Evaluating and Improving Multilingual Comprehension and Generation with Large Language Models

Niyati Bafna, Advisor: David Yarowksy

1 Introduction

The vast capabilities of large language models (LLMs) today, including multilingual LLMs (Lin et al., 2022; Workshop et al., 2023; Aryabumi et al., 2024b) are unfortunately restricted to a handful of high-resource languages in the world such as English and Chinese (Joshi et al., 2020), given that web-crawled pre-training corpora are dominated by high-resource languages (Xue et al., 2021; Chung et al., 2023; Li et al., 2025a; Joshi et al., 2024). We would to extend these capabilities to the rest of the world's languages.

Making LLMs multilingual is a broad endeavour, with several facets. We consider three dimensions as follows. Firstly, **evaluation / diagnosis** of multilingual capabilities and failures is the first and necessary step; we also want to **find solutions**. Secondly, multilingual capabilities can be separated into **comprehension** and **generation** capabilities, since these have very different problems and mitigation techniques. Finally, there are about 3800+ written languages in the world, and different groups of languages have varying problems, often associated with resourcedness. **Extremely low-resource languages**, or the long tail of languages constituting the majority category, have very scant training and evaluation resources, and may be nearly unseen to state-of-the-art models. Many languages are **dialects or variants**[1] of closely related high-resource language neighbours. Finally, we have **low-resource and low-to-mid resource languages**, which may have non-trivial presence in pretraining data.

Goal The proposed goal is to evaluate and improve multilingual comprehension and generation capabilities of LLMs, taking into account the needs of various languages as grouped by their resourcedness and relationships with other languages.

This document is structured in the following way: I provide background in this field in § 2, I provide an overview of completed work in § 3. I describe three of these works in more detail in § 4, § 5, and § 6. I summarize other completed works and introduce next directions in § 7. I describe planned work in § 8, § 9, and § 10. Finally, I provide a timeline in § 11.

2 Background

2.1 Understanding and evaluating multilinguality for comprehension and generation

Benchmarking Most existing benchmarks that LLMs are evaluated on focus on English and other high-resource languages (Grattafiori et al., 2024; Aryabumi et al., 2024a; Qwen et al., 2025). Several efforts attempts to extend benchmarking to other **low-resource** languages. Machine translation datasets have often shown the largest language coverage, such as `FLORES+` (NLLB Team et al., 2024) covering 212 languages, and `BOUQUET` (Andrews et al., 2025; Team et al., 2026) covering 266 languages and dialects. Other multilingual task datasets seek to provide multilingual coverage to a less extent, such as `BELEBELE` (Bandarkar et al., 2024) for machine reading comprehension dataset Aya Evaluation Suite (Singh et al., 2024) for multilingual instruction following, and XL-Sum (Hasan et al., 2021) for summarization. Recent work has also introduced benchmarks targeting multilingual retrieval and retrieval-augmented generation (RAG) (Lewis et al., 2020) such as MMTEB (Enevoldsen et al., 2025) covering 250+ languages, Mr. TyDi (Zhang et al., 2021),

[1] We use the term "dialect" from a linguistic perspective, whereby French is as much a dialect of Occitan as the other way round.

BordIRLines (Li et al., 2024a), XRAG (Liu et al., 2025b), and MIRAGE-Bench (Thakur et al., 2025), which expose challenges in reasoning, consistency, and answer generation under multilingual evidence.

Evaluating LLM performance in low-resource settings Several works demonstrate the degradation of LLMs on **low-resource languages** for extrinsic task performance. This is commonly studied on machine translation (Jiao et al., 2023; Hendy et al., 2023; Robinson et al., 2023), but has also been observed for other tasks like POS, NER, and summarization, among others (Lai et al., 2023; Bang et al., 2023; Asai et al., 2023; Ziems et al., 2023; Cahyawijaya et al., 2024). Kantharuban et al. (2023) attempt to identify economic, social, and linguistic correlates of MT performance in LLMs for **dialects**; they find positive correlations for dataset size and lexical similarity among other factors. In multilingual retrieval-augmented generation (RAG), prior work shows that models exhibit biases against low-resource languages, sometimes preferring English documents even when less relevant (Ki et al., 2025; Park and Lee, 2025), and that cross-lingual retrieval quality itself can vary significantly depending on query–document language mismatch (Amiraz et al.). These challenges are closely related to long-tail knowledge effects, where low-frequency linguistic and cultural information is underrepresented and more prone to failure (Badhe et al., 2026). Degradation on **accented speech** is also well documented (Najafian and Russell, 2020; Styles et al., 2023; Feng et al., 2021; Sanabria et al., 2023; Shi et al., 2021), although largely only for English dialects and accents. The availability of task data is commonly the bottleneck for such studies.

Studies have also looked at robustness of LLMs to orthographic variants, typos, and other kinds of noise using synthetic data (Belinkov and Bisk, 2018; Heigold et al., 2018; Zhu et al., 2024; Peters and Martins, 2025). Moradi and Samwald (2021) perform a similar study of BERT-like models for sentiment analysis, QA, and NER, among other tasks, with the intent of stress-testing LMs against natural user-generated noise such as synonym replacement, common misspellings, and verb tense errors. Wang et al. (2023) discuss the robustness of ChatGPT against adversarial and out-of-distribution inputs.

Cross-lingual transfer and multilingual generalization There is a rich literature in understanding the phenomenon of cross-lingual transfer and multilingual generalization, by which existing capabilities in high-resource language transfer to related languages and **dialects** (Vaswani et al., 2017; Devlin et al., 2019; Conneau et al., 2020; Pires et al., 2019). This includes studying the conditions and factors facilitating transfer, such as typological relatedness, script, token overlap, among others (Chai et al., 2022; Ri and Tsuruoka, 2022; Muller et al., 2021a; Khemchandani et al., 2021; K et al., 2020; Deshpande et al., 2022; Cahyawijaya et al., 2021; Dhamecha et al., 2021; Lim et al., 2024).

Studying intrinsic properties of multilingual language models Previous work seeks to understand inherent limitations of multilingual models, including the "curse of multilinguality" (Conneau et al., 2020), and the internal structure of the representation space of neural language models with respect to language agnosticity and language specificity (Libovický et al., 2019; Singh et al., 2019). The need for transfer via shared processing while maintaining language-specific features and processing and avoiding negative interference is shown to be an inherent trade-off (Wang et al., 2020; Han et al., 2025).

Previous work, usually studying mid-resource languages, extended by prior work to **low-to-mid resource languages** demonstrates that multilingual LLMs process input in a "language-agnostic" space, often showed to be close to English, by demonstrating the results of early decoding from intermediate layers for a non-English input primarily consisting of English tokens of the correct semantics (Wendler et al., 2024; Foroutan et al., 2022; Tang et al., 2024; Kojima et al., 2024; Schut et al., 2025; Bafna et al., 2025b). Initial and final layers of the LLM are therefore shown to be processing more language-specific subspaces in order to comprehend inputs and produce correct-language outputs (Muller et al., 2021b; Zhao et al., 2024; Lim et al., 2025). Wang et al. (2024) analyzes how this "cross-lingual representation alignment" emerges throughout pre-training. Previous works also show that the performance in a certain language is correlated by the representation similarity to semantically equivalent English sentences (Li et al., 2024b; Zhang et al., 2023; Li et al., 2025c; Liu and Niehues, 2025), and lack of alignment affects cross-lingual consistency and accuracy for factuality recall (Lu et al., 2025; Wang et al., 2025a).

