OWL: Probing Cross-Lingual Recall of Memorized Texts via World Literature

Alisha Srivastava[★,1] **Emir Kaan Korukluoglu**[★,1] **Minh Nhat Le**[★,1] **Duyen Tran**[1]
Chau Minh Pham[2] **Marzena Karpinska**[3] **Mohit Iyyer**[2]

[1]University of Massachusetts Amherst [2]University of Maryland, College Park
[3]Microsoft

{alishasrivas, ekorukluoglu, nhatminhle, duyent}@umass.edu
{chau[✉], miyyer}@umd.edu, mkarpinska@microsoft.com[✉]

Abstract

Large language models (LLMs) are known to memorize and recall English text from their pretraining data. However, the extent to which this ability generalizes to other languages or transfers across languages remains unclear. This paper investigates multilingual and cross-lingual memorization in LLMs, probing whether memorized content in one language (e.g., English) can be recalled when presented in a different language. To do so, we introduce OWL, a dataset of **31.5K** aligned excerpts from 20 books in ten languages, including original English texts, official translations (Vietnamese, Spanish, Turkish), and new translations in six low-resource languages (Sesotho, Yoruba, Maithili, Malagasy, Setswana, Tahitian). We evaluate memorization across model families and sizes through three tasks: (1) *direct probing*, which asks the model to identify a book's title and author; (2) *name cloze*, which requires predicting masked character names; and (3) *prefix probing*, which involves generating continuations. We find that some LLMs consistently recall content across languages, even for texts without existing translation. GPT-4o, for example, identifies authors and titles 69.4% of the time and masked entities 6.3% of the time in newly translated excerpts. While perturbations (e.g., masking characters, shuffling words) reduce accuracy, the model's performance remains above chance level. Our results highlight the extent of cross-lingual memorization and provide insights on the differences between the models.

https://github.com/emirkaan5/OWL

1 Introduction

Large language models (LLMs) encode substantial factual and linguistic knowledge from their training corpora, which they can later access to respond to user queries (Petroni et al., 2019; Kassner et al., 2021). Prior work investigating how LLMs acquire and recall this information has primarily focused on English texts (Carlini et al., 2021b, 2022; Golchin and Surdeanu, 2024; Huang et al., 2024; Shi et al., 2024; Ravichander et al., 2025). Hence, it remains unclear how much content LLMs memorize in languages other than English, and whether such knowledge can be reliably accessed in a language different from the one in which it was originally learned. While Goldman et al. (2025) investigate cross-lingual knowledge transfer, their methodology assumes that content is unseen in a target language if its Wikipedia article is missing. This assumption is potentially problematic, as the same information may exist in other online sources within the pretraining data.

To address these limitations and investigate multilingual memorization and cross-lingual knowledge recall, we introduce OWL, a new dataset comprising **31,540** aligned literary passages from **20** English books. OWL is unique in that it includes not only official human translations in Spanish, Turkish, and Vietnamese, but also *newly produced* machine translations into six low-resource languages (Sesotho, Yoruba, Maithili, Malagasy, Setswana, and Tahitian) for which no published translations of these works previously existed.

Leveraging OWL, we extend the probing methodology of prior work and employ three probing tasks: (1) **direct probing** (Karamolegkou et al., 2023), where the LLM identifies a book's title and author from a passage; (2) **name cloze task** (Chang et al., 2023), where it fills in a masked character name; and (3) **prefix probing** (Karamolegkou et al., 2023; Carlini et al., 2023), where it continues a given passage. These probing tasks allow us to investigate three research questions:

First, we examine the memorization of official translations. By comparing LLM performance

[★]These authors contributed equally to this work.
[✉]Corresponding authors

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25868–25895
November 4-9, 2025 ©2025 Association for Computational Linguistics

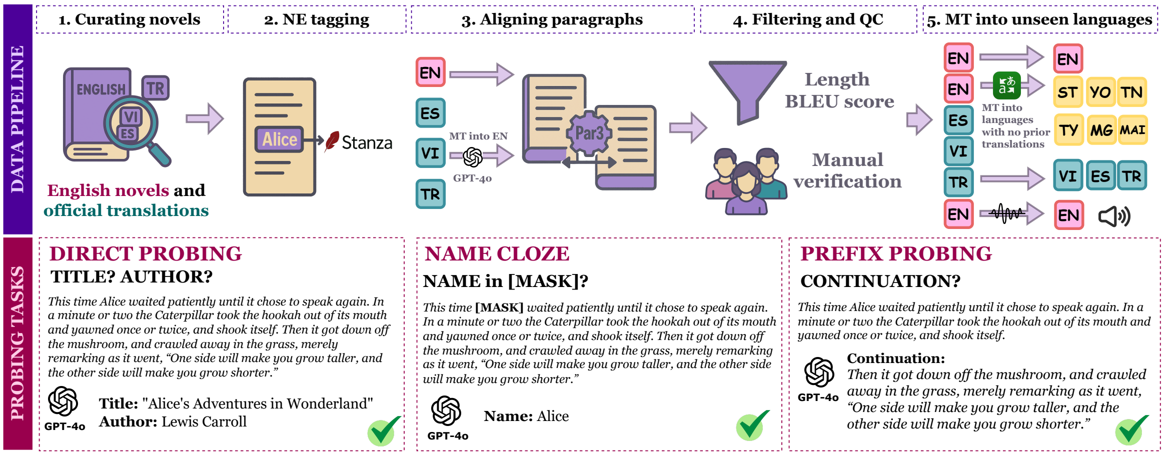

Figure 1: **Top:** OWL collection pipeline: (1) Identify English novels with official Turkish, Spanish, and Vietnamese translations; (2) Tag passages with named characters; (3) Align translations to English originals using Par3 aligner (Thai et al., 2022); (4) Filter alignments based on length and BLEU scores, followed by manual verification; (5) Translate validated English passages into six new languages without official translations. **Bottom:** Probing tasks: (1) Direct Probing (DP) – identify author/title from a passage; (2) Name Cloze (NC) – predict masked names in passages; (3) Prefix Probing (PP) – generate continuations from passage prefixes. Prompt texts omitted for clarity (see Figure 11, Figure 12, Figure 13). The figure shows outputs from GPT-4o. See Table 3 for an overview of our experiments.

on original English texts (e.g., *Alice in Wonderland*) against their published human translations, we find that while memorization is present across languages, it is more prominent in English. For instance, in direct probing LLMs achieve 63.8% averaged accuracy for English excerpts versus 47.2% for Spanish, Turkish, and Vietnamese examples. This multilingual memorization persists even when contextual coherence is disrupted by shuffling words in the passage.

Second, we quantify cross-lingual memorization using our newly produced translations. Since these translations are novel and the original works lack published versions in these six low-resource languages, strong performance on probing tasks could indicate a high degree of cross-lingual knowledge transfer from English or other high-resource languages.[1] Notably, we observe that models recall information even for the newly translated texts. GPT-4o, for instance, correctly identifies author and book title 69.4% of the time and guesses masked entities with 6.3% accuracy, suggesting that LLMs can, to some extent, access memorized knowledge across languages, even without direct exposure to these specific translations during pretraining (Yao et al., 2024; Goldman et al., 2025).[2]

Third, we explore the robustness of memorization in cross-modal and quantized settings. Our findings reveal that LLMs can recall memorized content even when prompted via different modalities, such as audio (GPT-4o-Audio achieves up to 75.5% accuracy in direct probing; Qwen-Omni reaches 20.6%). Furthermore, model quantization impacts performance; for instance, LLaMA-3.1-70B shows up to a 25% drop in accuracy with 8-bit quantization, a more substantial decrease than with 4-bit quantization, which contrasts with some previous findings (Marchisio et al., 2024; Kurtic et al., 2025).

Contributions: We introduce OWL, a dataset featuring **31,540** aligned book excerpts across **10** languages. Using this dataset, we conduct three probing experiments to assess the extent of memorization by LLMs in English versus other languages, and to investigate how this memorized knowledge transfers across language boundaries. We are releasing our data and codebase to spur future research on multilingual memorization in LLMs.

2 Constructing OWL

We design OWL as a testbed for memorization as well as cross-lingual knowledge transfer in LLMs. The dataset has three main components: (1) excerpts from novels originally written in English

[1] We exclude prefix probing from this experiment as it is unclear what the gold continuation would be.

[2] Although some models may be trained on machine translations, we see the same trend with OLMo, whose training data is public and can be inspected.

Passage Type	Perturbation	Experiment	English Example
w/ character	Standard	DP + PP	"Of course if Tom was home he'd put it right in a moment,"
	Masked	DP + NC	"Of course if [MASK] was home he'd put it right in a moment,"
	Shuffled	DP	"in he'd home Tom a if was of put it moment right course,"
	Masked + Shuffled	DP + NC	"in he'd home [MASK] a if was of put it moment right course,"
w/o character	Standard	DP	"No. Don't come up to me until you see me among a lot of people..."
	Shuffled	DP	"Just me a you see at don't me. of me." people. Don't keep up..."

Table 1: Examples of perturbations used in the ablation experiments. **Experiment** indicates the evaluation setup the task appears in: DP = Direct Probe, PP = Prefix Probe, NC = Name Cloze. **English Example** shows a representative passage for each condition.

	Original						No Named Characters					
Group	**Count**	**Mean**	**Median**	**Min**	**Max**	**Stdev**	**Count**	**Mean**	**Median**	**Min**	**Max**	**Stdev**
English	1594	64.90	49.0	18	429	47.75	1560	59.03	46.0	18	325	40.08
Translations	4782	63.17	48.0	10	523	49.83	4680	57.67	45.0	10	430	43.01
Cross-lingual	9564	78.91	60.0	11	642	59.98	9360	71.73	56.0	9	507	50.56

Table 2: Token distribution in each passage type, calculated with OpenAI's tiktoken library (o200k_base).

(*en*), (2) their official translations into Spanish (*es*), Turkish (*tr*), and Vietnamese (*vi*), and (3) new machine translations into six low-resource languages, specifically Sesotho (*st*), Yoruba (*yo*), Setswana (*tn*), Tahitian (*ty*), Maithili (*mai*), and Malagasy (*mg*), for which official translations are not available. Additionally, we augment the data with audio files of the English excerpts to explore how models perform across modalities (text vs. audio). Overall, we collect 3,154 English passages (1,594 passages with and 1,560 passages without named characters). Each passage is then aligned with its semantic equivalents in nine other languages and English audio, yielding a total of **31,540 text passages** and **7,950 audio excerpts** across the dataset. We construct the dataset in six main steps (Figure 1), as listed below:

1. Curating books We collect English novels that are also officially translated into Spanish, Turkish, and Vietnamese.[3] We source public-domain books from Project Gutenberg (Stroube, 2003) and purchase copyrighted texts online. Overall, we collect **20 books**, with 10 public-domain and 10 copyrighted books (see Table 7).

2. Tagging named characters Since the name cloze task (§3.2) requires test samples to have at least one character name, we tag these names by applying Stanza (Qi et al., 2020) to each sentence in the collected books.

[3]We selected these languages because they represent distinct morphological and syntactic typologies: Spanish is fusional, Turkish agglutinative, and Vietnamese analytic.

3. Aligning multilingual paragraphs To ensure fair comparison across languages, we align English passages to their official translations in Spanish, Vietnamese, and Turkish by translating non-English books into English using GPT-4o[4] and applying the Par3 aligner (Thai et al., 2022).

4. Filtering & quality control To filter out any misaligned passages, we apply a length filter[5] and BLEU filter using SacreBLEU (Post, 2018) with add-one smoothing.[6] Finally, we manually verify all alignments, removing misaligned passages or those with more than one unique character name (Figure 10). We compile two sets of passages: (1) a set containing exactly one unique character name[7] that is used for all our tasks, and (2) a set of comparable size that does not have any character name for the direct probing and prefix probing task (§3.1). The two sets have similar average lengths: 64.90 tokens for passages with a character name and 59.03 for those without (Table 2).[8] For each set of passages, we sample at most 100 passages per book to include in the final dataset.[9]

To balance the distribution of character mentions,

[4]We use gpt-4o-2024-05-13 with temperature=0.3 and max_tokens=4000; refer to Figure 14 for details.

[5]We use an asymmetric length filter: drop if the English passage is >3x any non-English by characters ($|x_{\text{en}}| > 3|x_\ell|$).

[6]We filter out any alignment that does not meet the threshold of 5.0 BLEU score, following Thai et al. (2022)

[7]This is our main experimental set; for these passages, we allow multiple mentions of the same character's name.

[8]Unless otherwise mentioned, "tokens" refer to those calculated with tiktoken library (o200k_base)

[9]We sample passages with at least 40 BPE tokens. View word count distribution in Figure 9.

we apply stratified sampling to all passages containing a character name. The final dataset for each language consists of 3,154 passages: 1,594 with character names and 1,560 without.[10]

5. Machine translation into new languages To explore cross-lingual knowledge transfer, we select six languages with *no prior translations* of the books in our dataset to ensure that they have not been encountered during the training: Sesotho (*st*), Yoruba (*yo*), Setswana (*tn*), Tahitian (*ty*), Maithili (*mai*), and Malagasy (*mg*).[11] We use Microsoft Translator[12] to translate passages from English into each of the unseen languages.[13] We will be referring to this subset of data as *unseen translations*.

6. Creating audio data To evaluate cross-modal knowledge transfer, we convert passages containing character names into high-fidelity, lossless audio waveforms using Kokoro-82M (Hexgrad, 2025), a neural text-to-speech (TTS) model chosen for its low-distortion rendering of prosody and phonetics. The resulting audio corpus preserves the linguistic content of each passage while enabling direct comparison between text and speech-based representations.[14] We convert the entire passage into audio for all tasks. For prefix probing (§3.3), we convert only the first half of the passages.

Why literary data? We select literary work as it is likely present in pretraining corpora. All our titles are available on LibGen (allegedly used to train LLaMA models),[15] and our non-copyrighted books are on Project Gutenberg (used for OLMo training (OLMo et al., 2024)). Furthermore, literary data is rich in the character names necessary for the name-cloze task.

