued need for human-in-the-loop evaluations and specialized techniques for culturally grounded, entity-focused machine translation.

UALberta: “*Prompting and Ensembling for Entity-Aware Translation*” (Shi et al., 2025). The authors develop a new strategy for entity-aware translation, focusing on large language model (LLM) prompting and ensemble methods to boost performance on named entities. First, they combine retrieval-augmented generation with in-context learning, ensuring that LLM outputs align with external knowledge bases (e.g., WikiData or BabelNet). Structured prompts include named entity translations, role alignment (an “expert translator” framing), and example source-target pairs. Second, they explore ensemble mechanisms that combine outputs from

multiple translation systems, including both LLM and commercial MT engines. The core ensemble approach prioritizes any candidate containing a valid named entity translation, while optional semantic overlap features also favor translations with improved word-level alignment. Experiments reveal that both carefully designed LLM prompts and ensembling yield significant gains in producing accurate entity translations.

VerbaNexAI: “*Enhancing Entity-Aware Translation with Wikidata-Enriched MarianMT*” (Peña Gnecco et al., 2025). The authors present a resource-efficient system for English–Spanish entity-aware translation that enriches MarianMT with a static collection of 240,432 Wikidata entity pairs. This setup aims to address named-entity coverage (e.g., “*Águila de San Juan*”) while maintaining stable fluency on general content. Despite achieving a solid COMET score (87.1), the proposed system underperforms on M-ETA (24.6)—a shortcoming traced to rigid, non-adaptive reliance on Wikidata and the inherent difficulty of exact-match scoring for rare or context-sensitive entities. In contrast, dynamic, retrieval-based large language model methods excel by integrating flexible external knowledge. Their findings emphasize that while static knowledge bases improve translation for well-documented entities, effective cross-domain entity accuracy likely requires adaptive retrieval-augmented or on-demand fine-tuning strategies.

YNU-HPCC: “*Local Cache and Online Retrieval-Based method for Entity-Aware Machine Translation*” (Li et al., 2025). The authors introduce a multi-faceted approach that leverages both traditional and large language model (LLM) architectures to improve entity-aware translation. Specifically, they propose four methods that integrate named entity recognition (NER) modules (BERT or Qwen-based), a local cache of entity translations, and an online retrieval mechanism for unseen entities. By systematically incorporating Wikidata lookups and performing careful prompt engineering for Qwen models, the system achieves higher entity-specific accuracy (M-ETA) and maintains strong overall translation quality (COMET). Notably, a ReAct-based framework further enhances interpretability by explicitly separating reasoning steps from execution, allowing the model to iteratively refine entity translations or query additional resources when encountering ambiguous cases. Ex-

perimental results across ten languages demonstrate the viability of these methods, underscoring that LLM-driven pipelines can surpass traditional MT systems in both robustness and adaptability for entity-centric text.

Zero: “*Entity-Aware Machine Translation: Fine-Tuning NLLB for Improved Named Entity Translation*” (Gundam et al., 2025). The authors address the challenge of translating named entities by fine-tuning a distilled NLLB-200 model with LoRA on a “silver dataset” derived from Google Translate outputs, focusing on efficient adaptation rather than relying on large general-purpose models. This methodology enables the system to learn entity-specific patterns while preserving overall translation quality, as demonstrated by improvements in BLEU, COMET and M-ETA. Notably, while certain languages (e.g., Spanish, Turkish) achieve robust performance, others (e.g., Chinese) remain difficult due to structural complexity and rare entities. Nevertheless, the work underscores that specialized training on moderately sized data can substantially enhance entity translation accuracy, suggesting a cost-effective alternative to massive language models for entity-aware machine translation.

B Extended Results

In this section, we provide additional results for the systems submitted to the EA-MT task. We include the full ranking of all systems across all the 10 languages, as well as the average ranking across all languages. The results are divided into two tables: one for the “gold” systems and one for the end-to-end systems. For all the numerical results, we redirect the reader to the official leaderboard of the task, which is available at <https://huggingface.co/spaces/sapienzanlp/ea-mt-leaderboard>.

