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

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

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 non-English languages or transfers across languages remains unclear. This paper investigates multilingual and cross-lingual memorization in LLMs, probing if memorized content in one language (e.g., English) can be recalled when presented in translation. To do so, we introduce OWL, a dataset of **31.5K** aligned excerpts from 20 books in ten languages, including English originals, 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 LLMs consistently recall content across languages, even for texts without direct translation in pretraining data. GPT-4o, for example, identifies authors and titles 69% of the time and masked entities 6% of the time in newly translated excerpts. Perturbations (e.g., masking characters, shuffling words) modestly reduce direct probing accuracy (7% drop for shuffled official translations). Our results highlight the extent of cross-lingual memorization and provide insights on the differences between the models.

[*]Equal contribution [1]Manning College of Computer Sciences, University of Massachusetts Amherst, Massachusetts, United States of America [2]University of Maryland College Park, Maryland, United States of America [3]Microsoft, Washington, United States of America. Correspondence to: Chau Minh Pham <chau@umd.edu>.

Published at ICML 2025 Workshop on the Impact of Memorization on Trustworthy Foundation Models, Vancouver, Canada. PMLR 267, 2025.

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 & 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 existing 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 exist.

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 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% aver-

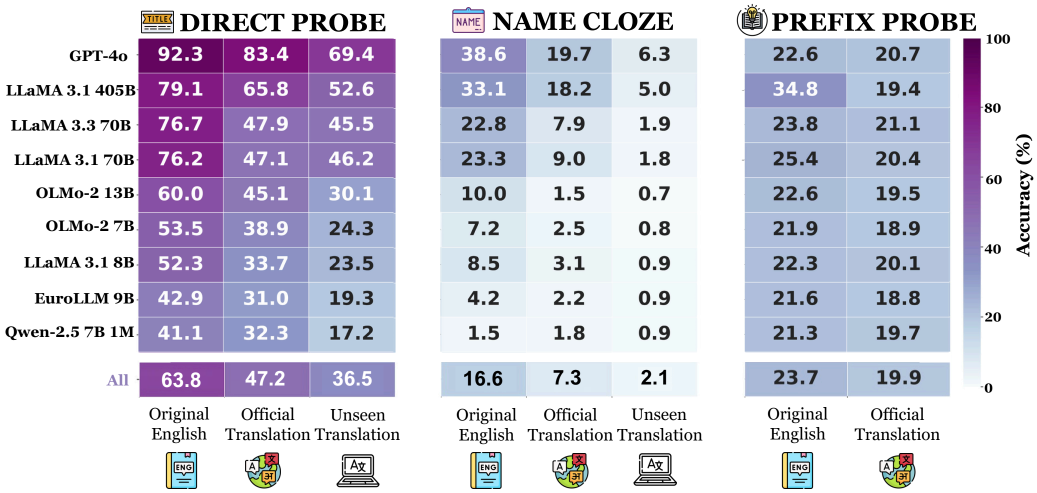


Figure 1. **Overall performance:** GPT-4o consistently outperforms other models in probing tasks, followed by LLaMA 405B. Direct and prefix probing use unmasked passages, while name cloze uses masked ones without named entities.[2]

aged 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 would indicate a high degree of cross-lingual knowledge transfer from English or other high-resource languages.[3] Notably, models demonstrate the ability to recall information even for the newly translated texts. GPT-4o, for instance, correctly identifies author and book information 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).

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%). 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).

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

2. Experiments

Test data: Unless noted otherwise, we evaluate on *(1) original English data* (recall of likely-seen content), *(2) official translations* (baseline on other high-resource languages), *(3) unseen translations* (cross-lingual transfer), and *(4) English audio data* (audio vs. text). We also run experiments on newly published books, as a baseline.

2.1. 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. In the cross-modal setup, we provide the audio of the passage.

Metric: We measure accuracy by comparing (`author`, `title`) predictions against ground truth.[4] For cross-lingual experiments, we prompt the model to respond in English.

Ablations: To measure performance robustness, we introduce three additional task variations (see Table 2):

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

[4]We allow for minor formatting/diacritic differences by normalizing special characters and applying 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) A prediction is considered correct if the model identifies the correct author and book title (either in English or the passage language).

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 (§2.2), 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 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.

2.2. Experiment 2: Name Cloze

Task: In the name cloze task, we reuse the same passages from §2.1, each containing exactly one character name, and replace that name with `[MASK]` token to test recall (Chang et al., 2023).[5] 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.[6] Ground-truth entities 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 test the robustness of models by shuffling the words within each passage (as in §2.1), 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.

2.3. 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.

[5] Unlike Chang et al. (2023), we do not restrict passages to have only one character name or limit the passage length to allow for more realistic texts and analysis of passage-length effects.

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

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

2.4. Audio ablation

To explore whether models can recall memorized textual content when presented with different modalities (audio), we adapt three core experiments above. Text-specific ablations are excluded. Due to higher text-to-speech (TTS) quality, all audio experiments are limited to English, with models receiving textual instructions and providing textual responses.

2.5. Models

We test a set of open-weights and closed-source models: 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-7B and 13-B (OLMo et al., 2024), EuroLLM (Martins et al., 2025), as well as GPT-4o (OpenAI, 2024).[7] For audio experiments, we use GPT-4o-audio and Qwen2.5-Omni-7B (Xu et al., 2025). We also run our experiments on the quantized versions of LLama-3.1-70B-Instruct and Llama-3.1-8B-Instruct (Table 5).[8]

3. Results

LLMs can recognize official translations LLMs can identify passages from English novels with an average accuracy of 63.8%, though performance varies by model (Figure 1). While recognition is lower for translated texts, the performance remains substantial, especially among larger models. GPT-4o, for instance, reaches 83.4% accuracy in direct probing of translations. This high 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 9 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 a considerable degree of 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 pretraining. In contrast, the non-trivial performance on OWL suggests that the models are exposed to the content of the original books during training.

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

[8] Quantized models are obtained from NeuralMagic.

Cross-lingual knowledge transfers without explicit translation supervision Despite not being pretrained on book translations in low-resource languages, models show non-trivial performance on previously unseen excerpts in these languages (Table D). In direct probing, Sesotho yields the strongest results: GPT-4o achieves 76.9% accuracy, while Qwen-2.5-7B-1M scores above 18%. Even with Maithili, the lowest-performing language, GPT-4o reaches 66.5% accuracy, with LLaMA-3.1-405B close behind at 46.7% Name cloze results are lower but still above zero, ranging from 0.1% (OLMo-2-13B on Maithili) to 10.5% (GPT-4o on the same language). Notably, OLMo scores above 22% on st, yo, mg, tn, and ty for direct probing even though its pretraining reportedly uses only English. This performance implies that a meaningful degree of crosslingual transfer can emerge even when the target languages are scarcely represented, if at all, in the pretraining data.[9]

LLMs can recall knowledge even when probed in a different modality Qwen-Omni and GPT-4o-Audio show none-zero performance on the direct probing and name cloze tasks, even when the book content is presented as audio (§A.1). Compared to the textual data, performance by both models degrade only by a small amount. Specifically, GPT-4o-Audio achieves up to 75.5% accuracy on the direct probing task, while Qwen-Omni reaches 20.6%. 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 recall information across modalities.[10]

Shuffling inputs only moderately reduce 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, declines around 6-7% for official translation and 3-8% for unseen translations across all excerpt types. A similar trend can be observed for name cloze (Figure 5), where the gap between standard and standard performance can be as low as 1.1% for unseen translation and as high as 11.7% for English texts. The moderate drop shows models handle minor edits but still stumble on superficial rewordings.