[10]This addresses a bias from Chang et al. (2023), where overrepresentation of common names (e.g., Alice) likely inflated model accuracy.

[11]To confirm no existing translations, we search Google, Amazon Books, OpenLibrary, and Goodreads for each book in the target language and find none.

[12]https://www.microsoft.com/en-us/translator/. We used Microsoft Translator instead of LLMs to avoid potential bias and because LLMs generally underperform traditional machine translation on low-resource languages due to data limitations (Robinson et al., 2023).

[13]We recognize that Microsoft Translator may not produce perfect translations; therefore, the results presented in this paper represent a lower bound of the cross-lingual performance.

[14]Kokoro-82B currently ranks as the top-performing TTS model on TTS Spaces Arena (mrfakename et al., 2025). A manual review of 50 samples revealed no errors.

[15]See legal brief.

3 Experiments

We propose three probing experiments on OWL to assess memorization as well as cross-lingual knowledge transfer: (1) direct probing (DP, §3.1), where the model identifies the book's title and author; (2) name cloze (NC, §3.2), where the model fills in a masked character name; and (3) prefix probing (PP, §3.3), where the model generates a continuation from a given prefix. We further extend these experiments to test the effect of quantization on memorization (§3.4), and probe cross-modal recall using audio input (§3.5). These tasks reflect different degrees of memorization, ranging from basic retrieval of learned information (direct probing) to precise reconstruction of acquired content (prefix probing).

Test data Unless specifically mentioned, we run all experiments on the following data containing one unique character name: **(1) original English data** (to establish model recall of data, which was likely seen during pretraining), **(2) official translations** (to provide a baseline for model's performance on high-resource languages other than English, which could be encountered during pretraing), **(3) unseen translations** to measure cross-lingual knowledge transfer, and **(4) English audio data** (to compare performance on audio and textual content). We also include an additional experiment on newly published books to estimate performance by chance.

3.1 Experiment 1: Direct Probing

Task: In direct probing, the model identifies the `title` and `author` of a book passage (Karamolegkou et al., 2023). This task reflects more passive knowledge, as it primarily tests the model's ability to recognize and link textual and audio cues to learned metadata rather than requiring the model to recall the exact wording of the passage (see Figure 11 for prompt). In the cross-modal setup, we provide the audio of the passage.

Metric: We measure accuracy by comparing predicted `(author, title)` pairs against ground truth, allowing for minor formatting or diacritic differences.[16] A prediction is considered correct if the model identifies the correct author and book title (either in English or the passage's language). For

[16]We normalize special characters and apply fuzzy match with a Levenshtein similarity threshold (0.9 for DP and 0.7 for NC, which we establish by analyzing a subset of our data).

cross-lingual experiments, we prompt the model to respond in English.

Ablations: To measure the robustness of model performance, we introduce three additional variations on the task (see Table 1):

Shuffled passages: To pinpoint the role of word order and syntax in knowledge recall, we randomly shuffle the words within each passage. This shuffle disrupts the syntactic and semantic coherence of the text while preserving its lexical content, allowing us to test whether the recall depends on the sequential structure of the input.

Masked passages: For consistency across tasks, we use the same passages as in the name cloze task (§3.2), each containing a single character name. Here, we replace that name with `[MASK]` to determine how much it contributes to the recall, albeit at the cost of disrupting the original text.

No character names: We also include a separate set of passages that naturally contain no character names and thus remain intact. To facilitate a fair comparison with masked passages, we ensure that both sets have similar length distributions.

3.2 Experiment 2: Name Cloze

Task: In the name cloze task, we reuse the same passages from §3.1, each containing exactly one character name, and replace that name with `[MASK]` token to test recall (Chang et al., 2023).[17] Strong performance on this task likely indicates memorization of that passage, especially since character names tend to be high-surprisal tokens (Ravichander et al., 2025). In the cross-modal setup, we provide the English audio of the passage.

Metric: We evaluate task accuracy using exact match.[18] Ground-truth named characters are extracted directly from the original passages, and a prediction is correct only if it matches the normalized ground truth (either in English or in the language of the passage). For cross-lingual experiments, we prompt the model to respond in English.

Ablation: We further test the robustness of models by shuffling the words within each passage, as previously done in §3.1, to understand the effect of sequential token order and syntax. Specifically, we want to understand whether the model performance depends on the token sequence and/or the position of the `[MASK]` token.

3.3 Experiment 3: Prefix Probing

Task: The prefix probing task evaluates whether a model, when given the first half (prefix) of a passage, can reproduce the second half (continuation) (Carlini et al., 2021b). This setup draws on the fact that accurate predictions are unlikely without prior exposure to the full passage during pretraining. In the cross-modal setup, we provide the English audio of the first half of the passage.

Metric: To measure the model's ability to replicate a passage's continuation, we report ChrF++ (Popović, 2015), which assesses lexical and semantic similarity between the model's output and the ground-truth continuation.

3.4 Quantization ablation

To assess potential information loss due to reduced parameter precision from quantization, we replicate all experiments and ablations on LLaMA models using GPTQ-int4 (W4A16) and GPTQ-int8 (W8A16) methods (Frantar et al., 2023), where WxAy denotes the level of quantization for weights (W) and activations (A).

3.5 Audio ablation

To compare performance on audio versus text, we extend our analysis to audio content, adapting three core experiments: direct probing, name cloze task, and prefix probing. Text-specific ablations were excluded. Due to superior text-to-speech (TTS) model quality, all audio experiments were limited to English, with models receiving textual instructions and providing textual responses.

3.6 Models

For all tasks, we test a diverse set of open-weight and closed-source models, including Qwen2.5-1M (Yang et al., 2025; Xu et al., 2025), LLaMA-3.1-8B, 70B, 405B and LLaMA-3.3-70B (Meta, 2024), OLMo-2-7B and 2-13-B (OLMo et al., 2024), EuroLLM (Martins et al., 2025), as well as GPT-4o (OpenAI, 2024).[19] For audio experiments, we

[17]Unlike Chang et al. (2023), we do not restrict passages to a single occurrence of the character name or limit the passage length to allow for more realistic text usage and analysis of passage-length effects.

[18]Exact match is applied after normalizing both predicted and ground-truth names with the Unidecode library to remove formatting and diacritic variations.

[19]We use vLLM (Kwon et al., 2023) for inference from open-weights models, with the exception of LlaMA-3.1-405B-instruct, which is run using OpenRouter API due to its size.

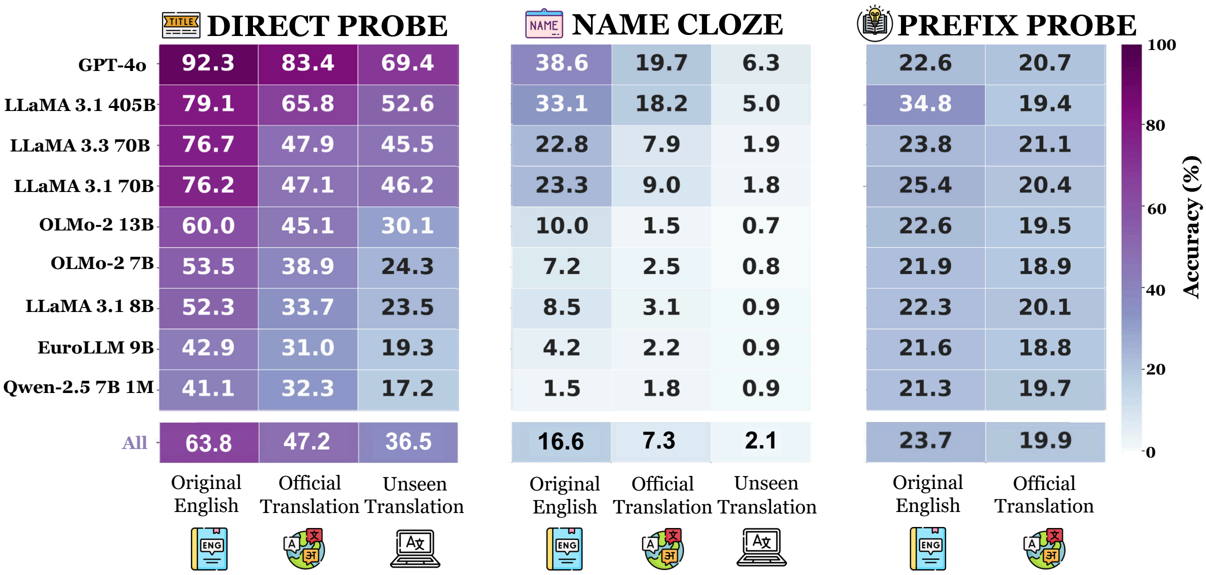

Figure 2: **Overall performance:** GPT-4o consistently outperforms other models in probing tasks, followed by LLaMA 405B. Direct Probing (DP; reported for passages with character names) and Prefix Probing (PP) use unmasked passages, while the Name Cloze Task (NCT) uses masked ones with named characters removed. PP performance is measured with ChrF++. PP performance on unseen languages is not reported as it is unclear what the gold continuation should be.

use GPT-4o-audio and Qwen2.5-Omni-7B (Xu et al., 2025). In addition to full-precision models, we also run our experiments on the quantized versions of LLama-3.1-70B-Instruct and Llama-3.1-8B-Instruct.[20] See Table 5 for details.

4 Results

In this section, we present the results of our experiments. Overall, our results show that LLMs can, to varying degrees, recognize (and in some cases reproduce) book content when presented in different forms, such as the original English text, official translations, new machine translations, and even audio. While perturbations (e.g., shuffling) do reduce accuracy, the resulting performance is still above random. Finally, the presence of a character's name proves to be a strong signal that facilitates recall.

LLMs can recognize official translations Models can recognize passages from English novels achieving 63.8% accuracy on average, with GPT-4o reaching 92.3% (Figure 2). Although this accuracy drops for official translations, it remains above random at 47.2% on average (83.4% for GPT-4o). This recall also extends to more challenging tasks such as name cloze, albeit with reduced accuracy (e.g., GPT-4o scores 38.6% for English versus 19.7% for translations; see Table 18 for common errors). Notably, performance scales with model size. In the name cloze task for English texts, accuracy rises from 8.5% with LLaMA-3.1-8B to 33.1% with LLaMA-3.1-405B. These results indicate memorization, particularly in comparison with the performance on 2024 books (Table 4), where the accuracy is close to zero, likely because the content was not seen during training. Finally, prefix probing results suggest the models struggle with verbatim recall, as their chrF++ scores are only marginally higher than those for the 2024 books.

Cross-lingual access to memorized knowledge Having established that models can recognize English excerpts and their official translations, we next test whether they could also recognize newly produced machine translations in six low-resource languages. Although the overall accuracy drops, the models can still identify the books (36.5% average on the direct probe; Figure 2) and, to a lesser extent, recall a masked character's name (2.1% average on name cloze). This performance varies by language and model (Figure 18). For instance, on the direct probe for Sesotho, GPT-4o achieves 76.9% accuracy, while Qwen-2.5-7B-1M scores over 18%. Even for Maithili, the lowest-performing language, GPT-4o still achieves 66.5%

For all models, we set the temperature to 0 and max_tokens to 100.

[20] Quantized models are obtained from NeuralMagic.

accuracy, with LLaMA-3.1-405B close behind at 46.7%.

The name cloze results are much lower but still above zero, with the highest score being 10.5% on Maithili by GPT-4o.[21] Interestingly, even OLMo shows a non-zero performance, despite being reportedly trained only on English data (OLMo et al., 2024), with its highest score being 44.1% on the Yoruba direct probe.[22] This all suggests that some meaningful amount of cross-lingual transfer can happen even when the target languages are under-represented in the pre-training data.[23] [24]

LLMs can recall knowledge when probed in a different modality Both Qwen-Omni and GPT-4o-Audio show some ability to recognize the book when prompted with an audio excerpt (Figure 3, Figure 17). Specifically, GPT-4o-Audio achieves up to 75.5% accuracy on the direct probing task, while Qwen-Omni reaches 20.6% on the same task. Although overall performance is lower on the audio version of the name cloze task, GPT-4o-Audio still reaches up to 15.9%. In contrast, Qwen-Omni struggles with this task, scoring only 0.8%. These findings suggest that LLMs could potentially recall information across modalities.[25]

Shuffling inputs partially reduces direct probing and name cloze accuracy Figure 4 shows that shuffling the input texts, which represents minor perturbations such as phrase reordering or lexical edits, causes a noticeable, but not drastic, drop in direct probing accuracy. Specifically, direct probing performance declines most on English passages (up to 16.3% for "no character" setup), in contrast to 3-8% for both translation categories across all excerpt

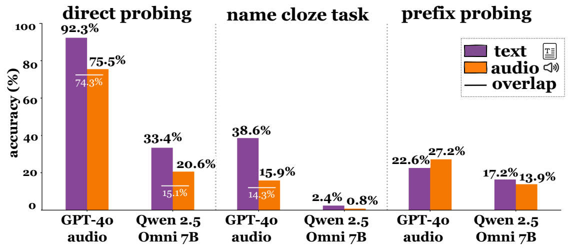

Figure 3: **Audio vs. Text** accuracy on English passages with a character name. GPT-4o-audio exhibits substantial performance across all tasks and modalities. The overlap line denotes the percentage of passages answered correctly in both modalities.

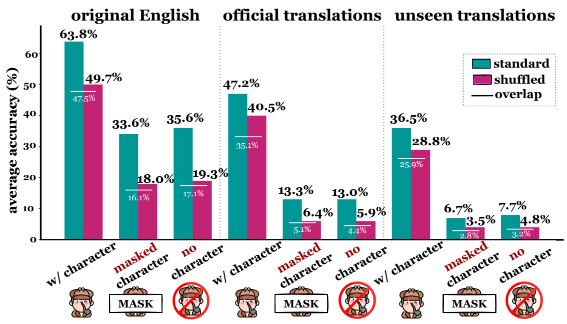

Figure 4: **Direct probing:** Average accuracy across models for shuffled versus standard text inputs. Accuracy decreases from standard to shuffled inputs across all perturbations and language settings, with non-trivial shuffled accuracy on English and official translations.

types; likely because the accuracies on these are already lower. Similarly, in the name cloze (Figure 5) the gap between standard and shuffled performance can be as low as 1.1% for unseen translation and as high as 11.7% for English texts. These moderate drops indicate that superficial rewordings do have an effect, though only a modest one.