B.1 Gold Systems

Table 4 shows the ranking of the “gold” systems submitted to EA-MT for each language pair and on average across languages. The systems are sorted by average ranking, with the best-performing system at the top. We can observe significant cross-lingual performance variations among systems that leverage gold entity information, highlighting that even with perfect entity identification, translation quality remains language-dependent.

The Qwen2.5-based systems from Pingan Team consistently outperform other approaches, achieving top average rankings (3.70 and 4.20). How-

ever, their performance exhibits substantial cross-lingual variance—particularly for Korean (9th and 8th place) versus Germanic languages (1st and 4th place). This pattern suggests that despite using identical model architectures and training methodologies, certain linguistic families present inherently different challenges for entity name translation. The effectiveness of parameter-efficient fine-tuning methods is also evident, as LoRA-based approaches occupy three of the top five positions. Notably, these approaches maintain representational capacity across languages while specifically adapting to entity-translation tasks, outperforming both full fine-tuning and zero-shot prompting strategies, even though different participants employed different underlying models, which may have contributed to the observed performance differences.

Knowledge Integration Mechanisms. The systems submitted to the EA-MT task employed various knowledge integration mechanisms, including retrieval-augmented generation (RAG), ensemble methods, and LLM-only approaches. The performance of these systems varied significantly across languages, indicating that the choice of knowledge integration mechanism plays a crucial role in entity-aware translation.

- **RAG-based systems** (Deerlu, RAGthoven, CHILL) demonstrate strong average performance (4.80–6.30) but with high variance across languages. For instance, RAGthoven’s system ranks 1st for Chinese but 9th for Japanese, despite their writing system similarities. This suggests that knowledge retrieval quality varies significantly by language, potentially reflecting disparities in Wikidata coverage or retrieval accuracy.
- **Ensemble methods** (UAlberta’s WikiEnsemble) show particular strength for Arabic (1st) but mediocre performance for Romance languages like Italian (10th) and French (10th). This pattern indicates that ensemble advantages are most pronounced for languages with greater structural divergence from English, where aggregating multiple translation hypotheses proves valuable.

Language-Specific Challenges. The rank distribution reveals three distinct language clusters with different system behaviors:

- **Germanic and Romance languages** (German, French, Spanish, Italian) show relatively

Team Name	Citation	Publication Title
Arancini	Zenotto et al. (2025)	Arancini at SemEval-2025 Task 2: Multilingual Translation Enhanced by Lightweight LLMs, NER, and RAG for Named Entities
CHILL	Lee et al. (2025b)	CHILL at SemEval-2025 Task 2: You Can’t Just Throw Entities and Hope—Make Your LLM to Get Them Right
Deerlu	Xu (2025)	Deerlu at SemEval-2025 Task 2: Wikidata-Driven Entity-Aware Translation—Boosting LLMs with External Knowledge
FII the Best	Grigorita et al. (2025)	FII the Best at SemEval-2025 Task 2: Steering State-of-the-art Machine Translation Models with Strategically Engineered Pipelines for Enhanced Entity Translation
GinGer	Naebzadeh (2025)	GinGer at SemEval-2025 Task 2: Challenges in Entity-Aware Machine Translation with Fine-Tuning and Zero-Shot Prompting
HausaNLP	Abubakar et al. (2025)	HausaNLP at SemEval-2025 Task 2: Entity-Aware Fine-tuning vs. Prompt Engineering in Entity-Aware Machine Translation
Howard University-AI4PC	Aryal and Agyemang-Prempeh (2025)	Howard University-AI4PC at SemEval-2025 Task 2: Improving Machine Translation With Context-Aware Entity-Only Pre-translations with GPT4o
Pingan Team	Chen (2025)	pingan-team at SemEval-2025 Task 2: LoRA-Augmented Qwen2.5 with Wikidata-Driven Entity Translation
RAGthoven	Skottis et al. (2025)	RAGthoven at SemEval-2025 Task 2: Enhancing Entity-Aware Machine Translation with Large Language Models, Retrieval Augmented Generation and Function Calling
Sakura	Poncelas and Htun (2025)	Sakura at SemEval-2025 Task 2: Enhancing Named Entity Translation with Fine-Tuning and Preference Optimization
SALT	Völker et al. (2025)	SALT at SemEval-2025 Task 2: A SQL-based Approach for LLM-Free Entity-Aware-Translation
SheffieldGATE	Yang et al. (2025)	SheffieldGATE at SemEval-2025 Task 2: Multi-Stage Reasoning with Knowledge Fusion for Entity Translation
silp_nlp	Singh et al. (2025)	silp_nlp at SemEval-2025 Task 2: An effect of Entity Awareness in Machine Translation using LLM
Team ACK	Lee et al. (2025a)	Team ACK at SemEval-2025 Task 2: Beyond Word-for-Word Machine Translation for English-Korean Pairs
UAlberta	Shi et al. (2025)	UAlberta at SemEval-2025 Task 2: Prompting and Ensembling for Entity-Aware Translation
VerbaNexAI	Peña Gnecco et al. (2025)	VerbaNexAI at SemEval-2025 Task 2: Enhancing Entity-Aware Translation with Wikidata-Enriched MarianMT
YNU-HPCC	Li et al. (2025)	YNU-HPCC at SemEval-2025 Task 2: Local Cache and Online Retrieval-Based method for Entity-Aware Machine Translation
Zero	Gundam et al. (2025)	Zero at SemEval-2025 Task 2: Entity-Aware Machine Translation: Fine-Tuning NLLB for Improved Named Entity Translation