LLaMA-3.1-70B's performance degrades more under 8-bit than under 4-bit quantization While LLaMA-3.1-70B maintains relatively stable accuracy at 4-bit precision, it experiences notable performance drops (up to -25%) when quantized to 8 bits (Figure 17).[11] The smaller LLaMA-3.1-8B's performance remains within 1% of the BF16 baseline at 8-bit precision, with noticeable degradation appearing only under 4-bit quantization. These results 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 (see §E).

4. 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 & Schwartz, 2022), improbable token prediction (Lee et al., 2022; Radhakrishnan et al., 2019), and membership inference attacks (Carlini et al., 2021a; Golchin & Surdeanu, 2024; Song & Shmatikov, 2019; Shokri et al., 2017; Asai et al., 2020; Stoehr et al., 2024). Early probing experiments, which are largely monolingual and clozestyle (Tirumala et al., 2022; Chang et al., 2023), have been complemented by theoretical work showing that memorized outliers can steer the model's learning trajectory (Allen-Zhu & Li, 2024).

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 & Lee, 2020) show that models can achieve high performance by exploiting shallow or dataset-specific cues. Our work relates closest to Goldman et al. (2025), who measures cross-lingual transfer by evaluating LLMs through Wikipedia entries across languages.

5. Conclusion

In this study, we demonstrate that LLMs exhibit substantial multilingual and cross-lingual memorization capabilities through probing experiments on aligned book excerpts across ten languages. We also find that performance is only modestly impacted by input perturbations such as word shuffling, audio-formatted passages, and character masking. We release our data and code to spur further research on cross-lingual generalization and LLM memorization.

[9] Recent court documents show that Llama models were likely trained on LibGen book data (all of our English and official translation books can be found in LibGen).

[10] §E shows greater in detail of overlapping correct answers in both modalities.

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

Impact Statement

Our study explicitly evaluates whether LLMs recall specific passages from copyrighted books, using translated variants to test the boundaries of memorization across languages. This analysis 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,[12] but acknowledge that probing for memorization can implicate intellectual property rights, underscoring the need for transparency in training data sources and greater scrutiny of how multilingual capabilities may amplify copyright risks.

Legal implications We empirically characterize memorization patterns, but 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.

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, or named entity handling could affect the model's ability to recover factual content, especially in low-resource languages.

References

Allen-Zhu, Z. and Li, Y. 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, 2024.

Asai, A., Kasai, J., Clark, J., Lee, K., Choi, E., and Hajishirzi, H. Xor qa: Cross-lingual open-retrieval question answering. In *North American Chapter of the Association for Computational Linguistics*, 2020. URL https://api.semanticscholar.org/CorpusID:225040672.

Asai, A., Kasai, J., Clark, J., Lee, K., Choi, E., and Hajishirzi, H. XOR QA: Cross-lingual open-retrieval question answering. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y. (eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 547–564, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.46. URL https://aclanthology.org/2021.naacl-main.46/.

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

Baltrusaitis, T., Ahuja, C., and Morency, L.-P. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423443, February 2019. ISSN 0162-8828. doi: 10.1109/TPAMI.2018.2798607. URL https://doi.org/10.1109/TPAMI.2018.2798607.

Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., and Song, D. X. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *USENIX Security Symposium*, 2018. URL https://api.semanticscholar.org/CorpusID:170076423.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. Membership inference attacks from first principles. *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2021a. URL https://api.semanticscholar.org/CorpusID:244920593.

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., and Raffel, C. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pp. 2633–2650. USENIX Association, August 2021b. ISBN 978-1-939133-24-3. URL https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting.

Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., and Tramèr, F. Membership inference attacks from first principles. In *2022 IEEE Symposium on Security and Privacy (SP)*, pp. 1897–1914, 2022. doi: 10.1109/SP46214.2022.9833649.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., and Zhang, C. Quantifying memorization across neural language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=TatRHT_1cK.

Chang, K., Cramer, M., Soni, S., and Bamman, D. Speak, memory: An archaeology of books known to Chatgpt/GPT-4. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7312–7327, 2023.

Elangovan, A., He, J., and Verspoor, K. Memorization vs. generalization : Quantifying data leakage in NLP performance evaluation. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 1325–1335, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.113. URL https://aclanthology.org/2021.eacl-main.113/.

Elliott, D., Kiela, D., and Lazaridou, A. Multimodal learning and reasoning. In Birch, A. and Zuidema, W. (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, Berlin, Germany, August 2016. Association for Computational Linguistics. URL https://aclanthology.org/P16-5001/.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. OPTQ: Accurate quantization for generative pre-trained transformers. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=tcbBPnfwxS.

Golchin, S. and Surdeanu, M. Time travel in LLMs: Tracing data contamination in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=2Rwq6c3tvr.

Goldman, O., Shaham, U., Malkin, D., Eiger, S., Hassidim, A., Matias, Y., Maynez, J., Gilady, A. M., Riesa, J., Rijhwani, S., Rimell, L., Szpektor, I., Tsarfaty, R., and Eyal, M. Eclektic: a novel challenge set for evaluation of cross-lingual knowledge transfer. *ArXiv*, abs/2502.21228, 2025. URL https://api.semanticscholar.org/CorpusID:276725165.

Hartmann, V., Suri, A., Bindschaedler, V., Evans, D., Tople, S., and West, R. Sok: Memorization in general-purpose large language models, 2023. URL https://arxiv.org/abs/2310.18362.

Hessel, J. and Lee, L. Does my multimodal model learn cross-modal interactions? it‘s harder to tell than you might think! In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 861–877, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.62. URL https://aclanthology.org/2020.emnlp-main.62/.

Hexgrad. Kokoro-82m (revision d8b4fc7), 2025. URL https://huggingface.co/hexgrad/Kokoro-82M.

Huang, J., Yang, D., and Potts, C. Demystifying verbatim memorization in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10711–10732, 2024.

Huang, Z., Yu, P., and Allan, J. 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, pp. 10481056, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9781450394079. doi: 10.1145/3539597.3570468. URL https://doi.org/10.1145/3539597.3570468.

Jiang, Z., Anastasopoulos, A., Araki, J., Ding, H., and Neubig, G. X-FACTR: Multilingual factual knowledge retrieval from pretrained language models. In Webber, B., Cohn, T., He, Y., and Liu, Y. (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5943–5959, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.479. URL https://aclanthology.org/2020.emnlp-main.479/.

Karamolegkou, A., Li, J., Zhou, L., and Søgaard, A. Copyright violations and large language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.458. URL https://aclanthology.org/2023.emnlp-main.458/.

Kassner, N., Dufter, P., and Schütze, H. Multilingual LAMA: Investigating knowledge in multilingual pretrained language models. In Merlo, P., Tiedemann, J., and Tsarfaty, R. (eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3250–3258, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.284. URL https://aclanthology.org/2021.eacl-main.284/.

Kurtic, E., Marques, A., Pandit, S., Kurtz, M., and Alistarh, D. "give me bf16 or give me death"? accuracy-performance trade-offs in llm quantization, 2025. URL https://arxiv.org/abs/2411.02355.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626, 2023.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C., and Carlini, N. Deduplicating training data makes language models better. In Muresan, S., Nakov, P., and Villavicencio, A. (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8424–8445, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.577. URL https://aclanthology.org/2022.acl-long.577/.

