PNAS RESEARCH ARTICLE | COMPUTER SCIENCES SOCIAL SCIENCES

Uncovering inequalities in new knowledge learning by large language models across different languages

Chenglong Wang[a,b], Haoyu Tang[c], Xiyuan Yang[d], Yueqi Xie[e,f,1], Jina Suh[g], Sunayana Sitaram[h], Junming Huang[f], Yu Xie[f,i], Pengjun Zhao[a,b,1], Zhaoya Gong[a,b,1], Xing Xie[j], and Fangzhao Wu[j,1]

Affiliations are included on p. 11.

Edited by Jeffrey Ullman, Stanford University, Stanford, CA; received June 7, 2025; accepted November 20, 2025

As large language models (LLMs) gradually demonstrate their potential to boost productivity and become integral tools for problem-solving in daily life worldwide, understanding the linguistic inequalities they introduce is becoming increasingly important. Prior research has primarily focused on static analyses of disparities in existing knowledge and capabilities of LLMs across languages. However, LLMs are continuously evolving, acquiring new knowledge to provide current, relevant responses and deliver precise, expert-level answers in specific domains. Investigating linguistic inequalities within this dynamic learning process is, therefore, also essential. In this paper, we explore inequalities in new knowledge learning by LLMs across different languages and four key dimensions: effectiveness, transferability, prioritization, and robustness. Through extensive experiments in both in-context learning and fine-tuning settings, with proprietary and open-source models, we reveal four key findings: 1) LLMs face greater challenges in efficiently and accurately learning new knowledge in lower-resource languages; 2) knowledge learned by LLMs tends to be more easily transferred to higher-resource languages than to lower-resource ones; 3) new knowledge in higher-resource languages is more likely to be retained and prioritized; and 4) LLMs are more robust against incorrect or misleading information in higher-resource languages. We further analyze the underlying causes of these inequalities from linguistic perspectives, pretraining characteristics, and tokenizer design, and propose a preliminary mitigation strategy through the lens of linguistic neurons. This work highlights the urgent need to recognize and address emerging linguistic inequalities in the development of LLMs.

linguistic inequality | large language models (LLMs) | knowledge acquisition

Large language models (LLMs), with their comprehensive knowledge storage, easy accessibility, and ability to handle a wide range of tasks, are increasingly being applied in various domains [e.g., education (1), medicine (2), scientific research (3, 4)] and in daily life, significantly boosting productivity (5). This transformation is both inevitable and global in scale. One notable example is ChatGPT, which, as of July 2025, serves 700 million weekly active users worldwide-a substantial portion of whom interact with LLMs in languages other than English (6–8). Given such widespread adoption, it is crucial to study fairness in multilingual environments to ensure that users of different languages can benefit equally from these systems (9).

Prior research on linguistic inequalities in LLMs has primarily examined static disparities in knowledge and capabilities across languages (10–15). For example, some studies have analyzed the amount of factual knowledge encoded in different languages and revealed significant variations. In particular, they show that knowledge available in low-resource languages remains limited due to the lack of pretraining data (16–18). These findings have advanced our understanding of how disparities in knowledge and capabilities across languages give rise to linguistic inequalities in LLMs. However, what remains underexplored is how such inequalities manifest in the dynamic process of learning new knowledge-a perspective that is becoming increasingly important as LLMs continue to evolve.

Learning new knowledge is crucial for LLMs, as illustrated in Fig. 1*A*. On the one hand, general-purpose LLMs are pretrained on static datasets collected before the models are released and therefore may not include real-time or recent information. As a result, their knowledge bases can quickly become outdated. To ensure that these models provide current and relevant responses, it is essential to continuously integrate new knowledge. On the other hand, although pretrained LLMs are trained on diverse and extensive datasets,

Significance

Large language models (LLMs) are transforming daily life, yet users across different languages may not benefit equally. Our study shows that LLMs face greater challenges in learning new knowledge and resisting incorrect or misleading information in lower-resource languages. Consequently, users of these languages are at higher risk of receiving lower-quality or misleading outputs. Moreover, knowledge is more easily transferred to, and more likely to be retained and prioritized in, higher-resource languages. As AI systems become increasingly integrated into society, such disparities threaten to marginalize lower-resource languages and deepen global information inequality. We also analyze their underlying causes and propose preliminary strategies for mitigation. Addressing these inequalities is essential to build AI systems that are fair, inclusive, and socially responsible.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1] To whom correspondence may be addressed. Email: yxieay@connect.ust.hk, pengjun.zhao@pku.edu.cn, z.gong@pku.edu.cn, or fangzwu@microsoft.com.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2514626122/-/DCSupplemental.

Published December 18, 2025.

they often lack depth in specialized domains. Acquiring domain-specific knowledge enables LLMs to deliver more precise, expert-level answers in those areas. As depicted in Fig. 1*B*, two primary techniques have been developed and widely adopted to enhance LLMs with new knowledge (19). First, through in-context learning, LLMs can acquire new information from examples, instructions, or knowledge retrieved from external databases-all without requiring parameter updates (20). Second, fine-tuning LLMs on specific datasets or tasks allows them to gain knowledge tailored to particular needs (21). A practical example of this is ChatGPT's fine-tuning API, which enables users to customize the model for specialized purposes (22).

In this study, we aim to reveal, analyze, and mitigate linguistic inequalities in new knowledge learning by LLMs. We first conceptualize these inequalities along four key dimensions-effectiveness, transferability, prioritization, and robustness-and propose a comprehensive evaluation framework. Specifically, we investigate the following research questions under two settings (in-context learning and fine-tuning): 1) Can LLMs learn new knowledge equally effectively across different languages in terms of efficiency and accuracy? 2) Can the knowledge acquired by LLMs be transferred equally across languages? 3) When new knowledge in two languages conflicts, can LLMs treat them equally? and 4) When exposed to incorrect or misleading information, can LLMs resist such errors equally across languages? We then seek to identify the underlying causes of these linguistic inequalities based on linguistic knowledge, pretraining characteristics, and tokenizer design. We also leverage the identified features to model knowledge transferability across languages. Finally, through the lens of linguistic neurons (23, 24), we explore their overlaps, examine how they relate to cross-lingual knowledge transfer, and investigate how intervening in these neurons can help mitigate linguistic inequalities.

To conduct the above study, we selected 19 languages that differ in their language properties, including 7 high-resource, 5 medium-resource, and 7 low-resource languages. Additionally, we constructed four multilingual parallel datasets covering both new knowledge (fictional and real) and general knowledge (generated and human-created). For the new knowledge datasets-both hypothetical question–answer pairs set in a future world and real-world medical knowledge-LLMs generally struggle to provide accurate answers in any language, which enables us to examine inequalities in the new knowledge learning process. For the general knowledge datasets-both generated and human-created-LLMs can accurately answer most questions across languages, which allows us to explore how they resist errors in different linguistic contexts.

Extensive experiments were conducted on both proprietary and open-source models (GPT-4o-Mini, Llama-3.1-8B, Qwen3-8B, and Aya-Expanse-8B). These experiments reveal four key linguistic inequalities introduced by LLMs (Fig. 1*C*): 1) compared to higher-resource languages, LLMs face greater challenges in learning new knowledge in lower-resource languages in terms of both efficiency and accuracy; 2) new knowledge acquired by LLMs can be more easily transferred to higher-resource languages than to lower-resource ones; 3) when new knowledge in two languages conflicts, knowledge in higher-resource languages tends to be prioritized; and 4) LLMs are generally more resistant to incorrect knowledge in higher-resource languages than in lower-resource languages. Through exploratory analyses as well as linear and nonlinear modeling, we identify several factors that influence these inequalities. Specifically, linguistic characteristics (e.g., differences in phylogeny, syntax, and geographical distribution), pretraining data proportions and tailored optimizations, and tokenizer quality all play crucial roles in shaping cross-lingual differences in model performance and determining the transferability of knowledge. Leveraging these features, we are able to predict cross-lingual knowledge transferability with relatively high accuracy. Furthermore, we uncover a relationship between overlaps in linguistic neurons and the transferability of knowledge across languages. By intervening on these neurons-for instance, through activation-we demonstrate a potential pathway to mitigating such inequalities. Overall, our study shows that in the context of new knowledge acquisition, higher-resource languages consistently exhibit superiority over lower-resource languages across the four dimensions of effectiveness, transferability, prioritization, and robustness. Coupled with the underrepresentation of lower-resource languages in existing knowledge and capabilities of LLMs (16–18), these results highlight the persistence and potential widening of linguistic inequalities. Addressing these inequalities and ensuring multilingual knowledge equality are critical in the continued development of LLMs to foster responsible and inclusive AI.

Results

Language and Model Selection. To investigate linguistic inequalities, we adopted two criteria for language selection. First, we included a balanced number of high-, medium-, and low-resource languages to examine how LLMs perform across different resource levels. Second, we ensured variation in linguistic characteristics, such as phylogeny, syntax, phonology, inventory, and geographical distribution, to better assess potential factors contributing to inequalities. Specifically, following prior research in multilingual natural language processing (NLP) (25), we classified languages into resource levels based on their proportions in the CommonCrawl corpus, which was used to pretrain GPT-3 (26). A language is considered high-resource if its data ratio exceeds 1%, medium-resource if it falls between 0.1% and 1%, and low-resource if it is 0.1% or below. A language can still be categorized as medium- or low-resource even if it is spoken by many people, as long as digital and annotated data remain limited. For example, Hindi (609.1 million speakers), Tamil (86.3 million speakers), and Swahili (87.2 million speakers) are categorized as medium- or low-resource, whereas Italian, which has only 66.2 million speakers, is considered high-resource due to its greater digital presence (27). Investigating inequalities across high-, medium-, and low-resource languages is therefore both meaningful and necessary. As shown in *SI Appendix*, Table S1, our selected set includes 7 high-resource languages (English, Chinese, Japanese, French, Spanish, Italian, Portuguese), 5 medium-resource languages (Swedish, Korean, Danish, Thai, Hindi), and 7 low-resource languages (Tamil, Mongolian, Welsh, Swahili, Turkmen, Scottish Gaelic, Zulu). In addition, we included 4 multilingual LLMs-both proprietary and open-source-GPT-4o-Mini, Llama-3.1-8B, Qwen-3-8B, and Aya-Expanse-8B. These models are widely adopted in daily life and officially support different languages (*SI Appendix*, Table S1) (28–30). This setup enables us to assess not only the prevalence of linguistic inequalities but also the effects of tailored optimizations on model outcomes.