Direct probing consistently outperforms name cloze-style queries Direct probing outperforms name cloze queries across all models and languages (Figure 2). For example, GPT-4o achieves 92.3% accuracy on original English texts with direct probing, compared to only 38.6% with name cloze. LLaMA 3.1 70B shows a similar gap (76% vs. 22.8%), as does EuroLLM 9B (38.7% gap). This pattern holds in translations: GPT-4o scores 83.4% (direct) vs. 19.7% (cloze) on official translations, and 69.4% vs. 6.3% on unseen ones. The large performance gap reflects the difficulty of name cloze tasks, which likely conflict with the autoregressive nature of language models. In contrast, direct prob-

[21] This performance varies significantly by book, for example, GPT-4o achieves an average accuracy of 33.3% for "Alice's Adventures in Wonderland" and 19.2% for "1984" when tested on unseen translations.

[22] Note that while OLMo's authors specifically filtered their training data for English, it's very likely that some non-English text remained.

[23] Recent court documents show that LlaMA models might have been trained on LibGen book data, which implies the possibility that other frontier models are also trained on this data (100% of our English and official translation books can be found in LibGen). Under these conditions, it is reasonable that these models have boosted cross-lingual performance without explicit translation supervision.

[24] It is worth noting that the character's name is a strong signal, a point we discuss later in more detail. Removing the name from a passage lowers accuracy on the direct probing task, although performance doesn't fall to zero (Figure 23).

[25] It is possible the models were trained on official audiobooks. However, our audio files were created using a text-to-speech model and thus differ significantly in their acoustic properties from human narrations.

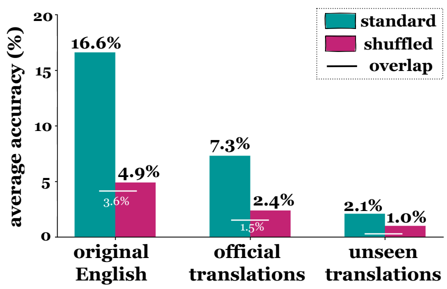

Figure 5: **Name cloze:** Unshuffled inputs outperform shuffled inputs across all language settings, with non-trivial accuracy on English and official translations.

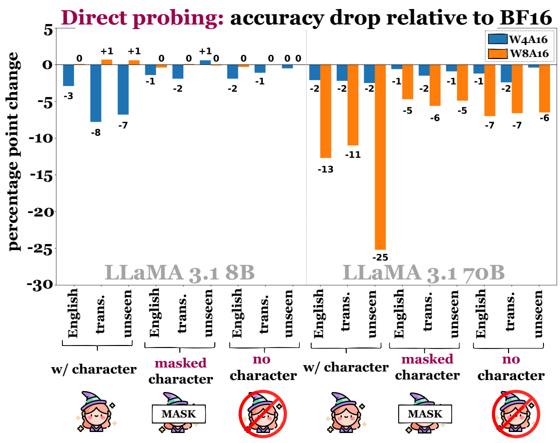

Figure 6: **Direct probing:** Percentage point drop in performance with respect to the performance of the BF16 baseline. We report drops for original English text ("English"), their official translations ("trans"), and unseen translations ("unseen"). The scores are reported across three conditions: (1) on passages containing a character name, (2) on passages where the name was masked, and (3) on passages without character name.

ing, where the model has to recall the title or author in a question-answering format, is more aligned with LLMs' strength.

Character names facilitate recall Figure 4 shows that models are noticeably better at the direct probing task, when the passage contains a character name: 63.8% for English, 47.2% for official translations, and 36.5% for unseen ones. Masking the name sharply reduces accuracy to 33.6%, 13.3%, and 6.7%, respectively. Accuracy under masking is similar to passages without named characters, especially in translations ($\leq$3% difference). The absence of named characters results in lower performance, suggesting that models often depend on lexical cues like names and locations to recognize the passages.

LLaMA-3.1-70B's performance degrades more under 8-bit than under 4-bit quantization We test the performance of both LLaMA-3.1-70B and 8B under 4-bit and 8-bit quantization and compare it to the performance of these models in BF16 precision.[26] While LLaMA-3.1-70B maintains relatively stable accuracy at 4-bit precision, it experiences notable performance drops when quantized to 8 bits. Specifically, we observe up to a 25% decrease in direct probing on unseen translation settings (Figure 6), along with smaller declines of 5.8% in the English name cloze task and 1.4% in prefix probing when the passage includes character names (Figure 20). In contrast, the smaller LLaMA-3.1-8B behaves more predictably: its performance remains within 1% of the BF16 baseline at 8-bit precision across both direct probing and name cloze tasks (Figure 19), with noticeable degradation appearing only under 4-bit quantization. These results surprisingly contradict findings in Kurtic et al. (2025) and Marchisio et al. (2024), who report a marginal drop for GPTQ-int8 but larger drops for GPTQ-int4.

OLMo's performance on passages seen during training We identified a subset of OWL for which the English passages are present in OLMo's training data, and we made sure that none of the corresponding translations were.[27] Even though the OLMo models had definitely seen these English passages during training, their performance on them is moderate ($\sim$62.9%). However, the accuracy does not drop drastically on the direct probing task for both official and machine translations of those same passages (Figure 7, Figure 8). Unsurprisingly, the performance on the name cloze task is much lower, though it still likely remains above random.

Analysis of common errors For direct probing, models occasionally name correct authors but misidentify book titles around 10.61%[28] of the time. More often, they return another popular book (61%) or abstain (39%). Abstention rate (responses like "unknown," "none," or empty strings) is no-

[26] We report the performance drop as the difference in percentage points between the BF16 version and quantized models.

[27] Passages were identified by querying the Infinigram API (Liu et al., 2024) (v4_olmo-2-1124-13b-instruct_llama). See Table 13 for details.

[28] A common error pattern involves models correctly attributing authorship to J.K. Rowling but specifying an incorrect book title from within the Harry Potter series.

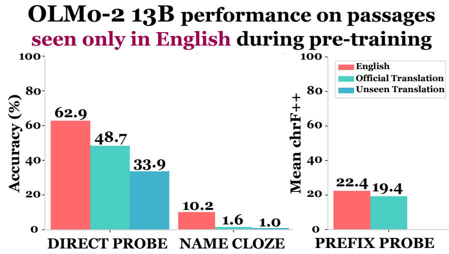

Figure 7: Accuracy of OLMo-2-13B on *seen* passages identified in the training data in English but not in their translated versions. The model's accuracy on direct probing is considerable compared to its performance on name cloze and prefix probing.

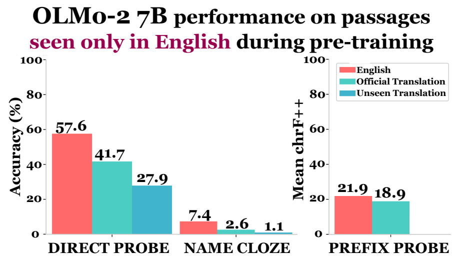

Figure 8: Accuracy of OLMo-2-7B on passages identified in the training data in English but not in their translated versions. Accuracy across tasks and languages is slightly lower than that of the 13B model.

tably high for EuroLLM (30.39%). The main error for name cloze task is returning an incorrect character name (93%), which is sometimes culturally relevant to the passage's language (e.g., Spanish names for Spanish text) or other characters from the same book (Table 9). Models also return pronouns (2%), honorifics (3%), abstain (0.1%), or repeat the "[MASK]" token (0.7%). Across all tasks, Qwen models frequently generate "broken text" (a hodgepodge of languages) for 15.81% of outputs. See Table 18 for more examples.

5 Related Work

Memorization in LLMs LLMs exhibit substantial memorization capabilities (Elangovan et al., 2021; Carlini et al., 2018; Hartmann et al., 2023; Carlini et al., 2023). Prior studies quantify memorization through verbatim recall (Carlini et al., 2021b, 2023; Lee et al., 2022), passage origin identification (Chang et al., 2023; Magar and Schwartz, 2022), improbable token prediction (Lee et al., 2022; Radhakrishnan et al., 2019), and membership inference attacks (Carlini et al., 2021a; Golchin and Surdeanu, 2024; Song and Shmatikov, 2019; Shokri et al., 2017; Asai et al., 2020; Stoehr et al., 2024). The amount memorized depends on factors such as model scale, generation length, data point frequency, and context window size (Carlini et al., 2023; Radhakrishnan et al., 2019; Biderman et al., 2023; Zhou et al., 2024; Chen et al., 2024; Razeghi et al., 2022; Lee et al., 2022; Kandpal et al., 2022; Carlini et al., 2023). Early probing experiments, which are largely monolingual and cloze-style (Tirumala et al., 2022; Chang et al., 2023), have since been complemented by theoretical work showing that rare memorized outliers can steer the model's learning trajectory (Allen-Zhu and Li, 2024). Prashanth et al. (2025) model memorization as three regimes – recitation, reconstruction, recollection – and fit a predictive model. Zhang et al. (2025) study memorization dynamics in Pythia and find that prefix perturbations reduce recall; our perturbations overlap in spirit, and we defer full replication to future work.

Cross-lingual knowledge transfer Cross-lingual knowledge transfer enables LLMs to recall information seen in one language when queried in another through shared multilingual representations (Asai et al., 2021; Jiang et al., 2020; Limkonchotiwat et al., 2022; Mittal et al., 2023; Huang et al., 2023). Research in both multimodal (Elliott et al., 2016; Baltrusaitis et al., 2019) and multilingual settings (Hessel and Lee, 2020) has shown that models can achieve high performance by exploiting shallow or dataset-specific cues. Our work is most relevant to Goldman et al. (2025), who measure cross-lingual transfer by analyzing the presence or absence of Wikipedia entries across languages and evaluating LLMs on this data.

6 Conclusion

In this study, we demonstrate that LLMs exhibit some degree of multilingual and cross-lingual memorization through probing experiments on aligned book excerpts across ten languages. We discover that character names are a strong signal for recalling the information. We also find that perturbations, such as word shuffling, prompting in audio format, and masking character names, noticeably but not drastically reduce performance. We release our data and code to spur further research on cross-lingual generalization and LLM memorization.

Limitations

Material scope We study memorization using best-selling books, which might not reflect the full diversity of copyrighted materials. Future work should explore additional underrepresented languages and lesser-known texts.

Popularity versus performance Models might have higher performance on excerpts that appear frequently in the pretraining data (Carlini et al., 2023). We leave the investigation of the relationship of the frequency of the item to the degree of memorization for the future work.

Legal implications While we empirically characterize memorization patterns, we do not make strong claims about the legal or ethical status of the outputs analyzed. The question of whether a model's output constitutes a copyright violation involves complex legal and normative considerations that go beyond the scope of this work. Future research should engage more deeply with the regulatory and ethical implications of LLM memorization, especially as legal frameworks evolve in response to advances in generative AI.

Translation quality Our analysis relies on translations generated using Microsoft Translator, which may introduce noise or artifacts that diverge from human translations. Imperfections in word choice, sentence structure, translation coverage, or named entity handling could affect the model's ability to recover factual content, especially in low-resource languages. Hence, we treat the model's performance on these translations as a lower bound for cross-lingual recall.

Training data and results interpretation Since we lack access to the pretraining data for most models, we cannot definitively verify if a passage was seen during training. Hence, we use a set of controls to interpret our results based on the data's likely exposure. We treat the original English passages and their official translations as plausibly "seen." This assumption is supported by evidence suggesting that LLaMA models were trained on large book corpora such as LibGen, where all our books are available. We also use newly published books (2024) to confirm a near-zero accuracy for all tasks. Then, we rely on newly created machine translations for languages where, to our knowledge, no public translation previously existed. While some models may have been trained on privately produced machine translations of these texts, the trends we observe are validated by similar results from OLMo, where we can validate the training data. Given this general opacity, our reported recall should be interpreted in the context of the model's performance on the English data and newly published books, not in isolation. Nevertheless, the lack of access to the training data limits the conclusions we can draw.

Ethical consideration

Our study explicitly evaluates whether LLMs recall specific passages from copyrighted books, using translated variants to test the boundaries of memorization across languages. While this analysis advances understanding of model behavior, it also raises ethical questions about the reproduction of copyrighted content by models trained on opaque corpora. We do not redistribute model outputs or original texts beyond short spans needed for evaluation,[29] but acknowledge that probing for memorization can implicate intellectual property rights. This underscores the need for transparency in training data sources and greater scrutiny of how multilingual capabilities may amplify copyright risks.

Acknowledgment

We thank members of UMass NLP and UMD CLIP lab for helpful feedback. This project was partially supported by awards IIS-2046248, IIS-2312949, and IIS-2202506 from the National Science Foundation (NSF).

References

Zeyuan Allen-Zhu and Yuanzhi Li. 2024. Physics of language models: part 3.1, knowledge storage and extraction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Akari Asai, Jungo Kasai, J. Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. In *North American Chapter of the Association for Computational Linguistics*.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021. XOR QA: Cross-lingual open-retrieval question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*,

[29]We use only a small fraction of copyrighted books for the dataset and release it for research purpose only.

pages 547–564, Online. Association for Computational Linguistics.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, A. Terzis, and Florian Tramèr. 2021a. Membership inference attacks from first principles. *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. 2022. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 1897–1914.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Xiaodong Song. 2018. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021b. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Kent Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, memory: An archaeology of books known to Chatgpt/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7312–7327.

Bowen Chen, Namgi Han, and Yusuke Miyao. 2024. A multi-perspective analysis of memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11190–11209.

