

Massive Multilinguality with Large Language Models

Niyati Bafna

1 Introduction: LLMs for everyone

Large language models (LLMs) today, including multilingual LLMs (Lin et al., 2022; et al, 2023; Aryabumi et al., 2024), often degrade in their performance in multilingual scenarios (Joshi et al., 2020, 2024). Improving the multilingual performance of LLMs over the range of their reasoning and retrieval-augmented generation capabilities is crucial for supporting language communities around the world. This is a broad endeavour, which I organize along two dimensions. Firstly, **evaluating / diagnosing** multilingual capabilities and failures is the first and necessary step; we also want to find mitigating or **improving solutions**. Secondly, different groups of languages have varying problem areas and needs, and the goal is to synergistically advance performance across these groups. We are interested in achieving parity for **mid-resource languages** with high-resource languages for reasoning, retrieval-augmented generation (RAG), tool use, and other frontier capabilities of LLMs. Many languages are **dialects or variants** of closely related high-resource language neighbours, and we aim to leverage cross-lingual transfer for robust comprehension for these languages. Finally, we have the long tail of **low-resource languages**, which have very scant training and evaluation resources, and for which we aim for basic fluency and machine translation.

My previous work focuses on evaluation and training protocols for low-resource languages and dialects, and diagnosing generation failures for mid-resource languages. In the remaining years of my PhD, I plan to explore fundamental questions in end-to-end as well as translation-assisted multilinguality for **mid-resource languages**, focusing on retrieval-augmented generation and reasoning as application scenarios. **The goal of my work is to develop robust, effective, and scalable solutions to multilinguality with LLMs.**

2 Low-resource languages: Benchmarking, Dataset design, and Experimentation Frameworks

For the vast majority of the world’s 3800+ languages, data availability drops off sharply: many low-resource language are nearly unseen from LLMs, with very evaluation or training data (Joshi et al., 2020; Wang et al., 2022). The first pain point is *benchmarking*: we have no way to assess even basic LLM capabilities for these languages, and as a consequence, they are left out of the iterative loop of advancement and mainstream evaluation in LLM technologies. This motivated my work on **ChiKhaPo**, a massively multilingual benchmark that **evaluates** basic lexical understanding and generation across thousands of languages using four tasks in both directions (Chang and Bafna, 2025). Next, *curating training data* for these languages has unique practical and scientific challenges. Given a fixed budget and feasible dataset size, it is crucial to include diverse and representative examples in it, so as to gain as much information as possible about comprehension and generation in the language. This motivates research in protocols for **curating training data** in extremely low-resource languages while maintaining domain, length, and grammatical diversity in curated samples, which I explored as part of the **Omnilingual MT project** (Omnilingual MT Team et al., 2026). Given pretraining data scarcity, a dominant paradigm in this regime is in-context language learning with LLMs using a variety of external materials like lexicons, grammar rules, and other types of linguistic processing in the context of an LLM (Zhang et al., 2024, 2025; Hus and Anastasopoulos, 2024). However, *scientific exploration and experimentation* remains a problem in this resourcedness regime, given the core problem of resource scarcity in terms of basic NLP tools, metalinguistic materials, and researcher expertise, for cheaply exploring new strategies, impeding cheap and fast progress for these languages. Bafna et al. (2026) introduces a general-purpose framework for systematically **assessing** and **exploring new strategies** in in-context learning for unseen languages, leveraging artificial cipher-constructed languages. **This line of work seeks to put hundreds of previously invisible languages on the radar of LLM-driven NLP.**

3 Dialects, variants, and accents: Robustness across the Language Continuum

Large populations of the world speak dialectal variants of high-resource languages. Many such low-resource dialects occur in language and dialect families, exhibiting continuous and structured differences from each

other (Hovy and Purschke, 2018; Bergman and Diab, 2022). There is accordingly a rich and long tradition of investigating *cross-lingual transfer*: training in one language benefiting others (Vaswani et al., 2017; Pires et al., 2019; Devlin et al., 2019; Conneau et al., 2020; Dhamecha et al., 2021; Cahyawijaya et al., 2021; Bafna et al., 2023, *inter alia*). In general, these works still treat dialects and closely-related languages as discrete monoliths: however, language in casual use is fluid and dynamic across dialect boundaries, calling for *principled general approaches to linguistic variation*. It is therefore important to **evaluate** the robustness of LLMs to fluid dialectal variation (Bafna et al., 2024). Further, this model of language as a continuum inspired DialUp, a dual technique to **induce robustness** across the continuum rather than for particular dialects (Bafna et al., 2025a). This conception is also relevant and critical in speech processing, where accent and dialect variation still breaks SOTA models (Bafna and Wiesner, 2025). Future work may explore porting these robustness methods to speech models. **This line of work seeks to construct principled strategies for handling fluidity and variation in language.**