Table 3: Overview of the systems submitted to EA-MT, sorted by team name.

Team	System Name	ar_AE	de_DE	es_ES	fr_FR	it_IT	ja_JP	ko_KR	th_TH	tr_TR	zh_TW	AVG
Pingan Team	Qwen2.5-72B-LoRA	8	1	3	1	2	4	9	5	2	2	3.70
Pingan Team	Qwen2.5-72B-LoRA + zhconv	7	4	2	2	1	5	8	4	3	6	4.20
Deerlu	Qwen2.5-Max-Wiki	6	3	1	6	5	1	7	6	6	7	4.80
RAGthoven	GPT-4o + WikiData + RAG	4	8	5	5	3	9	5	2	8	1	5.00
Pingan Team	Phi4-FullIFT	9	5	4	3	4	2	2	11	4	8	5.20
Lunar	LLaMA-RAFT-Plus-Gold	12	2	7	7	7	3	1	3	5	12	5.90
CHILL	GPT4o-RAG-Refine	5	9	8	11	8	6	3	1	1	11	6.30
UAlberta	WikiEnsemble	1	6	9	10	10	7	4	7	7	5	6.60
RAGthoven	GPT-4o + Wikidata	3	7	6	4	6	13	11	10	12	3	7.50
UAlberta	WikiGPT4o	2	10	10	9	11	8	6	8	10	4	7.80
Arancini	WikiGemmaMT	13	11	11	8	9	10	10	9	11	14	10.60
YNU-HPCC	LLaMA + MT	10	12	12	12	12	11	12	12	14	9	11.60
YNU-HPCC	Qwen2.5-32B	11	13	13	13	13	12	13	13	15	10	12.60
Lunar	LLaMA-RAFT-Gold	16	15	14	16	14	14	14	15	13	13	14.40
SALT \square	Salt-Full-Pipeline + Gold	14	14	16	14	15	16	16	16	9	15	14.50
Howard University-AI	DoubleGPT	15	16	15	15	16	15	15	14	16	16	15.30
HausaNLP	Gemini-few-shot	17	17	17	17	17	17	17	17	17	17	17.00
HausaNLP	FT-NLLB	18	18	18	18	18	18	-	-	-	-	18.00
VerbaNexAI Lab	TransNER-SpEn	-	-	19	-	-	-	-	-	-	-	19.00
silp_nlp	NER-M2M100	-	-	-	-	-	19	-	-	-	-	19.00
silp_nlp	T5-MT-Instruct	19	19	20	19	19	20	-	-	-	-	19.33

Table 4: Ranking of “gold” systems submitted to EA-MT for each language pair and on average across languages.