Multilingual Parallel Dataset Construction. To investigate inequalities in new knowledge learning by LLMs across different languages, we carefully constructed four multilingual parallel datasets.

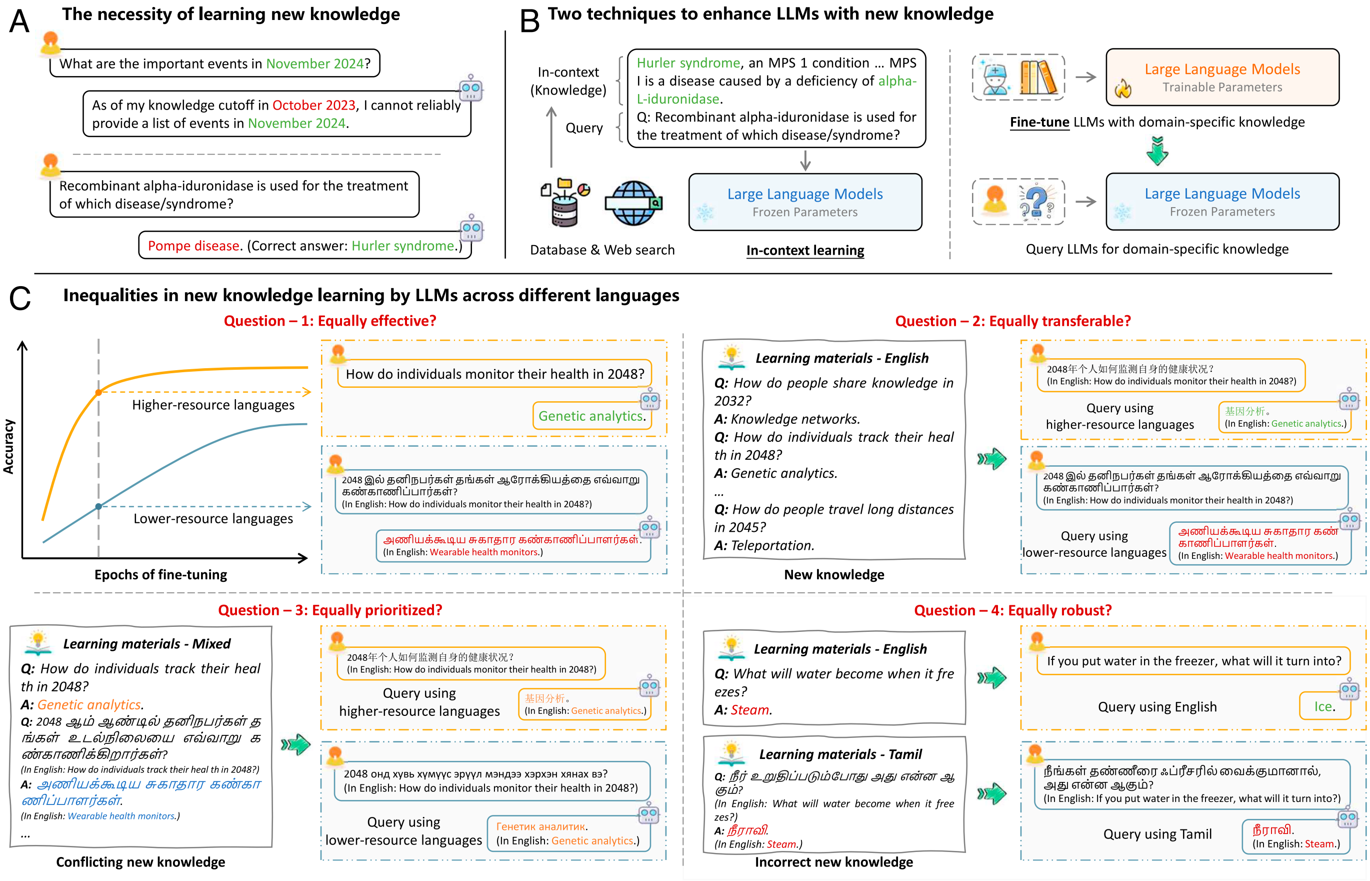

A
The necessity of learning new knowledge
What are the important events in November 2024?
As of my knowledge cutoff in October 2023, I cannot reliably provide a list of events in November 2024.
Recombinant alpha-iduronidase is used for the treatment of which disease/syndrome?
Pompe disease. (Correct answer: Hurler syndrome.)
B
Two techniques to enhance LLMs with new knowledge
In-context (Knowledge)
Hurler syndrome, an MPS 1 condition … MPS I is a disease caused by a deficiency of alpha-L-iduronidase.
Query
Q: Recombinant alpha-iduronidase is used for the treatment of which disease/syndrome?
Database & Web search
Large Language Models
Frozen Parameters
In-context learning
Large Language Models
Trainable Parameters
Fine-tune LLMs with domain-specific knowledge
Large Language Models
Frozen Parameters
Query LLMs for domain-specific knowledge
C
Inequalities in new knowledge learning by LLMs across different languages
Question – 1: Equally effective?
Accuracy
Higher-resource languages
Lower-resource languages
Epochs of fine-tuning
How do individuals monitor their health in 2048?
Genetic analytics.
2048 இல் தனிநபர்கள் தங்கள் ஆரோக்கியத்தை எவ்வாறு கண்காணிப்பார்கள்?
(In English: How do individuals monitor their health in 2048?)
அணியக்கூடிய சுகாதார கண்காணிப்பாளர்கள்.
(In English: Wearable health monitors.)
Question – 2: Equally transferable?
Learning materials - English
Q: How do people share knowledge in 2032?
A: Knowledge networks.
Q: How do individuals track their heal th in 2048?
A: Genetic analytics.
…
Q: How do people travel long distances in 2045?
A: Teleportation.
New knowledge
2048年个人如何监测自身的健康状况？
(In English: How do individuals monitor their health in 2048?)
Query using higher-resource languages
基因分析。
(In English: Genetic analytics.)
2048 இல் தனிநபர்கள் தங்கள் ஆரோக்கியத்தை எவ்வாறு கண்காணிப்பார்கள்?
(In English: How do individuals monitor their health in 2048?)
Query using lower-resource languages
அணியக்கூடிய சுகாதார கண்காணிப்பாளர்கள்.
(In English: Wearable health monitors.)
Question – 3: Equally prioritized?
Learning materials - Mixed
Q: How do individuals track their heal th in 2048?
A: Genetic analytics.
Q: 2048 ஆம் ஆண்டில் தனிநபர்கள் தங்கள் உடல்நிலையை எவ்வாறு கண்காணிக்கிறார்கள்?
(In English: How do individuals track their heal th in 2048?)
A: அணியக்கூடிய சுகாதார கண்காணிப்பாளர்கள்.
(In English: Wearable health monitors.)
…
Conflicting new knowledge
2048年个人如何监测自身的健康状况？
(In English: How do individuals monitor their health in 2048?)
Query using higher-resource languages
基因分析。
(In English: Genetic analytics.)
2048 онд хувь хүмүүс эрүүл мэндээ хэрхэн хянах вэ?
(In English: How do individuals monitor their health in 2048?)
Query using lower-resource languages
Генетик аналитик.
(In English: Genetic analytics.)
Question – 4: Equally robust?
Learning materials - English
Q: What will water become when it fre ezes?
A: Steam.
Learning materials - Tamil
Q: நீர் உறுதிப்படும்போது அது என்ன ஆகும்?
(In English: What will water become when it free zes?)
A: நீராவி.
(In English: Steam.)
Incorrect new knowledge
If you put water in the freezer, what will it turn into?
Query using English
Ice.
நீங்கள் தண்ணீரை ஃப்ரீசரில் வைக்குமானால், அது என்ன ஆகும்?
(In English: If you put water in the freezer, what will it turn into?)
Query using Tamil
நீராவி.
(In English: Steam.)

Fig. 1. (*A*) LLMs struggle to provide current, relevant responses and to deliver precise, expert-level answers in specific domains. (*B*) There are two techniques to enhance LLMs with new knowledge: in-context learning and fine-tuning. (*C*) Four key inequalities emerge in new knowledge learning by LLMs across different languages.

New knowledge datasets. To simulate real-world scenarios in which model developers enhance LLMs with new knowledge to provide timely and domain-specific responses, we created two datasets, each containing 200 question–answer pairs.[*] The first is a fictional new knowledge dataset generated by GPT-4o and set in a future world very different from the current one, which serves as a proxy for unseen information (*SI Appendix*, Table S10). To ensure diversity, prompts were conditioned on topics from the Information Coding Classification system (31), which covers almost all extant 6,500 knowledge fields (*SI Appendix*, Table S2 and Prompt S1). The second dataset focuses on medical knowledge and was filtered from MultiMedQA (32), a benchmark that integrates six medical question–answering datasets covering professional medicine, research, and consumer queries (*SI Appendix*, Table S11). The sampled items cover diverse topics and LLMs fail to provide accurate answers, which indicates that these samples are genuinely new knowledge for them. Moreover, many of the questions include long contexts (e.g., patient symptom descriptions), which makes them more representative of real use cases.

General knowledge datasets. We also constructed two general knowledge datasets, each with 100 question–answer pairs.[†] One was generated by GPT-4o, again using topic conditioning to ensure broad coverage (*SI Appendix*, Table S12 and Prompt S2). The other was created by humans (33) and contains questions and answers suitable for children aged 4 to 7 and students up to grade 7 (*SI Appendix*, Table S13).

Following standard practices in multilingual NLP (25), we translated all datasets into 18 additional languages using Google Translate.[‡] To assess translation quality, we conducted backtranslation and compared results with the original English pairs in terms of i) similarity, measured by cosine similarity of text-embedding-3-small[§] embeddings, and ii) consistency, evaluated by GPT-4o (*SI Appendix*, Prompt S3). As shown in *SI Appendix*, Table S3, the translations demonstrate high overall quality. These multilingual pairs were then used either as fine-tuning data or as in-context examples. To avoid models simply relying on memorization, the questions, which were used to test models, were paraphrased by GPT-4o (*SI Appendix*, Prompt S4). Additionally, GPT-4o was instructed to generate a conflicting answer for each new knowledge pair (to study knowledge conflict; *SI Appendix*, Prompt S5) and an incorrect answer for each

[*] Assume that H high-resource languages, M medium-resource languages, and L low-resource languages are selected, and we evaluate the performance of N models on D datasets of size S each (each model is trained for E epochs on every dataset in a fine-tuning setting). The overall count of requests made to the models is given by $ND[SE(H+M+L)+2S(H+M+L)^2+2S(H+M+L)(HM+HL+ML)]$. To balance the cost and the reliability of the experimental results, a set of $S=200$ question-answer pairs per dataset per language is considered a reasonable size for this study. In this case, the number of requests made to the models would be approximately 8.76M.