4 Mid-resource languages: End-to-end Multilinguality and the Translation Cascade

Currently, frontier capabilities of LLMs such as RAG (Lewis et al., 2020) and complex reasoning degrade for mid-resource languages like Arabic, Bengali, and Swahili, which may have large user bases (Hendy et al., 2023; Cahyawijaya et al., 2024; Jiao et al., 2023; Marchisio et al., 2024; Shen et al., 2024).

Previous work demonstrates that *natively multilingual* LLMs internally process input in a “high-resource language space”, shown to be close to English, then transitioning to the target language near final layers (Wendler et al., 2024; Foroutan et al., 2022; Tang et al., 2024; Kojima et al., 2024; Schut et al., 2025). This motivated the “translation barrier hypothesis” (Figure 1) (Bafna et al., 2025b), which presents the **diagnosis** that the implicit translation stage bears a large extent of blame for final generation failure. In fact, the *explicit translation cascade*, where we use machine translation models to convert language inputs into English, perform any LLM processing including complex reasoning, RAG, tool calls, etc. in English, and translate outputs back to the required language, is a very competitive approach, often outperforming end-to-end multilinguality (Qin et al., 2023; Etxaniz et al., 2024; Liu et al., 2025; Chiu and Bai, 2025).

The above studies motivate two directions of exploration for my planned work: firstly, we want to evaluate and improve end-to-end multilinguality, focusing on the robustness, alignment, and cultural grounding of multilingual representation spaces, and secondly, we want to assess the capabilities of and advance the paradigm of translation-assisted multilinguality. **This line of work seeks to achieve parity for mid-resource languages with high-resource in terms of frontier LLM capabilities.**

4.1 Robust Retrieval with Multilingual Text Embeddings

Retrieval is an important hurdle for multilingual and cross-lingual RAG, and relies on quality embedding-based retrieval (Goworek et al., 2025; Enevoldsen et al., 2025). Due to the core problem of data imbalance across languages, constructing robust and reliable multilingual semantic spaces is still an unsolved problem, despite various efforts in better multilingual alignment (Søgaard et al., 2018; Wada et al., 2019). In the context of retrieval, this can be understood 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

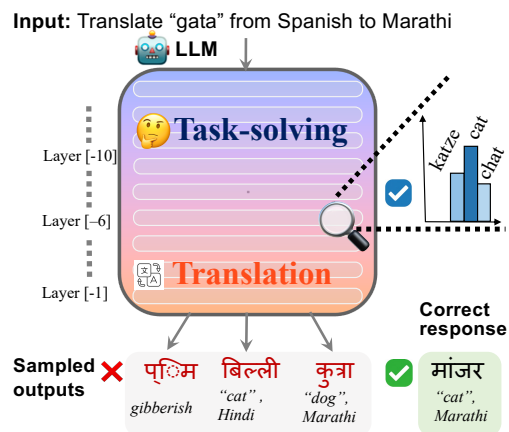


Figure 1: *Task-solving* succeeds with the correct answer concept (i.e. **cat**) discovered in intermediate layers, expressed in various high-resource languages, but the model fails to realize or *translate* the concept into the target language. These insights motivate my planned work in bettering multilingual representation spaces as well as explicit translation cascades.

documents in high-resource languages (Goworek et al., 2025; Nouri and Yangarber, 2014; Park and Lee, 2025). This motivates work in **exploring methods 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 fundamental insights on multilingual representation spaces trained on data-imbalanced corpora. First, I propose to 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 post-hoc corrective techniques, adjusted scoring mechanisms, and mechanisms to fallback to sparse hybrid retrieval systems (Karpukhin et al., 2020).

