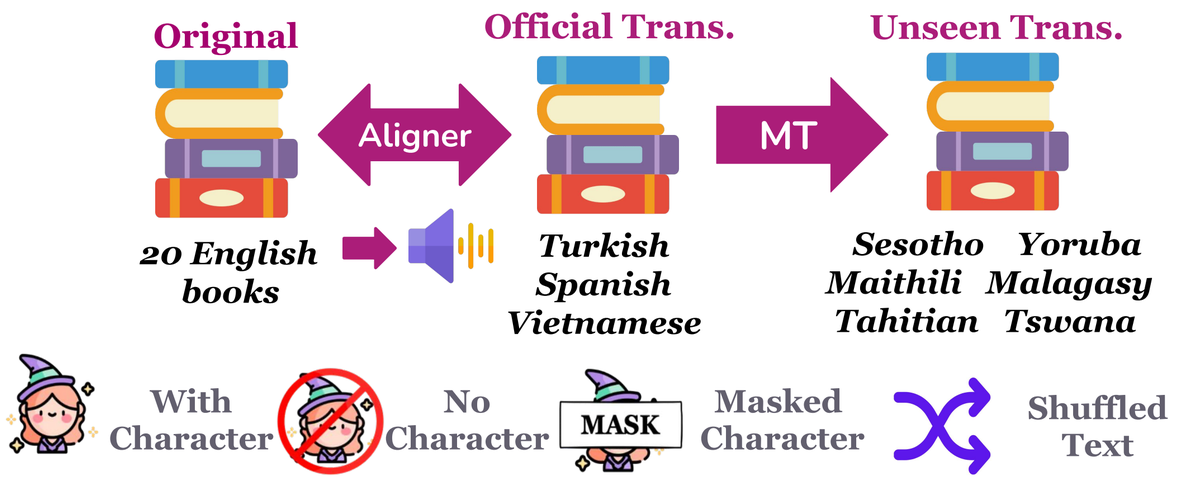
OWL: Probing Cross-Lingual Recall of Memorized Texts via World Literature

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

TL;DR Evaluating memorization of newly created translations & existing translations of English novels

