Are the LLMs Capable of Maintaining at Least the Language Genus?

Sandra Mitrović[*,1], David Kletz[*,1], Ljiljana Dolamic[2], Fabio Rinaldi[1]
[1] SUPSI, IDSIA, Switzerland
[2]armasuisse, Science & Technology, Switzerland
{sandra.mitrovic, david.kletz, fabio.rinaldi}@supsi.ch
ljiljana.dolamic@armasuisse.ch

Abstract

Large Language Models (LLMs) display notable variation in multilingual behavior, yet the role of genealogical language structure in shaping this variation remains underexplored. In this paper, we investigate whether LLMs exhibit sensitivity to linguistic genera by extending prior analyses on the MultiQ dataset. We first check if models prefer to switch to genealogically related languages when prompt language fidelity is not maintained. Next, we investigate whether knowledge consistency is better preserved within than across genera. We show that genus-level effects are present but strongly conditioned by training resource availability. We further observe distinct multilingual strategies across LLMs families. Our findings suggest that LLMs encode aspects of genus-level structure, but training data imbalances remain the primary factor shaping their multilingual performance.

Keywords: Multilinguality, LLM, Genus Sensitivity, QA

1. Introduction

Numerous studies have investigated how variations in the input prompt affect the outputs of Large Language Models (LLMs) (Habba et al., 2025; Liu et al., 2025a; Zhou et al., 2023). Even superficial modifications that leave the semantic content unchanged can yield substantial differences in model responses; for example, altering the order of proposed answers in multiple-choice benchmarks (Alzahrani et al., 2024) or reordering few-shot examples (Zhao et al., 2021). A particularly salient dimension of this phenomenon is the language of the prompt. For instance, Bandarkar et al. (2024) show that gpt3.5-turbo and Llama-2-chat perform dramatically better –by 40.8 and 25.4 points, respectively– on a QA task when questions are posed in English rather than Icelandic.

Beyond performance differences, research has highlighted a related phenomenon: infidelity to the prompt language. LLMs frequently generate responses in a language different from that of the input (Shaham et al., 2024; Liu et al., 2025b). When queried in Arabic, models may partially or fully switch to English (Chen et al., 2025). Even when posed mathematical questions in non-English languages, language-aligned LLMs often produce English chain-of-thought reasoning before providing the final answer (Tran et al., 2025; Zhu et al., 2024). Several studies have documented systematic patterns of such language-switching behaviors across different models and tasks (Wiśniewski et al., 2025; Almasi and Kristensen-McLachlan, 2025).

While existing research has primarily examined language-switching as a binary phenomenon (adherence vs. deviation from the prompt language), the linguistic structure underlying these behaviors remains underexplored. In this paper, we introduce a genealogical perspective on multilingual LLM behavior. For this investigation, we rely on a group of languages gathered by genealogical relations, which WALS (Dryer and Haspelmath, 2013) relying on Dryer (1989), calls a genus[1]. More precisely, we investigate whether linguistic proximity correlates with consistency in model outputs.

Our central hypothesis is that LLMs may encode a form of genus-level coherence, potentially leading to more stable behaviors within linguistic families than across them.

Research questions We explore this potential coherence through two complementary analyses:

1. Genus fidelity: When LLMs fail to respond in the prompt language, do they preferentially switch to another language of the same genus?

2. Knowledge sharing across a genus: If an LLM answers a question correctly in one language, is it also likely to answer correctly when the same question is asked in another language of the same genus?

Throughout the paper, we refer to the prompt's original language as the source language, its translation as the target language, and the LLM's response language as the generation language[2].

* Equal contribution.

[1]"A genus is a group of languages whose relatedness is fairly obvious without systematic comparative analysis."

[2]Our code is available on github at https://github.com/IDSIA-NLP/GenusPref

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8954–8970
11–16 May 2026. ©ELRA Language Resources Association (ELRA), 2026

2. Methodology and data

Our methodology builds primarily on the work of Holtermann et al. (2024), who introduced the MultiQ dataset for evaluating multilingual capabilities of LLMs. Rather than conducting new large-scale data collection, we leverage these existing resources to perform a targeted secondary analysis focused on genealogical effects: a dimension not explored in the original study. This approach allows us to benefit from MultiQ's extensive coverage and rigorous design while introducing our novel genealogical perspective on multilingual LLM behavior.

2.1. The MultiQ dataset (Holtermann et al., 2024)

Holtermann et al. (2024) developed MultiQ to investigate fundamental multilingual capabilities of LLMs through a large-scale parallel question-answering dataset comprising 27,400 questions across 137 languages[3]. The dataset covers diverse question types (open-ended, closed-ended, reasoning questions) and domains (chemistry, physics, astronomy, history, maths, geography, art, sports, music, animals). The original study evaluated LLMs along two primary dimensions:

Language fidelity: Whether the model generates its response in the same language as the input prompt.

Question-answering accuracy: Whether the generated response is factually correct.

Their findings revealed substantial variation both across models and across languages, highlighting critical gaps in multilingual alignment. However, their language grouping strategy, while methodologically sound for their research questions, was too coarse for our genealogical analysis.

Model Selection To ensure direct comparability with existing results, we analyze the same six models evaluated by Holtermann et al. (2024): Llama-2-(7|14|70)B-Chat (Touvron et al., 2023), Mistral-7B-Instruct-v0.1 (Jiang et al., 2023), Mixtral-8x7B-Instruct-v0.1 (Jiang et al., 2024), Qwen1.5-7B-Chat (Bai et al., 2023).

Apertus-8B The models evaluated in MultiQ are Open-Weight Models, but not Fully Open Models (Hernández-Cano et al., 2025), as their training data are not completely transparent. However, we consider it essential to include in our results at least one Fully Open Model. We therefore select Apertus-8B, a multilingual model whose pretraining data are fully disclosed (Hernández-Cano et al., 2025). According to its documentation, Apertus was trained on data covering approximately 1,800 languages, with around 40% of the corpus being non-English.