4.2 Managing Alignment and Diversity in Retrieval with Multilingual Representation Spaces

Multilingual transformer-based data-driven models face a inherent and critical tension between language *alignment* (cross-lingual access of information and consistency) and *diversity* (culturally and linguistically appropriate grounding) (Han et al., 2025). In the context of RAG, the corresponding observed failure modes can be understood as the **knowledge gap** (information in one language is not accessible in another) (Lu et al., 2025; Wang et al., 2025) and **localization bias** (the model fails to localize itself in diverse language and cultural scenarios) (Hershcovich et al., 2022; Naous et al., 2024). These phenomena have been largely discussed from an observational and extrinsic standpoint concerned with desirable outcomes from either side in isolation, despite the demonstrable trade-off between them (Han et al., 2025). In ongoing work, I present a unifying **fundamental understanding** of the capability of a multilingual representation space to encode both types of behaviours. This involves exploring the conditions under which a fixed capacity representation space can most efficiently encode alignment and diversity requirements: I prove that its efficiency is controlled by computable and interpretable characteristics such as the norm of the matrix representing diverse requirements, and the extent of orthogonality between constructed aligned and diverse subspaces. These findings present us with intrinsic metrics for multilingual representation spaces for evaluating state-of-the-art multilingual text embeddings, and inspiring **techniques for their improvement**. The insights and methods from the above two works can be adapted for general scenario of multilingual representation spaces as countering issues stemming from the core problem of data imbalance across multiple languages.

4.3 The Translation Cascade in the Era of LLMs

Using an MT cascade at various points in the LLM pipeline for a complex reasoning task (translating inputs, outputs, queries, external documents) has many advantages for multilingual processing, in terms of a practical solution for extending the powerful reasoning and RAG capabilities of LLMs to other languages (Liu et al., 2024), without needing to optimize for individual languages. In ongoing work, I show that this is particularly relevant for mid-resource languages, which may have decent machine translation support but may still suffer in complex reasoning and other scenarios with end-to-end processing. 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 output would be inadequate, and we would like our “MT” system to actually modify the content or to disregard it (MT+). I propose to systematize the use of translation in LLM workflows. In planned work, I conduct a comparison of the MT cascade with end-to-end multilingual approaches given a variety of input scenarios (source and target language, task, input, domain, model used), **extracting insights** on the failure modes of either paradigm. I propose to construct a lightweight and adaptable module that can dynamically route inputs to the optimal approach for **boosting performance**. Finally, I propose to **develop MT+ systems** tailored to an LLM cascade. This direction aims for practical and scientific insights and techniques promoting scalable multilinguality.

References

- 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. 2024. [Aya 23: Open weight releases to further multilingual progress](#).
- 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.
- Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. [LLMs are few-shot in-context low-resource language learners](#). 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 405–433, Mexico City, Mexico. Association for Computational Linguistics.
- 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.
- 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.
- 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.
- 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.
- 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*.
- BigScience Workshop et al. 2023. [Bloom: A 176b-parameter open-access multilingual language model](#).
- 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.

- Negar Foroutan, Mohammadreza Banaei, Rémi Lebre, 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.
- 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*.
- HyoJung Han, Sweta Agrawal, and Eleftheria Briakou, Rethinking cross-lingual alignment: Balancing transfer and cultural erasure in multilingual llms. *arXiv preprint arXiv:2510.26024*.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla, How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.
- 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.
- 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](#).
- 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.
- Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.
- 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.
- 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.

- 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. 2025. [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.
- 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*.
- Kelly Marchisio, Wei-Yin Ko, Alexandre Bérard, Théo Dehaze, and Sebastian Ruder. 2024. Understanding and mitigating language confusion in llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6653–6677.
- 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.
- 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.
- 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.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.
- Lingfeng Shen, Weiting Tan, Sihao Chen, Yunmo Chen, Jingyu Zhang, Haoran Xu, Boyuan Zheng, Philipp Koehn, and Daniel Khashabi. 2024. [The language barrier: Dissecting safety challenges of LLMs in multilingual contexts](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2668–2680, Bangkok, Thailand. Association for Computational Linguistics.
- 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.

- 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.
- 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 systematic investigation on 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** and Matthew Wiesner, Lid models are actually accent classifiers: Implications and solutions for lid on accented speech. *arXiv preprint arXiv:2506.00628*.
- 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](#).
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz 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.
- 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*.
- Xinyi Wang, Sebastian Ruder, and Graham Neubig. 2022. [Expanding pretrained models to thousands more languages via lexicon-based adaptation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 863–877, Dublin, Ireland. Association for Computational Linguistics.

- 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.
- 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.
- Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. 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.