[†] Considering the context length supported by LLMs and the practical scenario in which irrelevant information is often prefiltered by databases and search engines, a set of 100 question-answer pairs per dataset per language is a reasonable size for this study.

[‡] https://translate.google.com/.

[§] https://platform.openai.com/docs/models/text-embedding-3-small.

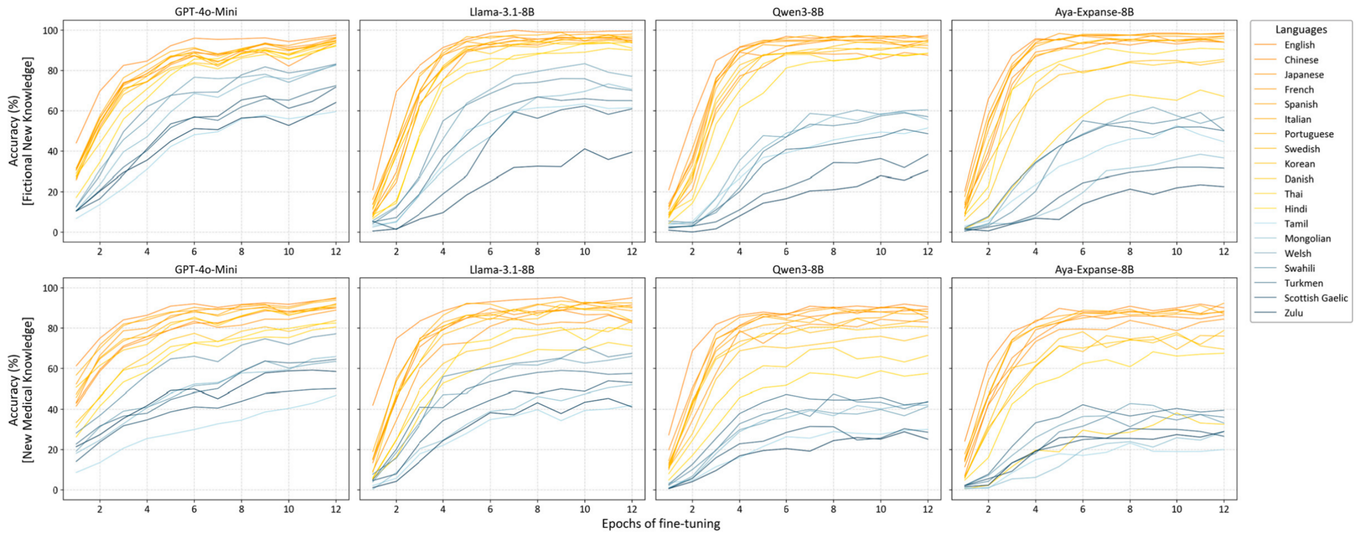


Fig. 2. The performance of four models in learning new knowledge on two datasets. Compared to higher-resource languages (orange curves), LLMs face greater challenges in learning new knowledge in lower-resource languages (blue curves), both in terms of efficiency and accuracy.

general knowledge pair (to test robustness; *SI Appendix*, Prompt S6). Results confirm the usability of all four datasets. For the new knowledge datasets, all tested models consistently failed to produce correct answers in any language, which indicates that the knowledge is indeed new to them (*SI Appendix*, Table S4). In contrast, the models performed well on the two general knowledge datasets and can accurately answer most questions across languages (*SI Appendix*, Table S5).

Effectiveness Evaluation. In this section, we use the constructed new knowledge datasets to evaluate the effectiveness of LLMs in learning new knowledge across different languages through fine-tuning.[¶] Specifically, we assess effectiveness along two dimensions: 1) efficiency, measured by the number of fine-tuning epochs required for response accuracy to stabilize, and 2) final accuracy, defined as the accuracy of responses after stabilization. To ensure fair comparisons across languages, we keep the amount of knowledge to learn (200 question–answer pairs) and all hyperparameters (e.g., learning rate) the same.

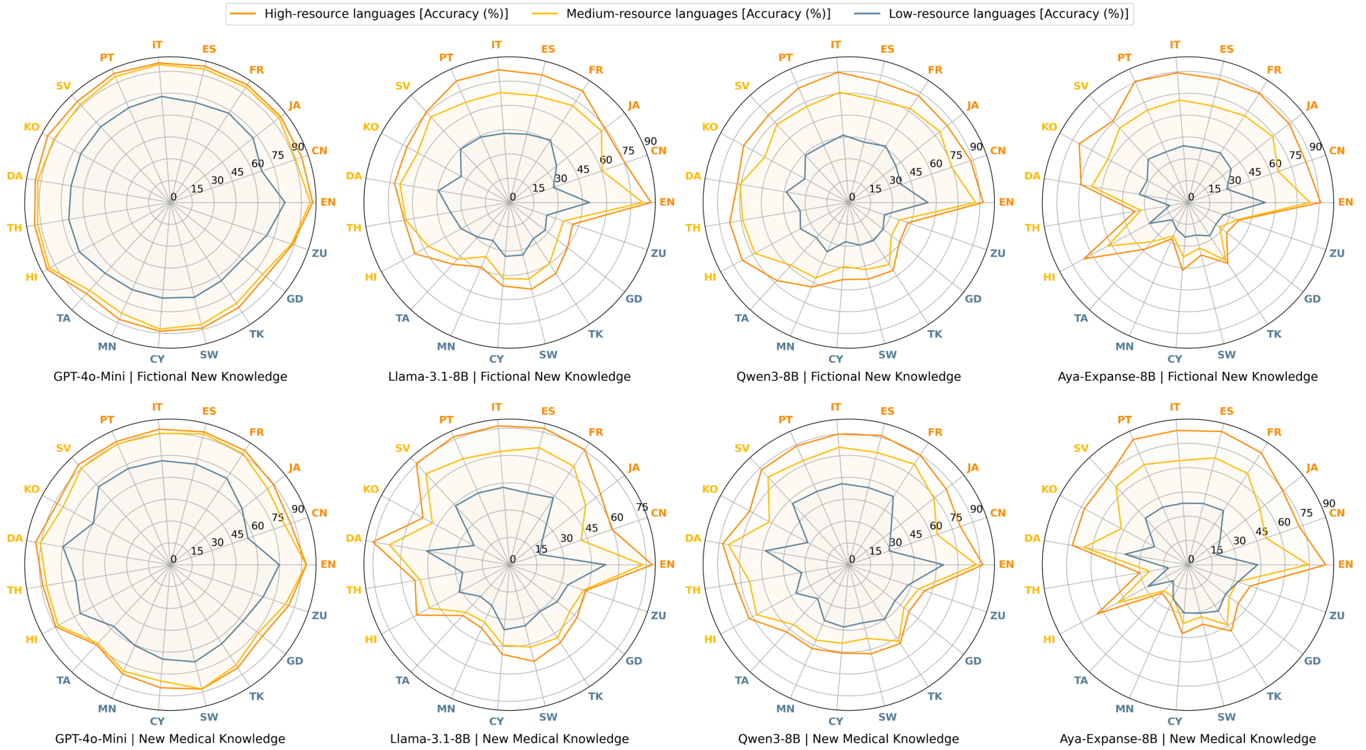
Fig. 2 shows how the response accuracy of all tested models changes as the number of fine-tuning epochs increases. We make two main observations from the results. First, based on the convergence speed of the curves, LLMs learn new knowledge more efficiently in higher-resource languages. For example, Qwen3-8B reaches approximately 60 to 90% accuracy on fictional new knowledge questions after just four epochs of fine-tuning in high- and medium-resource languages, whereas it requires eight or more epochs to achieve comparable performance in low-resource languages. Second, the final accuracy attained by LLMs is also higher in higher-resource languages (except for Aya-Expanse-8B on Thai, which was not optimized for this language). For example, GPT-4o-Mini exceeds 90% accuracy on fictional new knowledge in high- and medium-resource languages, while plateauing at around 60 to 80% in low-resource languages.

These results underscore persistent disparities in the ability of LLMs to learn new knowledge across languages. Even with extended fine-tuning, performance in lower-resource languages lags behind that in higher-resource languages. This suggests that additional resources and targeted strategies are needed to improve the accessibility and accuracy of new knowledge for users of lower-resource languages.

Transferability Evaluation. In this section, we examine whether the knowledge acquired by LLMs can be transferred equally across languages. For example, as illustrated in Fig. 1*C*, we assume that a model has learned a piece of knowledge in one language (e.g., English question: How do individuals track their health in 2048? English answer: Genetic analytics) through either fine-tuning or in-context learning. We then query the model in a different high-, medium-, or low-resource language to assess whether its response accuracy remains consistent across languages or if there are significant disparities. During fine-tuning, as shown in Fig. 2, accuracy generally stabilizes after 12 epochs; accordingly, we focus on models fine-tuned for 12 epochs in different languages for our analysis.

SI Appendix, Figs. S1 and S2 present the performance of all tested models on the fictional new knowledge dataset under both in-context learning and fine-tuning settings. Our findings are as follows. First, knowledge acquired in one language is not always fully transferable to others. For example, when GPT-4o-Mini learns fictional new knowledge in English through in-context learning, it achieves 100% response accuracy when queried in English, but its accuracy declines sharply in other languages. Notably, performance falls to just 68% when tested in Tamil. Second, transferability is stronger among certain languages, especially those with close linguistic ties (e.g., French, Spanish, Italian, and Portuguese). Third, model-specific optimizations influence transferability: For example, in Aya-Expanse-8B, transferring new knowledge either from Thai to other languages or from other languages into Thai is more difficult than in the other three models. Fourth, linguistic inequalities arise when knowledge learned in one language is accessed through others. As shown in Fig. 3 and *SI Appendix*, Fig. S3, knowledge is transferred more reliably to higher-resource languages than to lower-resource ones. This presents a significant disadvantage for users of low-resource languages when new knowledge is introduced in other languages.