To integrate Apertus into our evaluation, we generate responses to the 200 questions in 137 languages of the MultiQ benchmark, and assess them both for answer correctness and language identification. To ensure full comparability, we strictly replicate the configuration[4] of Holtermann et al. (2024): we run their publicly released code[5] and use the same evaluation models, namely `gpt-4-0125-preview` (OpenAI et al., 2024) for answer quality assessment and `cis-lmu/glotlid, model_v2`[6] (Kargaran et al., 2023) for language detection.

2.2. From Family-Level to Genus-Level Classification

The original MultiQ analysis grouped languages into three broad categories: English, Same (languages from the same family as the source language), and Other. We argue that for our RQs this classification proves insufficient as language families are too broad and mix together languages of different characteristics.

Consider, for example, the Indo-European family: it encompasses both English –which dominates LLM training corpora (Zhong et al., 2024; Csaki et al., 2024; Gupta et al., 2025) – and Hindi, which remains underrepresented in training data. These two languages are very different in many aspects. With respect to syntax, English heavily relies on subject-verb-object order, whereas Hindi uses subject-object-verb order; with respect to the writing system, English uses the Latin alphabet and Hindi uses Devanagari; with respect to vocabulary, English mostly borrows from Latin while Hindi has borrowings from Persian and Arabic (Masica, 1993; Shapiro, 1989). Therefore, as these languages differ substantially in their representation within LLM training corpora and their structural characteristics, making family-level grouping is potentially misleading.

To address this limitation, we adopt a more fine-grained, genus-level classification, which provides optimal granularity for our analysis. Genus represents an intermediate taxonomic level in linguistic classification, more specific than family but broader than individual languages, making it well-suited for

[3]Details about dataset statistics are provided in the AppendixTODO

[4]For more details including prompts, please refer to the MultiQ paper.

[5]Available at https://github.com/paul-rottger/multiq/tree/main

[6]https://huggingface.co/cis-lmu/glotlid

detecting systematic patterns while maintaining sufficient statistical power.

Language Coding and Genus Mapping Our genus-level analysis required careful alignment given that multiple coding systems were used in MultiQ. More specifically:

- Source languages are annotated with Google Translate IDs and WALS codes
- Generation languages are automatically identified using GlotLID (Kargaran et al., 2023), which assigns ISO 639-3 codes

We map all languages to their corresponding genera using the World Atlas of Language Structures (WALS) database Dryer and Haspelmath (2013), which contains 2,662 language entries, each annotated with genealogical information including genus classification. WALS provides both WALS-specific codes and ISO 639-3 codes, enabling consistent cross-referencing across the different identifier systems used in MultiQ. This mapping process involved: 1) a direct mapping: Languages with existing WALS codes were directly mapped to their genera and 2) ISO code alignment: Languages identified only by ISO 639-3 codes were matched to WALS entries.

The resulting genus mapping covers 47 genera across 21 language families, providing sufficient diversity for robust statistical analysis while maintaining genealogical precision.

Our analysis extends the original MultiQ evaluation framework in two key dimensions:

1. Genus Fidelity Analysis: We examine whether language-switching patterns respect genealogical boundaries by comparing within-genus vs. cross-genus switching rates;
2. Cross-Genus Consistency Analysis: We assess whether question-answering accuracy correlates more strongly within genealogical groups than across them.

3. Genus fidelity

In this section, we investigate whether LLMs display systematic biases toward or within specific language genera. Our primary focus is on genus-level fidelity, as this provides the most direct test of genealogical effects.

3.1. Methodology

We operationalize genus fidelity as the tendency for models to generate responses within the same genus as the one of the input prompt. For each source genus, we collect all model responses, classify their genera using our WALS-based mapping, and compute generation distributions.

Our primary metric is *cross-genus model fidelity* which captures the proportion of outputs in which the model faithfully maintains the linguistic genus of the input, thereby providing a quantitative measure of cross-genus consistency.

Formally, we define the FidelityScore for a given model as follows:

$$\begin{aligned}\text{FidelityScore}_{LLM} &= \frac{N_{\text{faithful}}}{N_{\text{total}}} \\ &= \frac{|\{p : G_p = G_{LLM(p)}\}|}{|\{p\}|}\end{aligned} \tag{1}$$

where p denotes an input prompt, $LLM(p)$ the corresponding model output, $|.|$ the cardinality of a set and G_x the genus associated with the text x. In other words:

$$\begin{aligned}N_{\text{faithful}} &= \text{Number of prompts where the output} \\ &\quad\ \text{genus matches the input genus,} \\ N_{\text{total}} &= \text{Total number of prompts.}\end{aligned}$$

3.2. Results

Model-centric Perspective Table 1 presents genus fidelity scores across all evaluated models.

Model	Fidelity
Llama-2-7b	17.3
Llama-2-14b	23.1
Llama-2-70b	27.9
Mixtral-8x7B	73.9
Mistral-7B	75.0
Qwen1.5-7B-Chat	70.5
Apertus-8B	**92.3**

Table 1: Genus fidelity score by model.

The results indicate substantial variation in genus fidelity between models, revealing a clear separation into three fidelity-score tiers. Models from the Llama family get the lowest fidelity scores, with Llama-2-7B achieving as only 0.17 genus consistency. While increasing the model size seems to have positive effect on genus fidelity score, even the largest LLama model Llama-2-70B only reaches 0.28 genus fidelity. Mistral-7B, Mixtral-8x7B and Qwen1.5-7B-Chat have remarkably higher fidelity scores ranging from 0.7 to 0.75 of genus fidelity. The highest performing in terms of genus fidelity is Apertus with 0.92. This suggests that some models have developed stronger genealogical coherence in their multilingual representations. However, it is