Limkonchotiwat, P., Ponwitayarat, W., Udomcharoenchaikit, C., Chuangsuwanich, E., and Nutanong, S. CL-ReLKT: Cross-lingual language knowledge transfer for multilingual retrieval question answering. In Carpuat, M., de Marneffe, M.-C., and Meza Ruiz, I. V. (eds.), *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 2141–2155, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-naacl.165. URL https://aclanthology.org/2022.findings-naacl.165/.

Liu, J., Min, S., Zettlemoyer, L., Choi, Y., and Hajishirzi, H. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=u2vAyMeLMm.

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

Marchisio, K., Dash, S., Chen, H., Aumiller, D., Üstün, A., Hooker, S., and Ruder, S. How does quantization affect multilingual LLMs? In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15928–15947, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.935. URL https://aclanthology.org/2024.findings-emnlp.935/.

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J. G., Birch, A., and Martins, A. F. Eurollm: Multilingual language models for europe. In *Proceedings of the Second EuroHPC user day*, volume 255, pp. 53–62, 2025. doi: https://doi.org/10.1016/j.procs.2025.02.260. URL https://www.sciencedirect.com/science/article/pii/S1877050925006210.

Meta. The llama 3 herd of models. *ArXiv*, abs/2407.21783, 2024. URL https://api.semanticscholar.org/CorpusID:271571434.

Mittal, S., Kolluru, K., Chakrabarti, S., and Mausam. mOKB6: A multilingual open knowledge base completion benchmark. In Rogers, A., Boyd-Graber, J., and Okazaki, N. (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–214, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.19. URL https://aclanthology.org/2023.acl-short.19/.

mrfakename, Srivastav, V., Fourrier, C., Pouget, L., Lacombe, Y., main, Gandhi, S., Passos, A., and Cuenca, P. Tts arena 2.0: Benchmarking text-to-speech models in the wild. https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2, 2025.

OLMo, T., Walsh, P., Soldaini, L., Groeneveld, D., Lo, K., Arora, S., Bhagia, A., Gu, Y., Huang, S., Jordan, M., et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

OpenAI. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Petroni, F., Rocktäschel, T., Riedel, S., Lewis, P., Bakhtin, A., Wu, Y., and Miller, A. Language models as knowledge bases? In Inui, K., Jiang, J., Ng, V., and Wan, X. (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1250. URL https://aclanthology.org/D19-1250/.

Popović, M. chrF: character n-gram F-score for automatic MT evaluation. In Bojar, O., Chatterjee, R., Federmann, C., Haddow, B., Hokamp, C., Huck, M., Logacheva, V., and Pecina, P. (eds.), *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pp. 392–395, Lisbon, Portugal, September 2015. Association for Computational Linguistics. doi: 10.18653/v1/W15-3049. URL https://aclanthology.org/W15-3049/.

Post, M. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Belgium, Brussels, October 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/W18-6319.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., and Manning, C. D. 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, 2020.

Radhakrishnan, A., Belkin, M., and Uhler, C. Memorization in overparameterized autoencoders. In *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*, 2019. URL https://openreview.net/forum?id=BJeZw4B334.

Ravichander, A., Ghela, S., Wadden, D., and Choi, Y. Halogen: Fantastic llm hallucinations and where to find them, 2025. URL https://arxiv.org/abs/2501.08292.

Robinson, N., Ogayo, P., Mortensen, D. R., and Neubig, G. ChatGPT MT: Competitive for high- (but not low-) resource languages. In Koehn, P., Haddow, B., Kocmi, T., and Monz, C. (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 392–418, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.40. URL https://aclanthology.org/2023.wmt-1.40/.

Shi, W., Ajith, A., Xia, M., Huang, Y., Liu, D., Blevins, T., Chen, D., and Zettlemoyer, L. Detecting pretraining data from large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=zWqr3MQuNs.

Shokri, R., Stronati, M., Song, C., and Shmatikov, V. Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2017. doi: 10.1109/SP.2017.41.

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

Stoehr, N., Gordon, M., Zhang, C., and Lewis, O. Localizing paragraph memorization in language models. *arXiv preprint arXiv:2403.19851*, 2024.

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

Team, Q. Qwen2.5-1m: Deploy your own qwen with context length up to 1m tokens, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-1m/.

Thai, K., Karpinska, M., Krishna, K., Ray, B., Inghilleri, M., Wieting, J., and Iyyer, M. Exploring document-level literary machine translation with parallel paragraphs from world literature. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9882–9902, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.672. URL https://aclanthology.org/2022.emnlp-main.672/.

Tirumala, K., Markosyan, A. H., Zettlemoyer, L., and Aghajanyan, A. Memorization without overfitting: Analyzing the training dynamics of large language models. In Oh, A. H., Agarwal, A., Belgrave, D., and Cho, K. (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=u3vEuRr08MT.

Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., and Lin, J. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Yang, A., Yu, B., Li, C., Liu, D., Huang, F., Huang, H., Jiang, J., Tu, J., Zhang, J., Zhou, J., Lin, J., Dang, K., Yang, K., Yu, L., Li, M., Sun, M., Zhu, Q., Men, R., He, T., Xu, W., Yin, W., Yu, W., Qiu, X., Ren, X., Yang, X., Li, Y., Xu, Z., and Zhang, Z. Qwen2.5-1m technical report. *arXiv preprint arXiv:2501.15383*, 2025.

Yao, F., Zhuang, Y., Sun, Z., Xu, S., Kumar, A., and Shang, J. Data contamination can cross language barriers. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17864–17875, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.990. URL https://aclanthology.org/2024.emnlp-main.990/.

A. Accuracy

A.1. Ablations

Character names facilitate recall Figure 4 shows that including character names significantly boosts direct probing accuracy: 63.8% for English, 47.2% for official translations, and 36.5% for unseen ones. Masking these names sharply reduces accuracy to 33.6%, 13.3%, and 6.7%, respectively, which highlighting models' reliance on lexical cues like names and locations. Accuracy under masking is similar to passages without named entities, especially in translations (≤3% difference). Name absence leads to more uniform, lower performance, indicating models often depend on surface-level patterns rather than deep understanding.

Direct name-inquiry prompts consistently outperform name cloze-style queries Direct probing significantly outperforms name cloze queries across all models and languages (Figure 1). 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 probing, where the model has to recall the title or author in a question-answering format, is more aligned with LLM's strength.

A.2. Accuracy by Context Length

Accuracy tends to increase with the number of tokens in the context. As shown in Figure 6, 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 050 token range to the 100400+ 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 7). In contrast, performance on official translations improves by roughly 14%, while unseen translations show only modest gains of around 7%.

B. 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 (*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 (1,595 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 2), as listed below:

1. Curating books We collect English novels that are also officially translated into Spanish, Turkish, and Vietnamese.[13] 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 (§subsection 2.2) requires test samples to have at least one character name, we identify named characters in English passages using a multilingual NER pipeline (details in §subsection B.1). This allows us to isolate passages with a single, uniquely identifiable name for downstream tasks.

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,[14] and applying the Par3 aligner (Thai et al., 2022).

4. Filtering & quality control To filter out any misaligned passages, we apply length filter[15] and BLEU filter using SacreBLEU (Post, 2018) with add-one smooth-

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

[14] We use gpt-4o-2024-05-13 with temperature=0.3 and max_tokens=4000; refer to Figure 15 for details.

