Cultural and Knowledge Biases in LLMs Through the Lens of Entity-Aware Machine Translation

Lu Xu[1], **Luca Moroni**[1], **Roberto Navigli**[1,2]
[1]Sapienza NLP Group, Sapienza University of Rome, Rome, Italy
[2]Babelscape, Rome, Italy
{xu, moroni}@diag.uniroma1.it, navigli@{diag.uniroma1.it, babelscape.com}

Abstract

Large Language Models (LLMs) demonstrate strong multilingual capabilities yet exhibit systematic cultural biases that affect entity-aware machine translation. While external knowledge integration improves translation accuracy, the extent of these benefits across varying degrees of cultural specificity remains unexplored. To fill this gap, we propose a three-level cultural specificity classification and framework (Culturally Agnostic, Culturally Sensitive, and Culturally Local Knowledge) to systematically analyze how cultural context affects entity translation difficulty and the utility of external knowledge. Through experiments spanning 11 LLMs and 10 languages, we demonstrate that external knowledge provides substantially greater improvements for culturally local entities (up to 70% in m-ETA) compared to culturally agnostic ones. Our analysis reveals distinct behavioral patterns across model tiers: closed and open-weight models show synergistic improvements in both entity accuracy and overall translation quality, while open-data models struggle with instruction-following despite improved entity accuracy.

Keywords: Machine Translation, Multilinguality, Cultural Specificity

1. Introduction

Large Language Models (LLMs) have demonstrated remarkable multilingual capabilities, yet they exhibit systematic cultural biases that reflect uneven knowledge across cultures. Recent benchmarks reveal that LLMs perform significantly better on globally prominent cultures than on underrepresented ones (Myung et al., 2024; Chiu et al.; Cao et al., 2023; Alhanai et al., 2025), often struggling with local cultural knowledge even when queries are posed in culturally appropriate languages (Etxaniz et al., 2024). This imbalance largely stems from skewed training data distributions: Western and high-resource language content dominates large-scale corpora, leaving models with limited parametric knowledge of culturally specific entities from less represented regions (Arora et al., 2023).

These knowledge gaps create tangible challenges in generation tasks that require cultural understanding. Machine Translation (MT), increasingly powered by LLMs, is particularly sensitive to the correct interpretation of word meaning in context. Prior work has shown that MT systems struggle with lower-frequency word senses (Campolungo et al., 2022; Martelli et al., 2025), and more broadly that Word Sense Disambiguation becomes significantly more difficult when models must resolve less common senses (Meconi et al., 2025). This challenge becomes even more pronounced when translating named entities that exhibit varying degrees of cultural specificity. While universally recognized entities tend to translate reliably across languages, culturally embedded entities, like traditional dishes, local festivals, regional landmarks, or historical figures, pose substantial difficulties (Budimir, 2025). Because such entities appear infrequently in training data, models are less likely to have encountered them or their context-dependent meanings, which increases the risk of mistranslation or incorrect interpretation. We thus need systematic methods to measure and address these knowledge gaps.

We argue that entity-aware machine translation (EA-MT) provides a systematic framework for studying these cultural knowledge gaps. The task focuses on accurately translating named entities within their sentential context (Conia et al., 2025). Entities are particularly diagnostic of cultural knowledge imbalances because they span a clear spectrum of cultural specificity: universally known entities like “United Nations” or “World Health Organization” are well represented across training data, whereas culturally embedded entities such as regional festivals, traditional dishes, or local historical figures may appear predominantly in specific cultural contexts. Integrating external knowledge from multilingual knowledge bases like Wikidata has proved capable of producing substantial improvements in translation accuracy (Xu, 2025). However, two critical questions remain unanswered: *Do LLMs exhibit systematic behavioral differences across varying levels of cultural specificity, and does the benefit of external knowledge vary accordingly?*

Cultural specificity has long been recognized in translation studies, where culture-specific items (CSIs) have been extensively analyzed in terms of translation strategies (Daghoughi and Hashemian, 2016). Recent MT research has begun address-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8794–8812
11–16 May 2026.

ing this challenge: work on cultural MT has developed benchmarks and datasets focusing on culturally embedded entities (Yao et al., 2024; Conia et al., 2024). However, existing approaches treat collected cultural entities uniformly without systematically distinguishing between different levels of cultural embeddedness. For instance, a globally recognized landmark like the "Eiffel Tower" and a local temple known only in a specific prefecture are both labeled as cultural entities, despite representing vastly different degrees of cultural specificity and likely requiring different levels of external knowledge support. Meanwhile, traditional MT difficulty metrics focus on linguistic factors such as sentence length, syntactic complexity, and word rarity (Araghi and Palangkaraya, 2024; Lim et al., 2024), which are largely culture-agnostic and do not capture the cultural dimension of translation difficulty. However, no systematic framework exists to categorize entities by their degree of cultural specificity and analyze how external knowledge benefits vary across this continuum.