Team	System Name	ar_AE	de_DE	es_ES	fr_FR	it_IT	ja_JP	ko_KR	th_TH	tr_TR	zh_TW	AVG
SheffieldGATE	Llama-Wiki-DeepSeek	-	1	1	1	1	-	-	-	-	-	1.00
SALT \square	Salt-Full-Pipeline	1	2	2	2	2	1	1	1	1	3	1.60
SALT \square	Salt-MT-Pipeline	2	3	3	3	3	2	2	3	2	4	2.70
FII-UAIC-SAI	Qwen2.5-Wiki-MT	5	4	4	4	4	3	3	4	4	1	3.60
Lunar	LLaMA-RAFT-Plus	3	7	7	6	5	8	5	2	3	7	5.30
FII the Best	mBERT-WikiNEuRal	4	5	5	5	7	4	4	6	7	9	5.60
YNU-HPCC	Qwen2.5 + M2M	6	6	6	7	8	5	6	5	6	2	5.70
Lunar	LLaMA-RAFT	7	8	8	8	6	9	7	7	5	8	7.30
The Five Forbidden E	MBart-KnowledgeAware	8	9	12	10	9	6	8	11	10	10	9.30
UAlberta	PromptGPT	10	11	9	9	11	10	9	9	9	6	9.30
RAGthoven	GPT-4o + RAG	11	12	10	12	12	11	11	8	8	5	10.00
Team ACK	Gemini-pro-11m	-	-	-	-	-	-	10	-	-	-	10.00
The Five Forbidden E	Embedded Entities	9	10	11	11	10	7	15	13	12	11	10.90
Team ACK	Chatgpt-4o-11m	-	-	-	-	-	-	12	-	-	-	12.00
Team ACK	Claude-sonnet-11m	-	-	-	-	-	-	13	-	-	-	13.00
HausaNLP	Gemini-Oshot	13	14	13	13	13	13	17	10	13	12	13.10
Zero	FineTuned-MT	12	13	14	15	14	12	16	12	11	13	13.20
Team ACK	Chatgpt-o1-11m	-	-	-	-	-	-	14	-	-	-	14.00
Muhandro_HSE	NER-LLM	14	15	15	14	16	17	24	14	14	14	15.70
JNLP	Multi-task-mT5	-	16	16	16	-	-	-	-	-	-	16.00
sakura	Rakuten7b-P010	-	-	-	-	-	16	-	-	-	-	16.00
silp_nlp	GPT-4o	15	17	18	18	17	14	21	16	15	16	16.70
silp_nlp	GPT-4o-mini	16	18	17	17	15	15	23	15	16	15	16.70
GinGer	LoRA-nllb-distilled-200-distil	17	-	-	-	18	18	-	-	-	-	17.67
Team ACK	Chatgpt-o1-mini-11m	-	-	-	-	-	-	18	-	-	-	18.00
Team ACK	Gemini-flash-11m	-	-	-	-	-	-	19	-	-	-	19.00
Team ACK	Chatgpt-4o-mini-11m	-	-	-	-	-	-	20	-	-	-	20.00
Team ACK	Claude-haiku-11m	-	-	-	-	-	-	22	-	-	-	22.00
Team ACK	Llama-11m	-	-	-	-	-	-	25	-	-	-	25.00

Table 5: Ranking of end-to-end systems submitted to EA-MT for each language pair and on average across languages.

consistent rankings across systems, suggesting more predictable entity translation patterns.

- **East Asian languages** exhibit the highest variability. For Japanese, the ranking difference between the best and worst performing systems (Deerlu’s Qwen2.5-Max-Wiki at 1st vs. Howard’s DoubleGPT at 15th) is striking. Similar patterns emerge for Korean and Chinese. This suggests that entity handling for languages with non-Latin scripts and different

naming conventions benefits differently from various knowledge integration strategies.

- **Turkish and Thai** display unique patterns where CHILL’s GPT4o-RAG-Refine performs exceptionally well (1st for both), despite middling performance on European languages. This system’s iterative refinement approach appears particularly effective for agglutinative languages (Turkish) and languages with unique script properties (Thai).