[15] We filter out alignments where one or more passages is significantly shorter/longer than the others

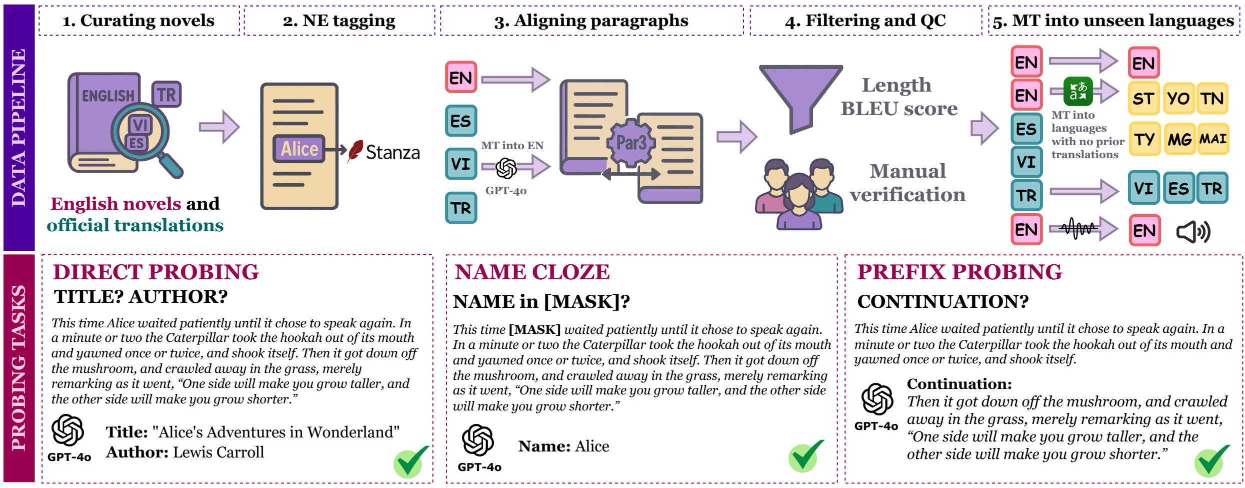


Figure 2. **Top:** OWL collection pipeline: **Bottom:** Probing tasksPrompt texts omitted for clarity (see Figure 12, Figure 13, Figure 14). Outputs are from GPT-4o. See Table 1 for an overview on our experiments.

Table 1. Overview of dataset splits, modalities, experiments, and expected outputs. "DP", "NC", and "PP" denote *direct probing*, *name cloze*, and *prefix probing* tasks, respectively.

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)

ing.[16] Finally, we manually verify all alignments, removing quadruples with misaligned passages or those with more than one unique character name (Figure 11). We compile two sets of passages: (1) a set containing exactly one unique character name[17] 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 (§subsection 2.1). Both sets have matching length distribution: original texts have 64.90 tokens on average, while texts without named characters have 59.03 tokens on average.[18] (Table 3). For each set of passages, we sample at most 100 passages per book to include in the final dataset.[19]

For the set with character names, we apply stratified sampling using character names to ensure a balanced distribution of character mentions. [20] Each language in the final dataset has 3,150 passages, including 1,594 passages with character names and 1,560 passages without character names.

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

[17]Character names can be repeated within the passage.

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

[19]We sample passages with at least 40 BPE tokens. View word count distribution in Figure 8.

[20]Chang et al. (2023)'s passages have an overrepresentation of common named characters like Alice, which makes it easier for models to get high accuracy. We address this bias by ensuring a more balanced distribution of character names.

Table 2. 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.

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 3. Token distribution in each passage type, calculated with OpenAI's `tiktoken` library (`o200k_base`).

Group	ORIGINAL						NO NAMED ENTITIES					
	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
Crosslingual	9564	78.91	60.0	11	642	59.98	9360	71.73	56.0	9	507	50.56

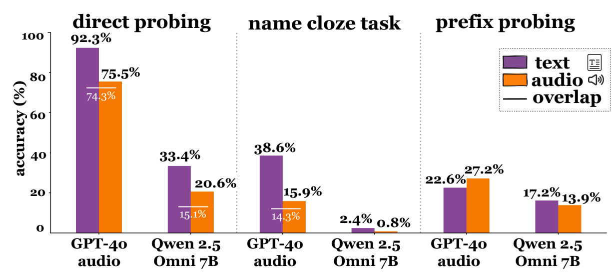


Figure 3. **Audio vs. Text:** Accuracy on text versus audio probing tasks in English standard setting. GPT-4o-audio exhibits substantial performance across all tasks in audio and text, Qwen Omni exhibits substantial performance on direct and prefix probing in audio and text.

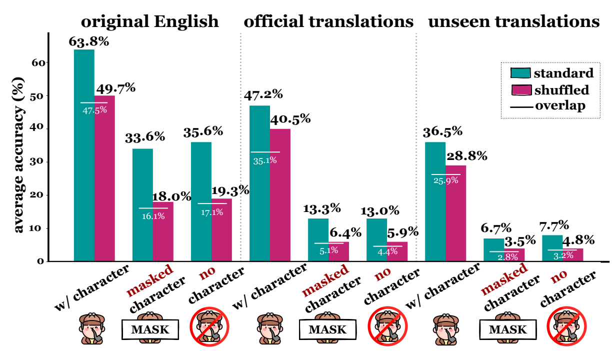


Figure 4. **Direct probing:** Average accuracy across models for shuffled versus standard text inputs. Accuracy decreases from standard to shuffled inputs across all perturbations and language settings, with non-trivial shuffled accuracy on English and official translations.

5. Machine translation into unseen 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*).[21] We use Microsoft Translator[22] to translate passages from English into each of the unseen languages.[23] We will be referring to this subset of data as *unseen translations*.

6. Creating audio data To evaluate cross-modal knowledge transfer, we first convert our textual data 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 prompt while enabling direct comparison between

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

[22]`https://www.microsoft.com/en-us/translator/`. We use Microsoft Translator rather than large language models (LLMs), as LLMs are unlikely to outperform traditional machine translation systems for low-resource languages due to limited training data (Robinson et al., 2023).

[23]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.

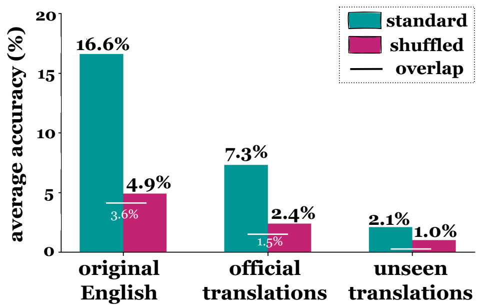


Figure 5. **Name cloze:** Unshuffled inputs outperform shuffled inputs across all language settings, with non-trivial accuracy on English and official translations.

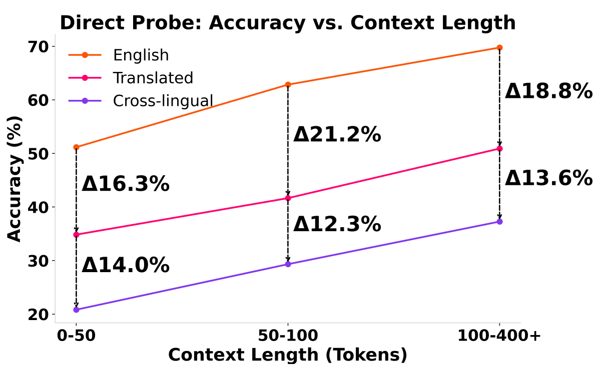


Figure 6. Direct probing accuracy across English texts, official translations, and unseen translations for different token ranges (0-50, 50-100, and 100-400+).

text and speech-based representations.[24] We convert the entire passage into audio for all tasks. For prefix probing (§subsection 2.3), we generate only the first half of the passages. We now expand on the extraction and alignment steps described above.