2.2 Improving multilingual performance for comprehension and generation

Curating training data One of the main thrusts in improving massive multilinguality is naturally training and fine-tuning data collection on a large scale, including **extremely low-resource languages**. This includes web-crawled monolingual and parallel datasets with low-resource languages such as MADLAD (Kudugunta et al., 2023), GLOT500 (Imani et al., 2023), NLLB (NLLB Team et al., 2024), and the SMOL suite (Caswell et al., 2025). Note that these may be noisy and of unclear quality due to the scarcity of high-quality LRL content on the web (Kreutzer et al.) as well as LID quality issues for LRLs (Kargaran et al., 2023). There have also been manual data collection efforts focusing on particular language groups, such as Masakhane (Nekoto et al., 2020), Turkish Interlingua (Mirzakhalov et al., 2021), Kreyol-MT (Robinson et al., 2024), HinDialect (Bafna et al., 2022), as well as efforts for particular languages, such as Bhojpuri (Kumar et al., 2023), Yoruba (Adelani et al., 2021; Akpobi, 2025), Quechua (Ahmed et al., 2023), among many others.

Data manipulation, synthetic data creation, and inference-time techniques Previous researchers have developed methods of injecting noise into training data, including orthographic variants or typos Belinkov and Bisk (2018); Heigold et al. (2018), grammatical errors (Anastasopoulos et al., 2019), learned noise types (Brahma et al., 2023), and bilingual lexicon-induced code-mixing (Jones et al., 2023; Xia et al., 2019) in order to induce robustness to typos or **dialectal variants**. These approaches require monolingual data or bilingual lexicons for from-scratch training.

Previous work also looks at leveraging resources and tools for **low-resource** or **unseen languages** in the absence of pretraining data presence. This includes using lexicons, morphological tools, such as morphological analysers and interlinear glossed text to augment the prompt with relevant information (Elsner and Needle, 2023; Ghazvininejad et al., 2023; Zhang et al., 2024b; Lu et al., 2024; Zhang et al., 2025; Ramos et al., 2025; Dimakis et al., 2024), as well as syntactic information, such as curated syntax rules and grammar documentations (Zhang et al., 2024a; Hus and Anastasopoulos, 2024; Hus et al., 2025). This also includes investigating the use of derivable features like phonotactic information for speech (Shahin et al., 2023; Liu et al., 2022). Li et al. (2025b) compares few-shot ICL with provided word meanings with fine-tuning for low-resource languages.

Leveraging cross-lingual transfer Previous works looks at leveraging insights regarding cross-lingual transfer to induce performance on related languages or **dialects** by training on related high-resource languages (Cahyawijaya et al., 2021; Dhamecha et al., 2021; Bafna et al., 2023) among many others. Studies have also looked into attempting to overcome the trade-off between multilingual generalization and the need for language specific processing by introducing explicit modularity (Pfeiffer et al., 2022; Parović et al., 2022; Pfeiffer et al., 2020; Ansell et al., 2021). Other work presents general recipes in adapting a pretrained model to a new language (Csaki et al., 2023; Zamir et al., 2026; Team et al., 2026).

Multilingual alignment Previous also looks at explicit alignment of processing of multilingual inputs with dominant language process in order to take advantage of capabilities of the latter, usually for **low-resource languages** and some **mid-resource languages**. Several works explore a *cascaded system*, i.e. translating the input to English and performing process in English, translating outputs back to the target language if relevant (Qin et al., 2023; Etxaniz et al., 2024; Liu et al., 2025a; Chiu and Bai, 2025). Translation-based approaches are also very competitive in multilingual RAG pipelines (Moon et al., 2025; Ranaldi, 2026). Several of these works show that this consistently outperforms direct inference. Previous work has also using steering vectors to increase representational alignment, by explicitly steering internal vectors towards the shared space (Lu et al., 2025; Wang et al., 2025a; Lim et al., 2025). Wang et al. (2025b) align LLM representations for **mid-resource languages** at inference time at final layers.

3 Overview of work

Table 1 shows my previous and proposed work, categorized by 1) whether it seeks to *evaluate* current approaches, or develop strategies for *improving* them or mitigating discovered problems 2) whether it is concerned with *comprehension* or *generation*, and 3) the resource-range of languages that it deals in.

	Comprehension	**Generation**
Evaluating	**(1)** [a] Evaluating Large Language Models along Dimensions of Language Variation: A Systematik Invesdigatiom uv Cross-lingual Generalization [**EMNLP 2024**] **(4)** [b] ChiKhaPo: A Large-Scale Multilingual Benchmark for Evaluating Lexical Comprehension and Generation in Large Language Models [**ACL 2026**] **(5)** [a,c] LID Models are Actually Accent Classifiers: Implications and Solutions for LID on Accented Speech [**Interspeech 2025**] **(6)** [a] How Important is "Perfect" English for Machine Translation Prompts? [**EACL 2026 (Findings)**] **(7)** [a] RASHID: A Cipher-Based Framework for Exploring In-Context Language Learning [**Under review for EMNLP 2026**]	**(3)** [d] The Translation Barrier Hypothesis: Multilingual Generation with LLMs Suffers from Implicit Translation Failure [**IJCNLP-AACL 2025**] **(4)** [b] ChiKhaPo: A Large-Scale Multilingual Benchmark for Evaluating Lexical Comprehension and Generation in Large Language Models [**ACL 2026**] **(7)** [a] RASHID: A Cipher-Based Framework for Exploring In-Context Language Learning [**Under review for EMNLP 2026**] **(9)** Towards Multilingual Retrieval across Hundreds of Languages [**Proposed**] **(10)** Addressing the Agnosticity-Specificity Trade-off in Multilingual Text Embedddings [**Proposed**]
Improving	**(2)** [a,c] DialUp! Modeling the Language Continuum by Adapting Models to Dialects and Dialects to Models [**ACL 2025**] **(5)** [a,c] LID Models are Actually Accent Classifiers: Implications and Solutions for LID on Accented Speech [**Interspeech 2025**] **(8)** [b,c] Omnilingual MT: Machine Translation for 1,600 Languages [Preprint] **(11)** The Role of the Machine Translation Cascade in the era of LLMs [**Proposed**]	**(7)** [a] RASHID: A Cipher-Based Framework for Exploring In-Context Language Learning [**Under review for EMNLP 2026**] **(8)** [b,c] Omnilingual MT: Machine Translation for 1,600 Languages [Preprint] **(9)** Towards Multilingual Retrieval across Hundreds of Languages [**Proposed**] **(10)** Addressing the Agnosticity-Specificity Trade-off in Multilingual Text Embedddings [**Proposed**] **(11)** The Role of the Machine Translation Cascade in the era of LLMs [**Proposed**]

Table 1: Completed work and proposed directions. Completed papers are (co-)first-author, excepting (4), which I mentored as second / last author, and (8), which describes work I led as part of a larger effort and is not a standalone paper. Colours indicate the class of resourcedness of the languages that the paper deals in: extremely low-resource language, dialect/accent/variant, mid-to-low resource language. Letter superscripts indicate the nature of the work: [a] data manipulation strategy, [b] dataset design, [c] modeling/training, [d] interpretability.