To address this gap, we argue that entities should be understood through graduated levels of cultural specificity rather than as binary categories of "cultural" versus "non-cultural." This multi-level perspective is critical because LLMs' parametric knowledge is not uniformly distributed: models trained predominantly on high-resource languages and general-domain text may possess sufficient knowledge for universally recognized entities but progressively lack cultural knowledge as entities become more locally embedded. We operationalize this through a three-level cultural specificity framework that distinguishes between *Culturally Agnostic*, *Culturally Sensitive*, and *Culturally Local* entities. We hypothesize that external knowledge benefits will increase systematically along this dimension, providing the greatest improvements for culturally local entities where models' internal knowledge is most deficient. Our experiments across multiple models and language pairs strongly support this hypothesis.

We carry out entity-aware MT experiments spanning 10 languages and 11 LLMs, demonstrating consistent improvements when integrating external knowledge. To systematically understand these benefits, we conduct an in-depth cultural specificity analysis on two languages (Traditional Chinese and Italian), revealing that external knowledge provides substantially greater improvements for culturally local entities compared to culturally agnostic ones. This systematic pattern holds across different model sizes, though smaller models demonstrate greater overall reliance on external knowledge. Through extensive qualitative analysis on Traditional Chinese and Italian, we uncover how and why models exhibit distinct behavioral patterns when handling entities at different cultural specificity levels. These findings reveal that cultural specificity is a critical dimension for understanding when external knowledge benefits entity translation, with important implications for designing more culturally-aware MT systems.[1]

In this paper we put forward the following contributions:

- We propose a novel perspective on entity cultural specificity and operationalize it through a three-level annotation framework, providing a systematic approach for analyzing how cultural context affects translation difficulty and the utility of external knowledge.
- We conduct the first comprehensive study across several LLMs demonstrating that external knowledge benefits vary substantially and systematically across the three-tier cultural specificity.
- We release our cultural specificity annotations to facilitate future research on culturally-aware machine translation.

2. Related Work

Entity-Aware Machine Translation The accurate translation of named entities has long been recognized as a fundamental challenge in machine translation, as these entities often require specialized handling beyond literal word-to-word translation. Multi-task training frameworks that jointly optimize named entity recognition and machine translation objectives (Rikters and Miwa, 2024), and end-to-end entity-aware systems incorporating entity classifiers within encoders and decoders (Xie et al., 2022), have demonstrated improved accuracy through architectural innovations. However, these approaches do not systematically investigate when and why external knowledge benefits entity translation across varying degrees of cultural specificity. The integration of external knowledge sources has emerged as a complementary paradigm. Knowledge graphs enhance semantic feature extraction for rare entities and terminological expressions (Zhang et al., 2023), and have proven successful in question-answering systems where entity correctness is crucial (Srivastava et al., 2023). With Large Language Models, RAG-based frameworks using knowledge graphs as non-parametric sources have achieved substantial improvements through multi-task training that teaches models to refine and utilize multilingual knowledge (Wang et al., 2024). Our work addresses this gap by proposing a cultural specificity framework that systematically categorizes entities

[1]Available at: Github/SapienzaNLP/cultural-ea-mt.

and quantifies how external knowledge benefits vary across this spectrum.

Cultural and Knowledge Biases in LLMs Recent systematic evaluations across 107 countries reveal that LLMs consistently align with Western cultural values, especially those of English-speaking and Protestant European countries, with cultural distance from these reference points correlating strongly with model misalignment (Tao et al., 2024). This bias manifests across multiple dimensions: LLMs struggle with non-Western cultural nuances even when trained on the languages of those non-Western cultures (Aksoy, 2025), exhibit acute knowledge deficiencies for low-resource cultures (Ochieng et al., 2024), show limited symbolic diversity for non-Western regions (Li et al., 2024), and fail at culturally-grounded reasoning tasks (Cecilia Liu et al., 2024). These biases pose significant challenges for entity-aware translation, where accurate rendering of culturally-specific named entities requires deep cultural knowledge that current LLMs often lack. Our work directly addresses this gap by systematically quantifying how cultural specificity affects translation quality and the degree of external knowledge augmentation required. Unlike previous approaches, which treat entities uniformly, we demonstrate that knowledge integration benefits vary substantially from culturally agnostic to culturally local entities, and therefore provide a nuanced understanding of when and why external knowledge proves most useful, with particular attention to the interplay between model capabilities and entity cultural specificity.

3. Cultural-Level Annotation

Entity-aware machine translation requires handling entities with varying degrees of cultural specificity. We hypothesize that both the benefit of external knowledge and the difficulty of translation correlate with an entity's cultural specificity. To test this hypothesis, we develop a three-level annotation scheme:

- **Level 0 – Culturally Agnostic:** Entities whose recognition does not require cultural knowledge. Bilingual speakers familiar with the entity in their native language can recognize it in another language through direct lexical mapping. For instance, a native English speaker familiar with "Introduction to Algorithms" can identify "算法导论" via compositional transparency – i.e., "算法"(Algorithms) + "导论"(Introduction) without – knowing the cultural background of the target language.
- **Level 1 – Culturally Sensitive:** Entities with clear cultural origin but widely recognized internationally. Bilingual speakers familiar with the entity in their native language may not directly recognize it through its name in another language, but can identify it by leveraging cultural knowledge of both linguistic contexts. For instance, a native English speaker familiar with "the Forbidden City" may not immediately recognize "紫禁城" from lexical mapping alone, but by connecting these cultural concepts "紫" (purple) signifying royalty, "禁" (forbidden) denoting restriction, and "城" (city), the speaker can deduce this refers to the correct entity, which is China's most famous ancient imperial palace.
- **Level 2 – Culturally Local:** Entities requiring deep local or insider cultural knowledge, meaningful primarily within their own cultural context. Bilingual speakers familiar with the entity in their native language cannot recognize it through its name in another language without possessing insider cultural knowledge typically unavailable to non-native speakers. For instance, a native English speaker familiar with the TV series "Love Like The Galaxy" would not recognize "星汉灿烂，月升沧海" without a deep understanding of Chinese literary culture, specifically the classical poetry underlying "星汉灿烂" (stars splendid in the Milky Way) and "月升沧海" (moon rising over the vast sea), which form the poetic title that a typical bilingual speaker would not be able to connect to the English name.