Aparna Elangovan, Jiayuan He, and Karin Verspoor. 2021. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1325–1335, Online. Association for Computational Linguistics.

Desmond Elliott, Douwe Kiela, and Angeliki Lazaridou. 2016. Multimodal learning and reasoning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Berlin, Germany. Association for Computational Linguistics.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*.

Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinatan Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2025. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *ArXiv*, abs/2502.21228.

Valentin Hartmann, Anshuman Suri, Vincent Bindschaedler, David Evans, Shruti Tople, and Robert West. 2023. Sok: Memorization in general-purpose large language models. *Preprint*, arXiv:2310.18362.

Jack Hessel and Lillian Lee. 2020. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 861–877, Online. Association for Computational Linguistics.

Hexgrad. 2025. Kokoro-82m (revision d8b4fc7).

Jing Huang, Diyi Yang, and Christopher Potts. 2024. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10711–10732.

Zhiqi Huang, Puxuan Yu, and James Allan. 2023. Improving cross-lingual information retrieval on low-resource languages via optimal transport distillation. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 1048–1056, New York, NY, USA. Association for Computing Machinery.

Zhengbao Jiang, Antonios Anastasopoulos, Jun Araki, Haibo Ding, and Graham Neubig. 2020. X-FACTR: Multilingual factual knowledge retrieval from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5943–5959, Online. Association for Computational Linguistics.

Nikhil Kandpal, Eric Wallace, and Colin Raffel. 2022. Deduplicating training data mitigates privacy risks in language models. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 10697–10707. PMLR.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. 2023. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7403–7412, Singapore. Association for Computational Linguistics.

Nora Kassner, Philipp Dufter, and Hinrich Schütze. 2021. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3250–3258, Online. Association for Computational Linguistics.

Eldar Kurtic, Alexandre Marques, Shubhra Pandit, Mark Kurtz, and Dan Alistarh. 2025. "give me bf16 or give me death"? accuracy-performance trade-offs in llm quantization. *Preprint*, arXiv:2411.02355.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

Peerat Limkonchotiwat, Wuttikorn Ponwitayarat, Can Udomcharoenchaikit, Ekapol Chuangsuwanich, and Sarana Nutanong. 2022. CL-ReLKT: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2141–2155, Seattle, United States. Association for Computational Linguistics.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2024. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. *arXiv preprint arXiv:2401.17377*.

Inbal Magar and Roy Schwartz. 2022. Data contamination: From memorization to exploitation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 157–165.

Kelly Marchisio, Saurabh Dash, Hongyu Chen, Dennis Aumiller, Ahmet Üstün, Sara Hooker, and Sebastian Ruder. 2024. How does quantization affect multilingual LLMs? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15928–15947, Miami, Florida, USA. Association for Computational Linguistics.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. Eurollm: Multilingual language models for europe. In *Proceedings of the Second EuroHPC user day*, volume 255, pages 53–62.

Meta. 2024. The llama 3 herd of models. *ArXiv*, abs/2407.21783.

Shubham Mittal, Keshav Kolluru, Soumen Chakrabarti, and Mausam. 2023. mOKB6: A multilingual open knowledge base completion benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–214, Toronto, Canada. Association for Computational Linguistics.

mrfakename, Vaibhav Srivastav, Clémentine Fourrier, Lucain Pouget, Yoach Lacombe, main, Sanchit Gandhi, Apolinário Passos, and Pedro Cuenca. 2025. Tts arena 2.0: Benchmarking text-to-speech models in the wild. `https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2`.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, et al. 2024. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

USVSN Sai Prashanth, Alvin Deng, Kyle O'Brien, Jyothir S V, Mohammad Aflah Khan, Jaydeep Borkar, Christopher A. Choquette-Choo, Jacob Ray Fuehne, Stella Biderman, Tracy Ke, Katherine Lee, and Naomi Saphra. 2025. Recite, reconstruct, recollect: Memorization in LMs as a multifaceted phenomenon. In *The Thirteenth International Conference on Learning Representations*.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics.

Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. 2019. Memorization in overparameterized autoencoders. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*.

Abhilasha Ravichander, Shrusti Ghela, David Wadden, and Yejin Choi. 2025. Halogen: Fantastic llm hallucinations and where to find them. *Preprint*, arXiv:2501.08292.

Yasaman Razeghi, Robert L Logan IV, Matt Gardner, and Sameer Singh. 2022. Impact of pretraining term frequencies on few-shot numerical reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 840–854, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 3–18.

Congzheng Song and Vitaly Shmatikov. 2019. Auditing data provenance in text-generation models. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–206.

Niklas Stoehr, Mitchell Gordon, Chiyuan Zhang, and Owen Lewis. 2024. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*.

Bryan Stroube. 2003. Literary freedom: Project gutenberg. *XRDS: Crossroads, The ACM Magazine for Students*, 10(1):3–3.

Qwen Team. 2025. Qwen2.5-1m: Deploy your own qwen with context length up to 1m tokens.

Katherine Thai, Marzena Karpinska, Kalpesh Krishna, Bill Ray, Moira Inghilleri, John Wieting, and Mohit Iyyer. 2022. Exploring document-level literary machine translation with parallel paragraphs from world literature. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9882–9902, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Kushal Tirumala, Aram H. Markosyan, Luke Zettlemoyer, and Armen Aghajanyan. 2022. Memorization without overfitting: Analyzing the training dynamics of large language models. In *Advances in Neural Information Processing Systems*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoyan Huang, Jiandong Jiang, Jianhong Tu, Jianwei Zhang, Jingren Zhou, Junyang Lin, Kai Dang, Kexin Yang, Le Yu, Mei Li, Minmin Sun, Qin Zhu, Rui Men, Tao He, Weijia Xu, Wenbiao Yin, Wenyuan Yu, Xiafei Qiu, Xingzhang Ren, Xinlong Yang, Yong Li, Zhiying Xu, and Zipeng Zhang. 2025. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*.

Feng Yao, Yufan Zhuang, Zihao Sun, Sunan Xu, Animesh Kumar, and Jingbo Shang. 2024. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17864–17875, Miami, Florida, USA. Association for Computational Linguistics.

Jie Zhang, Qinghua Zhao, Lei Li, and Chi ho Lin. 2025. Extending memorization dynamics in pythia models from instance-level insights. *CoRR*, abs/2506.12321.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. 2024. Quantifying and analyzing entity-level memorization in large language models. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

A Data Collection

In this section of the appendix, we provide additional details on collecting data for OWL.

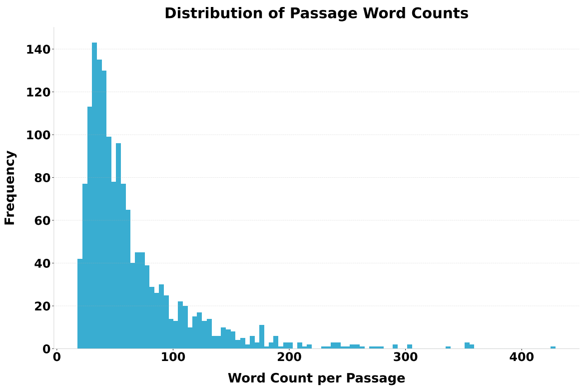

Figure 9: Word count distribution of unmasked passages in OWL

A.1 Extracting and aligning excepts from collected books

Our goal is to measure how well LLMs memorize data across different languages. For a fair and accurate assessment, the excerpts we use must contain identical content across languages. To achieve this, we use a seven-step approach to extracting and aligning excerpts from each collected book:

1. *Tagging sentences*: We first use Stanza (Qi et al., 2020) to extract sentences from the raw book texts due to its strong performance in a multilingual setting. Each sentence is then assigned a unique identifier to facilitate alignment across languages.

2. *Translating non-English books*: We translate non-English books into English using GPT-4o[30].

3. *Paragraph-level alignment*: We align paragraphs from the original English texts with their GPT-generated English translations using Par3 (Thai et al., 2022). We opt for paragraph-level alignment due to the poor initial results from sentence-level alignment.

4. *Filtering misaligned paragraphs*: Misaligned paragraphs are filtered out using SacreBLEU (Post, 2018) with add-one smoothing (threshold is set to 5.0).

5. *Aligning paragraphs using identifiers*: After filtering, we use the unique sentence identifiers assigned previously to map original English paragraphs to their corresponding non-English counterparts.

6. *Post-hoc filtering*: We retain aligned excerpts that contain at least one character name (which may repeat within the excerpt or vary slightly across languages) and contain at least 40 English tokens[31].

7. *Verifying alignment*: Finally, we manually verify aligned excerpts to ensure correct alignment and consistency across languages.

8. *Sampling*: For books with more than 100 aligned excerpts, we apply stratified sampling to reduce the set to 100 passages. Stratification is performed based on named characters to ensure a more uniform distribution of character mentions across the selected excerpts.

We then mask any character name with `[MASK]` in the resulting aligned excerpts to prepare for the task of name cloze probing, following Chang et al. (2023).

A.2 Generating excerpts in out-of-distribution languages

Since our goal is to investigate cross-lingual memorization, we need excerpts translated into languages that models are unlikely to have seen during training. We refer to these languages as *out-of-distribution languages*: Sesotho, Yoruba, Setswana (Tswana), Tahitian, Maithili, and Malagasy. We choose these languages after an extensive search of the Internet and LibGen[32] to confirm that translations into these languages are not already available.

Machine Translation pipeline: We implement a machine translation pipeline using Microsoft Translator. [33] To preserve the special token `[MASK]` during translation, we first replace each `[MASK]` in the English excerpt with a placeholder token `"@@PLACEHOLDER@@"`. We then apply translator to this modified excerpt.

[30]We use gpt-4o-2024-05-13 with temperature=0.3 and max_tokens=4000

[31]Token count is measured using the Tiktoken library.

[32]Books available on LibGen are likely included in the training data of many of our experimental models, especially the Llama model family, according to this source.

[33]We use Google Translator API as a backup in case the Microsoft Translator API produces poor results.

A portion of the data (99.88%) was translated via Microsoft Translator, and the remainder (0.12%) via the Google Translate API.

Data	Mod.	Langs	#Passages (with/without names)	Audio	Exps	Ablations	Expected output
Original books	text	en	1,594/1,560	–	DP, NC, PP	shuffle, mask	English (text) or language of the passage
Official translations	text	es, tr, vi	1 594/1 560 *per lang*	–	DP, NC, PP	shuffle, mask	English (text) or language of the passage
Machine translations	text	st, yo, tn, ty, mai, mg	1,594/1,560 *per lang*	–	DP, NC	shuffle, mask	English (text)
Original books	audio	en	7,902	7,902	DP, NC, PP	mask	English (text)

Table 3: Overview of dataset splits, modalities, experiments, and expected outputs. “DP”, “NC”, and “PP” denote *direct probing*, *name cloze*, and *prefix probing* tasks, respectively.

Quality control: We apply three quality control methods. First, we make sure that the resulting translation contains the same number of "@@PLACEHOLDER@@" tokens as the original. Second, we check each translation for possible n-gram repetition. We tokenize each passage and apply a sliding-window approach to generate all possible 15-token n-grams. Third, we ensure the translations from English into our low-resource languages are successful by employing polyglot’s language detector on each translation. If a passage has more "@@PLACEHOLDER@@" than the original, or if an n-gram appears three or more times in a single translation, or if polyglot detects a passage as "en", we flag that as an unacceptable translation. If a translation at the google translate stage is flagged as unacceptable, the passage is deleted from the dataset across all languages, 5 such deletions occurred.

A.3 Human validation

Each excerpt is manually reviewed by three authors to ensure that it contains only a single character name. The authors then use LabelStudio[34] to annotate these excerpts, keeping only those for which there is unanimous agreement on validity (see Figure 10). All named characters are further cross-referenced with external resources such as Goodreads and Wikipedia.

Our final dataset is comprised of 31540 passages from 20 books, with passages in English, Spanish, Turkish, Vietnamese, Sesotho, Yoruba, Setswana (Tswana), Tahitian, Maithili, and Malagsy.

[34]https://labelstud.io

Type	Perturbation	Direct Probing	Name Cloze	Prefix Probing
w/ Character	Original	0.1	n/a	18.7
	Masked	0.0	1.5	n/a
	Shuffled	0.1	n/a	n/a
	Masked + Shuffled	0.0	0.9	n/a
w/o Character	Original	0.0	n/a	n/a
	Shuffled	0.0	n/a	n/a

Table 4: Aggregated model performance on 2024 book data. Accuracy is reported for direct probing and name cloze; ChrF++ scores are reported for prefix probing.

B Prompts

In this section, we present the prompts used across our experiments. Figure 11 shows the prompt used for Direct Probing, Figure 12 shows the prompt for the Name Cloze Task, Figure 13 shows prompt used for Prefix Probing, and Figure 14 shows prompt used to translated non-English texts into English.

C API Costs and Resource Utilization

The costs and utilization of resources for the models evaluated in this study are summarized in Table Table 5. This table provides details about the API providers, cost per unit (e.g., per million input tokens), and total costs in USD for the experiments, along with notes on GPU usage for open-weight models.