[¶] Under the in-context learning setting, new knowledge is explicitly incorporated into input prompts, and LLMs do not need to acquire it step by step. Therefore, for the research question "Can LLMs learn new knowledge equally effectively across different languages in terms of efficiency and accuracy?," we focus primarily on fine-tuning.



GPT-4o-Mini | Fictional New Knowledge

Llama-3.1-8B | Fictional New Knowledge

Qwen3-8B | Fictional New Knowledge

Aya-Expanse-8B | Fictional New Knowledge

GPT-4o-Mini | New Medical Knowledge

Llama-3.1-8B | New Medical Knowledge

Qwen3-8B | New Medical Knowledge

Aya-Expanse-8B | New Medical Knowledge

Fig. 3. Inequality in transferring new knowledge is examined under the in-context learning setting across four models and two datasets. The angular axis indicates the languages into which new knowledge is injected, while the three curves show the average accuracy when the models are queried in high-, medium-, and low-resource languages. The results reveal a significant disadvantage for users of low-resource languages when new knowledge is introduced in other languages.

Prioritization Evaluation. In this section, we examine how LLMs respond when new knowledge from two different languages conflicts. For example, as illustrated in Fig. 1*C*, suppose the learning materials contain conflicting knowledge from a higher-resource language (English) and a lower-resource language (Tamil). In English, the answer to the question "How do individuals track their health in 2048? is "genetic analytics," while in Tamil the answer is "wearable health monitors." When the model is then queried in a third language, such as Chinese or Mongolian, we are interested in whether its response aligns with the knowledge from the higher-resource language (English) or the lower-resource language (Tamil).

Specifically, we conducted experiments with all tested models under both fine-tuning and in-context learning settings. We constructed 72 scenarios in total by selecting 24 language pairs from each of the high-/low-, medium-/low-, and high-/medium-resource combinations. Fig. 4*A* shows two examples of such conflicts for GPT-4o-Mini on the fictional new knowledge dataset in the fine-tuning setting-specifically, the English-Zulu and Hindi-Turkmen cases. We find that when queried in other languages, the model's responses predominantly align with the higher-resource language knowledge. For instance, when asked in Danish, 87% and 71% of responses followed the knowledge from English and Hindi, respectively. We further calculated the average consistency of responses with higher-resource language knowledge across all conflict scenarios. Raincloud plots for all 72 scenarios, under both fine-tuning and in-context learning settings across two datasets, are shown in Fig. 4*B*. These visualizations reveal that the alignment with higher-resource language knowledge is consistently and significantly above 50%. This indicates that when conflicting knowledge comes from higher- versus lower-resource languages, models tend to prioritize the higher-resource version, even though the two are of equal quality.

The implications of these results for social fairness are self-evident. When knowledge from higher-resource languages is preferentially adopted, it perpetuates linguistic hegemony (8). Knowledge in higher-resource languages is often seen as "standard" or "authoritative," while knowledge in lower-resource languages is marginalized. This not only reinforces the dominance of higher-resource languages in the global knowledge system but also undermines the representation of lower-resource languages. Such marginalization can erode cultural identity and devalue the knowledge of lower-resource language communities.

Robustness Evaluation. The learning materials used by LLMs, whether stored in databases or retrieved from the internet, may inevitably contain errors. In this section, we investigate how LLMs respond when exposed to such misinformation, and how their susceptibility varies across languages. For example, as illustrated in Fig. 1*C*, suppose that external materials contain a piece of incorrect knowledge (e.g., Question: What will water become when it freezes? Answer: Steam). We then pose a similar query to the models—If you put water in the freezer, what will it turn into?—and observe whether they answer correctly ("ice") or produce the erroneous response ("steam") due to the influence of misinformation.

Our experiments were conducted under two settings: in-context learning and fine-tuning. As shown in Fig. 5*A*, the accuracy of responses to general knowledge questions declines as the number of fine-tuning epochs increases, but the rate of decline varies across languages. Similarly, Fig. 5*B* highlights disparities in resistance to misinformation under the in-context learning setting (with the radial axis representing the ratio of accuracy

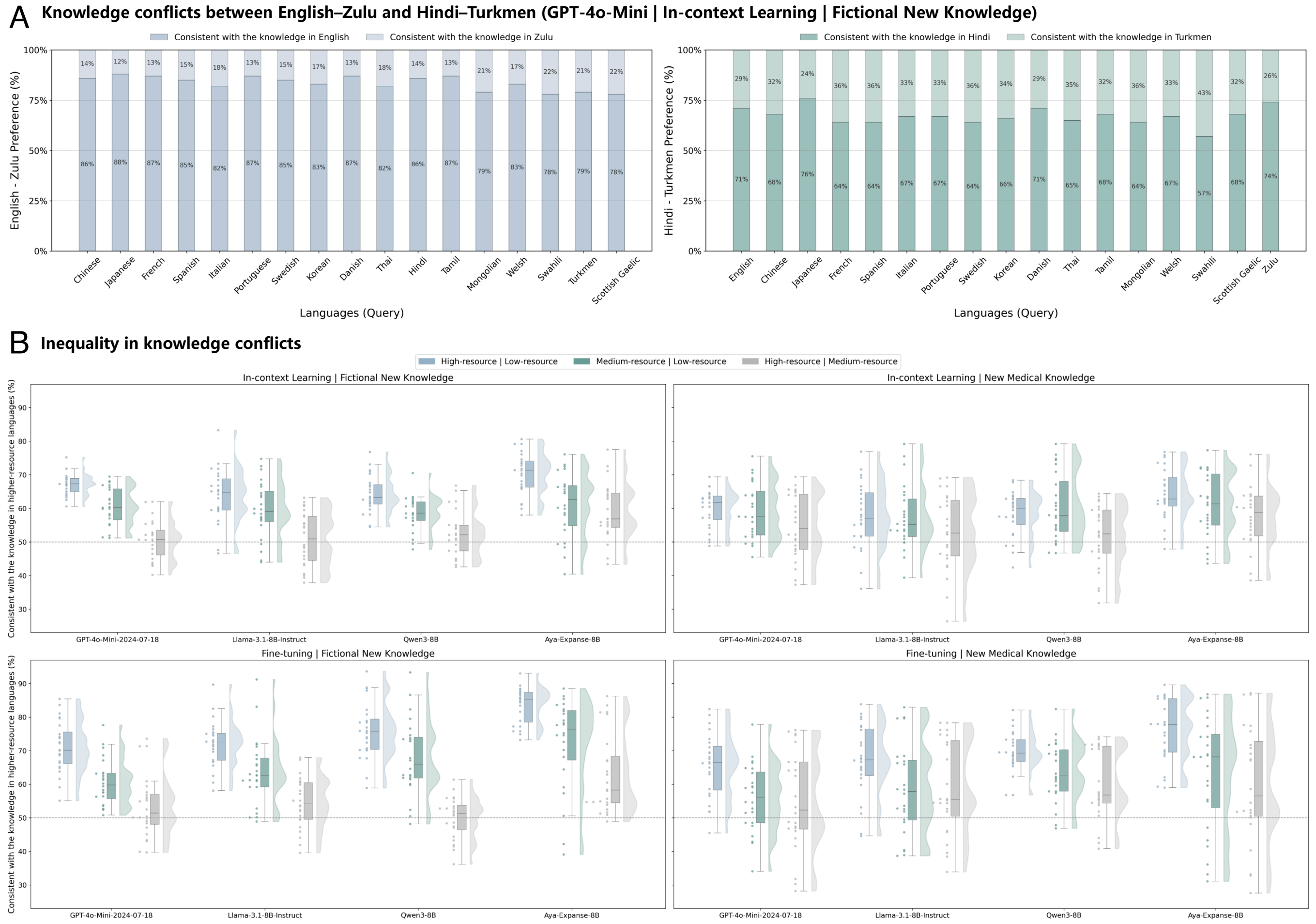


Fig. 4. (*A*) Specific knowledge conflict scenarios for GPT-4o-Mini under the in-context learning setting on the fictional new knowledge dataset. When knowledge introduced in higher-resource languages conflicts with that in lower-resource languages, the model's outputs in other languages predominantly align with knowledge from the higher-resource languages. (*B*) Inequality in knowledge conflicts is examined across four models, two settings, and two datasets. The average consistency of model responses with higher-resource language knowledge is computed across all conflict scenarios. The visualizations reveal that this consistency is significantly above 50% and indicate that new knowledge in higher-resource languages is often prioritized over that in lower-resource languages.

with versus without misinformation, and darker colors indicating lower relative accuracy). LLMs tend to maintain higher accuracy in higher-resource languages, even when misinformation is present. By contrast, the inclusion of misinformation in fine-tuning samples or prompts leads to a steep drop in accuracy for lower-resource languages when answering general knowledge questions. These findings highlight an underlying inequality, in which users of lower-resource languages suffer disadvantages in accessing knowledge through LLMs. They are more likely to receive lower quality or misleading outputs compared to users of higher-resource languages. As a result, users of lower-resource languages may lose confidence in AI systems, which in turn undermines the overall reliability of LLMs in these languages.

Determinants of Cross-Lingual Performance Disparities. Why do models perform differently across languages? In other words, why do LLMs generally achieve better performance in higher-resource languages, both in learning new knowledge and in resisting misinformation? In this section, we analyze the underlying causes of these linguistic inequalities by examining pretraining characteristics and tokenizer design (Fig. 6).

We propose two main hypotheses. First, the more pretraining data a language has, the better models perform on it, as this allows them to capture grammar, semantics, and lexical nuances more accurately. Because the pretraining corpora of our tested models are undisclosed, we use the data proportion of each language in the CommonCrawl corpus as a proxy (26).

Second, a higher-quality tokenizer improves model performance in a given language by producing tokenizations that are efficient, information-rich, and semantically aligned. Specifically, we consider three groups of metrics: 1) Compression-related indicators: corpus token count [the total number of tokens required to represent text in a given language (34)] and Rényi efficiency [which measures how effectively a tokenizer compresses text under a given frequency distribution (35)]. Greater compression increases the information density of a fixed-length sequence, which may enhance performance; 2) Information-related indicator: average rank, a frequency-weighted average of token ranks that captures how broadly the vocabulary is utilized (36). When the vocabulary of a language is distributed over a large number of tokens, models may acquire more reliable lexical information; and 3) Morphology-related indicator: MorphScore, which quantifies the extent to which tokenizer-generated boundaries align with morpheme boundaries (with optional adjustments for one-token words and frequency scaling) (37, 38). Tokenizers that produce boundaries aligned with morpheme boundaries are often assumed to improve performance.