4. Experimental Setup

4.1. Entity-Aware Machine Translation Dataset

We build upon XCTranslate, the Entity-Aware Machine Translation dataset introduced by Conia et al. (2024). Given a sentence s in English containing a mention of an entity e, the task is to translate s into a target language while adapting the name of e to the target language in such a way as to preserve the original meaning of the sentence.

Each input sample consists of:

- `wikidata_id`: the identifier of the entity e in the input text.
- `entity_types`: the types of which entity e is an instance (chosen from 14 available types, reported in Table 9).
- `source`: the source text (in English).
- `targets`: a list of possible translations, each containing the translated text in one of the 10

target languages and, as a `mention`, the correct translation of the entity e in the target language.

We translate from English into 10 languages using 7,278 samples from XCTranslate (over 700 samples per language). The target languages are classified into two tiers: German, Spanish, Chinese, Italian, French, and Japanese (high-resource languages), and Korean, Turkish, Arabic, and Thai (mid-resource languages). For our cultural analysis, we further annotate the entities appearing in the Italian and Chinese subsets, resulting in 306 and 267 annotated entities, respectively. We classify each language tier based on its presence in NLP resources related to Machine Translation, following the rationale of Joshi et al. (2020). This categorization helps highlight our findings more clearly by separating results according to language resource availability.

4.2. Models

We evaluate a diverse set of instruction-tuned LLMs, organizing them into three categories based on their openness and accessibility:

- (i) **Closed Models**: These models are accessible exclusively through proprietary APIs. We include GPT-5-mini (OpenAI) (OpenAI et al., 2024), Gemini-2.5-flash (Google) (Team et al., 2025a), and Qwen-plus (Qwen et al., 2025). Although Qwen-plus is derived from an open-source base model, it is only accessible via an API.
- (ii) **Open-Weight Models**: These models release their weights, but their entire training pipeline remains closed, including information about the data used. We evaluate Gemma-3-4B (Team et al., 2025b), Llama-3.1-8B (Grattafiori et al., 2024), and Qwen2.5-7B (Qwen et al., 2025).
- (iii) **Open-Data Models**: These models exemplify full open-science practices, both model weights and training data are publicly available, and the entire development pipeline is documented. From this category, we include Minerva-7B (Orlando et al., 2024b), Occiglot[2] (both Occiglot-de-7B and Occiglot-eu-7B variants), EuroLLM-9B (Martins et al., 2025), and Salamandra-7B (Gonzalez-Agirre et al., 2025).

4.3. Experimental prompt settings

To address the Entity-Aware Machine Translation task, we define three different prompt strategies, following the approach of Xu (2025):

[2] https://huggingface.co/occiglot

SYSTEM
`You are an expert translator. Translate from English to` **`{target_language}`**`. Only provide the translation without explanations.`
USER
`sentence:` **`{source_text}`**

Table 1: Prompt for Baseline setting.

- **Baseline:** Simple prompt to perform the machine translation task, without external knowledge.
- **Relik:** Extension of the Baseline prompt using entity information automatically retrieved via Relik (Orlando et al., 2024a), state-of-the-art Entity Linking system, which identifies entities and links them to Wikidata IDs in order to obtain source and target language names.
- **Gold:** An extension of the baseline prompt, enriched with Wikidata entity information retrieved using gold Wikidata IDs from the XCTranslate benchmark, including entity names in both the source and target languages.

In Table 1 we show the prompt used in the Baseline setting, while in Table 2 we show the prompt used in the Relik and Gold settings, which rely on the same prompt structure.

The bold variables in the prompts shown in the tables are placeholders for the input data and external knowledge.

4.4. Evaluation Metric

We rely on two different metrics to assess models' outputs. We use COMET (Rei et al., 2020) as an overall quality estimator metric, which is one of the most widely used metrics for the Machine Translation task. We also use m-ETA (mean Entity Translation Accuracy) as our task-specific metric, which directly measures whether entities are correctly translated through string matching.

5. Dataset Annotation

To enable our cultural study we annotate the XCTranslate data using our three-tier annotation scheme (see Section 3). We focus on two languages of XCTranslate's entity set: Chinese and Italian. The annotation was performed by three annotators: one native Italian speaker and one native Chinese speaker from our research group, plus one native Chinese-speaking PhD student from a

SYSTEM

```
You are an expert translator.
Translate from English to {target_language}. Only provide the
translation without explanations.
There is an entity in the sentence,
the Wikipedia titles of the en-
tity in both source language and
target language are given. When
you translate the sentence, please
use the given mention in the tar-
get language. If the mention in
the target language is 'Label not
found', then translate it word by
word.
```

USER

```
sentence: {source_text}.

The entity in English is
{source_title}.

In {target_language} it is {tar-
get_title}.
```

Table 2: Prompt for Relik and Gold settings.