Key Findings and Research Directions. These findings have several methodological implications for entity-aware translation research:

1. **Language-specific optimization** appears necessary even when using gold entity information, as no system achieved consistent top-tier performance across all languages.
2. **Knowledge retrieval quality** likely varies substantially across languages, suggesting the need for language-specific retrieval strategies rather than one-size-fits-all approaches.
3. **Script adaptation mechanisms** deserve focused attention, as the most dramatic performance variations occur between languages with different writing systems.

These insights indicate that accurate entity translation remains challenging even with gold entity information, reflecting deeper cross-cultural and linguistic adaptation issues that extend beyond simply retrieving correct entity mappings.

B.2 End-to-End Systems

Table 5 shows the ranking of end-to-end systems submitted to EA-MT for each language pair and on average across languages. The systems are sorted by average ranking, with the best-performing system at the top. The results indicate that the best-performing end-to-end systems (SALT, SheffieldGATE, and FII-UAIC-SAI) achieve high average rankings (1.60–2.70), demonstrating the effectiveness of their approaches in entity-aware translation. However, there is still significant room for improvement, as the average ranking across languages remains relatively high.

The systems submitted to the EA-MT task employed various approaches, including fine-tuning, prompting, and ensemble methods. The performance of these systems varied significantly across languages, indicating that the choice of approach plays a crucial role in entity-aware translation.

- **Retrieval-augmented generation (RAG) and function calling** (SheffieldGATE, SALT, LUnar) demonstrate strong average performance. For instance, SheffieldGATE’s system leveraged an agentic approach to enhance entity translation, achieving the best average ranking (1.00) across all languages, while SALT’s system used a SQL-based approach to achieve the second-best average ranking (1.60).

Lunar also performed well with RAG and function calling, achieving an average ranking of 5.30.

- **Fine-tuning approaches** (SALT, FII-UAIC-SAI, Lunar, UAlberta) show that fine-tuning models with entity-specific data can significantly improve translation quality, but the question on how to efficiently adapt the model to the task and how to produce high-quality entity-specific data remains open.

Non-LLM vs. LLM-based Approaches. A fascinating trend in the results is the competitive performance of specifically engineered non-LLM systems against larger language models:

- **SALT’s SQL-based approach** consistently outperforms many LLM-based systems across almost all languages (ranking 1st or 2nd in 9 out of 10 language pairs), demonstrating that lightweight, specialized pipelines can be highly effective when explicitly designed for entity handling. This challenges the assumption that ever-larger models are necessary for complex cross-lingual tasks.
- **FII-UAIC-SAI’s Qwen2.5-Wiki-MT** shows remarkable language-specific adaptability, ranking 1st for Chinese while maintaining strong performance (3rd-5th) across other languages. This suggests that targeted knowledge integration can offset raw model size advantages.

Language-Specific Observations. The end-to-end systems exhibit distinct patterns across language families:

- **Chinese** shows the highest divergence from patterns observed in other languages. FII-UAIC-SAI’s system ranks 1st for Chinese but only 4th overall, while SALT’s top-performing system ranks only 3rd for Chinese despite leading in most other languages. This suggests unique challenges in Chinese entity translation that benefit from specialized approaches.
- **Thai** yields particularly strong results for Lunar’s LLaMA-RAFT-Plus (2nd place), significantly outperforming its average ranking (5.30). This contrasts with Romance languages where the system performs less effectively (6th-7th places), indicating that the

proposed approach might have specific advantages for non-latin script and linguistic properties.

- **Korean** demonstrates a significant variability across systems. Team ACK’s extensive Korean-specific analysis produced a detailed error taxonomy, highlighting how language-specific insights can inform system design.

Key Findings and Research Directions. The results from the end-to-end systems provide several key insights and directions for future research in entity-aware translation:

1. **Entity detection quality** appears to be the critical bottleneck in end-to-end performance, as the gap between gold-information systems and end-to-end systems remains substantial (best M-ETA of 89.1 vs. 77.1).
2. **Computational efficiency tradeoffs** deserve more attention, as lightweight systems like SALT demonstrate that clever architectural choices can outperform resource-intensive approaches in specialized tasks.
3. **Cross-lingual consistency** remains elusive, with the top-5 systems showing performance variations across languages. This suggests that truly universal entity-aware translation systems may require language-family-specific components rather than pure monolithic approaches.