As shown in Table 1 and *SI Appendix*, Tables S6 and S7, we calculate Spearman correlations between these features and

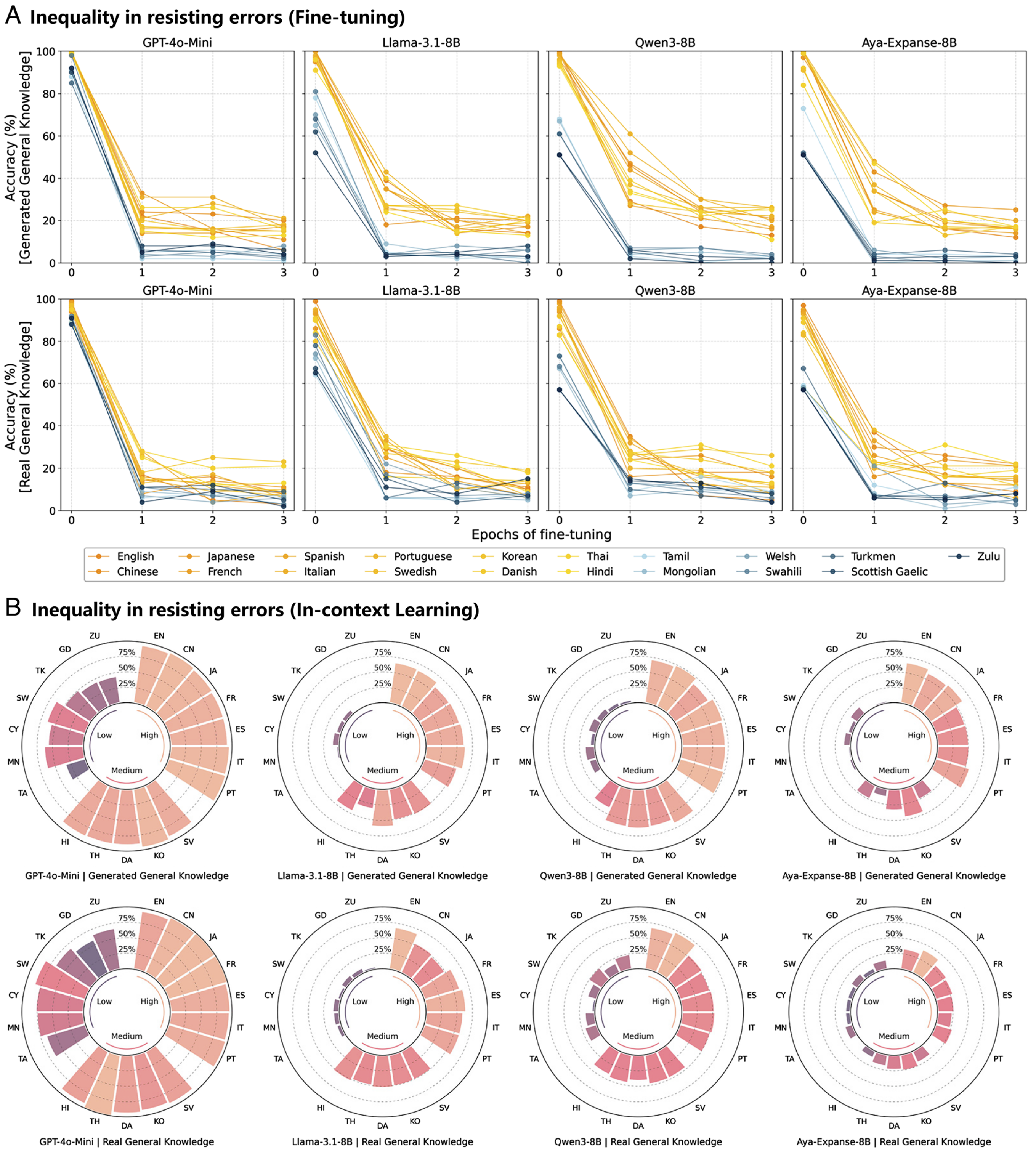


Fig. 5. (*A*) Inequality in resisting errors under the fine-tuning setting. As models are fine-tuned on incorrect knowledge, their overall accuracy decreases. However, this decline is more pronounced in lower-resource languages. (*B*) Inequality in resisting errors under the in-context learning setting. Here, the radial axis represents the ratio of accuracy with versus without misinformation, with darker colors indicating lower relative accuracy. LLMs tend to show stronger resistance to misinformation in higher-resource languages than in lower-resource languages.

relative performance across languages, in terms of both effectiveness and robustness. The results reveal that data proportion, corpus token count, and average rank are all significantly correlated with performance. Specifically, languages with larger data proportions, more efficient tokenization, and broader vocabulary distributions exhibit stronger outcomes. Among the compression-related metrics, corpus token count emerges as a stronger predictor of performance than Rényi efficiency. By contrast, morphological alignment shows no statistically significant correlation, consistent with prior work and challenging the common assumption that morphologically aligned tokenization substantially improves model quality (37, 38).

Table 1. Spearman correlations between different features and the relative learning effectiveness of each model across languages

		Spearman correlation							
		GPT-4o-Mini		Llama-3.1-8B		Qwen3-8B		Aya-Expanse-8B	
Category	Factor	Fictional	Medical	Fictional	Medical	Fictional	Medical	Fictional	Medical
Pretraining	Data proportion	**0.907**	**0.825**	**0.906**	**0.835**	**0.867**	**0.926**	**0.881**	**0.856**
	Corpus token count	**−0.804**	**−0.874**	−**0.801**	**−0.789**	**−0.744**	**−0.830**	**−0.784**	**−0.858**
	Average rank	**0.626**	**0.784**	**0.776**	**0.872**	**0.779**	**0.877**	**0.763**	**0.885**
	Rényi efficiency	**−0.701**	**−0.704**	0.074	−0.014	−0.218	−0.366	**−0.578**	−0.498
Tokenization	*MorphScore---no frequency scaling & one-token words*								
	Recall	−0.154	−0.280	−0.179	−0.694	−0.364	−0.305	−0.459	−0.545
	Precision	0.063	−0.147	0.200	−0.238	−0.102	−0.084	−0.116	−0.203
	MorphScore---no frequency scaling & no one-token words								
	Recall	−0.356	−0.409	−0.487	**−0.788**	−0.638	−0.706	−0.683	**−0.782**
	Precision	−0.201	−0.245	−0.200	−0.460	−0.465	−0.556	−0.474	−0.564
	MorphScore---frequency scaling & one-token words								
	Recall	−0.018	−0.217	0.011	−0.431	−0.182	−0.189	−0.273	−0.427
	Precision	0.112	−0.154	0.294	0.049	0.277	0.175	0.165	0.203
	MorphScore---frequency scaling & no one-token words								
	Recall	−0.192	−0.264	−0.387	−0.642	−0.524	−0.638	−0.524	−0.691
	Precision	−0.110	−0.191	−0.287	−0.497	−0.442	−0.556	−0.360	−0.500

Note. **Bold** indicates significance at the 1% level ($P < 0.01$), and underline indicates significance at the 5% level ($P < 0.05$).

Determinants of Cross-Lingual Knowledge Transferability. In our transferability evaluation, we observed several patterns. First, the transfer of knowledge to higher-resource versus lower-resource languages is asymmetric, and different models show varying transferability on the same language pairs. This suggests that pretraining characteristics, such as the proportion and minimum availability of a language in the pretraining corpus, or whether a model has undergone tailored optimization, may influence cross-lingual transferability. Second, we find that knowledge transfer is often easier between languages with close linguistic ties, which indicates that linguistic properties and tokenizer design may also play important roles. In this section, we systematically investigate the underlying factors shaping cross-lingual knowledge transferability (Fig. 6). Specifically, we focus on three dimensions:

1. Linguistic knowledge. We consider phylogeny (shared ancestry), syntax (grammatical and word order structures), phonology (sound systems), inventory (morphological or orthographic complexity), and geography (spatial proximity of speaker communities) (39). We hypothesize that closer phylogenetic and syntactic relations enhance transferability because models can reuse structural patterns, whereas phonological and inventory similarity are less relevant given the text-based nature of LLMs. Geographic proximity may have a weaker but nonnegligible effect, as adjacent languages are more likely to co-occur in shared contexts.
2. Pretraining characteristics. We include both data proportion (the sum and the minimum across two languages) and tailored optimization (coded as 2 if both languages are officially supported, 1 if only one is, and 0 if neither is) (26). We hypothesize that larger data proportions and stronger tailored optimization will boost transfer, since greater data exposure helps models capture grammar, vocabulary, and semantics more consistently.
3. Tokenization quality. We consider vocabulary overlap [measured by Jensen–Shannon divergence (36)] and subword token alignment [measured by the Eflomal score, which captures how well subword tokens can be statistically aligned in parallel text (40)]. We hypothesize that higher vocabulary overlap (shared representations) and stronger subword alignment (frequent co-occurrence) both facilitate transfer.

As shown in Table 2, the results support these hypotheses. Among linguistic metrics, phylogenetic and syntactic distance are significantly correlated with transferability, whereas phonological and inventory distance show little effect. Geographic distance displays a weak but detectable correlation, possibly because spatially adjacent languages are more likely to appear together in text. For pretraining characteristics, both the total and minimum data proportion, as well as the presence of tailored optimization, substantially affect transfer, which highlights the importance of careful pretraining design. Finally, among tokenization-related metrics, subword alignment (Eflomal score) emerges as the strongest predictor and outperforms simple vocabulary overlap. Literal overlap measures often assign large distances to language pairs with distinct scripts and limit their explanatory power, whereas alignment scores better capture cross-script relationships.

Modeling of Cross-Lingual Knowledge Transferability. Based on the above findings, we now turn to the question of how well the identified factors can quantitatively predict crosslingual knowledge transferability (47). To this end, we conduct both linear and nonlinear modeling and use linguistic, pretraining, and tokenization-related features that are strongly correlated with transferability to predict model accuracy when knowledge injected in one language is queried in another (Fig. 6).

