Analyzing the Impact of Typological Features on Cross-lingual Transfer for Generative LLMs

Anonymous ACL submission

Abstract

Recent advances in large multilingual language models have enabled impressive zero-shot cross-lingual transfer capabilities. However, their performance remains uneven across languages, with underrepresented languages often lagging behind. While prior research has explored typological similarities to explain performance disparities, it has largely ignored external factors such as resource availability and has primarily focused on encoder-only models. In this study, we investigate the interplay between typological features and resource-related factors in zero-shot reading comprehension tasks using decoder-only models. Specifically, we evaluate the performance of GPT-3.5 Turbo, Qwen-2.5B, and Aya-Expanse on the Belebele benchmark. We conduct a series of correlation and regression analyses to examine: (1) the influence of English similarity on transfer performance, (2) whether resource availability acts as a spurious or explanatory factor, and (3) which typological features most significantly predict multilingual model performance. Our findings offer deeper insight into the factors that drive cross-lingual generalization, with implications for improving model equity across languages.

1 Introduction

With large generative models improving every week, ZSCL (Zero-Shot Cross-Lingual) transfer has become the norm for low-resource language tasks. Despite significant progress in the development of large-scale multilingual language models, the mechanisms behind cross-lingual transfer are not yet fully understood, consequently, understanding the principles behind multilinguality remains a profound challenge. Many existing models demonstrate excellent performance on high-resource languages while exhibiting substantial deficiencies when applied to underrepresented languages.

Understanding the reasons behind the drop in performance is crucial for democratizing the use of LLMs among low-resource languages. Previous works addressing this issue have heavily focused on analyzing similarities among topological features of the languages, without considering other relevant factors (such as the amount of available data for those languages on the internet). These factors could not only better explain the drop in performance, but also affect the real impact of the features studied in previous works. In this study, we aim to analyze how the impact of the previously studied topological features changes when adding such confounding factors. Moreover, in previous studies, the main focus has been on encoder-only models performing token classification tasks such a Part of Speech tagging or Named Entity Recognition. The recent paradigm shift to decoder-only LLMs has altered the way Natural Language Processing (NLP) tasks are framed (Min et al., 2023). This research is driven by the following fundamental research questions (RQs): (1) What are the key typological features that influence multilingual model performance in a ZSCL setting? (2) Does having similarities to the English language benefit cross-lingual transfer significantly? (3) does the resource availability of a language serve as a spurious

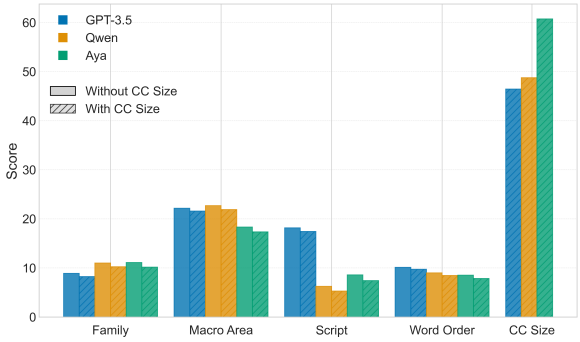

Figure 1: Impact of grouped typological features with and without considering the language size in Common Crawl as another factor. Language size outweighs all other features.

feature in the prior RQs and change the outcomes?

To answer our research questions, we evaluate three decoder-only language models: GPT 3.5 Turbo (Brown et al., 2020), Qwen-2.5 (Yang et al., 2025), and Aya-Expanse (Dang et al., 2024), using the Belebele benchmark (Bandarkar et al., 2024), a multilingual dataset designed to assess reading comprehension in a zero-shot setting.

2 Related Work

Prior research has explored various aspects of transfer learning and cross-lingual understanding in the past. A pivotal study by Conneau et al. (2018) introduced multilingual embeddings, demonstrating how shared linguistic structures can facilitate effective zero-shot learning. The study highlighted that languages with greater lexical and syntactic similarity benefit more from shared representation spaces. Building upon this foundation, Pires et al. (2019) examined multilingual BERT (mBERT) and found that it performs well across typologically similar languages but struggles with structurally distinct ones. This work provided empirical evidence that word order, morphological complexity, and script differences can hinder transfer learning. Similarly, Zubillaga et al. (2024) applied cross-lingual transfer learning to event extraction and conducted a typological analysis of the results, concluding that typological similarity significantly enhances transfer effectiveness in information extraction tasks.