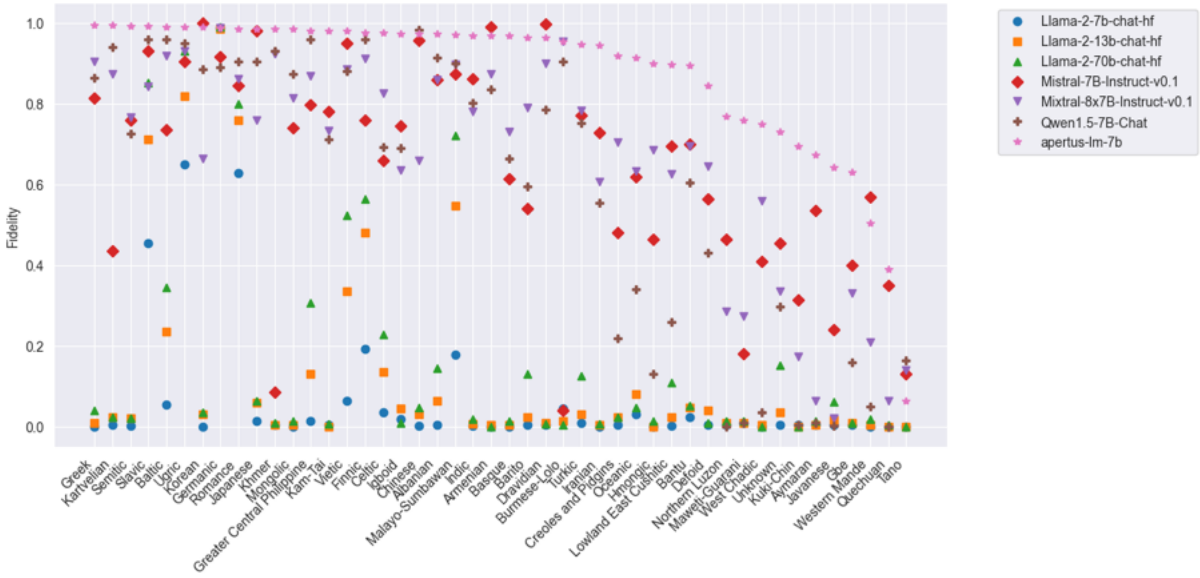


Figure 1: Fidelity scores across genera (existing in MultiQ) and models (existing in MultiQ + Apertus).

important to note that fidelity does not guarantee answer correctness.

Next, we take a closer look at each model separately. More specifically, for each model, and every single prompt p, we extract the input genus G_i (the genus of the input prompt) and the output genus G_o (the genus of the model-generated output $LLM(p)$). We present the genus fidelity across different genera (present in MultiQ) and models in Figure 1.

The results reveal clear differences in model behaviour across genera. Models from Llama family, although predominantly having low fidelity scores, have extremely high fidelity scores for Germanic genus (around 1, even for the smallest model) and notably high scores for Ugric (larger bigger models above 0.8), Romance and Slavic genera (above 0.7). We also observed a strong tendency of Llama models to default to English (a Germanic language) when confronted with non-English prompts. Mistral, Mixtral and Qwen show a ge nerally high fidelity, though their performances vary across genera –for example, significantly drops for the Tano, Quechuan and Kuki-Chin genera. Interestingly, Mixtral and Qwen exhibit more similar behaviour to each other than to Mistral. In particular, for Khmer and Burmese-Lolo genera, Mistral has a fidelity score ≤ 0.15 while Mixtral and Qwen have score ≥ 0.9. Similarly, for Vietic, Kam-Tai and Iranian genera, Mixtral and Qwen show closer results. Despite lacking explicit multilingual branding, Mistral often responds in the prompt genus, suggesting robust multilingual competence. Nevertheless, we noticed that on fallbacks these two models behave contrastingly: Mixtral almost always uses English while Qwen is more variable in fallbacks, sometimes producing outputs in unrelated languages (oftentimes in Chinese). Finally, Apertus demonstrates near-perfect genus fidelity, having score ≤ 0.5 only for Tano and Quechuan.

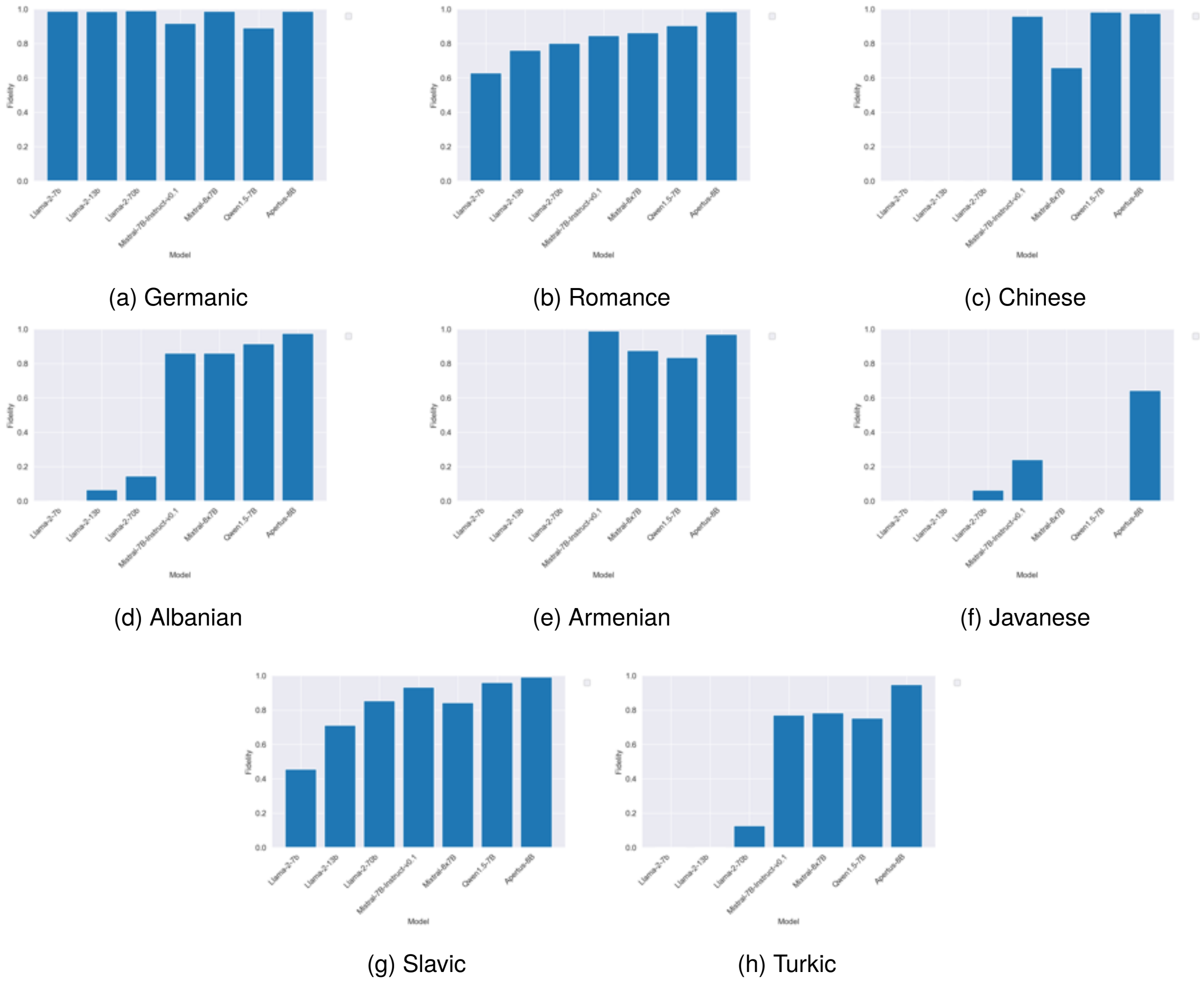
Finally, model descriptions do not always correspond to observed behaviour: despite multilingual claim, LLaMA exhibits a systematic English bias, whereas Mistral demonstrates stronger multilingual fidelity despite the absence of explicit multilingual positioning.