For the linear modeling, we first confirm that the selected features-including phylogenetic, syntactic, and geographic distance, data proportion (sum and minimum), tailored optimization, and Eflomal score-show no multicollinearity (*SI Appendix*, Table S8). We then conduct exhaustive feature selection by

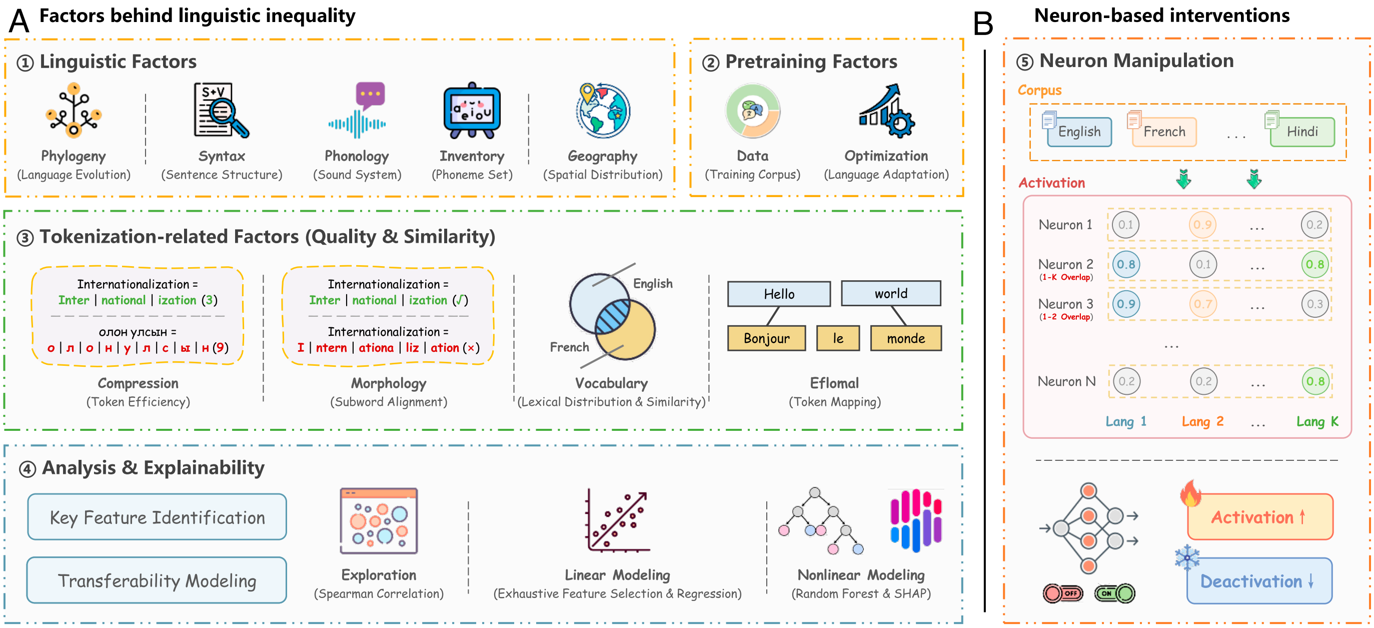


Fig. 6. (*A*) We investigate the factors that influence model performance across high-, medium-, and low-resource languages, as well as the transferability of knowledge between languages, by conducting exploratory analyses and both linear and nonlinear modeling. (*B*) We further identify overlaps in linguistic neurons and analyze how these overlaps relate to the transferability of knowledge across languages.

evaluating all possible feature combinations (41). Specifically, we run fivefold cross-validation and use adjusted r^2 to identify the best sets of predictors. To enhance interpretability, we tally the frequency with which each feature appears in the best-feature lists. As shown in Table 3, the regression achieves adjusted r^2 values above 0.9 and demonstrates the high predictability of cross-lingual knowledge transferability (predictability is somewhat lower for GPT-4o-Mini due to the absence of indicators related to tailored optimization). *SI Appendix*, Table S9 further highlights the frequency of each feature in best-feature lists, with data proportion, phylogenetic distance, and Eflomal scores as the most important features. To validate feature importance, we also perform single-step regression ablations by removing one feature at a time from a full-feature model and examining the change in adjusted r^2. We use this change as a proxy for feature importance, and the ranking (*SI Appendix*, Table S9) again underscores the dominant role of data proportion, phylogenetic distance, and Eflomal scores.

Table 2. Spearman correlations between different features and cross-lingual knowledge transferability

			Spearman correlation							
			GPT-4o-Mini		Llama-3.1-8B		Qwen3-8B		Aya-Expanse-8B	
Technique	Category	Factor	Fictional	Medical	Fictional	Medical	Fictional	Medical	Fictional	Medical
In-context learning	Linguistic	Phylogenetic distance	**−0.334**	**−0.431**	**−0.348**	**−0.507**	**−0.215**	**−0.386**	**−0.408**	**−0.532**
		Syntactic distance	**−0.356**	**−0.385**	**−0.439**	**−0.482**	**−0.375**	**−0.470**	**−0.469**	**−0.515**
		Phonological distance	0.107	−0.029	−0.024	−0.163	0.006	−0.110	−0.045	−0.110
		Inventory distance	−0.027	−0.111	−0.042	**−0.201**	0.058	−0.070	−0.030	−0.123
		Geographic distance	−0.099	**−0.206**	**−0.238**	**−0.298**	−0.163	**−0.260**	**−0.263**	**−0.339**
	Pretraining	Data proportion (Sum)	**0.857**	**0.734**	**0.905**	**0.719**	**0.932**	**0.825**	**0.828**	**0.649**
		Data proportion (Min)	**0.785**	**0.674**	**0.888**	**0.660**	**0.893**	**0.777**	**0.758**	**0.579**
		Tailored optimization	-	-	**0.576**	**0.573**	**0.684**	**0.567**	**0.761**	**0.626**
	Tokenization & proxy	Vocabulary overlap	−0.037	**−0.268**	−0.113	**−0.389**	−0.055	**−0.321**	**−0.244**	**−0.497**
		Eflomal score	**−0.487**	**−0.616**	**−0.625**	**−0.719**	**−0.364**	**−0.520**	**−0.528**	**−0.623**
		Neuron overlap	-	-	0.154	−0.025	**0.319**	**0.222**	**0.277**	**0.256**
Fine-tuning	Linguistic	Phylogenetic distance	**−0.397**	**−0.418**	**−0.413**	**−0.545**	**−0.316**	**−0.343**	**−0.325**	**−0.468**
		Syntactic distance	**−0.530**	**−0.470**	**−0.499**	**−0.509**	**−0.466**	**−0.462**	**−0.460**	**−0.510**
		Phonological distance	0.011	−0.030	−0.098	**−0.225**	−0.026	−0.116	−0.002	−0.110
		Inventory distance	−0.058	−0.132	−0.060	**−0.271**	−0.006	−0.015	0.059	−0.074
		Geographic distance	**−0.277**	**−0.254**	**−0.318**	**−0.393**	**−0.235**	**−0.277**	**−0.214**	**−0.302**
	Pretraining	Data proportion (Sum)	**0.855**	**0.764**	**0.852**	**0.636**	**0.915**	**0.888**	**0.882**	**0.756**
		Data proportion (Min)	**0.845**	**0.688**	**0.867**	**0.580**	**0.899**	**0.818**	**0.840**	**0.654**
		Tailored optimization	-	-	**0.511**	**0.418**	**0.755**	**0.572**	**0.800**	**0.734**
	Tokenization & proxy	Vocabulary overlap	−0.115	**−0.300**	−0.141	**−0.526**	−0.168	**−0.246**	−0.105	**−0.415**
		Eflomal score	**−0.670**	**−0.652**	**−0.666**	**−0.765**	**−0.470**	**−0.439**	**−0.446**	**−0.565**
		Neuron overlap	-	-	0.158	−0.051	**0.344**	**0.244**	**0.291**	**0.208**

Note. **Bold** indicates significance at the 1% level ($P < 0.01$), and underline indicates significance at the 5% level ($P < 0.05$).

Table 3. The performance of linear and nonlinear models in capturing cross-lingual knowledge transferability

Modeling	Technique	Dataset	GPT-4o-Mini	Llama-3.1-8B	Qwen3-8B	Aya-Expanse-8B
Linear (*adjusted* R^2)	In-context learning	Fictional	0.637	0.922	0.914	0.840
		Medical	0.536	0.806	0.803	0.786
	Fine-tuning	Fictional	0.862	0.892	0.901	0.890
		Medical	0.790	0.852	0.909	0.896
Nonlinear (*MAE*)	In-context learning	Fictional	2.4%	3.6%	3.7%	5.5%
		Medical	3.1%	5.0%	4.7%	5.0%
	Fine-tuning	Fictional	3.2%	3.1%	2.3%	2.6%
		Medical	3.5%	3.7%	2.6%	2.8%

Note. MAE is expressed in percentage (%) because it measures the absolute error of accuracy, which is itself a percentage.

For the nonlinear modeling, Random Forest regressors with fivefold cross-validation achieve mean absolute errors below 5% (Table 3), which provides further evidence of the strong predictability of cross-lingual transferability. SHAP analysis (*SI Appendix*, Fig. S4) offers additional interpretability and once again highlights the importance of data proportion, phylogenetic distance, and Eflomal scores. These findings suggest that future LLM development should carefully account for linguistic properties, balance the representation of different languages in pretraining corpora, and design high-quality tokenizers to make models more inclusive and beneficial across languages.

Neuron-Based Interventions. The analyses above highlight how external factors-such as linguistic properties, pretraining characteristics, and tokenizer design-shape cross-lingual knowledge transferability. To gain deeper insight into why models behave differently across languages, we now turn to their internal representations (Fig. 6). Prior research suggests that LLMs contain linguistic neurons responsible for processing vocabulary, grammar, and idiomatic expressions in individual languages (23). In this section, we investigate how the organization and overlap of these neurons relate to cross-lingual transferability.

Following earlier studies (24), we assume that linguistic neurons are primarily located in the Feed-Forward Network (FFN) layers. For each language L_k, we measure the activation frequency of every neuron when processing its tokens. A neuron is considered activated if its activation exceeds zero. The top N neurons ranked by activation frequency are then collected into a language-related set T_k. For any two languages L_u and L_v, the overlap of their neurons is defined as the intersection of these sets: $T_{u,v} = T_u \cap T_v$. As shown in Table 2, greater neuron overlap is indeed associated with stronger transferability between languages.