4 DialUp! Modeling the Language Continuum by Adapting Models to Dialects and Dialects to Models [completed]

Profile: We want to *improve* the *comprehension* of *dialectal variants* of high-resource languages.

4.1 Motivation

The roughly 6800 languages in the world can be grouped into a few hundred families. Closely-related languages (CRLs) and dialects within a family largely exhibit continuous and structured differences rather than being discrete monoliths (Hovy and Purschke, 2018; Bergman and Diab, 2022). However, it is unfeasible to collect training data for every lect; further, language is fluid and dynamic across and between dialect boundaries, calling for principled general approaches to dialectal variation.

We seek to expand models proficient in high-resource languages (HRLs) to their lower-resource CRLs, by inducing robustness on the input side of models to unseen varieties across language continua. We do this with a dual method DialUp: adapting models from an HRL to its language relatives, and adapting those relatives' data towards the HRL model, as in Figure 1 (Bafna et al., 2025a).

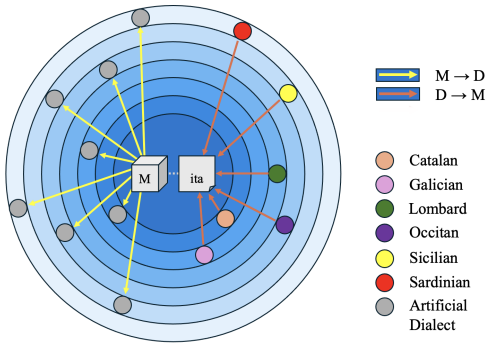

Figure 1: Two paradigms for robustness to dialects on a continuum of distances from an HRL. DialUp involves `M→D`: training the **model** on artificial dialectal variation, and `D→M`: bringing dialectal **data** closer to model expectations (HRL-like input) at inference.

4.2 Method

`M→D` We focus on an important aspect of dialectal variation: cognate lexical change induced by sound shift. We employ the approach of Bafna et al. (2024) to simulate language variation in terms of commonly observed sound shifts within dialects of a language family, thus creating artificial languages that mimic dialectal phenomena. Each artificial language is defined as the map of changes made from the HRL.

`D→M` In this paradigm, given CRL-HRL bilingual lexicons, we swap divergent words of CRL input with known HRL equivalents to pull data towards the model's proficiency distribution. We explore three `D→M` settings: `func`, `cont`, and `all`; in which we swap out only function words, only content words, and both, respectively.

Baselines We compare all proposed approaches to three baselines: (1) evaluating the model on CRL→`eng` MT without any adaptation (`off-the-shelf`) (2) evaluating on CRL→`eng` after fine-tuning on ordinary HRL→`eng` bitext without any simulated linguistic variation (`fthrl`), and (3) fine-tuning and evaluation after augmenting the HRL→`eng` bitext with completely random (i.e. not linguistically motivated or plausible) variations (`randaug`).

We work with X→`eng` MT for 49 languages in six (sub)families.

4.3 Results

Results show that our approach gives gains across the board, as compared to off-the-shelf baselines as well as other kinds of non-linguistically-motivated data augmentation, with varied trends by language family.

`D→M` introduces gains for many low-performing varieties where replacing dialectal words with HRL equivalents is helpful. In particular, we note that `D→M-func` consistently outperforms `D→M-cont`, highlighting the importance of treating dialectal function words in enabling model comprehension, and the utility of collecting CRL-HRL function word maps for low-resource languages.

Analysis We conduct an analysis to understand the features that determine the impact that DialUp has for a language. We find that languages with poorer baseline scores, which tend to be the low-performing varieties we aim to assist in this work, benefit more from DialUp, while languages with better baselines show small or negative gains. Proximity to the HRL, measured by string overlap in bitext, also appears to play a role: CRLs distant from their HRL, such as Samoan and Tagalog, as well as those extremely close to the

HRL, such as Malay (zsm), Azerbaijani (aze), and Saudi Arabic (ars), appear to benefit little from DialUp. We also see that M2M typically benefits more. This may be because aya-23 was likely exposed to some CRLs in pre-training, despite not being trained explicitly on them, given the heterogenity and poor documentation of LLM pre-training sets.

DialUp introduces a promising paradigm for making NLP flexible to potentially unseen and undocumented dialectal variation. We release the code for DialUp as a Python package.[2]

5 The Translation Barrier Hypothesis: Multilingual Generation with LLMs Suffers from Implicit Translation Failure [completed]

Profile: We want to *diagnose* multilingual *generation* for *mid-resource languages*.

5.1 Motivation

We lack a systematic understanding of the reasons for and mechanisms of the failure of multilingual generation with LLMs. Building upon intuition of previous work in mechanistic interpretability (Wendler et al., 2024; Foroutan et al., 2022; Tang et al., 2024; Kojima et al., 2024), we posit a *model-internal task-solving→translation* cascade, with the middle layers of the LLM responsible for *task-solving*, or discovering the required output concepts, in a target-language-agnostic manner, and the last few layers responsible for realizing those concepts in the target language, through *translation*.

We posit the **Translation Barrier Hypothesis**, which states that *translation failure in LLMs accounts for a large proportion of poor quality final outputs for multilingual generation, where translation refers to model-internal transfer of answer concepts into the intended output language, given an implicit* task-solving→translation *cascade.* See this hypothesis depicted in Figure 2 (Bafna et al., 2025b).

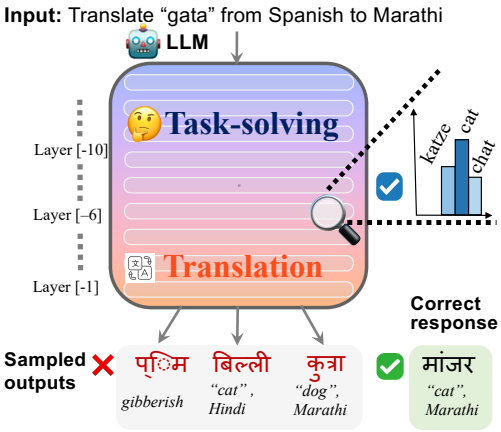

Figure 2: *Task-solving* succeeds with the correct answer concept (i.e. cat) discovered in intermediate layers, expressed in various HRLs, but the model fails to realize or *translate* the concept into the target LRL.

5.2 Method

Word translation We work with a word translation task, and curate a dataset of lexical equivalents and synonyms parallel across all studied languages. *Task-solving success* here consists of realizing the semantics of the source word in any non-source language in any non-final layer, and *task-solving → translation success* consists of producing the correct final answer in the intended source language at the final layer. Success is measured as a binary accuracy metric per word with exact string match over lexical translation equivalents.