Another relevant study by Artetxe et al. (2020) investigated ZSCL transfer without parallel data, demonstrating that pretrained language models inherently capture linguistic properties that allow them to generalize across languages. However, their research also noted limitations, particularly for languages with unique syntactic or phonological features. Subsequent analyses by Lauscher et al. (2020) and Chau et al. (2020) examined performance disparities in multilingual models, revealing that language model effectiveness is often correlated with both linguistic similarity to English and the amount of available training data. This raises important concerns about fairness and inclusivity in multilingual NLP.

The XTREME benchmark (Hu et al., 2020) and its follow-up XGLUE (Liang et al., 2020) introduced large-scale evaluation datasets for multilingual models, underscoring the persistent gap in performance between high-resource and low-resource languages. These studies often found that zero-shot transfer performance could be predicted by language family and script similarity to English, but typically lacked in-depth typological analysis. Recent work by Ponti et al. (2020) explored language transferability in the context of typological features. Their findings emphasized that structural similarities play a crucial role in transfer success. However, they primarily focused on encoder-based models and syntactic features such as grammatical tense, leaving other typological dimensions like script, morphology, etc., underexplored.

Lastly, Bandarkar et al. (2024) analyzed zero-shot performance on the Belebele benchmark, establishing baseline performance for several large language models and highlighting the English-centric bias in models like GPT-3.5. However, their study aggregated the performance by typological dimensions, leaving open questions about the features that matter the most. In contrast to prior work, our study explicitly quantifies the impact of four major typological features: script, language family, word order, and geographical proximity, on zero-shot cross-lingual performance using decoder-based LLMs. Additionally, we attempt to disentangle the influence of typology from resource availability, offering a more nuanced understanding of transfer mechanisms in multilingual models.

3 Experimental Setup

We divide our experiments and analysis into three sections, each pertaining to a research question on the potential factors affecting the ZSCL performance of a language: Typological Features, Similarity to English, and Resource Availability. We select the following typological features for our analysis: Script, Word Order, Language Family, and Geolocation. All features were directly obtained from the World Atlas of Language Studies (WALS) [1]. Due to additional features like morphosyntax not being available for a majority of under-resourced languages, we opted for the four features mentioned. We also use the Common Crawl size of each language [2] as a confounding feature for the final research questions.

We test each of these aspects on languages in the Belebele benchmark. While the benchmark consists of parallel data for 122 languages, detailed typological information is available for 73 languages

[1] https://wals.info/

[2] https://commoncrawl.github.io/cc-crawl-statistics/plots/crawlsize

on the WALS encyclopedia. Therefore, we have restricted our analysis to these 73 languages. We also select 3 state-of-the-art decoder models to test on, namely, GPT-3.5T, Qwen-2.5-7B, and Aya-Expanse-8B. These models were selected to have a wide coverage of training data ratios across languages. GPT-3.5T is expected to be more English-centric, Qwen2.5 could likely be more Chinese-driven, whereas Aya-Expanse claims to have highly advanced multilingual capabilities. For each model we first obtain ZSCL transfer results for each of the 73 languages on the Belebele benchmark. For GPT-3.5T we directly use the numbers reported by the Bandarkar et al. 2024. While for Aya-Expanse and Qwen2.5, we perform the experiments using the lm-evaluation-harness[3]. All regression tests were performed with a Ridge Regressor due to having many highly correlated predictors.

4 Results

4.1 Role of Typological Features

For the first research question we visualize the correlation between each of the four typological factors i.e., script, word order, geo-location and language family, and the ZSCL performance of a model. We plot each model independently to highlight the nuances in ZSCL performance due to the training data compositions.

Figure 1 demonstrates the analysis averaged over each group of typological features. While Figure 2 shows the impact of each individual classification as a binarized feature. Starting with language family, to no surprise, Indo-European performs consistently best across the models. The Uralic family, performs surprisingly well, while the performance of the Sino-Tibetan and Dravidian families is shockingly low despite fairly medium resource availability. GPT-3.5T starts strong with excellent performances for the Uralic, Indo-European, and other fairly resourced families but quickly falls off for various other well-resourced families like Dravidian and Afro-Asiatic. Even in the Indo-European family, GPT-3.5T seems the most inconsistent with the largest inter-quartile range out of all the 3 models. While Aya-Expanse seems the most consistent across families from the 3 models, it also seems to be the worst-performing model on average.