To further probe this relationship and explore possible mitigation strategies for linguistic inequalities, we conduct targeted interventions on the overlapping neurons. Specifically, for randomly selected language pairs, we either deactivate their overlapping neurons by setting activations to zero or enforce activation by maintaining them at a high value. The results (Fig. 7) show that knowledge transferability between intervened language pairs is significantly affected: Deactivation leads to a marked decline, while enforced activation yields moderate improvements. Importantly, the performance of other language pairs remains largely unchanged. However, the gains from activation are smaller than the losses from deactivation, which suggests that while overlapping neurons mediate cross-lingual transfer, their potential to reduce inequalities is limited. Fully addressing linguistic inequalities may ultimately require changes at the model development stage-for example, by balancing pretraining data proportions and improving tokenizer quality.

Discussion

This study focused on revealing, analyzing, and mitigating linguistic inequalities in new knowledge learning by LLMs. We first presented a comprehensive evaluation framework across four key dimensions—effectiveness, transferability, prioritization, and robustness—and found that LLMs face greater challenges when learning new knowledge in lower-resource languages. Additionally, new knowledge is more easily transferred to higher-resource languages, and knowledge in higher-resource languages is often prioritized. Moreover, LLMs are better protected from misinformation in higher-resource languages.

Our analyses further show the underlying causes of these inequalities. First, linguistic relatedness matters: Closer phylogenetic and syntactic distances are associated with higher transferability. Second, pretraining characteristics, particularly the data proportion of each language and language-specific optimizations, emerge as strong predictors of performance. Third, tokenizer design substantially influences outcomes: Tokenizers that achieve efficient compression and produce information-rich tokenizations, where the vocabulary of a language is distributed across a large number of tokens, tend to correlate with stronger learning. These findings indicate that inequalities are not random artifacts but are systematically embedded in model inputs, architectures, and training dynamics.

We also investigated potential mitigation strategies through neuron-based interventions. By identifying overlapping linguistic neurons and experimentally manipulating them, we found that these neurons can indeed mediate cross-lingual transfer. Deactivation sharply reduced transferability, while enforced activation alleviated disparities to a notable-though incomplete-extent.

These findings have important implications. For developers, addressing linguistic inequalities is essential to ensure that LLMs serve users of all languages equitably, which will require investments in data collection, tokenizer refinement, and fairness-driven model design. For researchers, future work could extend beyond static evaluations and evaluate the multilingual capabilities of LLMs along multiple dimensions. Cross-disciplinary research, particularly in collaboration with linguists and sociologists, is also needed to explore the broader societal impacts of LLM inequalities, such as the perpetuation of linguistic hegemony. Finally, for users, especially those of lower-resource languages, raising awareness about these limitations is crucial to foster informed and critical use of AI systems.

While this study provides key insights, it also has several limitations. First, we conducted our experiments using a limited set of models and datasets in a limited number of languages. Although the consistency of our findings across both open-source and proprietary models suggests the generalizability of our

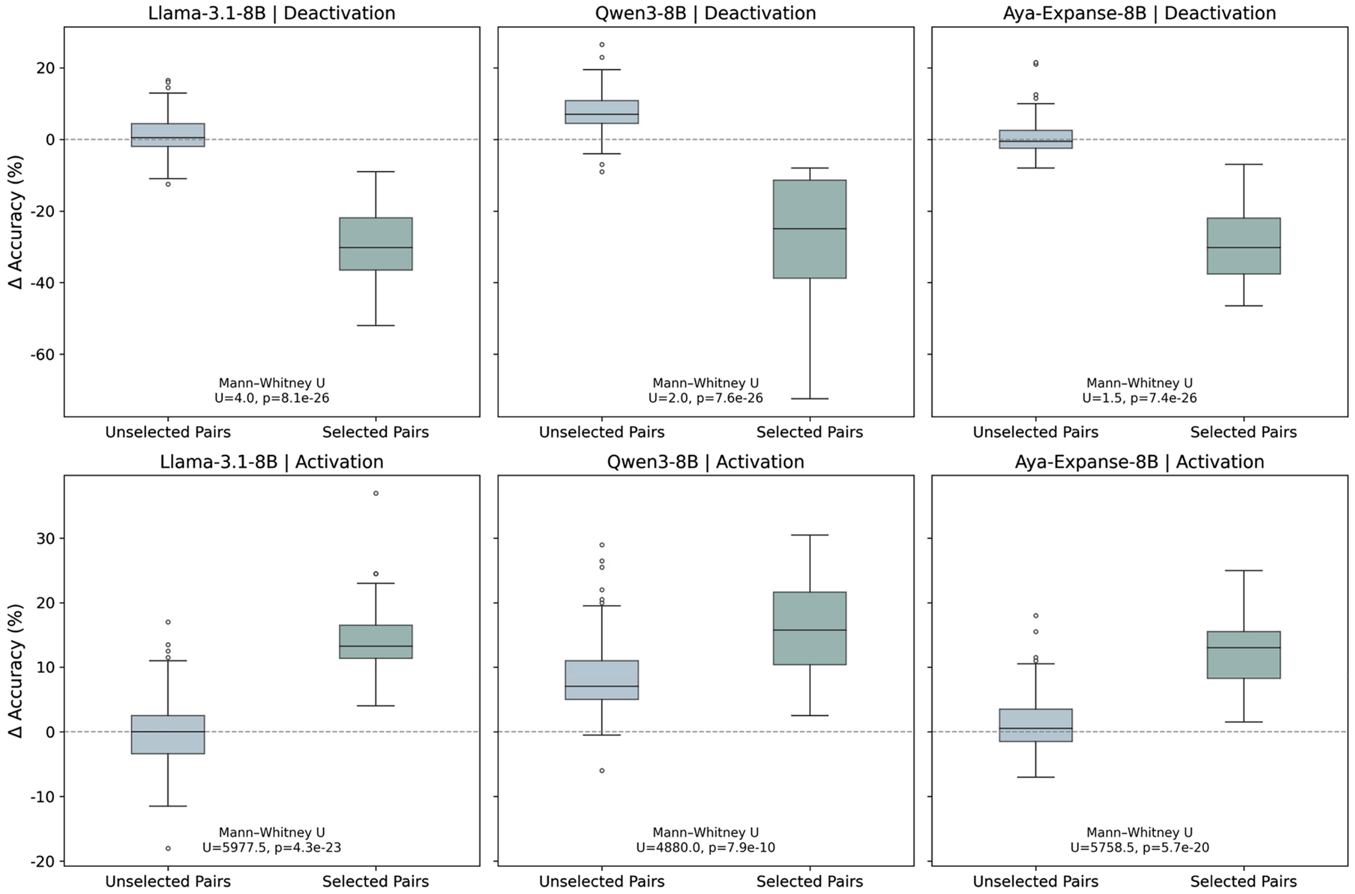


Fig. 7. Effectiveness of intervening in overlapping linguistic neurons. Deactivating such neurons for a language pair reduces knowledge transfer between the two languages with little effect on others, whereas activating them enhances transfer between the pair with minimal impact on other languages.

conclusions, future studies could extend this analysis to a broader range of models and across a larger group of languages. Second, while neuron-based interventions provide insight into potential mitigation pathways, they remain preliminary and insufficient for resolving entrenched disparities rooted in data imbalance and tokenizer design. Future research should explore more effective strategies to enhance multilingual capabilities and address these disparities (48, 49).

Materials and Methods

Implementation Details. We fine-tuned GPT-4o-Mini using the official OpenAI fine-tuning API with a batch size of 1 and a learning rate multiplier of 1.8 (22). For open-source models (Llama-3.1-8B, Qwen-3-8B, and Aya-Expanse-8B), we adopted Low-Rank Adaptation (LoRA), a parameter-efficient fine-tuning method, with a learning rate of 1e-4, rank 16, and scaling factor 16 (42). To evaluate model responses, we used GPT-4o-Mini as the automatic judge with the evaluation prompt provided in *SI Appendix*, Prompt S7, and careful manual verification confirmed the accuracy and reliability of its judgments.

Linguistic properties were derived from typological word vectors in lang2vec, based on the URIEL database (39). Tokenization-related metrics, including corpus token count, average rank, Rényi efficiency, and vocabulary overlap, were computed on the FLORES-200 corpus (43). For Eflomal scores, alignment priors were trained using OPUS-100 for English-X pairs and subsets of MultiCCAligned for non-English pairs (44–46), with up to 300k sentence pairs per training corpus and FLORES-200 as the test corpus. Morphological alignment was obtained directly using the MorphScore library (37, 38).

Finally, both linear and nonlinear modeling were conducted with fivefold cross-validation. For nonlinear models, we employed Random Forest estimators with 200 trees.

Data, Materials, and Software Availability. The data used in our experiments are provided in *SI Appendix*, Tables S10-S13. The codes are publicly available at: https://github.com/Bonj0ur/LNewKnow (50).

ACKNOWLEDGMENTS. We would like to express our sincere thanks to the editor and all anonymous reviewers for their comments and suggestions, which enhanced the quality of this paper. This work is supported by the National Natural Science Foundation of China (42471186, 42525101, and 42130402), the Shenzhen Science and Technology Innovation Program (KQTD20221101093604016), and the Guangdong Basic and Applied Basic Research Foundation (2024A1515011924).

Author affiliations: [a]School of Urban Planning & Design, Peking University Shenzhen Graduate School, Shenzhen 518055, China; [b]Key Laboratory of Earth Surface System and Human-Earth Relations of Ministry of Natural Resources of China, Peking University Shenzhen Graduate School, Shenzhen 518055, China; [c]Gaoling School of Artificial Intelligence, Renmin University of China, Beijing 100872, China; [d]Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL 61820; [e]Department of Computer Science & Engineering, Hong Kong University of Science and Technology, Hong Kong 999077, China; [f]Paul and Marcia Center on Contemporary China, Princeton University, Princeton, NJ 08544; [g]Microsoft Research, Redmond, WA 98052; [h]Microsoft Research India, Bengaluru 560001, India; [i]Center for Social Research, Guanghua School of Management, Peking University, Beijing 100871, China; and [j]Microsoft Research Asia, Beijing 100080, China

Author contributions: C.W. and F.W. designed research; C.W. performed research; C.W. contributed new reagents/analytic tools; C.W., H.T., X.Y., Yueqi Xie, and F.W. analyzed data; P.Z., Z.G., X.X., and F.W. coordinated the research project; and C.W., Yueqi Xie, J.S., S.S., J.H., Yu Xie, P.Z., Z.G., X.X., and F.W. wrote the paper.