B.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[25].

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 previ-

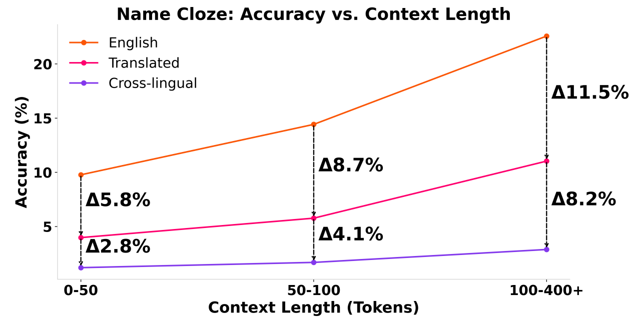


Figure 7. Name cloze accuracy across English texts, official translations, and unseen translations for different token ranges (0-50, 50-100, and 100-400+).

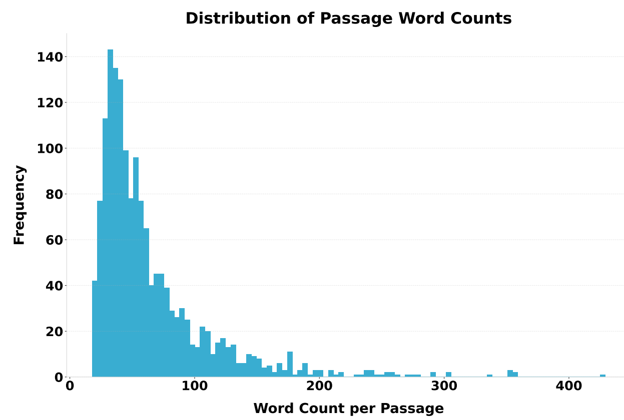


Figure 8. Word count distribution of unmasked passages in OWL

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

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

ously 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[26].

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 entities 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).

B.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[27] to confirm that translations into these languages are not already available.

Machine Translation pipeline: We implement a machine translation pipeline using Microsoft Translator. [28] 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.

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.

B.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[29] to annotate these excerpts, keeping only those for which there is unanimous agreement on validity (see Figure 11). All named entities 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 Malagasy.

B.4. OWL presence in training data

To study how the presence of OWL passages in the pretraining corpus of the model affects memorization, we searched the released corpus of OLMo using infinigram (Liu et al., 2024).

Out of 1,594 English OWL passages, 1,012 (63.5%) were found as exact matches in OLMo's corpus, with 292 (18.3%) partially found. In total, 82% of passages had a degree of presence, while 290 passages (18.2%) had no match (see Figure 9).

On direct probing, OLMo achieves 58.9% accuracy on English seen passages, compared to 47.2% on those unseen. This trend is consistent in translated passages: accuracy drops from 45.3% seen English to 29.1% unseen English for official translations, and from 31.1% seen English to 18.2% unseen English for machine translations (see Figure 10. Accuracy on unseen data may be due to partial exposure, such as seeing other passages from the same book, supporting recall. At the same time, OLMo achieves nontrivial accuracy on machine-translated passages that were not found in its pre-training data, strengthening our claim of cross-lingual knowledge transfer.

Each English passage was queried in the v4_olmo-2-1124-13b-instruct_llama index of OLMo's corpus using the Infini-Gram API. Each passage was first submitted in full, and a nonzero count was taken as an exact match. If no

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

[27] 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.

[28] 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.

[29] https://labelstud.io

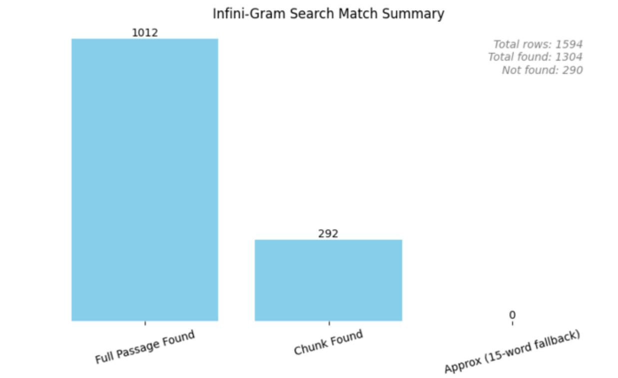


Figure 9. Distribution of match types for 1,594 English OWL passages searched in OLMos pretraining corpus using Infini-Gram

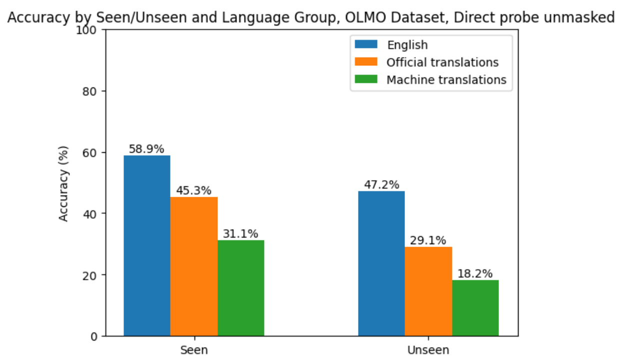


Figure 10. Direct probing accuracy of OLMo on OWL passages, grouped by whether the specific passage was found in the pretraining corpus ("Seen") or not ("Unseen")

exact match was found, a fallback sliding window procedure was employed, searching for chunks of the passage at varying word lengths: 40, 30, and 20 words with a stride of 5.

C. Prompts

In this section, we present the prompts used across our experiments. Figure 12 shows the prompt used for Direct Probing, Figure 13 shows the prompt for the Name Cloze Task, Figure 14 shows prompt used for Prefix Probing, and Figure 15 shows prompt used to translated non-English texts into English.

D. 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

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.

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

(e.g., per million input tokens), and total costs in USD for the experiments, along with notes on GPU usage for open-weight models.

E. Comparison of Quantized and Full-Precision Models

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). 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.

8-bit quantization (w8a16) causes significant degradation for LLaMA 3.170B, 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 3.18B exhibits relatively stable behavior across tasks and quantization settings. In Direct 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