Quantifying translation loss proportion We can study the translation barrier hypothesis in the context of a given task by quantifying the extent to which translation failure is responsible for poor final outputs, as opposed to task-solving failure. We use logit lens (Nostalgebraist, 2020) to obtain potentially multi-token text outputs from intermediate layers. We define *translation loss proportion* (*TLP*) as the proportion of incorrect final outputs that had correct task-solving (implying failed translation). We study 36 target languages, including a mix of mid-to-low resource languages.

[2] https://pypi.org/project/dialup/

5.3 Results

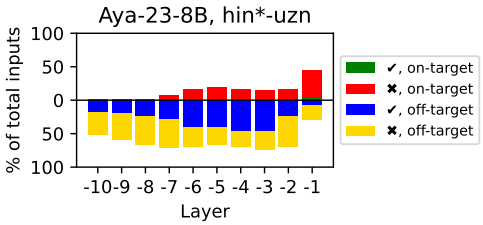

Figure 3: Layerwise evolution of on-target and off-target accuracy and inaccurate model responses. * supported language

Layerwise analysis We plot on-target and off-target accuracy for all language pairs over all outputs for which we found a reliable language tag (see Figure 3 for an example pair). We consistently observe high off-targetness in middle layers, including for Spanish and other supported or high-resource target languages, with considerable off-target accuracy, which subsequently decreases as the model pivots to the desired language. For supported or HRL targets, there is a corresponding increase in on-target outputs and accuracy in later layers, indicating successful translation. On the other hand, for Uzbek and other unsupported or low-resource languages, we simply observe a total drop in correct (on-target or off-target) intermediate outputs as the model moves away from off-target equivalents but also fails to translate correctly, resulting in poor final outputs. *This provides intuition for the translation barrier for LRL generation with LLMs.*

Quantifying the translation barrier We compute: (a) task-solving success or intermediate accuracy (the percentage of task outputs that were correct at *some* intermediate layer in *some* language) (b) final success (on-target accuracy at the final layer), and (c) *TLP*(the proportion of total failure that can be attributed to translation failure.) We look at these three quantities averaged over all target languages for each source language. We see that intermediate accuracy is much higher than final accuracy in all cases, with much less variance, indicating that the former is stabler and less dependent on the target language than the latter. This ties in with our understanding of the relatively target-language agnostic task-solving step, followed by a target-language dependent translation step. *Overall, translation loss accounts for a high percentage of total failure (more than* 50% *in all cases except one). This lends evidence to the translation barrier hypothesis.*

6 Rashid: A Cipher-Based Framework for Exploring In-Context Language Learning [completed]

Profile: We want to *evaluate* the *comprehension* and *generation* of *unseen languages* by LLMs, and introduce strategies targeted at *improving* generation.

6.1 Motivation

The task of processing unseen or extremely low-resource languages with large language models (LLMs), first introduced by Tanzer et al. (2023), holds the potential to make important strides in extending the benefits of LLMs to the long tail of languages. This is typically done by augmenting the prompt context with information about the language, using materials such as lexicons, grammar books and other tools for linguistic processing, as explored by several subsequent works (Zhang et al., 2024b, 2025; Hus and Anastasopoulos, 2024).

However, progress in the field of in-context language learning (ICLL), by its nature, is hindered by resource scarcity. The unseen or truly low-resource languages of relevance to ICLL are the very languages for which we lack a) quality evaluation strategies including access to human expertise, b) high-quality resources for cheaply exploring a range of new strategies without significant investment, and c) task data from frontier benchmarks, for testing the performance of these strategies in various application scenarios. Further, insights from these setups regarding language learning are often confounded by the model's differing existing capabilities in various low-resource languages.

6.2 Approach

We introduce RASHID, a general-purpose framework for systematically assessing and exploring the phenomenon of ICLL (Bafna et al., 2026). We construct truly unseen languages by reversibly ciphering high-resource languages (HRLs), preserving underlying linguistic characteristics, and providing us access to a wide array of available resources in these languages for experimentation. We use bijective monoalphabetic substitution ciphers for this purpose. Crucially, this allows us to bypass the above problems of resource scarcity and factors confounding insights, enabling sophisticated evaluation as well as the assessment of techniques otherwise requiring potentially expensive resource development for true LRLs.

Thus, we 1) construct a sandbox for exploring in-context language learning for languages guaranteed to be unseen, with access to the range of NLP tools and resources available for HRLs. We use it for 2) assessing current methods for ICLL for machine translation (MT) in both directions 3) exploring new research directions for improving ICLL, focusing on generation 4) expanding ICLL strategy evaluation beyond machine translation to rich downstream tasks.

6.3 Findings

We find that current strategies show intended benefits for *comprehension* of unseen languages, although they remain ineffective for *generation*, and that outputs remain low-quality for both directions. We present a schema of challenges for this paradigm of language learning. Next, our exploration of new directions shows that near-perfect lexicon coverage would be helpful, and that cascading through a related high-resource language for generation makes considerable progress on the problems of syntactic coherence and translation of idiomatic language. This promotes the development of rich parallel materials between related languages rather than with English. Finally, we find that added materials may not be directly beneficial for downstream tasks, but that a cascaded approach (MT with ICLL $\rightarrow$ task-solving) improves performance across the board. This underlines the need for task-oriented exploration of ICLL.

This work aims to provide a grounded assessment of current techniques in in-context language learning, as well as enable further progress despite the many challenges of the field.

7 Other completed work and next directions

Here are brief summaries of other work shown in Table 1.

(1) Systematik This work aims to understand how various aspects of dialectal divergence impact zero-shot performance of LLMs by using controlled Bayesian "noising" procedures mimicking dialectal variation (Bafna et al., 2024).

(4) ChiKhaPo This work curates a massively multilingual benchmark that evaluates basic lexical understanding and generation across thousands of languages using 4 subtasks in both directions (Chang and Bafna, 2025).

(5) LID identifies a major failure mode of speech language identification models on accented speech in misclassifying it as native language of speakers, and proposes a solution using explicit phonological features (Bafna and Wiesner, 2025).

(8) MeDLEy develops a protocol for curating training bitext for low-resource languages from scratch by ensuring grammatical, length, and domain diversity. This work was done as part of the larger effort towards omnilingual machine translation at Meta, and is described in Section 4 and Appendix A of the combined paper (Team et al., 2026).

Summary of prior work In my prior work, I explored low-data regime techniques for linguistic variants and extremely low-resource languages in the absence of data, dataset design for training and benchmarking for extremely low-resource languages, and interpreting the failure of LLMs on low-to-mid-resource languages.

Proposed work: General directions In general, my planned projects focus on *diagnosing and improving* LLM performance for **low-to-mid resource** languages, as shown in Table 1. My near-term projects focus on (a) a critical functionality of LLMs: their ability to leverage external information in any given situation i.e. gaining access to required materials for an informed response, as well as (b) an important technique for multilinguality that is widely opted in various scenarios: the use of a machine translation cascade. The following sections describe directions and may be split into more than one project each.