Genus-level Perspective We further provide a detailed breakdown of results for eight representative genera spanning diverse resource levels and linguistic families: Slavic, Germanic, Romance, Javanese, Albanian, Turkic, Armenian, and Chinese. We consider Javanese, Albanian and Armenian to be low-resource genera since their corresponding languages (one per genus) are low-resource languages according to the literature (Nuci et al. (2024), Goyal et al. (2022)).

Results for these eight genera are shown in the Figure 2. We additionally show Apertus-8B and Llama-2-7b (having the highest contrasted performances) in Figures 3a and 3b, with the other models detailed in Appendix B.1.

Across all models, Germanic languages consistently achieve high genus fidelity (≥ 0.75). A similar trend is observed for Romance languages, except for the smallest Llama model. Even lower-resource languages within these genera benefit from this stability (except from Llama models which show systematic bias towards English and thus Germanic genus as a fallback).

For Chinese languages, fidelity significantly drops for Llama models and Mixtral. In particular, Llama models again show a pronounced Germanic



(a) Germanic (b) Romance (c) Chinese (d) Albanian (e) Armenian (f) Javanese (g) Slavic (h) Turkic

Figure 2: Genus-level fidelity across models. For each representative genus, we report the proportion of model outputs that remain within the same genus as the prompt language.

bias in over 60% of non-faithful cases. This pattern suggests weaker multilingual competence for these genera, with models reverting to training-dominant languages.

Fidelity to Slavic languages overall remains strong (over 0.8) apart from two smaller models from Llama family.

For low-resource genera, fidelity varies substantially across models. While models maintain a minimal fidelity to Albanian (even Llama), only non-Llama models preserve fidelity for Armenian, Mistral and Apertus achieving particularly high fidelity scores. In contrast, for Javanese most models struggle and even Apertus barely reaches 0.6 fidelity score.

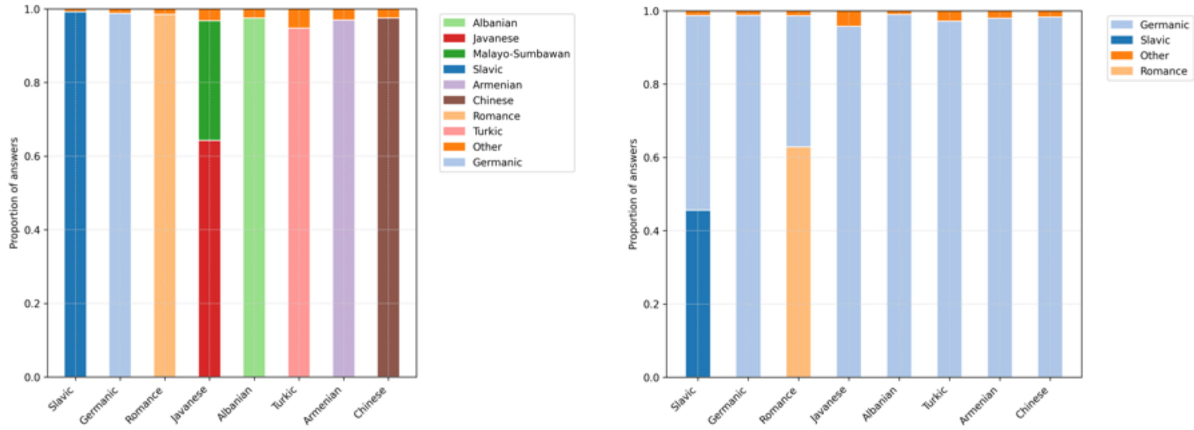
Turkic languages also exhibit complex patterns. Mistral and Qwen maintain a fidelity above 0.75 , whereas Llama produces Germanic outputs in over 80% of cases. Detailed inspection reveals substantial intra-genus variation: Turkish prompts yield relatively faithful responses, while Kazakh frequently triggers English outputs. This disparity likely reflects multiple factors: resource imbalance (Turkish being better represented in training data), script effects, and contact phenomena.

4. Genus switch

If an LLM answers a question correctly in one language, is it more likely to answer correctly when the same question is posed in another language of the same genus? We investigate whether genus consistency facilitates knowledge transfer across languages.

4.1. Methology

If a model demonstrates knowledge by answering correctly in one language, changing only the prompt language should not impede correct responses – assuming sufficient multilingual competence. We test whether genealogical proximity preserves this knowledge consistency better than genealogically distant language pairs.



(a) Apertus

(b) Llama-7b

Figure 3: Genus-level output distribution by model. For each prompt genus, we indicate the genus of the model's generated response. Remaining models are reported in Appendix B.1.

Setup Using MultiQ, we identify questions answered correctly in a source language, then evaluate the same questions across all available target languages. This controlled design isolates language effects from knowledge availability, since the model has already demonstrated requisite knowledge.