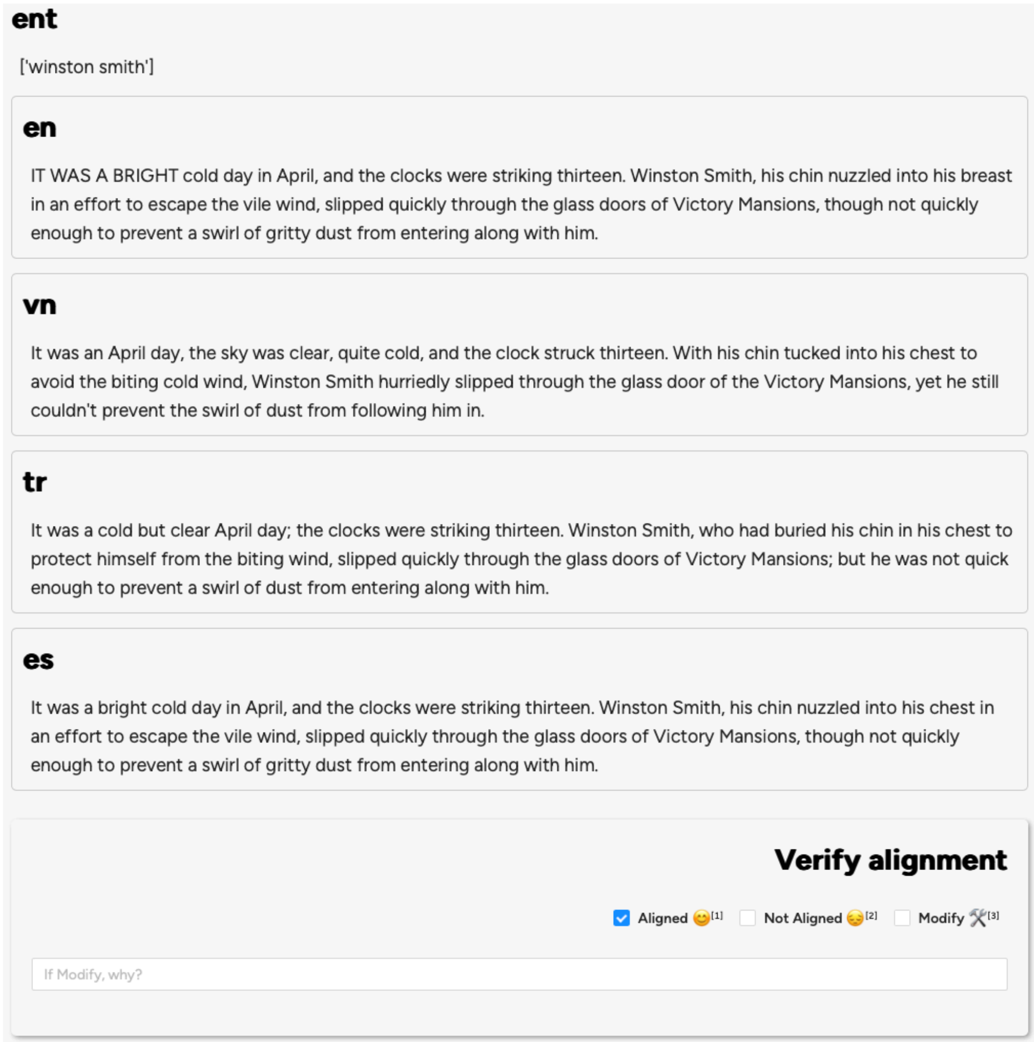


Figure 11. Annotation interface shown to annotators for verifying named entities and alignment consistency across languages. The passages shown under non-English languages are translated using gpt4o

(0.7 points) across all language groups. These results suggest that smaller models are more robust to quantization, and that 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 3.18B 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 degradation, particularly on English passages (-1.4).

F. Analysis of common errors

In direct probing, models occasionally named correct authors but misidentified book titles (10.61% [31] More often, they returned another popular book.

Abstention (responses like "unknown," "none," or empty strings) was notably high for EuroLLM (30.39%). For the name cloze task, the main error was returning an incorrect name (93%). These were sometimes culturally relevant to the passage's language (e.g., Spanish names for Spanish text) or other characters from the same book (Table 17). Models also returned pronouns (2%), honorifics (3%), abstained (0.1%), or repeated the "[MASK]" token (0.7%).

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

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 12. 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 13. 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 14. Prefix probing prompt. Given the beginning (prefix) of a passage, the model is prompted to generate its continuation (suffix).

Across all tasks, Qwen models frequently generated "broken text" (incoherently mixed languages) in 15.81% of outputs. More examples are given in Table 9.

G. Book-Level Accuracy Visualizations

To better understand how memorization patterns vary across individual titles, we visualize model performance

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 15. Prompt used to translate Vi, Es, and Tr book excerpts into English.

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.

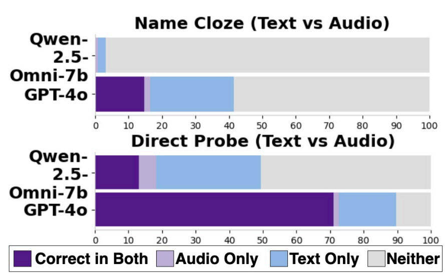
Model	Open Weights?	Inference Environment	Cost per Unit	Total Cost (USD)
GPT-4o (OpenAI, 2024)	🔒	OpenAI API	$2.50 / 1M input tokens	$156
GPT-4o-audio-preview (OpenAI, 2024)	🔒	OpenAI API	$40.00 / 1M audio tokens	$98
LLama-3.1-405b (Meta, 2024)	🔓	OpenRouter API	$2.50 / 1M input tokens	$300
LLama-3.1-8b (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 6. Metadata for books included in our OWL dataset

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**

at the book level for each probing task and setting. Figure 20, Figure 21, Figure 22, and Figure 23 display accuracy heatmaps for Direct Probing and Name Cloze, broken down by book title, language group, and model.

- **Figure 20** shows Direct Probing accuracy on unmasked passages containing character name.
- **Figure 21** reports accuracy when the named entity is masked from the passage.
- **Figure 22** displays Direct Probing accuracy on passages without named entities.
- **Figure 23** visualizes Name Cloze accuracy, where the model must recover the correct character name in a multiple-choice setting.



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*

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	Title (EN)	ES_Title	VI_Title	TR_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 Mt Tri Rc R	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	Tri Súc Vt	1945	1945	1950	1954	Yes	30164	42318	37072	56390	22398	48808	36580	47561
Bram Stoker	Dracula	Dracula	Dracula	Bá Tc Dracula	1897	1897	2006	1998	Yes	160277	215728	164910	255498	115279	221357	219100	266098
Frank Herbert	Dune	Dune	Dune	X cát	1965	1965	2009	1997	No	186476	304265	199058	354614	136096	328180	261793	407896
Ray Bradbury	Fahrenheit 451	Fahrenheit 451	Fahrenheit 451	451 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	Ca Chut và ca Ngi	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	Trm Nm Cô n	1967	1967	2003	1982	No	144517	158812	137795	164491	99790	211833	186705	198778
John Green	Paper Towns	Ciudades de papel	Kagittan Kentler	Nhng Thành Ph Giy	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	K Cp Tia Chp	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 sc	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	Bc Tranh Dorian Gray	1890	1891	2008	1971	Yes	78545	110952	77617	128029	57829	120590	100219	129334

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.

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

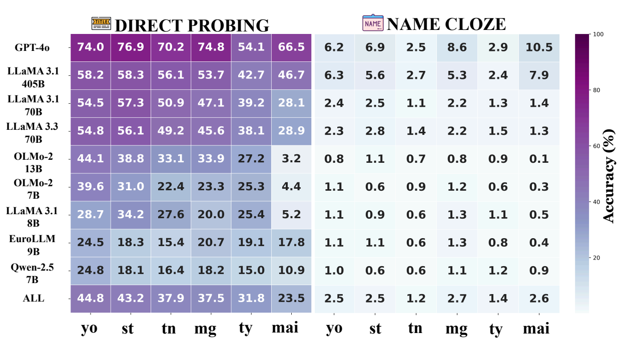


Figure 16. **Cross-lingual:** Accuracy on unseen translations by language. GPT-4o consistently outperforms other models on direct probing, followed by LLaMA 405B.

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.

H. Additional Limitations

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

Popularity versus performance Models might have higher performance on excerpts that appear frequently in the pretraining data. In a future iteration of this paper, we will analyze the occurrence frequency in the pretraining data in relation to model performance.