8 Towards multilingual retrieval across hundreds of languages [proposed]

Motivation LLMs increasingly serve as intelligent interfaces with the web or with the digital information infrastructure, in order to solve a variety of reasoning, search, and other tasks. This requires functionality for searching and using information external to that stored in model parametric memory. This is the goal of retrieval-augmented generation (RAG) (Lewis et al., 2020).

Retrieval is an important hurdle for multilingual and cross-lingual RAG, and relies on good text embeddings for embedding-based retrieval (Goworek et al., 2025; Enevoldsen et al., 2025). However, constructing robust and reliable multilingual semantic spaces is still an unsolved problem, from the times of static embeddings, to contextual embeddings with mBERT, through LLMs, despite various efforts in better multilingual alignment (Søgaard et al., 2018; Wada et al., 2019). This can be understoood as the root cause of poor retrieval of documents in low-resource languages during cross-lingual retrieval, and the related observed phenomenon where models prefer to retrieve documents in high-resource languages (Goworek et al., 2025; Nouri and Yangarber, 2014).

Goal We want to mitigate the effects of poor multilingual embedding spaces on multilingual retrieval, in a post-hoc and training-free manner. I propose to achieve this via adjustments in multilingual retrieval based on observed insights on the embedding space. I propose to first investigate whether we can formalize and quantify the phenomenon of quality degradation in an embedding space, given a language and a "concept" to be embedded. I then propose to explore methods of predicting this effect based on language resourcedness and concept rareness. Finally, I propose to leverage this understanding to design interventions for multilingual retrieval, such as adjusted scoring over multilingual documents.

Plan This study requires the following components. (1) *Formulation of a "concept"*, which may be at the document level, sentence level, or a more atomic fact or claim; (2) *formulation of the "rareness"* of a concept: one simple approach is using n-gram frequency over the wording of the concept from a large monolingual corpus, and more complex methods may consider several rewordings of the phrase, (3) formulation of the extent to which a particular concept is *"well-embedded"*, potentially using notions such as confidence or spatial distortion, (4) a *retrieval corpus* that is decomposed into concepts and contains concepts with *different rareness*, where Wikipedia is the most commonly used retrieval corpus and naturally contains variously rare concepts, though we can also consider corpora that are parallel across studied languages for more controlled experiments, constructed by translation or quasi-parallel subsets of Wikipedia, along with queries requiring retrieval of variously rare concepts across several languages, either from existing datasets or synthetically constructed; (5) a *controlled study of extrinsic and intrinsic measures of embedding quality* of concepts of varying rareness across languages of varying resourcedness; and (6) *methods for better cross-lingual retrieval* e.g. language-specific adjustments to ranking or scoring according to our knowledge of embedding quality for a given concept across languages.

9 Addressing the agnosticity-specificity tradeoff for multilingual text embedding spaces [proposed]

Motivation Multilingual transformer-based data-driven models exhibit an inherent tension between language *agnosticity* (cross-lingual access of information and consistency) and *specificity* (culturally and linguistically appropriate alignment) (Han et al., 2025). The corresponding observed failure modes can be understood as the **knowledge gap** (information in one language is not accessible in another) and **localization bias** (the model tends to localize itself in dominant viewpoints) (Hershcovich et al., 2022; Naous et al., 2024). While this phenomenon has been extensively discussed, with several mitigation strategies proposed (e.g. Pfeiffer et al. (2022)), this is usually from an observational standpoint, often with lack of formalization of the desired outcome, given the inherent trade-off between and desirability of these behaviours.

Goal I propose to view this problem from a design standpoint. 1) Can we describe an ideal outcome with respect to various input scenarios? 2) Can we describe the characteristics of a representation space that would yield such as outcome? 3) Can we use this formulation to construct metrics for real representation spaces or understand implicit constraints, and propose strategies for training such a space in practice?

While I propose to use retrieval as a testbed for this study, note that these ideas are generally applicable to any multilingual Transformer-based data-driven system trained in a similar manner.

Plan I describe one line of attack as follows. (1) *Defining the ideal outcome*: first, we construct a schema of possible scenarios that may require agnostic, specific, or a combination of retrieval tactics in the ideal outcome; (2) *Constructing a synthetic dataset*: we construct a dataset containing all the above scenarios, which may be a toy dataset with controlled and minimal variation in features of interest (language, content), designed to stress test embedding spaces, and perfectly solvable by an "ideal" multilingual representation space; (3) *Constructing a representation space*: we search for *necessary* constraints that the above requirements impose on an "ideal" representation space (similarly to, for example, Weller et al. (2026)), and also explore *sufficient* conditions, e.g., some of many possible solutions to this problem; (4) constructing *metrics* for multilingual representation spaces using the above; (5) *evaluating existing text embedding models* with these metrics; and (6) *constructing intervention methods* inspired by these metrics to improve existing text embedding models.

10 The role of the machine translation cascade in the era of LLMs [proposed]

Motivation Using an MT cascade at various points in the LLM pipeline (translating inputs, outputs, queries, external documents) has many advantages, both for mid and low-resource languages, both in terms of performance, as well as in terms of ease of generalization of powerful reasoning and RAG capabilities of LLMs to other languages (Liu et al., 2024). This lets us perform complex reasoning in English, and rely on translation systems to interface with the desired language, thus avoiding the need to optimize performance on individual languages. However, this comes with the following problems. Firstly, MT may be fallible on the same edge cases for which end-to-end performance is weak, such as rare concepts. Secondly, the MT cascade may introduce errors that may actually degrade performance. Finally, there may be particular queries where access to culture- or language-specific information is required. In these cases, performing translation of the English content would be inadequate, and we would like our MT system to actually modify the content or to disregard it (MT+).

Goals and plan We aim to 1) conduct a comparison of the MT cascade with natively multilingual approaches given a variety of input scenarios (source and target language, task, input, domain, model used) 2) contribute recommendations on the usage of MT cascades at various points in an LLM pipeline given various input scenarios, and 3) (if time) develop MT+ systems tailored to an LLM cascade.

11 Timeline

- Spring and summer 2026: Addressing the agnosticity-specificity tradeoff for multilingual text embedding spaces
- Fall 2026: Towards multilingual retrieval across hundreds of languages
- Spring 2027: The role of the machine translation cascade in the era of LLMs
- Summer 2027: Industry internship
- Fall 2027: Filling holes in proposed research agenda, overthinking, and fun collaborations.
- Spring 2028: Thesis writing

References

David Ifeoluwa Adelani, Dana Ruiter, Jesujoba O. Alabi, Damilola Adebonojo, Adesina Ayeni, Mofe Adeyemi, Ayodele Esther Awokoya, and Cristina España-Bonet. 2021. The effect of domain and diacritics in Yoruba–English neural machine translation. In *Proceedings of Machine Translation Summit XVIII: Research Track*, pages 61–75, Virtual. Association for Machine Translation in the Americas.

Nouman Ahmed, Natalia Flechas Manrique, and Antonije Petrović. 2023. Enhancing Spanish-Quechua machine translation with pre-trained models and diverse data sources: LCT-EHU at AmericasNLP shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 156–162, Toronto, Canada. Association for Computational Linguistics.