Metrics We compute SwitchScores measuring the proportion of questions answered correctly in target genus g_t given correct answers in source genus g_i:

$$\text{SwitchScore}(g_i, g_t) = \frac{\mid \mathcal{Q}g_i, g_t \mid}{\mid \mathcal{Q}g_i \mid} \tag{2}$$

where $\mathcal{Q}_{g_i,g_t}$ represents questions answered correctly in both genera, and $\mathcal{Q}_{g_i}$ represents questions answered correctly in the source genus.

We distinguish:

- SwitchScore-In (within the same genus): SwitchScore(g_i, g_i) - within-genus consistency
- SwitchScore-Out (outside of input genus): Average performance when switching to other genera (different from g_i)

Question selection The original MultiQ evaluation used the complete dataset across all languages. However, to ensure that observed differences genuinely reflect language effects rather than artifacts of question difficulty or translation quality, we construct a filtered subset optimized for cross-genus comparison.

More specifically, a question is retained if it is answerable across the compared genera, i.e., the model produces a correct answer in at least one language within each genus. This prevents biases arising from inherently unanswerable questions.

Moreover, we only keep languages where the model achieves at least a minimal number of correct answers N_c to ensure statistical reliability. We use N_c values of 20, 50 and 100.

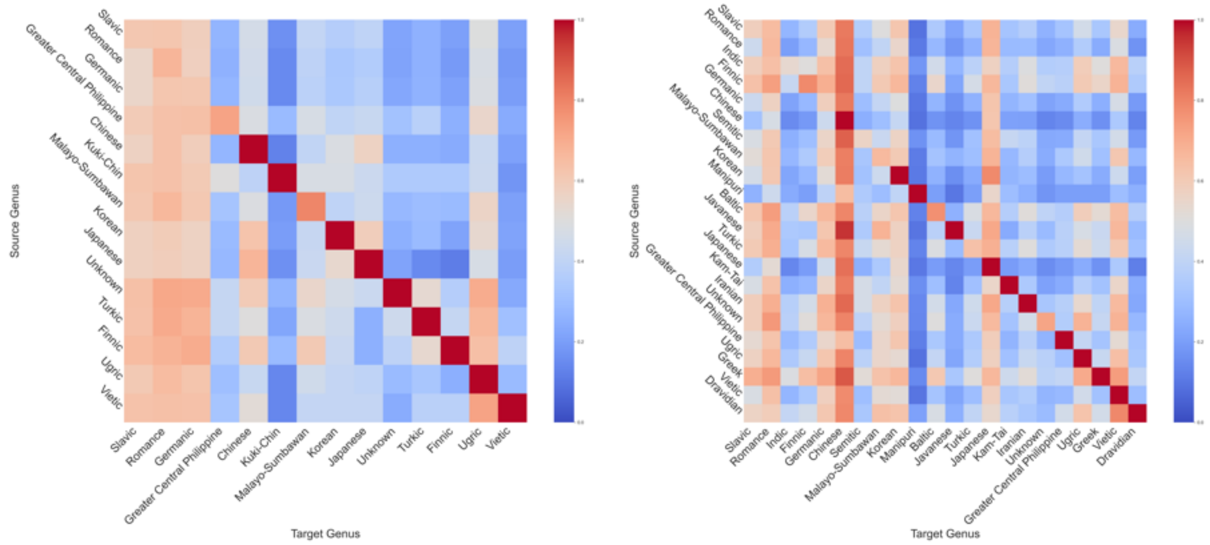
Resulting Dataset Characteristics Our filtering process yields a curated dataset optimized for genealogical analysis while maintaining the linguistic diversity required for robust conclusions. Table 2 presents the resulting dataset size under different filtering thresholds. With this approach, we aim at prioritizing interpretability and statistical validity over dataset size, ensuring that our genealogical findings reflect genuine linguistic patterns.

4.2. Results

Global SwitchScores appear in Table 3. To illustrate diversity (both in terms of LLMs and threshold), detailed genus-level scores are shown in Figure 4, for Mistral-7B(4a) and Qwen-1.5-7B (4b) with threshold 20, and Figure 5, for Llama-2-70b (5a) and Apertus-8B (5b) with threshold 50. Moreover, additional results can be found in Appendix C.

All models show substantially higher knowledge consistency within genera (80-90%) compared to cross-genus transfers (40-50%). This 35-40 percentage point advantage demonstrates that genealogical relatedness significantly facilitates knowledge preservation.

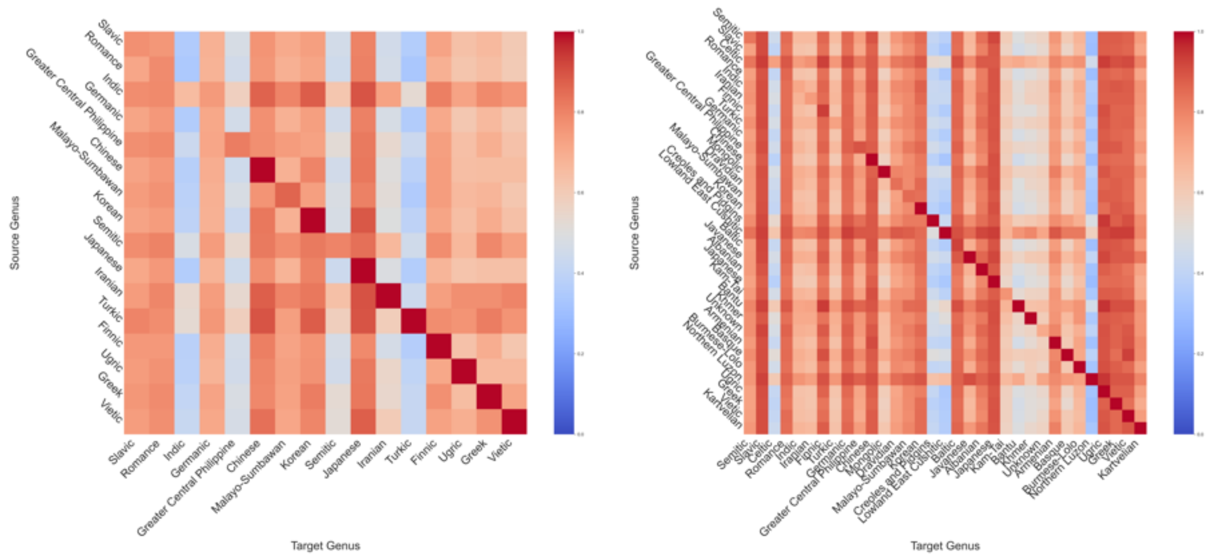
Detailed Switchscores A key finding is that performance depends critically on the target genus rather than the source (notice the red/blue *column* pattern across Figures 4) and 5). Well-resourced genera (Germanic, Romance) serve as robust targets regardless of source, while poorly resourced genera (e.g., Kuki-Chin) yield degraded performance even from high-resource sources.