Language	Level 0	Level 1	Level 2
it_IT	35	379	316
zh_TW	51	236	435

Table 3: Number of task samples for Italian (it_IT) and Chinese (zh_TW) divided by our three different cultural levels.

Chinese university. Each annotator independently labels each entity in the selected subset, and for each entity, the final cultural level is computed as the majority vote across all annotations.

We computed an averaged pairwise inter-annotator agreement (Cohen's kappa) of 0.81 for Chinese and 0.74 for Italian, which represent an "*almost perfect*" agreement in both languages.

Table 3 reports the number of samples per cultural category for both Italian and Chinese. We observe that culturally sensitive and culturally local entities make up the majority of the data in both languages, whereas culturally agnostic entities are underrepresented, accounting for only 4% of the Italian data and 7% of the Chinese data.

6. Results

6.1. Overall Performance

Table 4 reports the m-ETA and COMET scores across the three prompt settings, for 11 LLMs distinguished by their openness tier. The results are divided by language resourcedness tiers. Consistently with the findings of Xu (2025), we observe that the prompting strategy strongly affects m-ETA scores. The Baseline prompt yields consistently lower m-ETA values, with top performing models scoring around 40%, whereas knowledge-augmented prompts (Relik and Gold) lead to substantial improvements, leading models to reach quite good performances, around 80%-90%, showing an average improvement of 40%-50% across all models and languages.

Overall, the Relik prompt consistently trails the Gold prompt across all models for both m-ETA and COMET metrics. This highlights the critical role of accurate entity linking in leveraging external knowledge: even a state-of-the-art linker such as Relik underperforms relative to gold annotations in knowledge-enhanced, entity-aware translation. Nonetheless, Relik still provides a meaningful improvement over the Baseline, which fails to achieve competitive m-ETA scores without knowledge augmentation.

We also observe that the language resourcedness tier significantly influences performance, particularly for Open-Weight and Open-Data models. These models exhibit lower m-ETA and COMET scores for mid-resource languages compared to high-resource ones.

Finally, we identify distinct trends across model tiers. Closed and Open-Weight models show consistent gains in COMET when supplied with external knowledge, reflecting a synergistic improvement between entity accuracy (m-ETA) and overall translation quality (COMET). In contrast, Open-Data models exhibit the opposite pattern: while knowledge-augmented prompts raise m-ETA scores, COMET scores drop markedly. This suggests that Open-Science models, despite correctly incorporating external entities, may suffer from degraded overall translation quality, likely due to weaker instruction-following and limited capacity to integrate external knowledge effectively.

6.2. Performance Across Entity Types

To quantify the impact of the type of entity on translation difficulty, in Table 5 we report the m-ETA scores of Qwen-plus, one of the best performing models, prompted with the Baseline setting. We select a subset of entity types, choosing the ones that contain a sufficient number of samples per language. From our results we notice that some types like "TV series" and "Movies" are more difficult to be translated correctly, due to language specific and cultural aspects (e.g. names of movies that do not have the same meaning across languages: "*Eternal sunshine of a spotless mind*" (English) and "Se mi lasci ti cancello" (Italian) *EN: If you leave*

Model	Baseline		Relik		Gold	
	m-ETA	**COMET**	**m-ETA**	**COMET**	**m-ETA**	**COMET**
High-Resource Languages						
Closed Models						
GPT-5-mini	32.2	77.6	75.4	93.2	89.7	94.9
Gemini-2.5-flash	46.2	91.2	76.3	93.6	88.8	94.9
Qwen-Plus	38.7	90.9	75.0	93.4	89.2	94.8
Open-Weight Models						
Gemma-3-4B	23.8	86.1	71.7	87.1	87.6	89.8
Llama-3.1-8B	19.1	74.4	62.4	80.5	75.6	83.3
Qwen2.5-7B	17.7	83.0	68.5	83.4	84.5	87.7
Open-Data Models						
Minerva-7B	10.5	69.1	46.9	63.7	56.5	65.3
EuroLLM-9B	30.3	87.6	72.8	74.4	86.3	78.1
Occiglot-de-7B	14.5	56.5	55.8	51.6	68.5	52.9
Occiglot-eu-7B	16.8	55.5	52.2	54.3	64.3	55.6
Salamandra-7B	13.4	68.0	47.6	63.5	57.1	64.4
Mid-Resource Languages						
Closed Models						
GPT-5-mini	28.5	89.3	72.3	93.1	89.0	94.8
Gemini-2.5-flash	37.1	89.8	75.0	93.5	88.3	94.7
Qwen-Plus	25.0	88.9	72.0	92.7	87.8	94.2
Local Models						
Qwen2.5-7B	5.2	65.7	56.9	72.2	73.3	75.8
Gemma-3-4B	14.1	85.8	67.6	85.7	85.6	88.7
Llama-3.1-8B	7.5	58.9	52.1	67.8	65.8	70.0
Open-Data Models						
Minerva-7B	1.8	44.7	48.6	52.8	63.6	55.5
EuroLLM-9B	18.4	75.1	64.2	63.1	81.1	65.8
Occiglot-de-7B	2.7	41.9	47.5	42.7	59.1	43.5
Occiglot-eu-7B	2.3	39.2	45.9	39.7	60.8	40.9
Salamandra-7B	1.5	48.0	29.2	50.1	37.6	51.0

Table 4: m-ETA (%) and COMET scores averaged across high- and mid-resource languages. The results are reported separately for the three prompt settings (Baseline, Relik, and Gold).