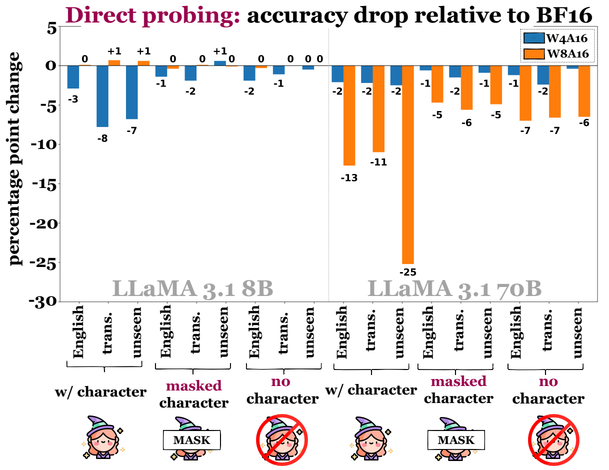
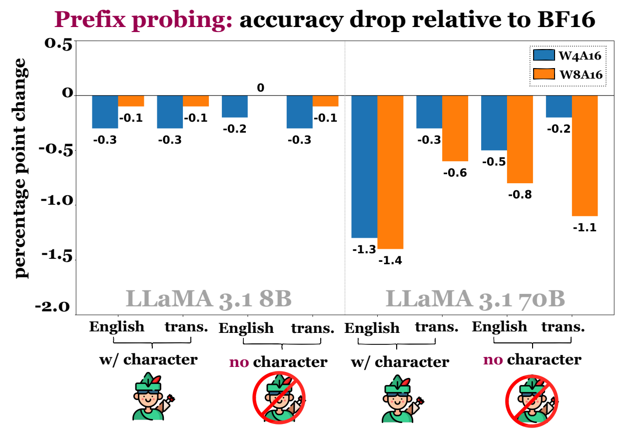


Figure 17. **Direct probing:** Percentage point drop in performance with respect to the performance of the BF16 baseline. We report drops for original English text ("English"), their official translations ("trans"), and unseen translations ("unseen"). The scores are reported across three conditions: (1) on passages containing a character name, (2) on passages where the name was masked, and (3) on passages without character name.



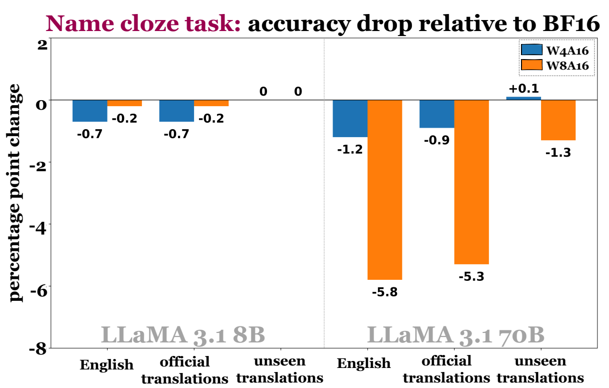


Figure 18. **Name cloze:** Percentage point drop relative to BF16 baseline. W8A16 quantization causes a significant accuracy drop in the name cloze task for the LLaMA 3.1 70B model, especially on English and officially translated data, compared to minimal impact on the 8B model.

Figure 19. **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.

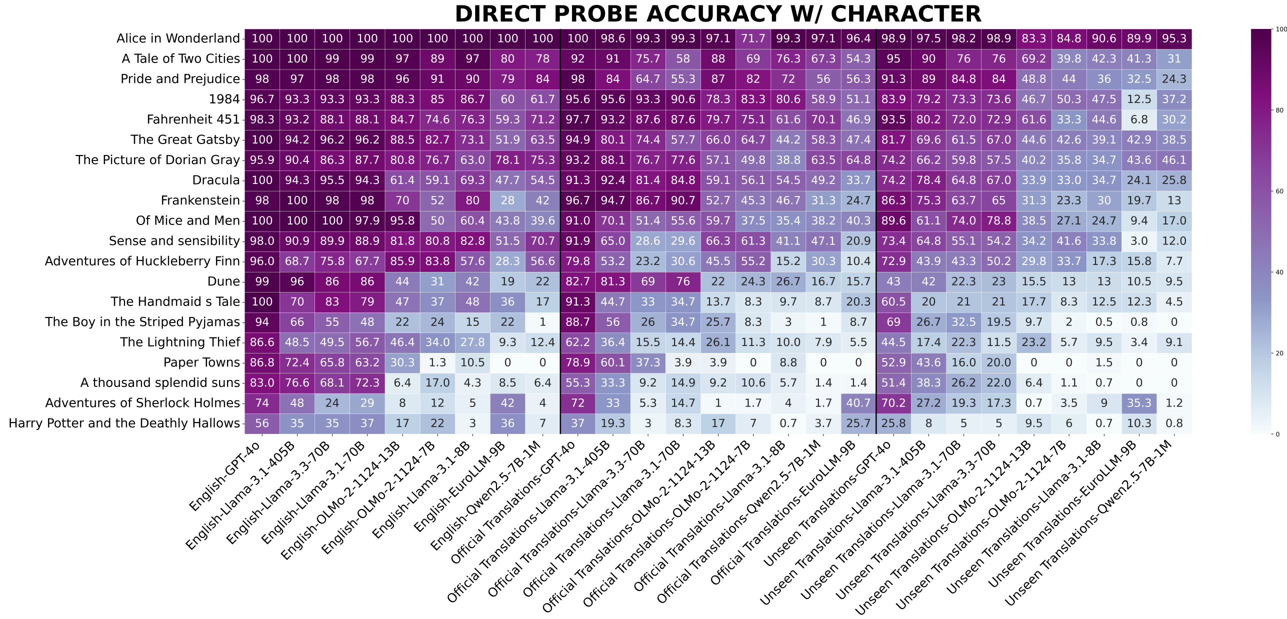


Figure 20. **Direct Probe** accuracy on unmasked passages containing named entities. Rows correspond to individual book titles, sorted top-to-bottom by average model performance. Columns represent language/model combinations grouped into three regions: English (left), Official Translations (center), and Unseen Translations (right). Accuracy is reported as a percentage.

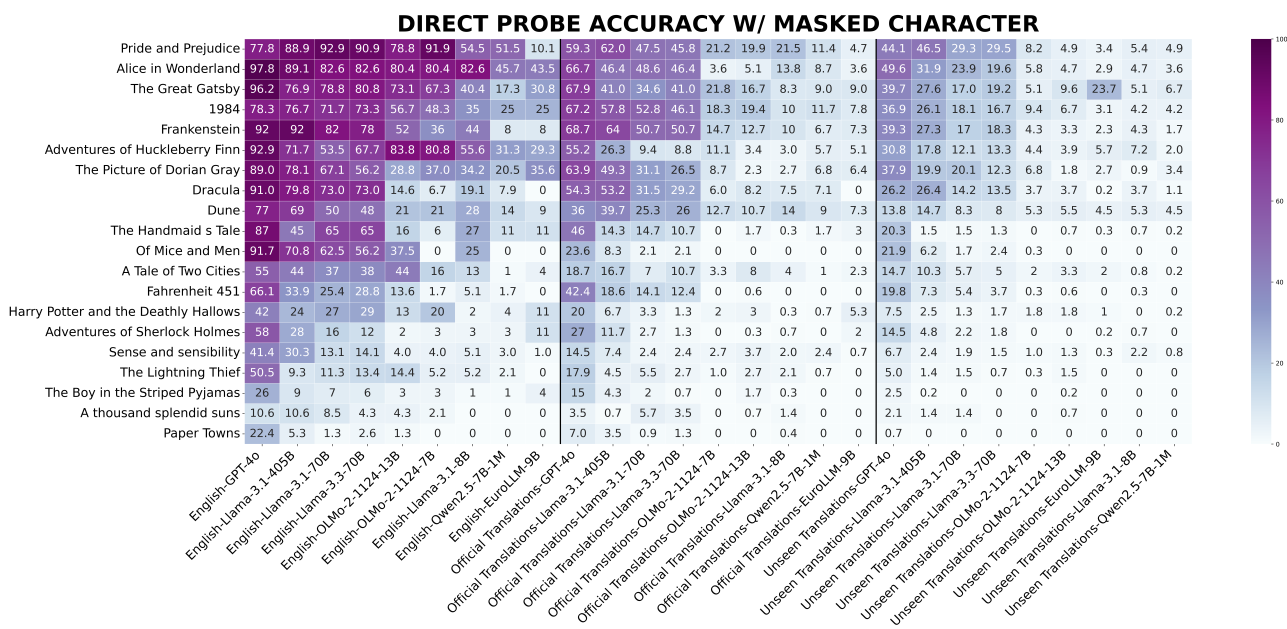


Figure 21. **Direct Probe** accuracy on masked passages where named entities have been replaced with a token. Books are sorted by overall average accuracy (top-to-bottom), and models are grouped by language setting: English, Official Translations, and Unseen Translations. Accuracy values are shown as percentages.