While GPT-3.5T shows the highest average performance, especially on high-resource Latin-script languages (e.g., English, French, Spanish). Aya-Expanse and Qwen-2.5 seem optimized for broad typological and script coverage, with consistent mid-to-high performance across scripts. Qwen-2.5B particularly excels with Hang, Japanese, Thai and other East Asian scripts. While Aya-Expanse seems to handle the Arabic script quite well showing the third highest median performance out of all scripts. Both models also out-pace GPT-3.5T for Devanagari and other Brahmic scripts, once again stressing the multilingual shortcomings of GPT-3.5T. Qwen-2.5 also exceeds expectations when analyzing for the Macro Area feature, showing the highest consistency in the Eurasia macro-area and the highest medians for Papunesia and Africa. While Aya-Expanse and GPT-3.5T are second-best, GPT-3.5T having higher median performances for the Africa and Papunesia region, but poorer consistency compared to Aya-Expanse and Qwen-2.5.

The final typological feature of word order (Figure 2 demonstrates that SVO languages like English consistently show better results compared to VSO and SOV languages. The box-plot demonstrates that models clearly have a preference for SVO word-order, while VSO languages like Arabic and other Afro-Asiatic languages come in second-best. Qwen outperforms both models once again, on all word orders, while GPT-3.5T usually comes in second-best.

4.2 Role of Similarity to English

For our second research question we wanted to confront the impact of similarity to English and its effect on the ZSCL performance for a language. To this end, we created binary features for all languages for each of the typological factors of script, macro-area, language family and word order, with *True* indicating similarity to English for that feature, while *False* indicating dissimilarity. We then proceed to train a simple linear model to predict the ZSCL performance for a language using binary features. We use the coefficients of each feature to quantify the impact of a particular typological factor in the ZSCL performance. Figure 2 summarizes these findings for all 3 models in question.

For all 3 models, having the same macro-area as English, i.e. Eurasia seems to be the most determining factor. For GPT-3.5T the second most determining factor by far is script similarity to English, i.e. Latin. This seems to not be the case for Qwen-2.5 and Aya-Expanse indicating that the former 2 models are more flexible w.r.t. script than GPT-3.5T,

[3]https://github.com/EleutherAI/lm-evaluation-harness

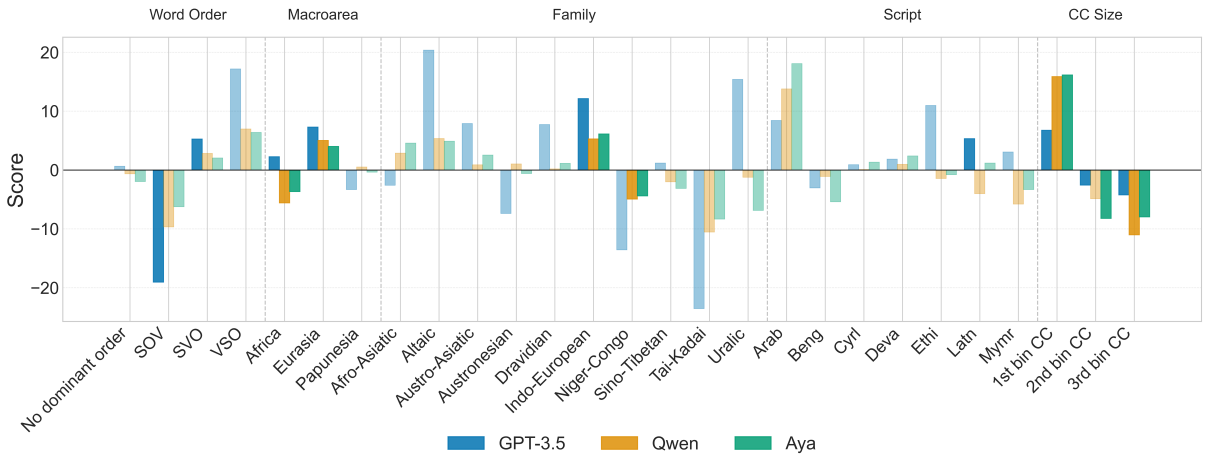

Figure 2: Impact of each individual typological factor on ZSCL performance including Common Crawl size as a confounding feature for GPT-3.5T, Qwen-2.5, and Aya-Expanse. Statistically significant features are bold.

while GPT-3.5T is heavily dependent on the Latin script. For Qwen-2.5 and Aya-Expanse, the second most-determining feature seems to be similar linguistic family to English, i.e. Indo-European. While script is the third-most important factor for Aya-Expanse, it is the least determinant factor for Qwen-2.5, suggesting that Qwen-2.5 is the most script flexible model.