Language	Musical	Artwork	Food	Book	Fictional	Landmark	Movie	Place of w.	TV series	Person
High-Resource	53.9	37.4	56.4	37.5	47.7	36.1	19.4	47.0	18.1	37.6
Mid-Resource	32.6	25.3	38.3	26.8	34.5	21.3	19.5	28.5	16.0	25.5

Table 5: m-ETA(%) scores for Qwen-plus using the Baseline prompt, for several entity types. Results are averaged across high-resource and mid-resource languages.

me, I'll erase you). Other entity types demonstrate higher accuracy; for example, "Musical" and "Food" entities achieve an average m-ETA higher than 50% across high-resource languages. This pattern likely arises because these entities are more commonly known and culturally shared.

6.3. Performance by Cultural Level

We hypothesize that *LLMs exhibit different behaviors when translating entities with varying degrees of cultural specificity, and external knowledge provides greater benefits for culturally specific entities*. To test this hypothesis, we utilize our annotation pipeline (Section 5) for Italian and Chinese, categorizing entities into three culturality levels: Level 0 (Culturally Agnostic), Level 1 (Culturally Sensitive), and Level 2 (Culturally Local). Table 6 reports m-ETA scores for all models across the three prompt settings and cultural levels, for Chinese and Italian.

Our results show a substantial degradation in m-ETA from Culturally Agnostic to the more culturally specific levels under the Baseline prompt setting. This trend is consistent across all models and both languages, highlighting a key limitation in how current LLMs handle culturally grounded information.

Model	Level 0 (Agnostic)			Level 1 (Sensitive)			Level 2 (Local)		
	Base	Relik	Gold	Base	Relik	Gold	Base	Relik	Gold
Chinese									
Closed Models									
GPT-5-mini	76.5	76.5	84.3	42.4	74.6	86.9	19.5	63.5	80.5
Gemini-2.5-flash	70.6	86.3	88.2	53.8	80.1	86.4	29.7	65.8	80.0
Qwen-Plus	60.8	72.6	84.3	52.1	77.5	86.9	25.8	63.0	80.2
Open-Weight Models									
GEMMA-3-4B	39.2	78.4	86.3	17.0	71.6	83.9	4.4	60.5	79.8
Llama-3.1-8B	33.3	74.5	74.5	8.1	62.3	74.2	3.2	52.6	64.1
Qwen2.5-7B	5.9	68.6	74.5	0.9	69.5	84.3	1.2	55.6	73.6
Open-Data Models									
Minerva-7B	7.8	60.8	62.8	0.0	46.2	62.7	0.7	44.1	58.2
EuroLLM-9B	39.2	80.4	92.2	23.7	71.2	83.9	12.9	61.8	74.9
Occiglot-de-7B	9.8	64.7	56.9	1.3	37.7	50.0	0.5	44.1	51.7
Occiglot-eu-7B	5.9	62.8	60.8	0.9	47.0	61.9	1.2	35.6	52.4
Salamandra-7B	3.9	47.1	43.1	1.3	38.1	47.9	0.2	29.7	39.5
Italian									
Closed Models									
GPT-5-mini	66.7	88.6	94.3	50.0	80.0	92.4	29.1	78.5	94.6
Gemini-2.5-flash	80.0	85.7	94.3	60.4	79.7	92.1	31.3	78.8	94.6
Qwen-Plus	60.0	82.9	91.4	55.9	78.9	92.6	21.8	77.5	94.3
Open-Weight Models									
GEMMA-3-4B	74.3	88.6	91.4	38.0	77.0	90.5	15.5	76.3	93.0
Llama-3.1-8B	51.4	80.0	88.6	31.1	68.9	82.6	13.0	64.6	82.9
Qwen2.5-7B	34.3	77.1	91.4	31.7	75.2	89.7	10.8	74.1	91.5
Open-Data Models									
Minerva-7B	54.3	62.9	88.6	30.6	64.4	70.7	13.9	51.0	56.7
Occiglot-de-7B	20.0	57.1	80.0	20.8	68.6	78.6	12.0	62.3	75.0
Occiglot-eu-7B	48.6	60.0	74.3	32.5	61.7	69.7	12.7	47.9	59.2
EuroLLM-9B	65.7	85.7	100.0	45.4	80.0	91.8	18.4	75.0	88.3
Salamandra-7B	48.6	60.0	54.3	22.4	53.0	58.1	9.2	50.6	62.3

Table 6: m-ETA (%) across three cultural levels and three prompt settings, for Traditional Chinese (zh_TW) (Top) and Italian (it_IT) (Bottom).