Maro Akpobi. 2025. Yankari: Monolingual Yoruba dataset. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 1–6, Vienna, Austria. Association for Computational Linguistics.

Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan, The cross-lingual cost: Retrieval biases in rag over arabic-english corpora.

Antonios Anastasopoulos, Alison Lui, Toan Q. Nguyen, and David Chiang. 2019. Neural machine translation of text from non-native speakers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3070–3080, Minneapolis, Minnesota. Association for Computational Linguistics.

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R. Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, Jean Maillard, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Eduardo Sánchez, Ioannis Tsiamas, Arina Turkatenko, Albert Ventayol-Boada, and Shireen Yates. 2025. BOUQuET : dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27515–27535, Suzhou, China. Association for Computational Linguistics.

Alan Ansell, Edoardo Maria Ponti, Jonas Pfeiffer, Sebastian Ruder, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2021. MAD-G: Multilingual adapter generation for efficient cross-lingual transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4762–4781, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Jon Ander Campos, Yi Chern Tan, Kelly Marchisio, Max Bartolo, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Aidan Gomez, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024a. Aya 23: Open weight releases to further multilingual progress.

Viraat Aryabumi, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh, Madeline Smith, Kelly Marchisio, Sebastian Ruder, Acyr Locatelli, Julia Kreutzer, Nick Frosst, Phil Blunsom, Marzieh Fadaee, Ahmet Üstün, and Sara Hooker. 2024b. Aya 23: Open weight releases to further multilingual progress.

Akari Asai, Sneha Kudugunta, Xinyan Velocity Yu, Terra Blevins, Hila Gonen, Machel Reid, Yulia Tsvetkov, Sebastian Ruder, and Hannaneh Hajishirzi, Buffet: Benchmarking large language models for few-shot cross-lingual transfer. *arXiv preprint arXiv:2305.14857*.

Sanket Badhe, Deep Shah, and Nehal Kathrotia, Long-tail knowledge in large language models: Taxonomy, mechanisms, interventions and implications. (arXiv:2602.16201). ArXiv:2602.16201 [cs].

Niyati Bafna, Emily Chang, Nathaniel R Robinson, David R Mortensen, Kenton Murray, David Yarowsky, and Hale Sirin, Dialup! modeling the language continuum by adapting models to dialects and dialects to models. *arXiv preprint arXiv:2501.16581*.

Niyati Bafna, Cristina España-Bonet, Josef Van Genabith, Benoît Sagot, and Rachel Bawden. 2023. Cross-lingual strategies for low-resource language modeling: A study on five Indic dialects. In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 28–42, Paris, France. ATALA.

Niyati Bafna, Tianjian Li, Kenton Murray, David R. Mortensen, David Yarowsky, Hale Sirin, and Daniel Khashabi. 2025b. The translation barrier hypothesis: Multilingual generation with large language models suffers from implicit translation failure. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1541–1568, Mumbai, India. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics.

Niyati Bafna, Kenton Murray, and David Yarowsky. 2024. Evaluating large language models along dimensions of language variation: A systematik invesdigatiom uv cross-lingual generalization. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18742–18762.

Niyati Bafna, Ryan Soh-Eun Shim, Barbara Plank, David Yarowsky, and Hale Sirin, Rashid: A cipher-based framework for exploring in-context language learning. *arXiv preprint arXiv:2603.22497*.

Niyati Bafna, Josef van Genabith, Cristina España-Bonet, and Zdeněk Žabokrtský. 2022. Combining noisy semantic signals with orthographic cues: Cognate induction for the Indic dialect continuum. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 110–131, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Niyati Bafna and Matthew Wiesner, Lid models are actually accent classifiers: Implications and solutions for lid on accented speech. *arXiv preprint arXiv:2506.00628*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al., A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Yonatan Belinkov and Yonatan Bisk. 2018. Synthetic and Natural Noise Both Break Neural Machine Translation.

A. Bergman and Mona Diab. 2022. Towards responsible natural language annotation for the varieties of Arabic. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 364–371, Dublin, Ireland. Association for Computational Linguistics.

Maharaj Brahma, Kaushal Maurya, and Maunendra Desarkar. 2023. SelectNoise: Unsupervised noise injection to enable zero-shot machine translation for extremely low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1615–1629, Singapore. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung, LLMs are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. Indonlg: Benchmark and resources for evaluating indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isaac Caswell, Elizabeth Nielsen, Jiaming Luo, Colin Cherry, Geza Kovacs, Hadar Shemtov, Partha Talukdar, Dinesh Tewari, Baba Mamadi Diane, Djibrila Diane, Solo Farabado Cissé, Koulako Moussa Doumbouya, Edoardo Ferrante, Alessandro Guasoni, Christopher Homan, Mamadou K. Keita, Sudhamoy DebBarma, Ali Kuzhuget, David Anugraha, Muhammad Ravi Shulthan Habibi, Sina Ahmadi, Anthony Munthali, Jonathan Mingfei Liu, and Jonathan Eng. 2025. SMOL: Professionally translated parallel data for 115 under-represented languages. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1103–1123, Suzhou, China. Association for Computational Linguistics.

Yuan Chai, Yaobo Liang, and Nan Duan. 2022. Cross-lingual ability of multilingual masked language models: A study of language structure. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 4702–4712, Dublin, Ireland. Association for Computational Linguistics.

Emily Chang and Niyati Bafna, Chikhapo: A large-scale multilingual benchmark for evaluating lexical comprehension and generation in large language models. *arXiv preprint arXiv:2510.16928*.

Yi-Ting Chiu and Zong-Han Bai, Translation or multilingual retrieval? evaluating cross-lingual search strategies for traditional chinese financial documents. *Proceedings of the FinTech in AI CUP Special Session, Tokyo, Japan*, 10.

Hyung Won Chung, Noah Constant, Xavier Garcia, Adam Roberts, Yi Tay, Sharan Narang, and Orhan Firat. 2023. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, page 8440–8451, Online. Association for Computational Linguistics.

Zoltan Csaki, Pian Pawakapan, Urmish Thakker, and Qiantong Xu, Efficiently adapting pretrained language models to new languages. *arXiv preprint arXiv:2311.05741*.

Ameet Deshpande, Partha Talukdar, and Karthik Narasimhan. 2022. When is bert multilingual? isolating crucial ingredients for cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 3610–3623, Seattle, United States. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, page 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tejas Dhamecha, Rudra Murthy, Samarth Bharadwaj, Karthik Sankaranarayanan, and Pushpak Bhattacharyya. 2021. Role of Language Relatedness in Multilingual Fine-tuning of Language Models: A Case Study in Indo-Aryan Languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8584–8595, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antonios Dimakis, Stella Markantonatou, and Antonios Anastasopoulos. 2024. Dictionary-aided translation for handling multi-word expressions in low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2588–2595, Bangkok, Thailand. Association for Computational Linguistics.

Micha Elsner and Jordan Needle. 2023. Translating a low-resource language using gpt-3 and a human-readable dictionary. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–13.

Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al., Mmteb: Massive multilingual text embedding benchmark. *arXiv preprint arXiv:2502.13595*.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. Do multilingual language models think better in English? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying Bias in Automatic Speech Recognition.

Negar Foroutan, Mohammadreza Banaei, Rémi Lebret, Antoine Bosselut, and Karl Aberer. 2022. Discovering language-neutral sub-networks in multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7560–7575, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Marjan Ghazvininejad, Hila Gonen, and Luke Zettlemoyer, Dictionary-based phrase-level prompting of large language models for machine translation. *arXiv preprint arXiv:2302.07856*.

Roksana Goworek, Olivia Macmillan-Scott, and Eda B Özyiğit, What drives cross-lingual ranking? retrieval approaches with multilingual language models. *arXiv preprint arXiv:2511.19324*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al., The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

HyoJung Han, Sweta Agrawal, and Eleftheria Briakou, Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms. *arXiv preprint arXiv:2510.26024*.

Tahmid Hasan, Abhik Bhattacharjee, Md. Saiful Islam, Kazi Mubasshir, Yuan-Fang Li, Yong-Bin Kang, M. Sohel Rahman, and Rifat Shahriyar. 2021. XL-sum: Large-scale multilingual abstractive summarization for 44 languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4693–4703, Online. Association for Computational Linguistics.

Georg Heigold, Stalin Varanasi, Günter Neumann, and Josef van Genabith. 2018. How Robust Are Character-Based Word Embeddings in Tagging and MT Against Wrod Scramlbing or Randdm Nouse? In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 68–80, Boston, MA. Association for Machine Translation in the Americas.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Christoph Purschke. 2018. Capturing regional variation with distributed place representations and geographic retrofitting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4383–4394, Brussels, Belgium. Association for Computational Linguistics.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219.

Jonathan Hus, Antonios Anastasopoulos, and Nathaniel Krasner. 2025. Machine translation using grammar materials for llm post-correction. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 92–99.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André FT Martins, François Yvon, et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, and Zhaopeng Tu. 2023. Is ChatGPT A Good Translator? Yes With GPT-4 As The Engine.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold, Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

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

Karthikeyan K, Zihan Wang, Stephen Mayhew, and Dan Roth, Cross-lingual ability of multilingual bert: An empirical study. (arXiv:1912.07840). ArXiv:1912.07840 [cs].

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7226–7245, Singapore. Association for Computational Linguistics.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Yash Khemchandani, Sarvesh Mehtani, Vaidehi Patil, Abhijeet Awasthi, Partha Talukdar, and Sunita Sarawagi. 2021. Exploiting language relatedness for low web-resource language model adaptation: An indic languages study. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, page 1312–1323, Online. Association for Computational Linguistics.

Dayeon Ki, Marine Carpuat, Paul McNamee, Daniel Khashabi, Eugene Yang, Dawn Lawrie, and Kevin Duh, Linguistic nepotism: Trading-off quality for language preference in multilingual rag. (arXiv:2509.13930). ArXiv:2509.13930 [cs].

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6919–6971, Mexico City, Mexico. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani10, Artem Sokolov, Claytone Sikasote12, et al., Quality at a glance: An audit of web-crawled multilingual datasets.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat, Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36:67284–67296.

Amit Kumar, Shantipriya Parida, Ajay Pratap, and Anil Kumar Singh, Machine translation by projecting text into the same phonetic-orthographic space using a common encoding. *Sādhanā*, 48(4):238.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT Beyond English: Towards a Comprehensive Evaluation of Large Language Models in Multilingual Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al., Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. 2024a. Bordirlines: A dataset for evaluating cross-lingual retrieval augmented generation. In *Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia*, page 1–13, Miami, Florida, USA. Association for Computational Linguistics.

Chong Li, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2024b. Improving in-context learning of multilingual generative language models with cross-lingual alignment. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8058–8076, Mexico City, Mexico. Association for Computational Linguistics.

Tianjian Li, Haoran Xu, Weiting Tan, Kenton Murray, and Daniel Khashabi. 2025a. Upsample or upweight? balanced training on heavily imbalanced datasets. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3325–3343, Albuquerque, New Mexico. Association for Computational Linguistics.

Yue Li, Zhixue Zhao, and Carolina Scarton. 2025b. It's all about in-context learning! teaching extremely low-resource languages to llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29532–29547.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du, Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28186–28194.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser, How language-neutral is multilingual bert? (arXiv:1911.03310). ArXiv:1911.03310 [cs].

Seonghoon Lim, Taejun Yun, Jinhyeon Kim, Jihun Choi, and Taeuk Kim. 2024. Analysis of multi-source language training in cross-lingual transfer. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 712–725, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Wei Lim, Alham Fikri Aji, and Trevor Cohn, Language-specific latent process hinders cross-lingual performance. *arXiv preprint arXiv:2505.13141*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing, Is translation all you need? a study on solving multilingual tasks with large language models. *arXiv preprint arXiv:2403.10258*.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025a. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Danni Liu and Jan Niehues. 2025. Middle-layer representation alignment for cross-lingual transfer in fine-tuned llms.

Hexin Liu, Leibny Paola Garcia Perera, Andy Khong, Suzy Styles, and Sanjeev Khudanpur. 2022. Pho-lid: A unified model incorporating acoustic-phonetic and phonotactic information for language identification. In *Interspeech 2022*, pages 2233–2237.

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. 2025b. Xrag: Cross-lingual retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, page 15669–15690, Suzhou, China. Association for Computational Linguistics.

Hongyuan Lu, Zixuan Li, and Wai Lam, Dictionary insertion prompting for multilingual reasoning on multilingual large language models. *arXiv preprint arXiv:2411.01141*.

Meng Lu, Ruochen Zhang, Ellie Pavlick, and Carsten Eickhoff, Paths not taken: Understanding and mending the multilingual factual recall pipeline. *arXiv preprint arXiv:2505.20546*.

Jamshidbek Mirzakhalov, Anoop Babu, Duygu Ataman, Sherzod Kariev, Francis Tyers, Otabek Abduraufov, Mammad Hajili, Sardana Ivanova, Abror Khaytbaev, Antonio Laverghetta Jr., Bekhzodbek Moydinboyev, Esra Onal, Shaxnoza Pulatova, Ahsan Wahab, Orhan Firat, and Sriram Chellappan. 2021. A large-scale study of machine translation in Turkic languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5876–5890, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hoyeon Moon, Byeolhee Kim, and Nikhil Verma. 2025. Quality-aware translation tagging in multilingual rag system. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, page 161–177, Suzhuo, China. Association for Computational Linguistics.

Milad Moradi and Matthias Samwald. 2021. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1558–1570, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021a. When being unseen from mbert is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 448–462, Online. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021b. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Maryam Najafian and Martin J. Russell, Automatic accent identification as an analytical tool for accent robust automatic speech recognition. *Speech Communication*, 122:44–55.

Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 16366–16393.

Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohungbe, Solomon Oluwole Akinola, Shamsuddeen Muhammad, Salomon Kabongo Kabenamualu, Salomey Osei, Freshia Sackey, et al. 2020. Participatory research for low-resourced machine translation: A case study in african languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang, Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. Accessed: 2025-03-29.