(a) Mistral-7B, threshold = 20

(b) Qwen-1.5-7B, threshold = 20

Figure 4: Switchscore distributions for Mistral-7B (left) and Qwen-1.5-7B (right) for **threshold = 20**. As can be seen from red/blue-column patterns, the performance critically depends on the target genus. Switchscore distributions for the rest of the models for threshold 20 can be found in Appendix C.



(a) Llama-70b, threshold = 50

(b) Apertus-8B, threshold = 50

Figure 5: Switchscore distributions for Llama-70B (left) and Apertus-8B (right) for **threshold = 50**. As can be seen from red/blue-column patterns, the performance critically depends on the target genus. Switchscore distributions for the rest of the models for threshold 50 can be found in Appendix C.

N_c	**Llama-7B**		**Llama-13B**		**Llama-70B**		**Mistral-7B**		**Mixtral-8x7**		**Qwen-7B**		**Apertus-8B**	
	# q	# G	# q	# G	# q	# G	# q	# G	# q	# G	# q	# G	# q	# G
20	7	26	5	23	4	24	8	33	3	14	3	22	**14**	**44**
50	6	16	4	17	3	15	7	20	3	8	2	9	**14**	**33**
100	5	11	3	7	1	6	5	11	2	3	1	4	**13**	**27**

Table 2: Number of questions (# q, expressed in thousands) and number of genera (# G) remaining after applying different filtering thresholds for each model.

Model	Switch-In	Switch-Out
Llama-2-7b	88.9	49.4
Llama-2-14b	82.8	51.6
Llama-2-70b	86.3	54.0
Mixtral-8x7B	83.6	47.7
Mistral-7B	88.8	41.8
Qwen1.5-7B-Chat	84.1	42.0
Apertus-8B	**90.4**	**60.6**

Table 3: Switch scores by model. Obtained with a threshold of 20.

Moreover, results are asymmetric: for Llama-70b switching from Javanese (Austronesian) to Germanic maintains high accuracy, whereas the reverse direction shows substantial degradation. This suggests that target language representation in training data dominates genealogical effects when resources are scarce.

Genealogical Boundaries Despite overall genus-level patterns, genealogical classification does not perfectly predict transfer success. Within Indo-European, Indic and Slavic genera exhibit markedly different behaviors despite shared family membership. Similarly, script overlap (Latin, Cyrillic) provides no guarantee of stable transfer performance. These exceptions highlight that while genealogical relatedness provides a useful organizational principle for understanding multilingual LLM behavior, it competes with training data distribution, script similarity, and other linguistic factors in determining cross-lingual knowledge consistency.

5. Conclusions

In this paper, we examined the genealogical sensitivity of Large Language Models through a genus-level analysis, extending the work of Holtermann et al. (2024). We found that LLMs exhibit higher fidelity and knowledge consistency within genealogical boundaries, but this effect is largely mediated by training resource availability. Distinct multilingual strategies also emerged across model families, with models defaulting to Germanic languages and others adopting more nuanced behaviors. Overall, our findings indicate that resource distribution, rather than genealogical structure, remains the primary driver of multilingual performance.

Impact statement

As we are witnessing the progressive usage of LLMs, also for the scopes of generating different benchmarks, we would like to remind that even these less-resource intensive activities contribute to high energy consumption and carbon emissions. To give our small contribution to the AI sustainability, we opted to use existing benchmark and intervene only as needed. We hope that this can inspire other LLM-related research to leverage existing resources at least equally optimally.

Acknowledgements

The work described in this paper has been supported by the "*Language Preference and Hallucination Investigation in Multilingual RAG (LPHI-mRAG)*" project, funded by armasuisse S&T, Switzerland.

6. Bibliographical References

Mina Almasi and Ross Kristensen-McLachlan. 2025. Alignment drift in CEFR-prompted LLMs for interactive Spanish tutoring. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 70–88, Vienna, Austria. Association for Computational Linguistics.

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Yiyi Chen, Qiongxiu Li, Russa Biswas, and Johannes Bjerva. 2025. Large language models are easily confused: A quantitative metric, security implications and typological analysis. In *Findings of the Association for Computational*

Linguistics: NAACL 2025, pages 3810–3827, Albuquerque, New Mexico. Association for Computational Linguistics.

Zoltan Csaki, Bo Li, Jonathan Lingjie Li, Qiantong Xu, Pian Pawakapan, Leon Zhang, Yun Du, Hengyu Zhao, Changran Hu, and Urmish Thakker. 2024. SambaLingo: Teaching large language models new languages. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 1–21, Miami, Florida, USA. Association for Computational Linguistics.

Matthew S. Dryer. 1989. Large linguistic areas and language sampling. *Studies in Language*, 13:257–292.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Vansh Gupta, Sankalan Pal Chowdhury, Vilém Zouhar, Donya Rooein, and Mrinmaya Sachan. 2025. Multilingual performance biases of large language models in education.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2025. DOVE: A large-scale multi-dimensional predictions dataset towards meaningful LLM evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11744–11763, Vienna, Austria. Association for Computational Linguistics.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Ines Altemir Marinas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendonççа, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments. `https://arxiv.org/abs/2509.14233`.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494, Bangkok, Thailand. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts.

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*.

Fengyuan Liu, Nouar AlDahoul, Gregory Eady, Yasir Zaki, and Talal Rahwan. 2025a. Self-reflection makes large language models safer, less biased, and ideologically neutral.