These findings underscore the need for future research in order to better address culturality and knowledge bias in multilingual and cross-cultural translation.

Moreover, knowledge-augmented prompts (Relik and Gold) largely mitigate this degradation. In most cases, they yield improvements of up to 70% in m-ETA over the Baseline, demonstrating that external knowledge integration can effectively bridge the gap in entity translation performance across different cultural levels.

The results underscore the value of our three-tier annotation framework. Across both languages and nearly all models, external knowledge produces the greatest improvements at Level 2 (Culturally Local) and the smallest at Level 0 (Culturally Agnostic). The differing behaviors of LLMs across the three levels of culturality highlight the importance of incorporating a third layer of cultural specificity, which is crucial for more effectively disentangling cultural factors.

Overall, our findings consistently support the hypothesis that LLMs behave differently across varying levels of culturality, and that the integration of external knowledge plays a crucial role, particularly for culturally local and culturally sensitive entities.

7. Qualitative Analysis

To further analyze the behavior of LLMs on the entity-aware machine translation task, we conduct a qualitative analysis on the output of LLMs for the Gold prompting setting, selecting GPT-5-mini as the LLM to analyze. We carry out our analysis on the Italian and Chinese languages, where GPT-5-mini with Gold prompt reaches an m-ETA of 93.4% Italian and 82.8% for Chinese.

Our manual analysis reveals that, across all in-

Input	Prediction	Wikidata Label	Gold Label
How many episodes are in the TV series Space Battleship Yamato II?	Quanti episodi ci sono nella serie TV Uchū senkan Yamato 2 ? (*EN: How many episodes are there in the TV series Space Battleship Yamato 2?*)	Uchū senkan Yamato 2 (*EN: Space Battleship Yamato 2*)	La corazzata Yamato (*EN: The Battleship Yamato*)
How many movements are there in Symphony No. 41?	Quanti movimenti ha la Sinfonia n. 41 ? (*EN: How many movements are there in Symphony No. 41?*)	Sinfonia n. 41 (*EN: Symphony No. 41*)	Symphony No. 41
What kind of animal is Perry the Platypus?	鴨嘴獸泰瑞是什麼動物? (*EN: What kind of animal is Perry the Platypus?*)	鴨嘴獸泰瑞 (*EN: Platypus Perry*)	鴨嘴獸佩里 (*EN: Platypus Perry*)

Table 7: Qualitative analysis of GPT-5-mini predictions. The table report input English text, model prediction, Wikidata retrieved labels and XCTranslate annotated labels. In the prediction the green text represent the correct usage of Wikidata injected knowledge.

Lang	level 0	level 1	level 2
it_IT	3 (8.5%)	29 (7.0%)	17 (5.3%)
zh_TW	8 (15.6%)	31 (13.1%)	86 (19.7%)

Table 8: Number of samples (and percentage %) where the retrieved Wikidata label in Italian or Chinese **does not correspond** to the gold annotated mention for entities in XCTranslate.

correctly translated entity names, a recurring issue arises from mismatches between the Wikidata labels and the corresponding entity names annotated in the XCTranslate dataset, which serves as the reference for computing the m-ETA score. Table 7 presents illustrative examples of such misalignments for Italian and Chinese, along with the corresponding predictions from GPT-5-mini. In these cases, the model fails to match the gold mention translation, despite correctly leveraging the retrieved Wikidata information.

For instance, in the first Italian example, the Wikidata label “Uchū senkan Yamato 2" does not align with the gold label “La corazzata Yamato", which is correctly annotated. In contrast, the second Italian case demonstrates that the Wikidata label actually provides the correct Italian translation, while the XCTranslate annotation is erroneous. A similar pattern is observed in the third example for Chinese, where the gold label is incorrect and fails to match the valid Wikidata label.

This analysis highlights inconsistencies and potential annotation errors in the XCTranslate dataset, suggesting the need for further verification and correction. Moreover, it underscores the inherent limitations of relying on external knowledge sources such as Wikidata, which may occasionally contain inaccuracies or mismatches that affect knowledge injection methodologies.

We quantify label misalignment in Table 8, where we show the percentage of mismatched Wikidata labels retrieved compared to the gold labels in XCTranslate. We see that the percentage of wrongly retrieved data is higher than the error reached by some models in Table 6. Additional manual analysis reveals that some misalignments between Wikidata and XCTranslate are due to some minor aspects, like misspelling, uncommon name associated with an entity, or plural version of the label. For example, LLMs can handle inconsistencies such as “cipolla d’inverno" (Welsh onion) in Wikidata and “cipolle d’inverno" (Welsh onions) in XCTranslate by exploiting the context of the input text.

Additionally, we analyze the predicted output of Open-Data Models under the Gold prompt setting. We notice that these models often struggle to correctly leverage external knowledge, resulting in translations that use incorrect entity names in the target language. Furthermore, our analysis highlights the limited instruction-following ability of Open-Data LLMs: in several cases, the models fail to adhere to the prompt requirement of producing only the translation, adding unrequested information. This observation helps explain the trend reported in Table 4, where both Relik and Gold prompts lead to improvements in the m-ETA score while at the same time degrading the overall translation quality, as reflected by lower COMET scores for Open-Data Models.