4.3 Role of resource availability

For our final research question, we examined the resource availability for a language and its possible impact on the ZSCL performance, which might over-shadow the typological factors discussed in the previous section. To this end, we repeat our analysis for the previous RQ with additional features pertaining to the size of each language's Common Crawl dump. We re-run the analysis to assess the impact of resource-availaibility and to observe if it significantly alters the results for other typological factors. Figures 1 and 2 demonstrates that the resource feature quickly becomes the most significant, reducing the coefficient scores for all typological features in proportional measure.

5 Conclusion

Our analysis highlights the multifaceted role of typological features, similarity to English, and resource availability in shaping ZSCL performance across multilingual language models. Typological factors such as language family, script, and word order demonstrate clear influence: models show systematic advantages for Indo-European languages, SVO word order, and Latin scripts. Among the evaluated models, Qwen-2.5B emerges as the most typologically robust, demonstrating consistent performance across scripts and word orders, particularly excelling in handling East Asian and Brahmic scripts, areas where GPT-3.5T struggles despite its overall higher average performance. Aya-Expanse, while being the most consistent across language families, shows a generally lower absolute performance. The second part of our study, modeling the influence of English similarity, confirms a structural bias towards English-typical features particularly for GPT-3.5T, which is highly dependent on shared script and macro-area. In contrast, Qwen-2.5 shows strong flexibility with respect to script and word order, suggesting deliberate design choices favoring typological generalization.

Finally, introducing resource availability as a factor reveals its overwhelming influence: the amount of data available for a language significantly diminishes the explanatory power of typological features. This underlines a key limitation in evaluating multilingual performance: high-resource languages benefit disproportionately, obscuring the effects of structural linguistic diversity. Overall, our findings suggest that while modern multilingual models have made strides in typological generalization, training data biases, both in terms of language proximity to English and raw data volume, still shape model behavior in fundamental ways. For truly equitable multilingual NLP, future efforts must balance data-driven advantages with architectural and training innovations that foreground linguistic diversity.

Limitations

Our study presents several limitations that warrant consideration. First, while our analysis incorporates a wide range of typological features (script, word order, macro-area, and language family), it does not capture finer-grained syntactic, morphological, or phonological characteristics that may influence model performance. This level of abstraction limits our ability to fully explain cross-linguistic variation in ZSCL behavior. Second, the resource availability feature used (Common Crawl size) serves only as a proxy for actual pretraining exposure. We do not have access to the proprietary training corpora of the evaluated models, and therefore cannot definitively establish causality between data availability and model performance. Similarly, our binary similarity-to-English features oversimplify a complex continuum of linguistic relatedness and may miss interactions between typological traits. Third, our analysis is limited to three publicly or semi-publicly accessible multilingual models. While these represent different design philosophies and coverage strategies, the findings may not generalize to other commercial or open-source models with different training regimes, architectures, or fine-tuning approaches.

Additionally, model evaluations are based on zero-shot performance without task-specific adaptation or finetuning. While this setup cleanly reveals model generalization, it does not reflect real-world usage conditions where models are often adapted or prompted more deliberately for specific languages or tasks. Finally, while our analysis includes over 70 languages, there is still an underrepresentation of certain low-resource or endangered languages, particularly from South America and parts of Africa and Oceania. This reflects both the limitations of available benchmarks and broader systemic biases in language technology development. We encourage future work to extend this typological analysis using more detailed linguistic features, broader and more representative datasets, and transparent access to model training data.

Acknowledgments

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4623–4637, Online. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Ethan C. Chau, Lucy H. Lin, and Noah A. Smith. 2020. Parsing with multilingual BERT, a small corpus, and a small treebank. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1324–1334, Online. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, and 5 others. 2020. XGLUE: A new benchmark dataset for cross-lingual pre-training, understanding and generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6008–6018, Online. Association for Computational Linguistics.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

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*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

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, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Mikel Zubillaga, Oscar Sainz, Ainara Estarrona, Oier Lopez de Lacalle, and Eneko Agirre. 2024. Event extraction in Basque: Typologically motivated cross-lingual transfer-learning analysis. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6607–6621, Torino, Italia. ELRA and ICCL.