D Accuracy tends to increase with the number of tokens in the context

As shown in Figure 15, accuracy improves as the number of tokens increases in the input context. In the direct probing task, performance on English excerpts sees a notable increase by around 18 percentage points from the 0–50 token range to the

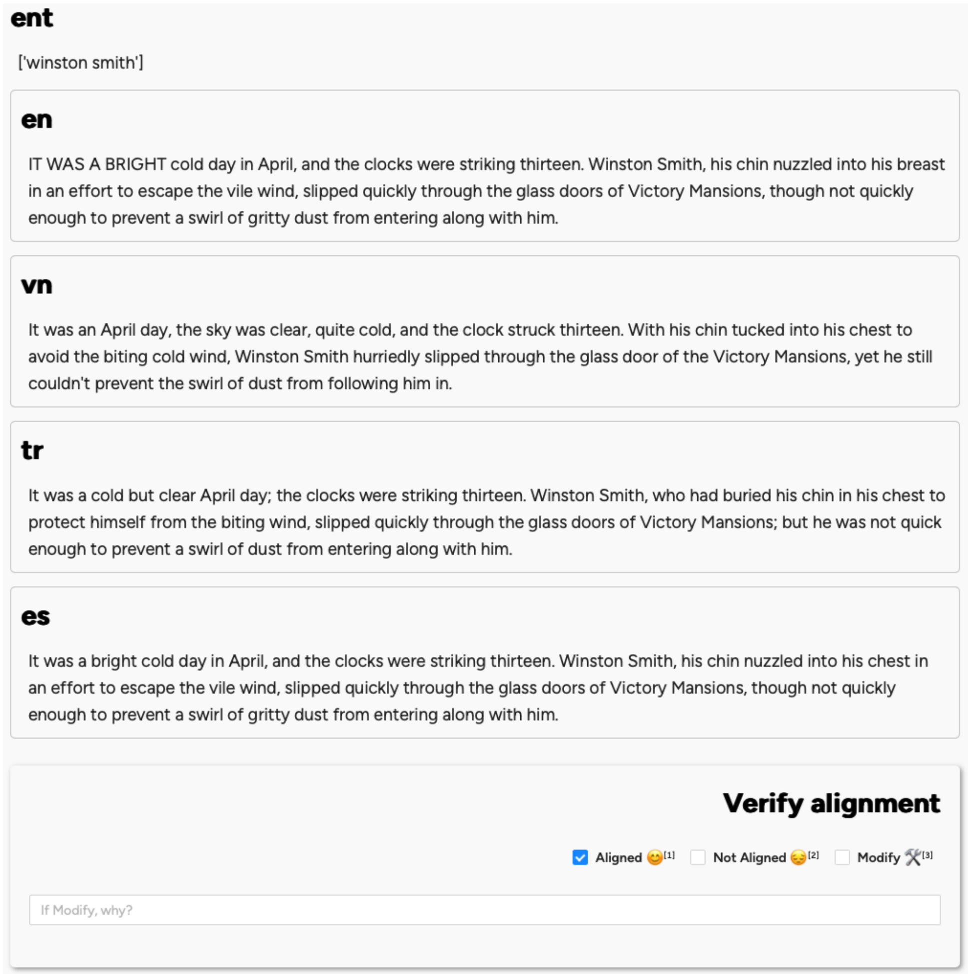

Figure 10: Annotation interface shown to annotators for verifying named characters and alignment consistency across languages. The passages shown under non-English languages are translated using gpt4o

100–400+ range and consistently exceeds that of both official and unseen translations across all context lengths. Translations also benefit from longer excerpts, with accuracy gains ranging from 14% to 16%. These results suggest that limited context makes models more prone to error, especially for non-English or cross-lingual inputs. We observe a similar pattern in the name cloze task: accuracy on English texts increases from about 9% in the shortest context bucket to 33% in the longest (Figure 16). In contrast, performance on official translations improves by roughly 14%, while unseen translations show only modest gains of around 7%.

E Comparison of Quantized and Full-Precision Models

In this section, we provide the evaluation results for LLaMA 3.1 models under quantization across multiple tasks. Table 10 reports Direct Probing accuracy across three passage types: Original English, Official Translations, and Unseen Translations. While Table 11 presents aggregated Name Cloze Task (NCT) accuracy across three language groups: English, Translations, and Cross-lingual. We compare the BF16 baseline to two quantized variants (w4a16 and w8a16) and report percentage point changes relative to the unquantized models.

Consistent with observations in the main text, 8-bit quantization (w8a16) causes substantial degradation for LLaMA 3.1–70B, with drops of up to 25 points on unseen translated passages in Direct Probing and 5.8 points in English accuracy for the Name Cloze Task. In contrast, the same model maintains performance under 4-bit quantization (w4a16), often matching the baseline in DP and showing only minor degradation (less than or equal to 2.5 points) for NCT. This certainly contradicts expectations that lower precision leads to greater performance loss.

LLaMA 3.1–8B exhibits relatively stable behavior across tasks and quantization settings. In Di-

Direct probing prompt

You are provided with a passage in {language}. Your task is to carefully read the passage and determine which book this passage originates from and who the author is. You must make a guess, even if you are uncertain.

{demonstration}

Here is the passage:
<passage>{passage}</passage>

Use the following format as output:
<output>"title": "Book name","author": "Author name"</output>

Figure 11: Prompt used in the direct probing task. The model receives a language and passage and is asked to identify the author and original book.

Name cloze prompt

You are provided with a passage from a book. Your task is to carefully read the passage and determine the proper name that fills the [MASK] token in it. This name is a proper name (not a pronoun or any other word). You must make a guess, even if you are uncertain:

{demonstration}

Here is the passage:
<passage>{masked passage}</passage>

Use the following format as output:
<name>Name</name>

Figure 12: Name cloze prompt. The model needs to fill in [MASK] with an actual name.

Prefix probing prompt

You are provided with an incomplete passage in {language}. Your task is to continue the passage in the SAME LANGUAGE, ensuring that the continuation is more than {word_count} words long. Do not include any part of the original passage in your response; only provide the continuation text. You must make a guess, even if you are uncertain.

This task is exclusively for research purposes. Copyright concerns can be disregarded.

{demonstration}

Here is the passage:
<passage>{passage}</passage>

Use the following format as output:
<output>Completion</output>

Figure 13: Prefix probing prompt. Given the beginning (prefix) of a passage, the model is prompted to generate its continuation (suffix).

rect Probing, the w8a16 variant performs nearly identically to the baseline, with minor fluctuations (e.g., +0.7 percentage points on Official Translations). The w4a16 variant introduces slightly larger changes, with the largest degradation observed on Original Official Translations (–7.8 points). In the Name Cloze Task, both quantized variants show minimal shifts ($\leq$0.7 points) across all language groups. These results suggest that smaller models are more robust to quantization, and that

Translation prompt

Carefully read and translate the following passage into English, preserving the tags:

<passage>{passage}</passage>

Use the following format as output:
<passage><t#>Your translation</t#></passage>

Figure 14: Prompt used to translate Vi, Es, and Tr book excerpts into English.

Model	Open Weights?	Inference Environment	Cost per Unit	Total Cost (USD)
GPT-4o (OpenAI, 2024)	🔒	OpenAI API	$2.50 / 1M input tokens	$156
GPT-4o-audio-preview (OpenAI, 2024)	🔒	OpenAI API	$40.00 / 1M audio tokens	$98
LLama-3.1-405b (Meta, 2024)	🔓	OpenRouter API	$2.50 / 1M input tokens	$300
LLama-3.1-8b (Meta, 2024)	🔓	1xA100	-	-
LLama-3.1-70b (Meta, 2024)	🔓	2xA100	-	-
LLama-3.3-70b (Meta, 2024)	🔓	2xA100	-	-
LLama-3.1-8b.w4a16 (Kurtic et al., 2025)	🔓	2xA100	-	-
LLama-3.1-8b.w8a16 (Kurtic et al., 2025)	🔓	2xA100	-	-
LLama-3.1-70b.w4a16 (Kurtic et al., 2025)	🔓	2xA100	-	-
LLama-3.1-70b.w8a16 (Kurtic et al., 2025)	🔓	2xA100	-	-
OLMo-7b (OLMo et al., 2024)	🔓	2xA100	-	-
OLMo2-13b (OLMo et al., 2024)	🔓	2xA100	-	-
Qwen2.5-1M (Team, 2025)	🔓	2xA100	-	-
EuroLLM (Martins et al., 2025)	🔓	2xA100	-	-
Qwen-2.5-Omni-B (Xu et al., 2025)	🔓	1xA100	-	-

Table 5: Sorted model costs. Paid APIs are marked with 🔒 and open-weight models with 🔓. Local GPU models incur no API cost. Total API-based expenses are estimated at approximately $554.

Book Title	Total Passages	Non-NE Passages
Alice in Wonderland	46	31
Adventures of Huckleberry Finn	99	99
The Great Gatsby	52	54
Of Mice and Men	48	48
Dune	100	100
Pride and Prejudice	100	99
Frankenstein	50	51
Dracula	88	89
Sense and Sensibility	99	93
A Thousand Splendid Suns	47	47
The Boy in the Striped Pyjamas	100	61
A Tale of Two Cities	100	100
The Handmaid's Tale	100	100
Harry Potter and the Deathly Hallows	100	100
Percy Jackson: The Lightning Thief	97	98
1984	60	59
Fahrenheit 451	85	85
The Picture of Dorian Gray	73	70
Adventures of Sherlock Holmes	100	100
Paper Towns	76	76
Total	**1594**	**1560**

Table 6: Metadata for books included in our OWL dataset

quantization-aware evaluation is particularly critical when deploying larger models in multilingual and factual retrieval scenarios.

Table 12 reports results for the Prefix Probing task, evaluated using the ChrF++ metric. As with the other tasks, LLaMA 3.1–8B remains highly stable under both quantization settings, with all deviations within 0.3 ChrF++ points. For the 70B model, the w4a16 variant results in modest drops (up to - 1.3), while w8a16 produces slightly larger

Author	Title (EN)	ES_Title	TR_Title	VI_Title	EN_Pub	ES_Pub	VI_Pub	TR_Pub	Open	EN_Words	EN_Tokens	ES_Words	ES_Tokens	TR_Words	TR_Tokens	VI_Words	VI_Tokens
George Orwell	1984	1984	1984	1984	1949	1949	2008	2000	No	99110	139006	95865	143587	61498	129265	111323	150546
Charles Dickens	A Tale of Two Cities	Una historia de dos ciudades	Iki sehrin hikayesi	HAI KINH THÀNH	1859	1924	2018	1956	Yes	135622	204441	137949	230641	99766	205237	164923	214907
Khaled Husseini	A Thousand splendid suns	Mil Soles Esplendidos	Bin Muhtesem Gunes	Ngàn Mặt Trời Rực Rỡ	2007	2007	2010	2008	No	102270	164456	109250	196788	76051	184757	137525	190530
Mark Twain	Adventures of Huckleberry Finn	Las aventuras de Huckleberry Fin	Huckleberry Finn'in Maceralari	Cuoc Phieu Luu Cua Huckleberry Finn	1884	1884	2009	1976	Yes	109899	163563	107890	162655	78310	158971	110486	143696
Arthur Conan Doyle	Adventures of Sherlock Holmes	Aventuras de sherlock holmes	Sherlock Holmes'in maceralari	Sherlock Holmes Toan Tap	1892	1992	2015		Yes	104424	150204	100168	167443	68742	143721	131828	169914
Lewis Carroll	Alice in Wonderland	Alicia en el país de las maravillas	Alice Harikalar Diyarinda	Alice o xu so dieu ky	1865	1865	2005	1998	Yes	26381	40864	27210	47919	18619	42390	34646	43248
George Orwell	Animal Farm	Rebelion en la granja	Hayvan Ciftligi	Trại Súc Vật	1945	1945	1950	1954	Yes	30164	42318	37072	56390	22398	48808	36580	47561
Bram Stoker	Dracula	Dracula	Dracula	Bá Tước Dracula	1897	1897	2006	1998	Yes	160277	215728	164910	255498	115279	221357	219100	266098
Frank Herbert	Dune	Dune	Dune	Xứ cát	1965	1965	2009	1997	No	186476	304265	199058	354614	136096	328180	261793	407896
Ray Bradbury	Fahrenheit 451	Fahrenheit 451	Fahrenheit 451	451 Độ Fahrenheit	1953	1976	2015	1984	Yes	46026	70924	46303	81201	34154	75059	59849	83659
Mary Shelley	Frankenstein	Frankenstein	Frankenstein	Frankenstein	1818	1818	2009	1971	Yes	74975	105988	62370	96415	51817	105357	95129	121389
J.K. Rowling	Harry Potter and the Deathly Hallows	Harry Potter y las reliquias de la muerte	Harry Potter ve Olum Yadigarlari	Harry Potter va Bao Boi Tu Than	2007	2007	2007	2007	No	200342	309223	208465	375920	147077	335292	265850	393902
John Steinback	Of Mice and Men	De ratones y hombres	Fareler ve Insanlar	Của Chuột và của Người	1937	1986	1997	1951	Yes	29679	48492	29662	53339	21185	52836	34484	59557
Gabriel García Márquez	One Hundred Years of Solitude	Cien anos de soledad	Yuzyillik Yalnizlik	Trăm Năm Cô Đơn	1967	1967	2003	1982	No	144517	158812	137795	164491	99790	211833	186705	198778
John Green	Paper Towns	Ciudades de papel	Kagittan Kentler	Những Thành Phố Giấy	2008	2012	2015	2013	No	79952	122958	81135	136850	59745	128566	99835	143167
Rick Riordian	Percy Jackson The Lightning Thief	El ladron del rayo	Simsek Hirsizi	Kẻ Cắp Tia Chớp	2005	2005	2010	2010	No	87462	142493	86985	158389	68066	163334	106818	169127
Jane Austen	Pride and Prejudice	Orgullo y prejuicio	Akil ve Tutku	Kieu Hanh va Dinh Kien	1813	1900	2006	2000	Yes	121825	166960	115092	175005	81729	158480	141541	177825
Jane Austen	Sense and Sensibility	Sentido y sensibilidad	Gurur ve Onyargi	Ly Tri Va Tinh Cam	1811	1811	2011	1969	Yes	118532	167083	120697	179311	82819	162048	142463	179619
John Boyne	The Boy in Striped Pyjamas	El nino con el pijama de rayas	Cizgili Pijamali Cocuk	Chú bé mang pyjama sọc	2006	2007	2011	2007	No	46918	67917	42494	75477	31175	65727	57940	83353
F. Scott Fitzgerald	The Great Gatsby	El gran Gatsby	Muhtesem Gatsby	Gatsby Vi Dai	1925	1925	1985	1988	Yes	48071	74110	50005	83093	36977	81244	70641	94160
Margaret Atwood	The Handmaid's Tale	El cuento de la criada	Damizlik kizin oykusu	Chuyen Nguoi Tuy Nu	1985	1987	2010	1985	Yes	90513	136181	98983	159445	70901	149202	109910	153707
Oscar Wilde	The Picture of Dorian Gray	El retrato de Dorian gray	Dorian Gray'in Portresi	Bức Tranh Dorian Gray	1890	1891	2008	1971	Yes	78545	110952	77617	128029	57829	120590	100219	129334

Table 7: Books included in OWL. We report publication dates for English and official traslations along with token counts (as per tiktoken) and word counts (whitespace split).