8. Conclusion

This work investigates the role of cultural specificity in entity-aware machine translation, addressing a critical gap in understanding when external knowledge integration provides the greatest benefit. Through systematic evaluation across 11 LLMs with varying degrees of openness, we demonstrate that cultural specificity fundamentally influences both the difficulty of entity translation, and the effectiveness of knowledge augmentation strategies in 10 high- and mid-resource languages.

We put forward a three-tier annotation strategy that shows substantial performance degradation from culturally agnostic to culturally local entities under baseline conditions. Our findings confirm that external knowledge provides the greatest benefit for culturally local entities and minimal improvement for culturally agnostic ones. Our qualitative analysis highlights important limitations in current evaluation paradigms, with misalignments between Wikidata labels and dataset annotations contributing to artificial evaluation penalties. These findings point to future research directions in developing culturally-aware knowledge sources and improving evaluation metrics for cross-cultural translation scenarios.

We hope our culturally-annotated dataset for Chinese and Italian, comprising annotations for over 500 entities across three cultural levels, will enable future research into cultural aspects of machine translation and language understanding.

9. Limitations

Although this paper presents an analysis of cultural and knowledge biases, it is not without limitations.

Due to the structure of the XCTranslate dataset, our analysis was limited to translations from English into the target languages. This constraint prevented us from examining translations in the opposite direction, which could provide valuable insights into additional forms of cultural bias in multilingual LLMs.

Our cultural specificity annotations are limited to two languages: Traditional Chinese and Italian. While these belong to distinct linguistic families, extending the annotations to additional languages – particularly mid- and low-resource languages such as Thai and Turkish – would strengthen the generalizability of our findings.

The annotation process, although achieving high inter-annotator agreement (κ = 0.81 for Chinese; κ = 0.74 for Italian), required substantial manual effort, which limited our ability to scale the analysis to all 10 languages in the XCTranslate dataset.

Although our three-level framework moves beyond a binary "cultural" versus "non-cultural" classification, cultural specificity lies fundamentally on a continuous spectrum. Our categorical approach, while more nuanced, still constitutes a discretization that obscures meaningful variation. Considerable within-level variation remains: two Level 2 entities may differ substantially in their degree of local embeddedness, ranging from regionally recognized to highly community-specific.

10. Acknowledgements

The authors gratefully acknowledge the support of the AI Factory IT4LIA project.

This work was carried out while Lu Xu was enrolled in the Italian National Doctorate on Artificial Intelligence at the Sapienza University of Rome.

11. References

Meltem Aksoy. 2025. Whose morality do they speak? unraveling cultural bias in multilingual language models. *Natural Language Processing Journal*, page 100172.

Tuka Alhanai, Adam Kasumovic, Mohammad M Ghassemi, Aven Zitzelberger, Jessica M Lundin, and Guillaume Chabot-Couture. 2025. Bridging the gap: enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812.

Sahar Araghi and Alfons Palangkaraya. 2024. The link between translation difficulty and the quality of machine translation: a literature review and empirical investigation. *Language Resources and Evaluation*, 58(4):1093–1114.

Arnav Arora, Lucie-Aimée Kaffee, and Isabelle Augenstein. 2023. Probing pre-trained language models for cross-cultural differences in values. *Cross-Cultural Considerations in NLP@ EACL*, page 114.

Bojana Budimir. 2025. The challenge of translating culture-specific items: Evaluating MT and LLMs compared to human translators. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 455–467, Geneva, Switzerland. European Association for Machine Translation.

Niccolò Campolungo, Federico Martelli, Francesco Saina, and Roberto Navigli. 2022. DiBiMT: A novel benchmark for measuring Word Sense Disambiguation biases in Machine Translation. In *Proceedings of the 60th Annual Meeting of the*

Association for Computational Linguistics (Volume 1: Long Papers), pages 4331–4352, Dublin, Ireland. Association for Computational Linguistics.

Yong Cao, Li Zhou, Seolhwa Lee, Laura Cabello, Min Chen, and Daniel Hershcovich. 2023. Assessing cross-cultural alignment between chatgpt and human societies: An empirical study. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 53–67.

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. Culturalbench: a robust, diverse and challenging benchmark on measuring (the lack of) cultural knowledge of llms.

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.

Simone Conia, Min Li, Roberto Navigli, and Saloni Potdar. 2025. Semeval-2025 task 2: Entity-aware machine translation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2535–2557.

Shekoufeh Daghoughi and Mahmood Hashemian. 2016. Analysis of culture-specific items and translation strategies applied in translating jalal al-ahmad's" by the pen". *English language teaching*, 9(4):171–185.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. Bertaqa: How much do language models know about local culture? *Advances in Neural Information Processing Systems*, 37:34077–34097.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, Daniel Tamayo, José Javier Saiz, Ferran Espuña, Jaume Prats, Javier Aula-Blasco, Mario Mina, Iñigo Pikabea, Adrián Rubio, Alexander Shvets, Anna Sallés, Iñaki Lacunza, Jorge Palomar, Júlia Falcão, Lucía Tormo, Luis Vasquez-Reina, Montserrat Marimon, Oriol Pareras, Valle Ruiz-Fernández, and Marta Villegas. 2025. Salamandra technical report.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Va-

sic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe,

Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. CULTURE-GEN: Revealing global cultural perception in language models through natural language prompting. In *First Conference on Language Modeling*.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Federico Martelli, Stefano Perrella, Niccolò Campolungo, Tina Munda, Svetla Koeva, Carole Tiberius, and Roberto Navigli. 2025. Dibimt: A gold evaluation benchmark for studying lexical ambiguity in machine translation. *Computational Linguistics*, 51(2):343–413.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M. Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. Eurollm-9b: Technical report.