Wei Liu, Sony Trenous, Leonardo F. R. Ribeiro, Bill Byrne, and Felix Hieber. 2025b. Xrag: Cross-lingual retrieval-augmented generation.

Colin P Masica. 1993. *The indo-aryan languages*. Cambridge University Press.

Krenare Pireva Nuci, Paul Landes, and Barbara Di Eugenio. 2024. RoBERTa low resource fine tuning for sentiment analysis in Albanian. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14146–14151, Torino, Italia. ELRA and ICCL.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual instruction tuning with just a pinch of multilinguality. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2304–2317, Bangkok, Thailand. Association for Computational Linguistics.

Michael C Shapiro. 1989. *A primer of modern standard Hindi*. Motilal Banarsidass Publ.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang D. Nguyen. 2025. Scaling test-time compute for low-resource languages: Multilingual reasoning in llms.

Dawid Wiśniewski, Antoni Solarski, and Artur Nowakowski. 2025. Exploring the feasibility of multilingual grammatical error correction with a single LLM up to 9B parameters: A comparative study of 17 models. In *Proceedings of Machine Translation Summit XX: Volume 1*, pages 231–247, Geneva, Switzerland. European Association for Machine Translation.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. 2024. Beyond english-centric llms: What language do multilingual language models think in?

Han Zhou, Xingchen Wan, Ivan Vulić, and Anna Korhonen. 2023. Survival of the most influential prompts: Efficient black-box prompt search via clustering and pruning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13064–13077, Singapore. Association for Computational Linguistics.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8411–8423, Bangkok, Thailand. Association for Computational Linguistics.

7. Language Resource References

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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The Flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4476–4494, Bangkok, Thailand. Association for Computational Linguistics.

A. Statistics

The languages in MultiQ span 47 genera, 18 of which are represented by multiple languages. Table 4 shows the distribution of these 18 genera.

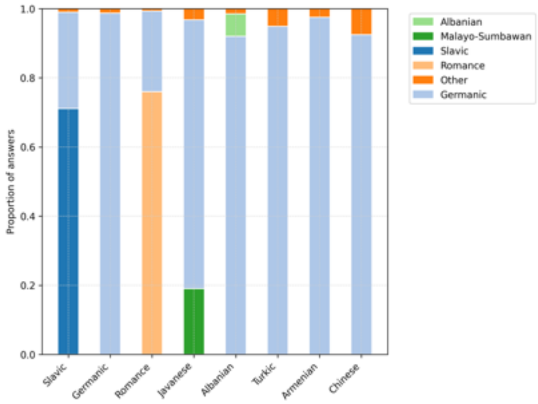
B. Genus output detail

B.1. Detail per model

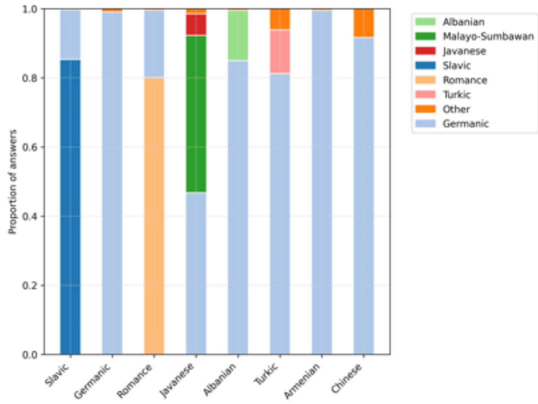
The details of output genera for eight selected genera per model can be seen in Figure 6.

Genus	Count	Genus	Count	Genus	Count
Baltic	2	Greater Central Philippine	2	Oceanic	3
Bantu	11	Indic	16	Romance	9
Celtic	3	Iranian	4	Semitic	5
Dravidian	4	Kam-Tai	2	Slavic	11
Finnic	2	Lowland East Cushitic	2	Turkic	8
Germanic	11	Malayo-Sumbawan	3	Unknown	3

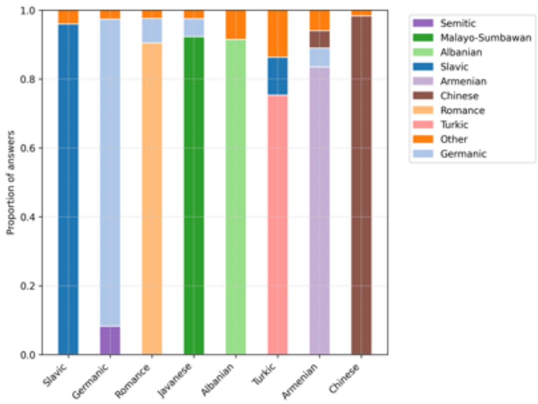
Table 4: Distribution of the genera with two or more languages in the dataset (N=18 genera, 101 languages)