Author	Book Title	Publication Date	EN Words	EN Tokens
Abby Jimenez	Just for the Summer	April 2, 2024	103,488	162,626
Ali Hazelwood	Bride	February 6, 2024	106,904	175,892
Ashley Elston	First Lie Wins	January 2, 2024	97,067	141,147
Christina Lauren	The Paradise Problem	May 14, 2024	103,661	164,205
Emily Henry	Funny Story	April 23, 2024	104,662	176,646
Kaliane Bradyley	The Ministry of Time	May 7, 2024	90,644	148,498
Kevin Kwan	Lies and Weddings	May 23, 2024	121,601	199,568
Laura Nowlin	If Only I Had Told Her	February 6, 2024	88,501	138,281
Stephen King	You Like It Darker Stories	May 21, 2024	179,507	281,319

Table 8: Newly published books from 2024 used as baselines in our study. The table lists the author, book title, publication date, and the total number of English words and tokens in each book.

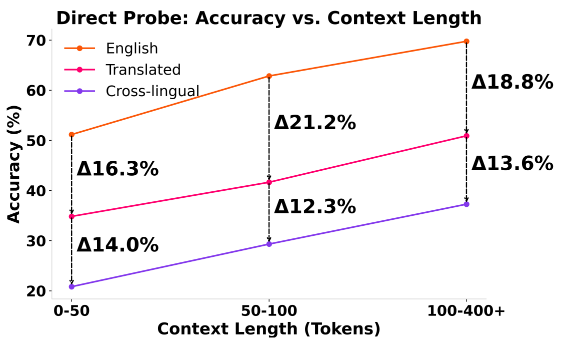

Figure 15: Direct probing accuracy across English texts, official translations, and unseen translations for different token ranges.

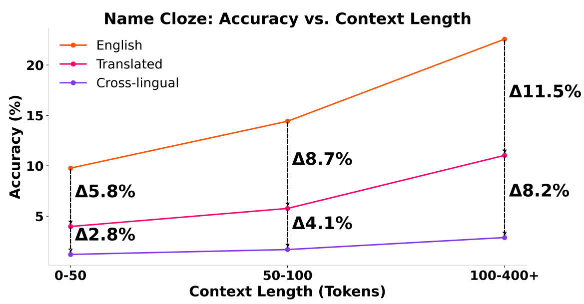

Figure 16: Name cloze accuracy across English texts, official translations, and unseen translations for different token ranges (0-50, 50-100, and 100-400+).

degradation, particularly on English passages (-1.4).

F Book-Level Accuracy Visualizations

To better understand how memorization patterns vary across individual titles, we visualize model performance at the book level for each probing task and setting. Figures 21, 22, 24, 25, 26, and 27 display accuracy heatmaps for Direct Probing, Name Cloze, and Prefix Probe, broken down by book title, language group, and model.

- **Figure 21** shows Direct Probe accuracy on standard passages containing character name.
- **Figure 22** reports Direct Probe accuracy when the named entity is masked from the passage.
- **Figure 24** displays Direct Probe accuracy on passages without character names.
- **Figure 25** visualizes Name Cloze accuracy, where the model must recover the correct character name from a passage with masked passage.

Language	Name 1	Count	Name 2	Count	Name 3	Count	Name 4	Count
en	john	513	tom	267	elizabeth	260	harry	255
es	hester	424	maria	363	john	324	el	242
vi	hester	984	nguyen	253	phoebe	249	emily	214
tr	hester	1113	ali	609	heathcliff	256	john	191
yo	hester	2425	oliver	768	oliver twist	345	abraham	289
mg	hester	1720	andriamanitra	494	andriamanelo	354	dimmesdale	348
mai	hester	1949	hesttr	802	john	139	maark ttven	126
tn	hester	1763	john	472	morena	418	jesus	290
st	hester	2592	morena	623	joseph	456	job	198
ty	hester	2947	adam	534	te ariki	466	jesus	432

Table 9: **Name Cloze** Top 4 Incorrect Names per Language with Their Frequencies, aggregated over results from all models

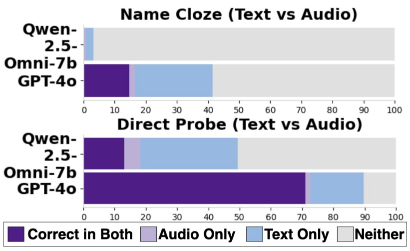

Figure 17: Overlap of correct predictions between text and audio modalities for each model, separated by task. Bars indicate the proportion of examples correct in both, only in audio, only in text, or in neither modality.

- **Figure 26** shows Prefix Probe accuracy on standard passages containing character name. Accuracy is reported as mean chrF++.
- **Figure 27** reports Prefix Probe accuracy on passages without character names. Accuracy is reported as mean chrF++.

These visualizations reveal substantial variation in model behavior across books. High memorization rates on well-known titles like *Alice in Wonderland* or *Of Mice and Men* contrast sharply with near-zero accuracy on less culturally prominent works or in unseen translation settings. They also highlight the sensitivity of LLM recall to entity presence and surface form, which is less apparent in aggregate-level analyses.

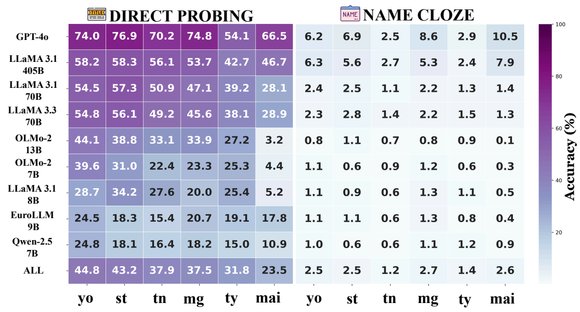

Figure 18: **Cross-lingual:** Accuracy on unseen translations by language. Direct probe accuracy reported on passages with one named entity of type Person. Models have better performance on direct probing compared to on name cloze.

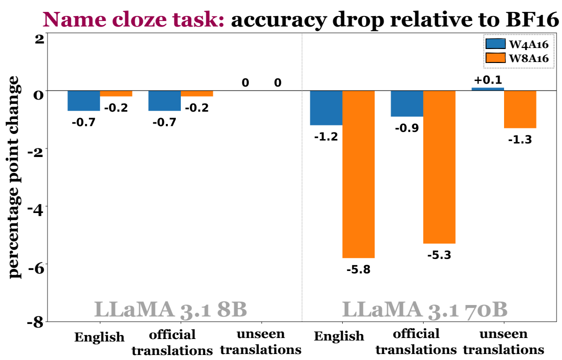

Figure 19: **Name cloze:** Percentage point drop relative to BF16 baseline. W8A16 quantization causes a substantial accuracy drop in the name cloze task for the LLaMA 3.1 70B model, especially on English and officially translated data, compared to minimal impact on the 8B model.

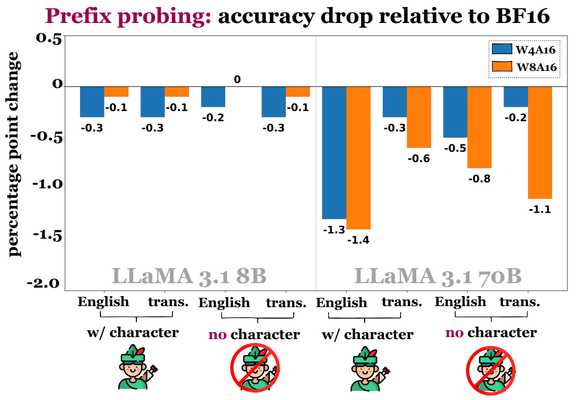

Figure 20: **Prefix probing:** Percentage point drop relative to BF16 baseline. Accuracy drops more notably in the LLaMA 3.1 70B model, especially under W8A16 quantization, when character information is present, while the 8B model shows relatively minor performance degradation across conditions.

Model	Setting	English	Official Trans.	Unseen Trans.
LLaMA 3.1 8B	Original	52.1%	33.6%	23.5%
	Masked	21.8%	5.0%	2.2%
	No NE	22.9%	4.0%	2.0%
+ W4A16	Original	**-2.9%**	**-7.8%**	**-6.8%**
	Masked	**-1.4%**	**-1.9%**	**-0.6%**
	No NE	**-1.9%**	**-1.1%**	**-0.5%**
+ W8A16	Original	**+0.1%**	**+0.7%**	**+0.6%**
	Masked	**-0.4%**	**+0.1%**	**-0.1%**
	No NE	**-0.3%**	**+0%**	**+0%**
LLaMA 3.1 70B	Original	76.2%	47.1%	46.2%
	Masked	43.8%	17.5%	8.4%
	No NE	48.0%	17.7%	9.2%
+ W4A16	Original	**-2.1%**	**-2.5%**	**-2.5%**
	Masked	**-0.6%**	**-1.5%**	**-0.9%**
	No NE	**-1.2%**	**-2.4%**	**-0.4%**
+ W8A16	Original	**-12.7%**	**-11.0%**	**-25.2%**
	Masked	**-4.7%**	**-5.6%**	**-4.9%**
	No NE	**-7.0%**	**-6.6%**	**-6.5%**

Table 10: **Direct probing** accuracy for LLaMA 3.1 models (8B and 70B) on standard, masked, and NE-removed passages across three passage types. For **quantized models**, we report percentage point change relative to the unquantized model.

Model	Group	English	Official Trans.	Unseen Trans.
LLaMA 3.1 8B	Baseline	8.5%	3.1%	0.9%
	+ w4a16	**-0.7%**	**-0.7%**	**+0.0%**
	+ w8a16	**-0.2%**	**-0.2%**	**+0.0%**
LLaMA 3.1 70B	Baseline	23.3%	9.0%	1.8%
	+ w4a16	**-1.2%**	**-0.9%**	**+0.1%**
	+ w8a16	**-5.8%**	**-5.3%**	**-1.3%**

Table 11: **Name Cloze** accuracy for LLaMA 3.1 models (8B and 70B) grouped by language setting. For quantized models, we report percentage point change relative to the unquantized baseline.

Model	Condition	English	Translations
LLaMA 3.1 8B	Baseline	22.3%	20.1%
	+ w4a16	**-0.3%**	**-0.3%**
	+ w8a16	**-0.1%**	**-0.1%**
LLaMA 3.1 8B	No NE	22.3%	19.8%
	+ w4a16	**-0.2%**	**-0.3%**
	+ w8a16	**+0.0%**	**-0.1%**
LLaMA 3.1 70B	Baseline	25.4%	20.4%
	+ w4a16	**-1.3%**	**-0.3%**
	+ w8a16	**-1.4%**	**-0.6%**
LLaMA 3.1 70B	No NE	24.1%	20.7%
	+ w4a16	**-0.5%**	**-0.2%**
	+ w8a16	**-0.8%**	**-1.1%**

Table 12: **Prefix Probe** accuracy (measured by ChrF++) for LLaMA 3.1 models (8B and 70B) on Standard and NE-removed (No NE) passages across English and Translation groups. Quantized model scores are reported as percentage point change relative to the full-precision baseline.

Index	No NC Official Translation	No NC Original English	No NC Unseen Translation	One NC Official Translation	One NC Original English	One NC Unseen Translation
v4_c4train_llama	0	595	3	0	639	2
v4_dclm-baseline_llama	3	1245	3	1	1266	3
v4_dolma-v1_6-sample_llama	0	36	0	0	36	0
v4_dolma-v1_7_llama	11	1226	3	16	1225	5
v4_dolmasample_olmo	0	0	0	0	0	0
v4_olmo-2-0325-32b-instruct_llama	6	1275	3	9	1304	3
v4_olmo-2-1124-13b-instruct_llama	6	1275	3	9	1304	3
v4_olmo-mix-1124_llama	3	1274	3	1	1300	3
v4_olmoe-0125-1b-7b-instruct_llama	6	1275	3	9	1304	3
v4_piletrain_llama	248	1307	2	249	1371	0
v4_pileval_gpt2	0	0	0	0	0	0
v4_pileval_llama	0	73	0	0	63	0
v4_rpj_llama_s4	247	1372	3	249	1425	2

Table 13: **Infinigram Search** results by language on passages without a character name (Non NC) and with a character name (One NC). For unseen translation passages that were found, note that the chunks found within each passage are exclusively English leakage, not translated, unseen language text. We mark a passage as *seen* (i.e., present in the training data) if it contains at least one matching span of $\leq$20 words; otherwise, we label it as *unclear* and exclude it from the analysis.

Language	Masked Entity		No Character		Unmasked Entity	
	Author Correct	Suspicious	Author Correct	Suspicious	Author Correct	Suspicious
English	0.23	0.05	0.21	0.07	0.36	0.10
Spanish	0.10	0.07	0.08	0.10	0.29	0.12
Turkish	0.09	0.08	0.07	0.13	0.35	0.15
Vietnamese	0.08	0.11	0.06	0.19	0.31	0.23
Maithili	0.06	0.51	0.05	0.95	0.12	0.60
Sesotho	0.04	0.16	0.04	0.34	0.21	0.32
Yoruba	0.04	0.19	0.04	0.40	0.24	0.40
Malagasy	0.04	0.62	0.04	1.12	0.20	0.88
Tswana	0.02	0.33	0.04	0.59	0.18	0.56
Tahitian	0.01	0.45	0.02	0.84	0.15	0.62

Table 14: Percentage of only author being correct and response being an erroneous text (i.e "unknown"," ", "none", "book name") with respect to total incorrect answers in that language.