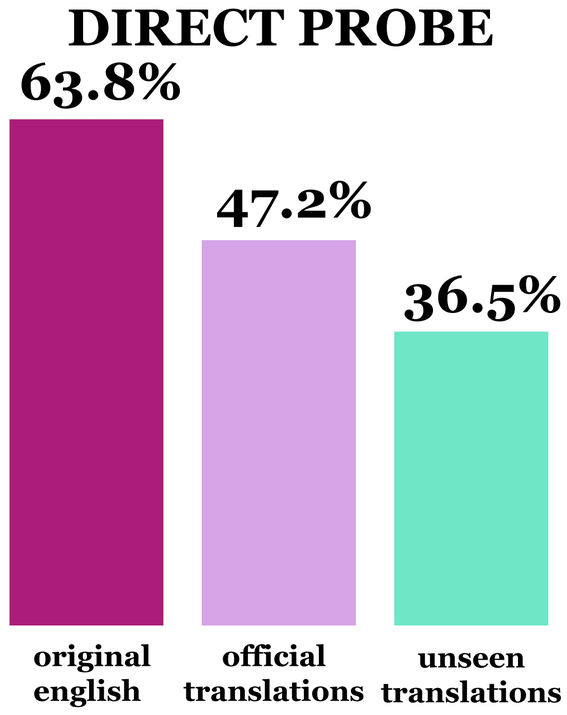
Data Pipeline

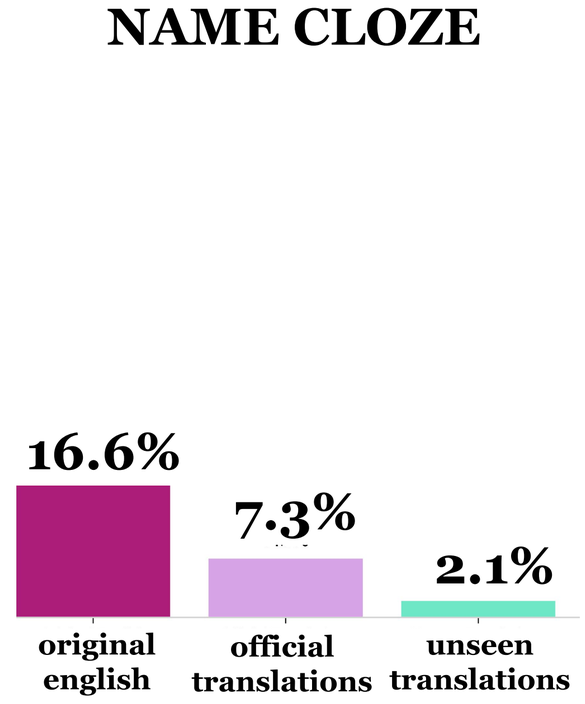
10 Languages | 31.5k Passages

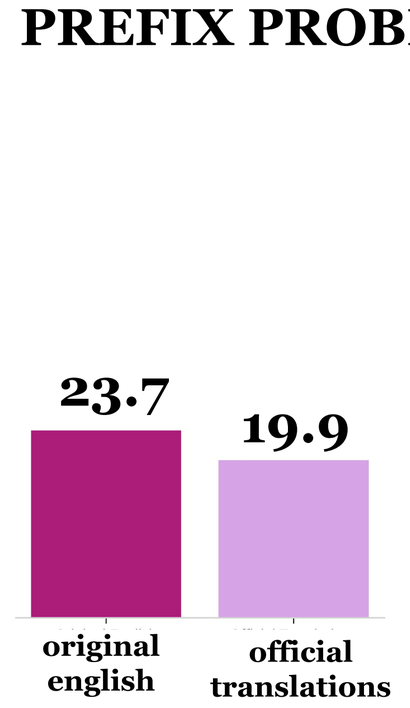


TASKS

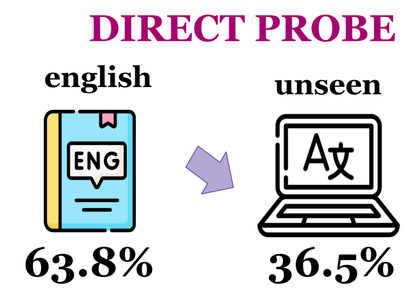
TASKS	INPUT	OUTPUT
DIRECT PROBE	*Stop him – stop him!" Hermione cried"...*	**Author:** J.K. Rowling **Book:** Harry Potter
NAME CLOZE	*"Stop him – stop him!"* **[MASK]** *cried...*	**Name:** Hermione
PREFIX PROBE	*"Stop him – stop him!"...*	**Continuation:** "Hermione cried..."







IN DATA unlikely to be seen



Passages in language where no published translation exist.

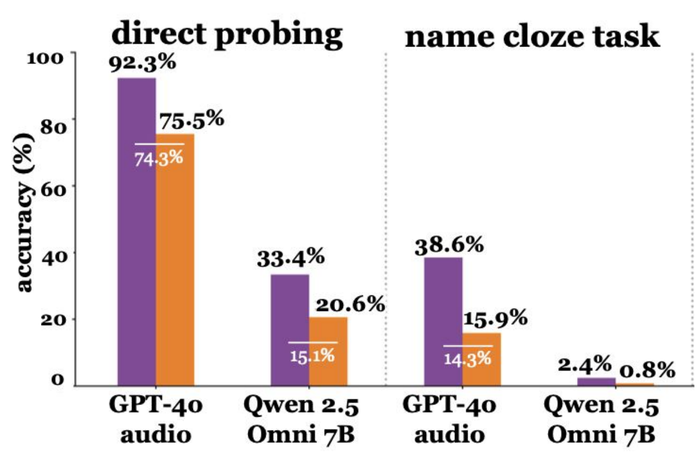
Non-trivial accuracy on such passages may indicate cross-lingual knowledge transfer.