Javad Nouri and Roman Yangarber. 2014. Measuring language closeness by modeling regularity. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 56–65, Doha, Qatar. Association for Computational Linguistics.

Jeonghyun Park and Hwanhee Lee. 2025. Investigating language preference of multilingual rag systems. In *Findings of the Association for Computational Linguistics: ACL 2025*, page 5647–5675, Vienna, Austria. Association for Computational Linguistics.

Marinela Parović, Goran Glavaš, Ivan Vulić, and Anna Korhonen. 2022. BAD-X: Bilingual adapters improve zero-shot cross-lingual transfer. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1791–1799, Seattle, United States. Association for Computational Linguistics.

Ben Peters and Andre Martins. 2025. Did translation models get more robust without anyone Even noticing? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2445–2458, Vienna, Austria. Association for Computational Linguistics.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, page 3479–3495, Seattle, United States. Association for Computational Linguistics.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 7654–7673.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual bert? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, page 4996–5001, Florence, Italy. Association for Computational Linguistics.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

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

Rita Ramos, Everlyn Asiko Chimoto, Maartje Ter Hoeve, and Natalie Schluter. 2025. Grammamt: Improving machine translation with grammar-informed in-context learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29920–29940.

Leonardo Ranaldi. 2026. Multilingual retrieval-augmented generation for knowledge-intensive question answering task. In *Findings of the Association for Computational Linguistics: EACL 2026*, page 697–716, Rabat, Morocco. Association for Computational Linguistics.

Ryokan Ri and Yoshimasa Tsuruoka. 2022. Pretraining with artificial language: Studying transferable knowledge in language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 7302–7315, Dublin, Ireland. Association for Computational Linguistics.

Nathaniel Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome Etori, et al. 2024. Kreyòl-mt: Building mt for latin american, caribbean and colonial african creole languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3083–3110.

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

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. 2023. The Edinburgh International Accents of English Corpus: Towards the Democratization of English ASR.

Lisa Schut, Yarin Gal, and Sebastian Farquhar, Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.

Mostafa Shahin, Zheng Nan, Vidhyasaharan Sethu, and Beena Ahmed. 2023. Improving wav2vec2-based spoken language identification by learning phonological features. In *Interspeech 2023*, pages 4119–4123.

Xian Shi, Fan Yu, Yizhou Lu, Yuhao Liang, Qiangze Feng, Daliang Wang, Yanmin Qian, and Lei Xie, The Accented English Speech Recognition Challenge 2020: Open Datasets, Tracks, Baselines, Results and Methods. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6918–6922.

Jasdeep Singh, Bryan McCann, Richard Socher, and Caiming Xiong. 2019. Bert is not an interlingua and the bias of tokenization. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, page 47–55, Hong Kong, China. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Minh Chien, Sebastian Ruder, Surya Guthikonda, Emad A. Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya dataset: An open-access collection for multilingual instruction tuning.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Suzy J. Styles, Victoria Y. H. Chua, Fei Ting Woon, Hexin Liu, Leibny Paola Garcia, Sanjeev Khudanpur, Andy W. H. Khong, and Justin Dauwels. 2023. Investigating model performance in language identification: beyond simple error statistics. In *Interspeech 2023*, pages 4129–4133.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. 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 (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi, A benchmark for learning to translate a new language from one grammar book. *arXiv preprint arXiv:2309.16575*.

The Omnilingual MT Team, Belen Alastruey, Niyati Bafna, Andrea Caciolai, Kevin Heffernan, Artyom Kozhevnikov, Christophe Ropers, Eduardo Sánchez, Charles-Eric Saint-James, Ioannis Tsiamas, Chierh Cheng, Joe Chuang, Paul-Ambroise Duquenne, Mark Duppenthaler, Nate Ekberg, Cynthia Gao, Pere Lluís Huguet Cabot, João Maria Janeiro, Jean Maillard, Gabriel Mejia Gonzalez, Holger Schwenk, Edan Toledo, Arina Turkatenko, Albert Ventayol-Boada, Rashel Moritz, Alexandre Mourachko, Surya Parimi, Mary Williamson, Shireen Yates, David Dale, and Marta R. Costa-jussà. 2026. Omnilingual MT: Machine translation for 1,600 languages.

Nandan Thakur, Suleman Kazi, Ge Luo, Jimmy Lin, and Amin Ahmad, Mirage-bench: Automatic multilingual benchmark arena for retrieval-augmented generation systems. (arXiv:2410.13716). ArXiv:2410.13716 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Takashi Wada, Tomoharu Iwata, and Yuji Matsumoto. 2019. Unsupervised multilingual word embedding with limited resources using neural language models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3113–3124, Florence, Italy. Association for Computational Linguistics.

Hetong Wang, Pasquale Minervini, and Edoardo Ponti. 2024. Probing the emergence of cross-lingual alignment during LLM training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12159–12173, Bangkok, Thailand. Association for Computational Linguistics.

Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al., On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze, Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025b. Bridging the language gaps in large language models with inference-time cross-lingual intervention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5418–5433, Vienna, Austria. Association for Computational Linguistics.

Zirui Wang, Zachary C Lipton, and Yulia Tsvetkov. 2020. On negative interference in multilingual models: Findings and a meta-learning treatment. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4438–4450.

Orion Weller, Michael Boratko, Iftekhar Naim, and Jinhyuk Lee, On the theoretical limitations of embedding-based retrieval. (arXiv:2508.21038). ArXiv:2508.21038 [cs].

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

BigScience Workshop, :, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala,

Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel McDuff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Mengzhou Xia, Xiang Kong, Antonios Anastasopoulos, and Graham Neubig. 2019. Generalized data augmentation for low-resource translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5786–5796.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings*

of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Syed Waqas Zamir, Wassim Hamidouche, Boulbaba Ben Amor, Luana Marotti, Inbal Becker-Reshef, and Juan Lavista Ferres, Byol: Bring your own language into llms. *arXiv preprint arXiv:2601.10804*.

Chen Zhang, Jiuheng Lin, Xiao Liu, Zekai Zhang, and Yansong Feng. 2025. Read it in two steps: Translating extremely low-resource languages with code-augmented grammar books. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3977–3997.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. Hire a linguist!: Learning endangered languages in llms with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15654–15669.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng, Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models. *arXiv preprint arXiv:2306.10968*.

Xinyu Zhang, Xueguang Ma, Peng Shi, and Jimmy Lin. 2021. Mr. tydi: A multi-lingual benchmark for dense retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, page 127–137, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do large language models handle multilingualism? In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, and Xing Xie. 2024. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM Workshop on Large AI Systems and Models with Privacy and Safety Analysis*, LAMPS '24, page 57–68, New York, NY, USA. Association for Computing Machinery.

Caleb Ziems, William Held, Jingfeng Yang, Jwala Dhamala, Rahul Gupta, and Diyi Yang. 2023. Multi-VALUE: A framework for cross-dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 744–768, Toronto, Canada. Association for Computational Linguistics.