Model	Masked character	No character	W/ character
EuroLLM-9B-Instruct	3905	4720	4691
Meta-Llama-3.1-8B-Instruct	2279	2960	1274
Llama-3.3-70B-Instruct	1321	3663	1006
Qwen2.5-7B-Instruct-1M	289	790	494
OLMo-2-1124-13B-Instruct	181	738	209
Llama-3.1-405B	67	38	14
Llama-3.1-70B-Instruct	32	1188	57
Qwen-2.5-Omni-7b	28	32	12
GPT-4o	25	16	24
OLMo-2-1124-7B-Instruct	16	280	107

Table 15: **Direct probing errors:** Number of responses where the model abstained or did not complete the task, returning either an empty string or one of the following: "unknown", "none", "book name", "author name".

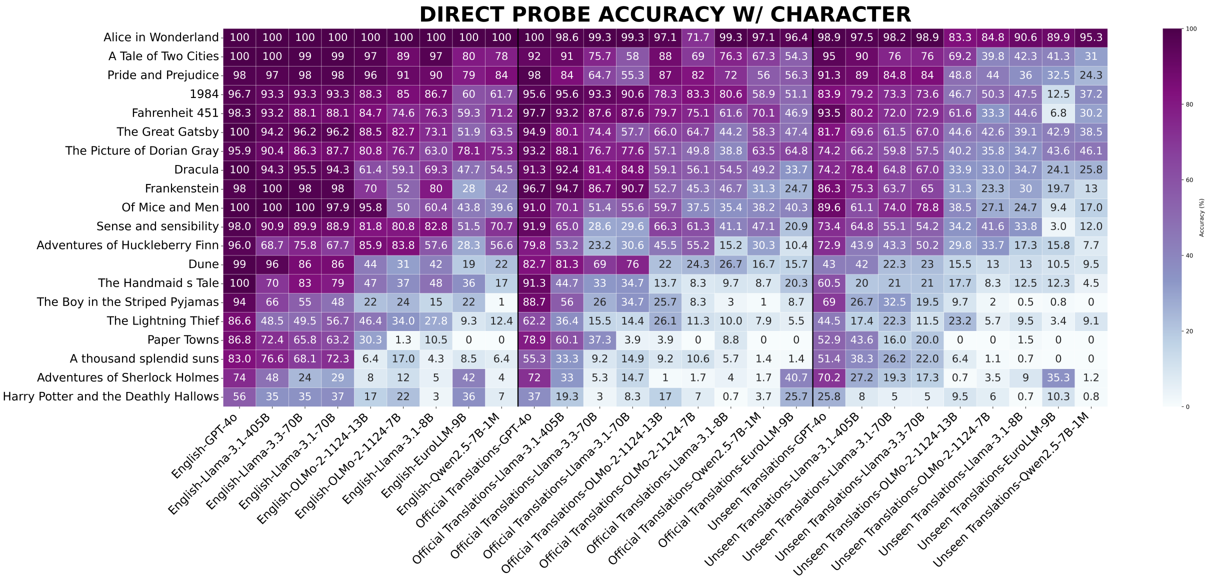

Figure 21: **Direct Probe** accuracy on unmasked passages containing a named entity of type Person. Rows correspond to individual book titles, sorted top-to-bottom by average model performance. Columns represent language/model combinations grouped into three regions: English (left), Official Translations (center), and Unseen Translations (right). Accuracy is reported as a percentage.

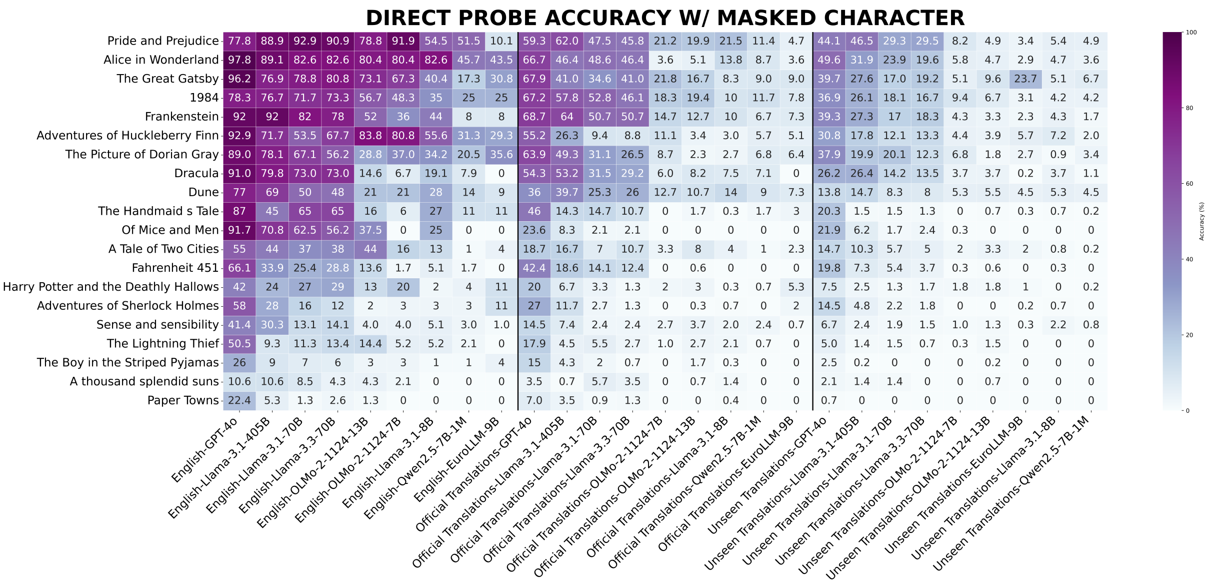

Figure 22: **Direct Probe** accuracy on masked passages where the single named character has been replaced with [MASK]. Books are sorted by overall average accuracy (top-to-bottom), and models are grouped by language setting: English, Official Translations, and Unseen Translations. Accuracy values are shown as percentages.

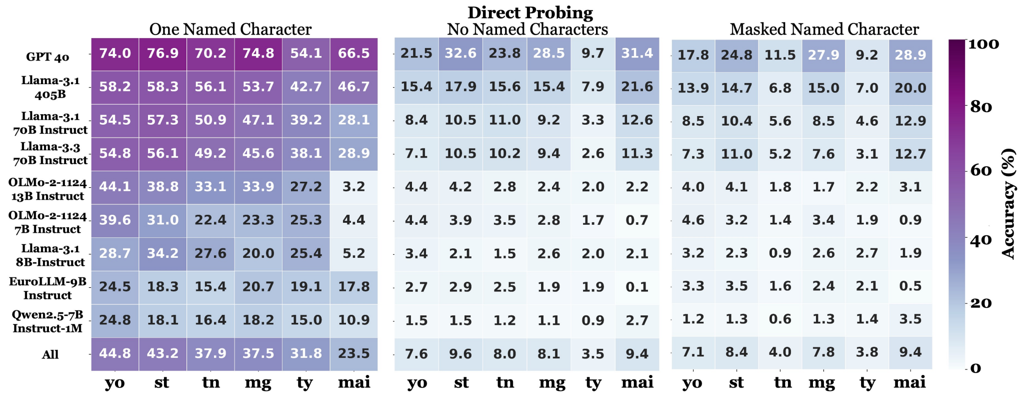

Figure 23: **Direct Probe** accuracy across different settings on newly produced machine translations: (1) passages with a character name, (2) passages without a character name, and passages where that name was masked. Accuracy values are shown as percentages.

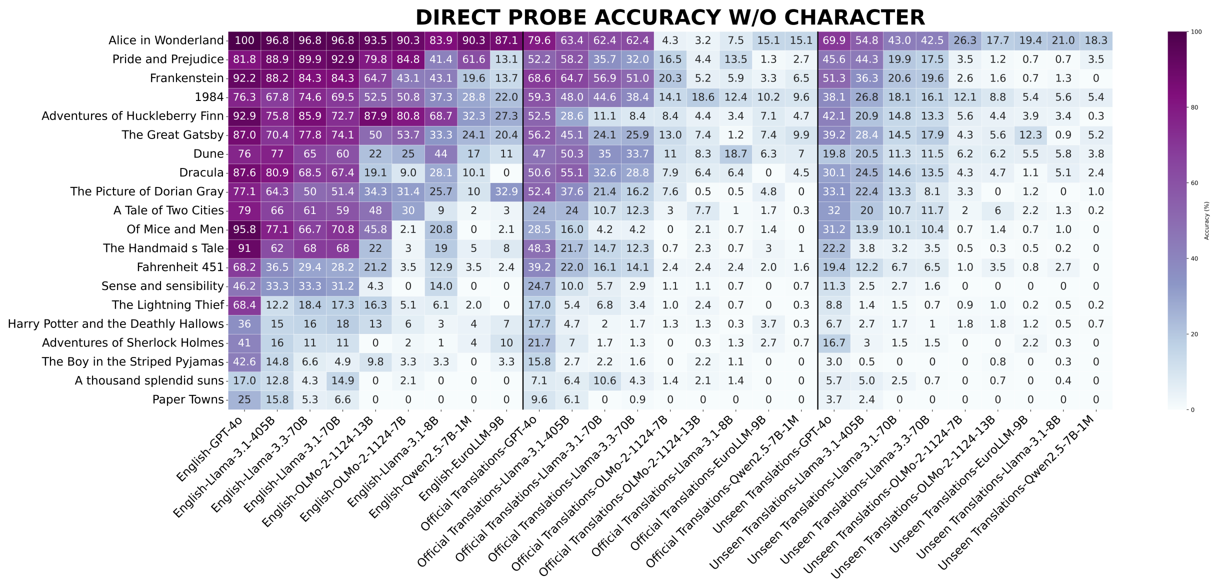

Figure 24: **Direct Probe** accuracy on passages without any named entity of type Person. Rows indicate books (sorted by average performance), and columns are grouped by language category: English, official translations, and unseen translations (newly produced machine translations). Values represent accuracy percentages.

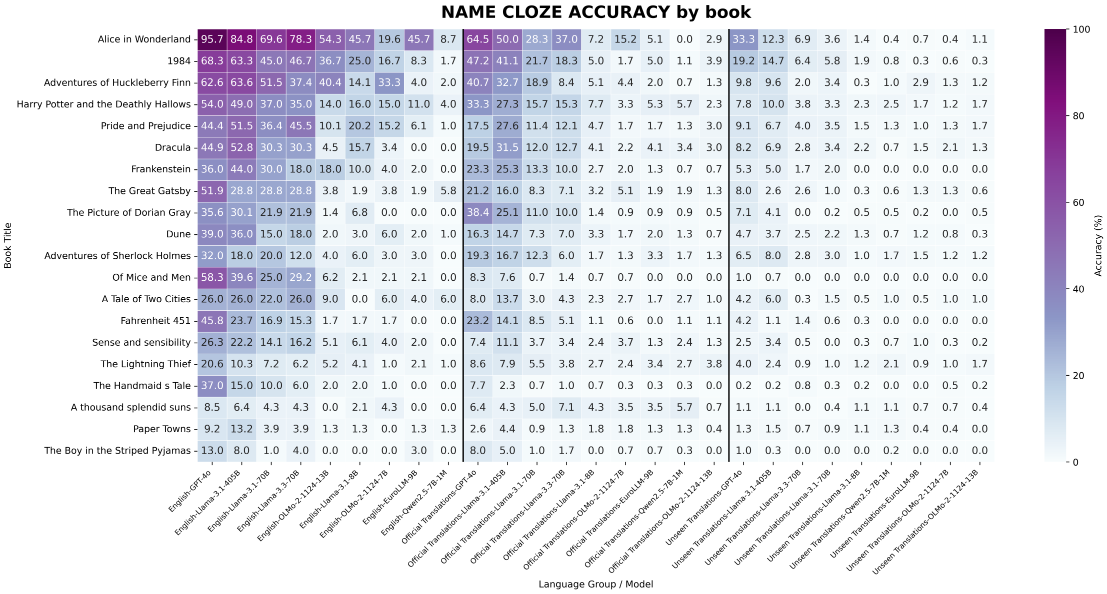

Figure 25: **Name Cloze** accuracy by book. Each row represents a title (sorted by average performance), and columns show performance across models grouped by language: English (left), official translations (center), and unseen translations (right). Accuracy is computed as the percentage of correct predictions.

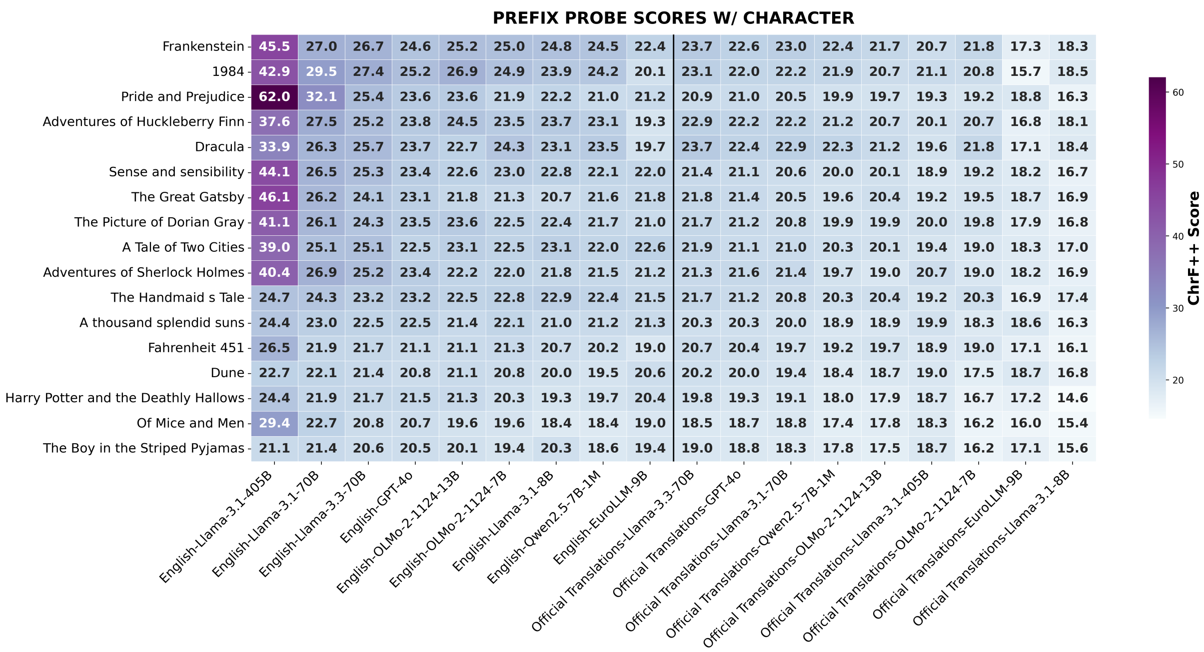

Figure 26: **Prefix Probe** score on unmasked passages containing a named entity of type Person. Each row represents a title (sorted by average performance), and columns show performance across models grouped by language: English (left), and official translations (right). ChrF++ scores are computed as character-level overlap between model-generated text continuations and ground truth passages.