accuracy by model

	DIRECT PROBE			NAME CLOZE			PREFIX PROBE	
	Original English	Official Translations	Unseen Translations	Original English	Official Translations	Unseen Translations	Original English	Official Translations
LLaMa 3.1 405B	79.1	65.8	52.6	33.1	18.2	5.0	34.8	19.4
GPT-4o	92.3	83.4	69.4	38.6	19.7	6.3	22.6	20.7
OLMo-2 13B	60.0	45.1	30.1	10.0	1.5	0.7	22.6	19.5
EuroLLM 9B	42.9	31.0	19.3	4.2	2.2	0.9	21.6	18.8
Qwen 2.5 7B 1M	41.1	32.3	17.2	1.5	1.8	0.9	21.3	19.7
All	63.8	47.2	36.5	16.6	7.3	2.1	23.7	19.9

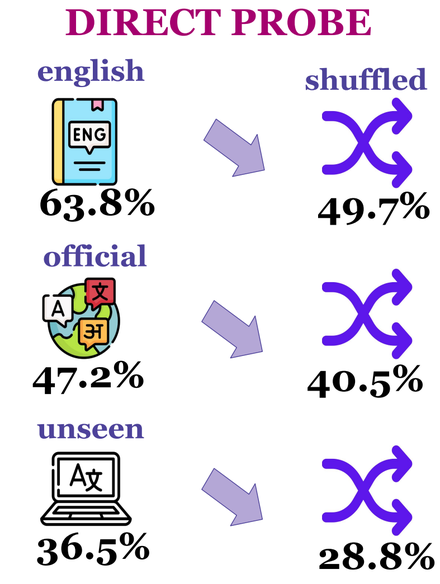
Accuracy (%)

across modalities



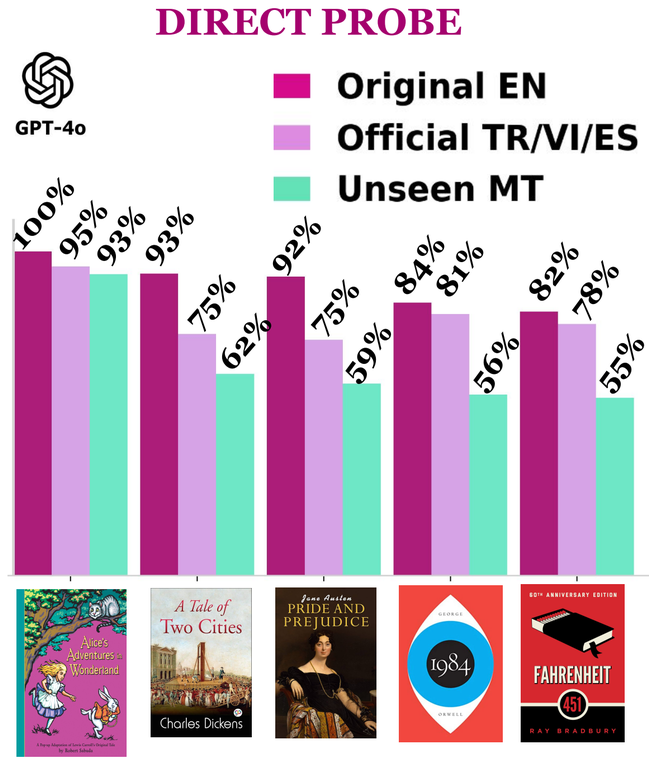
Audio Model accuracy drops from text, but remains **substantially** above zero.