Table 9. Defined error types with descriptions, examples, and applicable tasks (Direct Probe, Name Cloze, or Both)

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.[30] **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 absorbing riches.", "author": "(Balls to Become a Fishing Pro !)"` **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

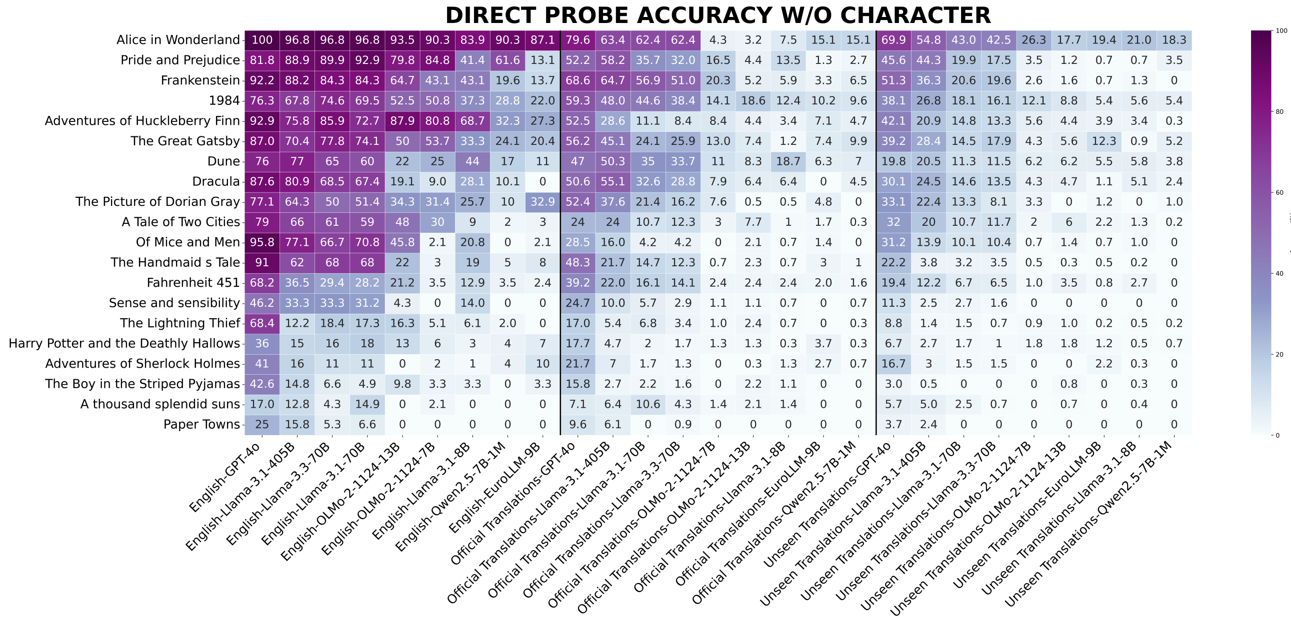


Figure 22. **Direct Probe** accuracy on passages with all named entities removed. Rows indicate books (sorted by average performance), and columns are grouped by language category: English, Offitial Translations, and Unseen Translations. Values represent accuracy percentages.

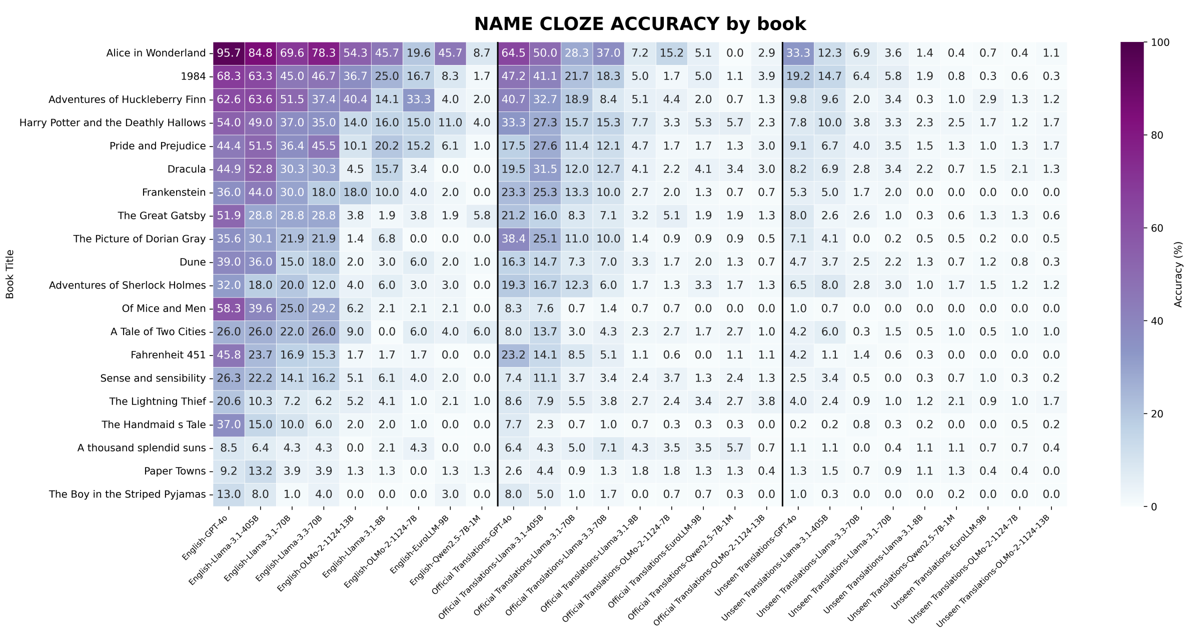


Figure 23. **Name Cloze** accuracy by book. Each row represents a title, and columns show performance across models grouped by language: English (left), Official Translations (center), and Unseen Translations (right). Accuracy is computed as the percentage of correct predictions in a multiple-choice setting.

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	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 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	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 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.

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 13. 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.

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. **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".

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

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	"Ak 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. **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.

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**: Top 4 incorrect names per language with their frequencies, aggregated over results from all models.

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	olivertwist	345	abraham	289
mg	hester	1720	andriamanitra	494	andriamanelo	354	dimmesdale	348
mai	hester	1949	hesttr	802	john	139	maarkttven	126
tn	hester	1763	john	472	morena	418	jesus	290
st	hester	2592	morena	623	joseph	456	job	198
ty	hester	2947	adam	534	teariki	466	jesus	432