

Robust and Effective Multilingual Knowledge Grounding

Niyati Bafna

1 Introduction

Recent benchmarks for multilingual retrieval and retrieval-augmented generation (RAG) highlight challenges in multilingual consistency, robustness, and grounding (Bandarkar et al., 2024; Singh et al., 2024; Enevoldsen et al., 2025; Li et al., 2024a; Liu et al., 2025b). A large body of work demonstrates that LLM performance degrades significantly for low-resource languages across tasks, and that these failures are driven by data scarcity leading to systematic biases such as preference for high-resource languages and instability under cross-lingual conditions (Cahyawijaya et al., 2024; Ki et al., 2025; Park and Lee, 2025; Amiraz et al.). These issues are closely related to long-tail knowledge effects and limitations in multilingual representation spaces, where models rely on partially language-agnostic internal representations that are often aligned toward English, leading to errors in cross-lingual transfer, factual recall, and generation (Conneau et al., 2020; Bafna et al., 2025b; Li et al., 2024b). Prior approaches for improving multilingual performance include curating and fine-tuning on multilingual data, leveraging cross-lingual transfer from high-resource languages, explicitly controlling the language of generation, and leveraging translation-based pipelines or cascaded systems, which remain strong baselines especially for low- and mid-resource languages and are widely used in multilingual RAG (NLLB Team et al., 2024; Nekoto et al., 2020; Dhamecha et al., 2021; Li et al., 2025; Liu et al., 2025a; Ranaldi, 2026). The above studies expose the consequences of poor multilingual representations, fundamental trade-offs between language-agnosticity and language-specificity in representation spaces, and the practicality of translation cascades. **This motivates a fundamental investigation of end-to-end as well as translation-assisted multilinguality for low- and mid-resource languages, where improving cross-lingual retrieval and generation remains an open problem central to practical multilingual LLM deployment.**

2 Prior Work

My prior work focuses on diagnosing generation failures for **mid-resource languages**, robust comprehension of **dialects**, and evaluation and training protocols for **low-resource languages**.

The first pain point for **low-resource languages** is benchmarking and lack of training data: as a consequence, they are excluded from the iterative loop of advancement and mainstream evaluation in LLM technologies. This motivated my work on ChiKhaPo, a *massively multilingual benchmark* for lexical comprehension and generation across thousands of languages (Chang and Bafna, 2025), as well as my work in protocols for *sample-efficient training dataset design* as part of the **Omnilingual MT project** (Omnilingual MT Team et al., 2026). Finally, given that in-context learning is a natural paradigm for the processing of low-resource languages (Zhang et al., 2024, 2025; Hus and Anastasopoulos, 2024), Bafna et al. (2026) introduces a *general-purpose framework for systematic experimentation* in in-context language learning for low-resource languages, seeking to accelerate progress in this field. Next: large populations of the world speak **dialectal variants** of high-resource languages. Previous literature in cross-lingual transfer (Vaswani et al., 2017; Pires et al., 2019; Cahyawijaya et al., 2021; Bafna et al., 2023, *inter alia*), treats 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*. In Bafna et al. (2024) and Bafna et al. (2025a), I present techniques to evaluate the dialectal robustness of LLMs and then induce dialectal robustness across the continuum, rather than for particular dialects. This idea is also 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.

Previous work in the field demonstrates that multilingual inputs, including **mid-resource language** inputs are internally processed by LLMs 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). My recent work shows the “translation barrier hypothesis” (Bafna et al., 2025b), which presents an *interpretable diagnosis for the failure of multilingual*

generation: i.e. that the implicit translation stage bears a large extent of blame for final generation failure. This motivates my planned work in fundamental advancements for end-to-end multilinguality as well as investigating the utility of the explicit translation cascade for tasks such as reasoning and RAG.

3 Managing Alignment and Diversity in Cross-Lingual and Multilingual Retrieval

Multilingual transformer-based data-driven models exhibit an inherent tension between language *agnosticity* (cross-lingual consistency of information and behaviour) 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) (Goworek et al., 2025) and **localization bias** (the model assumes the dominant viewpoint and is error-prone in diverse language and cultural settings) (Hershcovich et al., 2022). These phenomena have been largely discussed from an extrinsic standpoint concerned with desirable outcomes from either side in isolation, despite the demonstrable trade-off between them. In ongoing work, I present a unifying fundamental understanding of the capability of a multilingual representation space to encode both types of behaviours. I prove that the multilingual efficiency of a fixed capacity representation space is controlled by computable and interpretable characteristics, which can be considered intrinsic metrics for multilingual representation spaces as well as 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 Robust Retrieval with Multilingual Text Embeddings

Cross-lingual and multilingual RAG suffer from poor text embeddings for low-resource languages, leading to 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; Park and Lee, 2025; Enevoldsen et al., 2025; Ranaldi, 2026). This motivates work in exploring methods to mitigate the effects of poor multilingual embedding spaces on multilingual retrieval. I propose to achieve this in a post-hoc and training-free manner 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. I then propose to leverage this understanding to design interventions for multilingual retrieval, such as adjusted scoring mechanisms, post-hoc corrective techniques, or mechanisms to fallback to hybrid dense and sparse retrieval systems (Karpukhin et al., 2020).

5 The Translation Cascade for Reasoning and RAG 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, etc.) 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 problems of error propagation, MT failures on rare concepts, and inadequate treatment of inputs requiring cultural grounding. In these cases, it may be preferable to use the end-to-end approach, as well as “MT” systems that can intelligently adapt content to query-specific needs (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 design and develop MT+ systems tailored to an LLM cascade.

References

- Chen Amiraz, Yaroslav Fyodorov, Elad Haramaty, Zohar Karnin, and Liane Lewin-Eytan, The cross-lingual cost: Retrieval biases in rag over arabic-english corpora.
- 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.
- 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.
- 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*.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Mylé 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.
- 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*.
- 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.
- 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*.
- 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.

- 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.
- 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.
- 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.
- Bo Li, Zhenghua Xu, and Rui Xie, [Language drift in multilingual retrieval-augmented generation: Characterization and decoding-time mitigation](#). (arXiv:2511.09984). ArXiv:2511.09984 [cs].
- Bryan Li, Samar Haider, Fiona Luo, Adwait Agashe, and Chris Callison-Burch. 2024a. [Bordirines: 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.
- 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.
- 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.
- Wilhelmina Nekoto, Vukosi Marivate, Tshinondiwa Matsila, Timi Fasubaa, Taiwo Fagbohunge, 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.
- 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.
- 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.
- Lisa Schut, Yarin Gal, and Sebastian Farquhar. Do multilingual llms think in english? *arXiv preprint arXiv:2502.15603*.
- 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](#).
- 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, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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.