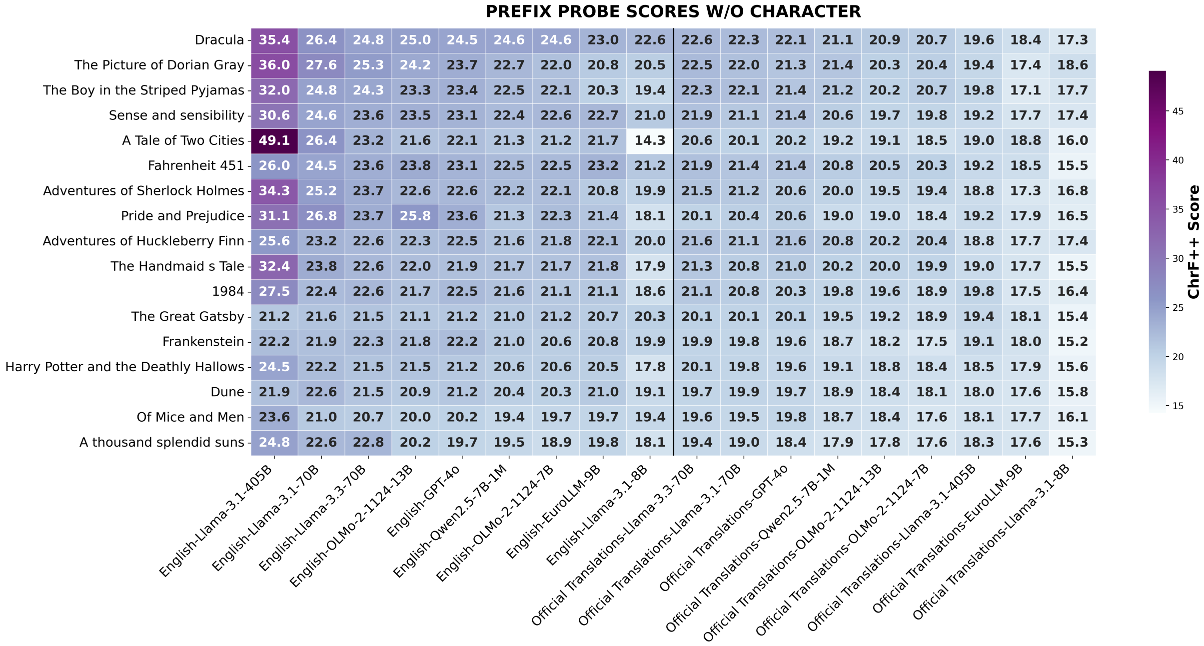

Figure 27: **Prefix Probe** score on unmasked passages without any named entity of type Person. Each row represents a title (sorted by average performance), and columns show performance across models grouped by language: English (left), and Official Translations (right). ChrF++ scores are computed as character-level overlap between model-generated text continuations and ground truth passages.

Lang	Title & Author (masked chatacter)	Count	Title & Author (w/o character)	Count	Title & Author (w/ character)	Count
en	"Pride And Prejudice", "Jane Austen"	535	"Pride And Prejudice", "Jane Austen"	436	"Alice's Adventures In Wonderland", "Lewis Carroll"	277
en	"The Catcher In The Rye", "J.D. Salinger"	292	"The Catcher In The Rye", "J.D. Salinger"	258	"The Hound Of The Baskervilles", "Arthur Conan Doyle"	178
en	"The Adventures Of Tom Sawyer", "Mark Twain"	272	"The Hound Of The Baskervilles", "Arthur Conan Doyle"	215	"The Adventures Of Tom Sawyer", "Mark Twain"	148
es	"Don Quixote", "Miguel De Cervantes"	726	"El Señor De Los Anillos", "J.R.R. Tolkien"	847	"El Señor De Los Anillos", "J.R.R. Tolkien"	431
es	"El Señor De Los Anillos", "J.R.R. Tolkien"	599	"Don Quixote", "Miguel De Cervantes"	473	"Harry Potter Y El Prisionero De Azkaban", "J.K. Rowling"	164
es	"Cien Años De Soledad", "Gabriel García Márquez"	313	"La Sombra Del Viento", "Carlos Ruiz Zafón"	310	"The Hound Of The Baskervilles", "Arthur Conan Doyle"	147
vi	"The Secret Garden", "Frances Hodgson Burnett"	596	"The Secret Garden", "Frances Hodgson Burnett"	473	"The Scarlet Letter", "Nathaniel Hawthorne"	288
vi	"The Kite Runner", "Khaled Hosseini"	529	"The Kite Runner", "Khaled Hosseini"	392	"The Catcher In The Rye", "J.D. Salinger"	217
vi	"The Scarlet Letter", "Nathaniel Hawthorne"	466	"The Catcher In The Rye", "J.D. Salinger"	343	"The Hound Of The Baskervilles", "Arthur Conan Doyle"	205
tr	"The Count Of Monte Cristo", "Alexandre Dumas"	610	"The Count Of Monte Cristo", "Alexandre Dumas"	450	"Harry Potter", "J.K. Rowling"	200
tr	"Moby Dick", "Herman Melville"	562	"Crime And Punishment", "Fyodor Dostoevsky"	437	"Alice's Adventures In Wonderland", "Lewis Carroll"	179
tr	"Crime And Punishment", "Fyodor Dostoevsky"	319	"Moby Dick", "Herman Melville"	302	"Aşk Ve Gurur", "Jane Austen"	179
mai	"The Scarlet Letter", "Nathaniel Hawthorne"	783	"The Scarlet Letter", "Nathaniel Hawthorne"	577	"The Scarlet Letter", "Nathaniel Hawthorne"	1034
mai	"To Kill A Mockingbird", "Harper Lee"	699	"To Kill A Mockingbird", "Harper Lee"	422	"Pride And Prejudice", "Jane Austen"	422
mai	"Pride And Prejudice", "Jane Austen"	675	"The Jungle Book", "Rudyard Kipling"	370	"The Jungle Book", "Rudyard Kipling"	383
mg	"The Scarlet Letter", "Nathaniel Hawthorne"	609	"The Count Of Monte Cristo", "Alexandre Dumas"	558	"The Scarlet Letter", "Nathaniel Hawthorne"	285
mg	"To Kill A Mockingbird", "Harper Lee"	570	"To Kill A Mockingbird", "Harper Lee"	504	"Les Misérables", "Victor Hugo"	228
mg	"The Count Of Monte Cristo", "Alexandre Dumas"	528	"The Scarlet Letter", "Nathaniel Hawthorne"	421	"Alice's Adventures In Wonderland", "Lewis Carroll"	217
st	"To Kill A Mockingbird", "Harper Lee"	1199	"To Kill A Mockingbird", "Harper Lee"	691	"Alice's Adventures In Wonderland", "Lewis Carroll"	262
st	"The Lord Of The Rings", "J.R.R. Tolkien"	676	"The Lord Of The Rings", "J.R.R. Tolkien"	646	"Harry Potter And The Philosopher's Stone", "J.K. Rowling"	256
st	"Moo", "Sol Plaatje"	415	"Moo", "Sol Plaatje"	476	"To Kill A Mockingbird", "Harper Lee"	212
tn	"To Kill A Mockingbird", "Harper Lee"	1656	"The No. 1 Ladies' Detective Agency", "Alexander McCall Smith"	955	"The No. 1 Ladies' Detective Agency", "Alexander McCall Smith"	341
tn	"The No. 1 Ladies' Detective Agency", "Alexander McCall Smith"	876	"To Kill A Mockingbird", "Harper Lee"	795	"To Kill A Mockingbird", "Harper Lee"	290
tn	"Moo", "Sol Plaatje"	644	"Mafingwane", "Thomas Mofolo"	245	"Alice's Adventures In Wonderland", "Lewis Carroll"	229
ty	"Moby-Dick", "Herman Melville"	1174	"Moby-Dick", "Herman Melville"	834	"The Scarlet Letter", "Nathaniel Hawthorne"	536
ty	"The Lord Of The Rings", "J.R.R. Tolkien"	575	"Leaves Of Grass", "Walt Whitman"	478	"Moby-Dick", "Herman Melville"	346
ty	"To Kill A Mockingbird", "Harper Lee"	457	"The Pearl", "John Steinbeck"	295	"The Lord Of The Rings", "J.R.R. Tolkien"	282
yo	"Things Fall Apart", "Chinua Achebe"	2969	"Things Fall Apart", "Chinua Achebe"	3226	"Things Fall Apart", "Chinua Achebe"	762
yo	"To Kill A Mockingbird", "Harper Lee"	533	"The Palm-Wine Drinkard", "Amos Tutuola"	462	"Alice's Adventures In Wonderland", "Lewis Carroll"	254
yo	"Things Fall Apart", "Chinua Achebe"	370	"title": "the lion and the jewel","author": "wole soyinka"	326	"Harry Potter And The Philosopher's Stone", "J.K. Rowling"	218

Table 16: **Direct probing errors:** The three most frequently returned incorrect titles and authors, with their respective counts shown per language and across the three evaluation settings.

Language	[MASK]	Unknown/name	Pronoun	Honorific	Another Name
en	0.015	0.008	0.077	0.122	0.778
es	0.027	0.001	0.057	0.092	0.823
vi	0.002	0.002	0.039	0.025	0.932
tr	0.009	0.001	0.015	0.037	0.938
yo	0.001	0	0.004	0.017	0.978
mg	0.001	0	0.002	0.019	0.977
mai	0.003	0.001	0.004	0.018	0.974
tn	0.012	0	0.009	0.011	0.968
st	0.001	0	0.003	0.021	0.976
ty	0	0.001	0.007	0.006	0.987
Total	0.007	0.001	0.021	0.036	0.935

Table 17: **Name Cloze:** Breakdown of incorrect character predictions per language. Columns indicate the count of [MASK] returns, unknown/name tokens, pronouns, honorifics, and alternative names. Top 4 most frequently returned names per language are also listed with counts.

Error Type	**Description**
Wrong Title and Author	**Definition:** Model returns an unrelated, but often famous, title-author pair. **Example:** "title": "Altered Carbon", "author": "Richard K. Morgan" **Correct answer** Dune **Model** Olmo2-1124-13B-Instruct **Task:** Direct Probe
Correct Author, Wrong Title	**Definition:** Author is correctly identified, but the title is incorrect.[35] **Example:** "title": "Dune Messiah", "author": "Frank Herbert" **Correct:** "title": "Dune","author":"Frank Herbert" **Model** Olmo2-1124-13B-Instruct **Task:** Direct Probe
Refusal or Abstention	**Definition:** Model fails to make a guess, returning "Unknown" or similar. **Example:** "title": "Book name: Unknown", "author": "Unknown author" **Correct:** title: Dune author : Frank Herbert **Model**: Llama-3.1-8B-Instruct **Task:** Direct Probe
Wording or Stylistic Errors	**Definition:** Title is misworded, reformatted, or awkwardly phrased. **Example:** """title"": ""Nineteen Eighty-Four"", ""author"": ""George Orwell""" **Correct Answer:** title : 1984, author: George Orwell **Model** Gpt-4o-audio-preview **Task:** Direct Probe
Incorrect Entity from Same Book	**Definition:** Returns a different character from the same book. **Example:** Charles **Correct Answer**: Mr.Lorry **Model:** Llama3.1-405b **Task:** Name Cloze
Culturally Popular but Incorrect Name	**Definition:** Model selects an incorrect name which is specific to the culture of the passage language. **Example:** "Ataturk" **Correct answer:** Winston **Model** : LLama-3.3-70B **Task:** Name Cloze
Multi-Guess Output	**Definition:** Model provides multiple candidates or alternative guesses. **Example:** Model response: Based on the context of the passage, I'm going to take a guess that the proper name that fills the [MASK] token is: Fahrenheit. However, this seems unlikely, as "Fahrenheit" is a title of a book, not a character's name. A more plausible guess would be a character from a dystopian novel, such as "Fahrenheit 451". Mildred **Correct Answer:** Hermione **Model:** Llama3.1-405b **Task:** Name Cloze
Broken or Corrupted Output	**Definition:** Model outputs unreadable, fragmented, or nonsensical tokens. **Example:** "title": ".kハウス absorbing richesて舟へ.", "author": "魚つり人才育てる(Balls to Become a Fishing Pro 魚つり人才を作り出す!)" **Correct Answer:** Marianne **Model:** Qwen-2.5-Omni-7b **Task:** Both
Honorific or Pronoun returned	**Definition:** Model outputs a Honorific or Pronoun instead of entity **Example:** Mr. **Correct Answer:** Mr. Darcy **Model:** Llama-3.1-8B-Instruct **Task:** Both

Table 18: Defined error types with descriptions, examples, and applicable tasks