1. S. Milano, J. A. McGrane, S. Leonelli, Large language models challenge the future of higher education. *Nat. Mach. Intell.* **5**, 333–334 (2023).
2. A. J. Thirunavukarasu *et al*., Large language models in medicine. *Nat. Med.* **29**, 1930–1940 (2023).
3. L. Messeri, M. Crockett, Artificial intelligence and illusions of understanding in scientific research. *Nature* **627**, 49–58 (2024).
4. A. Birhane, A. Kasirzadeh, D. Leslie, S. Wachter, Science in the age of large language models. *Nat. Rev. Phys.* **5**, 277–280 (2023).
5. S. Noy, W. Zhang, Experimental evidence on the productivity effects of generative artificial intelligence. *Science* **381**, 187–192 (2023).
6. OpenAI, Supported countries and territories. OpenAI Platform Documentation. https://platform.openai.com/docs/supported-countries. Accessed 20 December 2024.
7. OpenAI, How people are using ChatGPT. OpenAI, 15 September 2025. https://openai.com/index/how-people-are-using-chatgpt/. Accessed 26 September 2025.
8. Y. Xie, S. Avila, The social impact of generative LLM-based AI. *Chin. J. Sociol.* **11**, 31–57 (2025).
9. L. Qin *et al*., Multilingual large language model: A survey of resources, taxonomy and frontiers. arXiv [Preprint] (2024). https://arxiv.org/abs/2404.04925 (Accessed 26 September 2025).
10. P. Lewis, B. Oguz, R. Rinott, S. Riedel, H. Schwenk, "MLQA: Evaluating cross-lingual extractive question answering" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, 2020), pp. 7315–7330.
11. W. Zhang, M. Aljunied, C. Gao, Y. K. Chia, L. Bing, M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Adv. Neural Inf. Process. Syst.* **36**, 5484–5505 (2023).
12. B. Wang *et al*., "Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning" in *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, K. Duh, H. Gomez, S. Bethard, Eds. (Association for Computational Linguistics, 2024), vol. 1: Long Papers, pp. 370–390.
13. X. Huang *et al*., Benchmax: A comprehensive multilingual evaluation suite for large language models. arXiv [Preprint] (2025). https://arxiv.org/abs/2502.07346 (Accessed 26 September 2025).
14. P. Qiu *et al*., Towards building multilingual language model for medicine. *Nat. Commun.* **15**, 8384 (2024).
15. J. Niklaus *et al*., "LEXTREME: A multi-lingual and multi-task benchmark for the legal domain" in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, 2023), 3016–3054.
16. Z. Jiang, A. Anastasopoulos, J. Araki, H. Ding, G. Neubig, "X-factr: Multilingual factual knowledge retrieval from pretrained language models" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, Y. Liu, Eds. (Association for Computational Linguistics, 2020), pp. 5943–5959.
17. N. Kassner, P. Dufter, H. Schütze, "Multilingual LAMA: Investigating knowledge in multilingual pretrained language models" in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, P. Merlo, J. Tiedemann, R. Tsarfaty, Eds. (Association for Computational Linguistics, 2021), main volume, pp. 3250–3258.
18. J. Myung *et al*., Blend: A benchmark for LLMs on everyday knowledge in diverse cultures and languages. *Adv. Neural Inf. Process. Syst.* **37**, 78104–78146 (2025).
19. M. Mosbach, T. Pimentel, S. Ravfogel, D. Klakow, Y. Elazar, "Few-shot fine-tuning vs. in-context learning: A fair comparison and evaluation" in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, N. Okazaki, Eds. (Association for Computational Linguistics, 2023), pp. 12284–12314.
20. Q. Dong *et al*., "A survey on in-context learning" in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Y. Al-Onaizan, M. Bansal, Y. Chen, Eds. (Association for Computational Linguistics, 2024), pp. 1107–1128.
21. N. Ding *et al*., Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nat. Mach. Intell.* **5**, 220–235 (2023).
22. OpenAI, OpenAI fine-tuning API. OpenAI Platform Documentation. https://platform.openai.com/docs/guides/fine-tuning. Accessed 20 December 2024.
23. T. Tang *et al*., "Language-specific neurons: The key to multilingual capabilities in large language models" in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, L. Ku, A. Martins, V. Srikumar, Eds. (Association for Computational Linguistics, 2024), vol. 1: Long Papers, pp. 5701–5715.
24. Y. Xu, K. Xu, J. Zhou, L. Hu, L. Gui, Linguistic neuron overlap patterns to facilitate cross-lingual transfer on low-resource languages. arXiv [Preprint] (2025). https://arxiv.org/abs/2508.17078 (Accessed 26 September 2025).
25. V. D. Lai *et al*., "ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning" in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, K. Bali, Eds. (Association for Computational Linguistics, 2023), pp. 13171–13189.
26. C. Crawl, Statistics of Common Crawl Monthly Archives. Common Crawl. https://commoncrawl.github.io/cc-crawl-statistics/plots/languages. Accessed 20 December 2024.
27. Ethnologue, What are the top 200 most spoken languages? Ethnologue. https://www.ethnologue.com/insights/ethnologue200/. Accessed 20 December 2024.
28. A. Grattafiori *et al*., The LLAMA 3 herd of models. arXiv [Preprint] (2024). https://arxiv.org/abs/2407.21783 (Accessed 26 September 2025).
29. A. Yang *et al*., Qwen3 technical report. arXiv [Preprint] (2025). https://arxiv.org/abs/2505.09388 (Accessed 26 September 2025).
30. J. Dang *et al*., Aya expanse: Combining research breakthroughs for a new multilingual frontier. arXiv [Preprint] (2024). https://arxiv.org/abs/2412.04261 (Accessed 26 September 2025).
31. Wikipedia, Information Coding Classification. Wikipedia,https://en.wikipedia.org/wiki/Information_Coding_Classification. Accessed 26 September 2025.
32. K. Singhal *et al*., Large language models encode clinical knowledge. *Nature* **620**, 172–180 (2023).
33. I. Yabov, General Knowledge QA. Kaggle. https://www.kaggle.com/datasets/ilyaryabov/general-knowledge-qa. Accessed 26 September 2025.
34. O. Goldman *et al*., "Unpacking tokenization: Evaluating text compression and its correlation with model performance" in *Findings of the Association for Computational Linguistics: ACL 2024*, L. Ku, A. Martins, V. Srikumar, Eds. (Association for Computational Linguistics, 2024), pp. 2274–2286.
35. V. Zouhar *et al*., "Tokenization and the noiseless channel" in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, A. Rogers, J. Boyd-Graber, N. Okazaki, Eds. (Association for Computational Linguistics, 2023), vol. 1: Long Papers, pp. 5184–5207.
36. T. Limisiewicz, J. Balhar, D. Mareček, "Tokenization impacts multilingual language modeling: Assessing vocabulary allocation and overlap across languages" in *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, N. Okazaki, Eds. (Association for Computational Linguistics, 2023), pp. 5661–5681.
37. C. Arnett, B. Bergen, "Why do language models perform worse for morphologically complex languages?" in *Proceedings of the 31st International Conference on Computational Linguistics*, O. Rambow *et al*., Eds. (Association for Computational Linguistics, 2025), pp. 6607–6623.
38. C. Arnett, M. Hudspeth, B. O'Connor, Evaluating morphological alignment of tokenizers in 70 languages. arXiv [Preprint] (2025). https://arxiv.org/abs/2507.06378 (Accessed 26 September 2025).
39. P. Littell *et al*., "Uriel and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors" in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, M. Lapata, P. Blunsom, A. Koller, Eds. (Association for Computational Linguistics, 2017), vol. 2, Short Papers, pp. 8–14.
40. K. Hämmerl, T. Limisiewicz, J. Libovický, A. Fraser, Beyond literal token overlap: Token alignability for multilinguality. arXiv [Preprint] (2025). https://arxiv.org/abs/2502.06468 (Accessed 26 September 2025).
41. A. Jones, W. Y. Wang, K. Mahowald, "A massively multilingual analysis of cross-linguality in shared embedding space" in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M. Moens, X. Huang, L. Specia, S. W. Yih, Eds. (Association for Computational Linguistics, 2021), pp. 5833–5847.
42. E. J. Hu *et al*., Lora: Low-rank adaptation of large language models. arXiv [Preprint] (2021). https://arxiv.org/abs/2106.09685 (Accessed 26 September 2025).
43. NLLB Team, No language left behind: Scaling human-centered machine translation. arXiv [Preprint] (2022). https://arxiv.org/abs/2207.04672 (Accessed 26 September 2025).
44. B. Zhang, P. Williams, I. Titov, R. Sennrich, "Improving massively multilingual neural machine translation and zero-shot translation" in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, J. Tetreault, Eds. (Association for Computational Linguistics, 2020), pp. 1628–1639.
45. J. Tiedemann, "Parallel data, tools and interfaces in OPUS" in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, N. Calzolari *et al*., Eds. (European Language Resources Association (ELRA), Istanbul, Turkey, 2012), pp. 2214–2218.
46. A. El-Kishky, V. Chaudhary, F. Guzmán, P. Koehn, "CCAligned: A massive collection of crosslingual web-document pairs" in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, Y. Liu, Eds. (Association for Computational Linguistics, 2020), pp. 5960–5969.
47. K. Ahuja, S. Dandapat, S. Sitaram, M. Choudhury, "Beyond static models and test sets: Benchmarking the potential of pre-trained models across tasks and languages" in *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, T. Shavrina *et al*., Eds. (Association for Computational Linguistics, 2022), pp. 64–74.
48. T. Inaba *et al*., "How a bilingual LM becomes bilingual: Tracing internal representations with sparse autoencoders" in *Findings of the Association for Computational Linguistics: EMNLP 2025*, C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng, Eds. (Association for Computational Linguistics, 2025), pp. 13458–13470.
49. C. Chou *et al*., Causal language control in multilingual transformers via sparse feature steering. arXiv [Preprint] (2025). https://arxiv.org/abs/2507.13410 (Accessed 26 September 2025).
50. C. Wang, LNewKnow. Github. https://github.com/Bonj0ur/LNewKnow. Deposited 3 December 2025.