Domenico Meconi, Simone Stirpe, Federico Martelli, Leonardo Lavalle, and Roberto Navigli. 2025. Do large language models understand word senses? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33897–33916, Suzhou, China. Association for Computational Linguistics.

Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems*, 37:78104–78146.

Millicent Ochieng, Varun Gumma, Sunayana Sitaram, Jindong Wang, Vishrav Chaudhary, Keshet Ronen, Kalika Bali, and Jacki O'Neill. 2024. Beyond metrics: evaluating llms' effectiveness in culturally nuanced, low-resource real-world scenarios. *arXiv preprint arXiv:2406.00343*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz

Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Riccardo Orlando, Pere-Lluís Huguet Cabot, Edoardo Barba, and Roberto Navigli. 2024a. 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.

Riccardo Orlando, Luca Moroni, Pere-Lluís Huguet Cabot, Simone Conia, Edoardo Barba, Sergio Orlandini, Giuseppe Fiameni, and Roberto Navigli. 2024b. Minerva LLMs: The first family of large language models trained from scratch on Italian data. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 707–719, Pisa, Italy. CEUR Workshop Proceedings.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

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.

Matiss Rikters and Makoto Miwa. 2024. Entity-aware multi-task training helps rare word machine translation. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 47–54, Tokyo, Japan. Association for Computational Linguistics.

Nikit Srivastava, Aleksandr Perevalov, Denis Kuchelev, Diego Moussallem, Axel-Cyrille Ngonga Ngomo, and Andreas Both. 2023. Lingua franca–entity-aware machine translation approach for question answering over knowledge graphs. In *Proceedings of the 12th Knowledge Capture Conference 2023*, pages 122–130.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon

Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao,

Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Roopali Vij, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Antonio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie

Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Alek Andreev, Antoine He, Kevin Hui, Sheleem Kashem, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2025a. Gemini: A family of highly capable multimodal models.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan

Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. 2025b. Gemma 3 technical report.

Jiaan Wang, Fandong Meng, Yingxue Zhang, and Jie Zhou. 2024. Retrieval-augmented machine translation with unstructured knowledge. *arXiv preprint arXiv:2412.04342*.

Shufang Xie, Yingce Xia, Lijun Wu, Yiqing Huang, Yang Fan, and Tao Qin. 2022. End-to-end entity-aware neural machine translation. *Machine Learning*, 111(3):1181–1203.

Lu Xu. 2025. Wikidata-driven entity-aware translation: Boosting LLMs with external knowledge. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 1802–1809, Vienna, Austria. Association for Computational Linguistics.

Binwei Yao, Ming Jiang, Tara Bobinac, Diyi Yang, and Junjie Hu. 2024. Benchmarking machine translation with cultural awareness. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13078–13096.

Min Zhang, Limin Liu, Zhao Yanqing, Xiaosong Qiao, Su Chang, Xiaofeng Zhao, Junhao Zhu, Ming Zhu, Song Peng, Yinglu Li, et al. 2023. Leveraging multilingual knowledge graph to boost domain-specific entity translation of chatgpt. In *Proceedings of Machine Translation Summit XIX, Vol. 2: Users Track*, pages 77–87.

12. Appendices

The following table reports the average number of entities for each entity type across two language groups. The first group corresponds to high-resource languages (French, German, Italian, Japanese, Chinese, and Spanish), while the second group represents mid-resource languages (Korean, Thai, Turkish, and Arabic). Table 9 presents the average number of unique entities per type, highlighting differences in data coverage between the two groups. Table 10 then provides representative examples for each entity type.

Type	High-Resource	Mid-Resource
Animal	0.2	–
Artwork	44.0	61.0
Book	30.3	35.5
Book series	2.7	2.0
Fictional entity	26.2	17.5
Food	17.8	10.5
Landmark	22.5	11.8
Movie	28.0	36.3
Musical work	24.7	13.0
Natural place	1.0	–
Person	33.0	40.0
Place of worship	33.3	29.0
Plant	0.7	1.0
TV series	28.0	30.5

Table 9: Average entity count per entity type across language groups.

Type	Example
Animal	Sergey Lavrov
Artwork	The Gift of the Magi
Book	Pentaglot Dictionary
Book series	Tsubasa to Hotaru
Fictional entity	Noah's Ark
Food	Sichuan pepper
Landmark	St. Cecilia Cathedral
Movie	Night of the Kings
Musical work	Flight of the Bumblebee
Natural place	Sokolniki Park
Person	Jia Yingchun
Place of worship	Solovetsky Monastery
Plant	Welsh onion
TV series	Love Like The Galaxy

Table 10: Entity types in the dataset with representative examples, sampled across multiple language subsets.