These findings suggest that future research may focus on modular, knowledge-enhanced architectures that can specialize for different language families while maintaining computational efficiency. The success of lightweight but informed systems indicates that architectural innovation may yield more immediate benefits than simply scaling up model size for entity-aware translation.

C XC-Translate: Addendum

Here, we provide an overview of the contents of XC-Translate, our novel gold benchmark dataset for entity-aware machine translation.

C.1 Data Format

The data is provided in JSONL format, where each line in the file contains a JSON object.

The JSON object contains the following fields, as shown in Figure 4:

```
{
  "id": "Q2461698_0",
  "wikidata_id": "Q2461698",
  "entity_types": [
    "Fictional entity"
  ],
  "source":
    "Who are the main antagonistic forces
    in the World of Ice and Fire?",
  "targets": [{
    "translation":
      "Chi sono le principali forze
      antagoniste nel mondo delle
      Cronache del ghiaccio e
      del fuoco?",
    "mention":
      "mondo delle Cronache del ghiaccio
      e del fuoco"
  }],
  "source_locale": "en",
  "target_locale": "it"
}
```

Figure 4: Example of a data entry from XC-Translate showing the JSON structure. Note how the entity “World of Ice and Fire” is translated to “mondo delle Cronache del ghiaccio e del fuoco” in Italian, demonstrating the non-literal translation characteristic of the dataset.

- **id**: A unique identifier for the entry.
- **wikidata_id**: The Wikidata ID of the entity being translated.
- **entity_types**: A list of entity types associated with the entity.
- **source**: The source sentence containing the entity to be translated.
- **targets**: A list of target translations, each containing:
 - **translation**: The translated sentence in the target language.
 - **mention**: The mention of the entity in the translated sentence.
- **source_locale**: The locale of the source sentence (e.g., “en” for English).
- **target_locale**: The locale of the target sentence (e.g., “it” for Italian).

This format allows for easy parsing and processing of the data, making it suitable for training and evaluating machine translation systems.

Entity ID	Type(s)	Text	Mention(s)	Locale
Q746666	Musical work	Can you sing the chorus of the folk song Ring a Ring o' Roses ?	Ring a Ring o' Roses	English
		Puoi cantare il ritornello della canzone popolare Girotondo ?	Girotondo	Italian
Q157073	Person	How long was Mary of Burgundy mar- ried to Emperor Maximilian I?	Mary of Burgundy	English
		Per quanto tempo Maria di Borgogna è stata sposata con l'imperatore Massimil- iano I?	Maria di Borgogna	Italian
Q850522	Movie	Who are the main characters in the movie Little Women ?	Little Women	English
		¿Quiénes son los personajes principales de la película Mujercitas ?	Mujercitas	Spanish
Q1204366	Book	Who is the author of the book A Room of One's Own ?	A Room of One's Own	English
		¿Quién es el autor del libro Una habitación propia ?	Una habitación propia	Spanish

Table 6: Examples of Entity Translations in XC-Translate Dataset. For each example, we display the entity ID (Wikidata ID), the entity type(s), the source text with the entity mention highlighted in light blue, the target text with the translated entity mention highlighted in light peach, and the locale of the source and target texts. The examples illustrate the diversity of entities and their translations across different languages, even when the languages mostly share the same script.

C.2 Examples from XC-Translate

Table 6 shows some examples of entity translations in the XC-Translate dataset. The examples illustrate the diversity of entities and their translations across different languages, highlighting the challenges and complexities involved in entity-aware machine translation even when the languages are closely related and share mostly the same script. The examples also demonstrate the non-literal translations that are often required for proper entity translation, as seen in the translations of “Ring a Ring o’ Roses” to “Girotondo” and “Mary of Burgundy” to “Maria di Borgogna”.