upon shuffled data

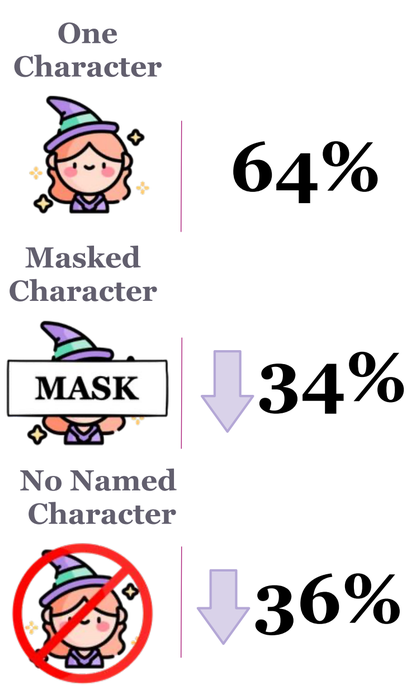


Low drop in recall despite the texts being shuffled.

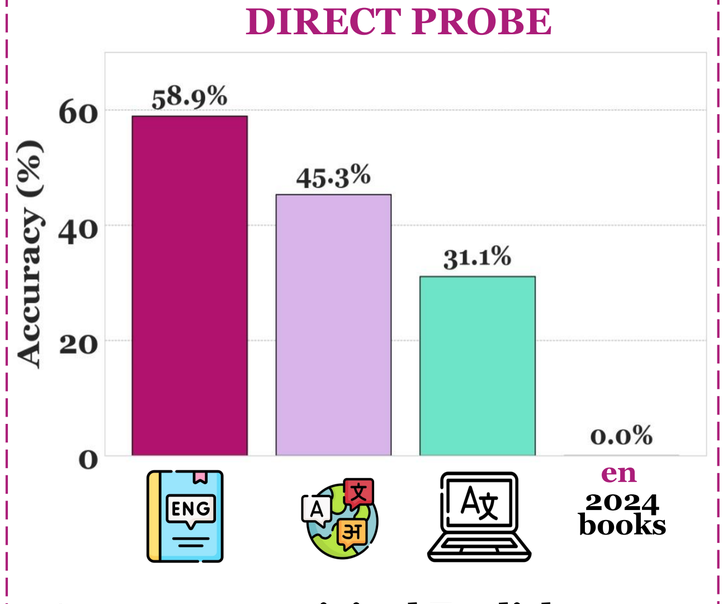
across books



w & w/o characters

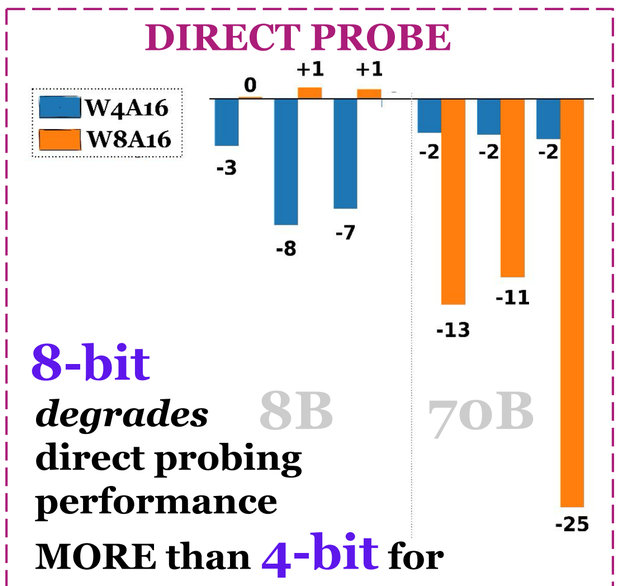


OLMo seen in corpus



Accuracy on original English passages from OLMo corpus vs their original and machine translations.

across quantizations



8-bit *degrades* direct probing performance MORE than **4-bit** for **LLaMA-3.1-70B**. Opposite to the trend seen LLaMA-8B model.

1 University of Massachusetts Amherst | 2 University of Maryland College Park | 3 Microsoft
* These authors contributed equally to this work.