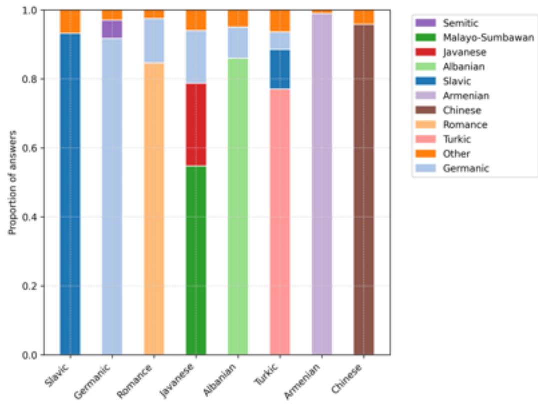
(a) Llama-13b



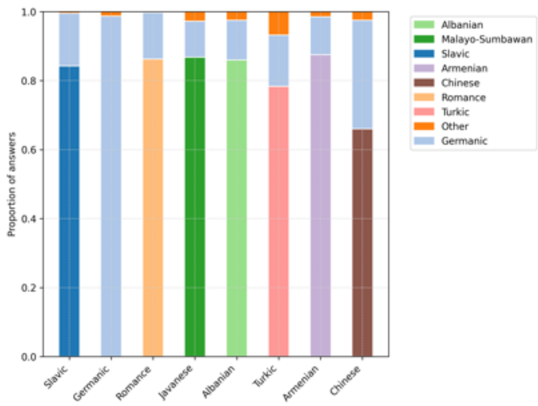
(b) Llama-70b



(c) Qwen1.5-7B



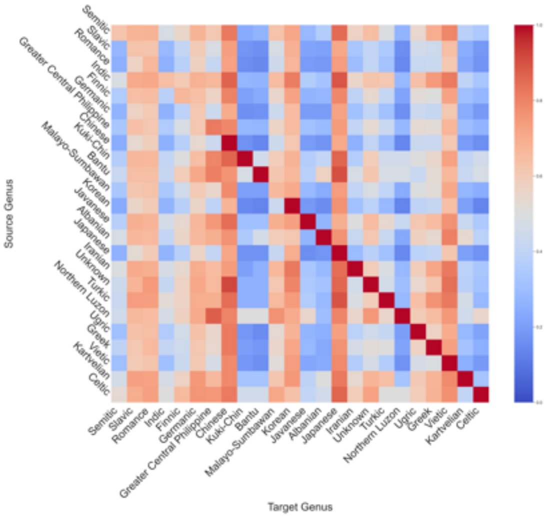
(d) Mistral-7B



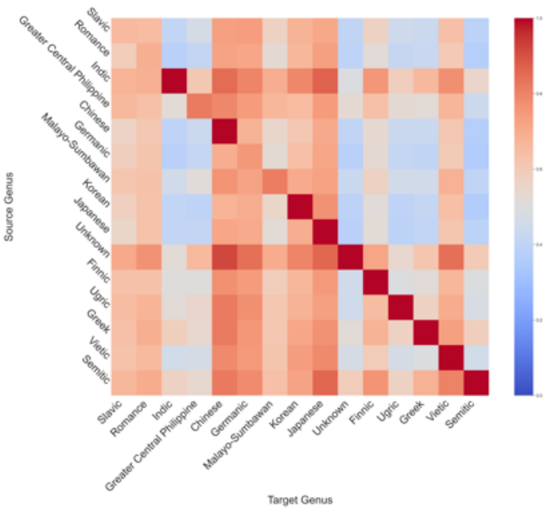
(e) Mixtral-8x7B

Figure 6: Genus-level output distribution by model. For each prompt genus, we indicate the genus of the model's generated response.

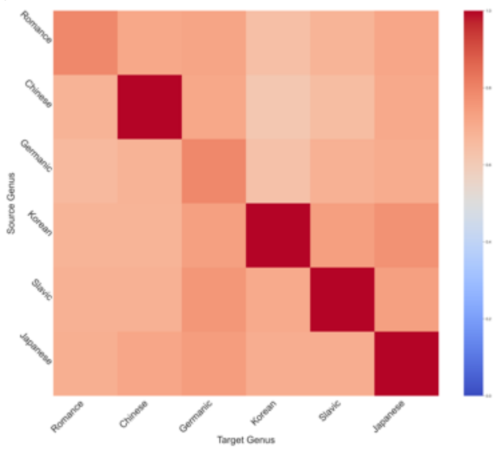
C. Switchscores



(a) Switchscores of model Llama-7b, threshold 20.

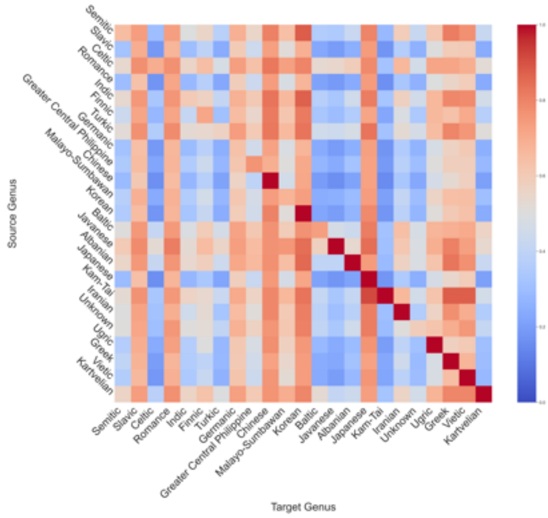


(b) Switchscores of model Llama-7b, threshold 50.

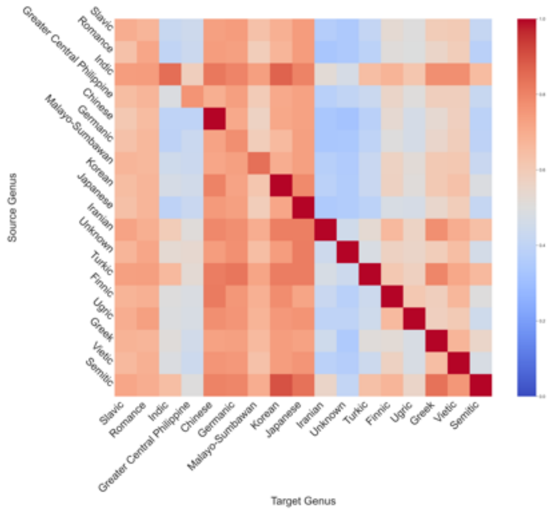


(c) Switchscores of model Llama-7b, threshold 100.

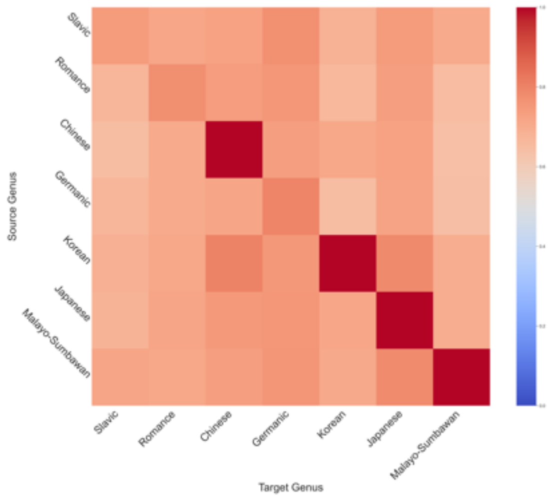
Figure 7: Switchscores Llama-7b.



(a) Switchscores of model Llama-13b, threshold 20.

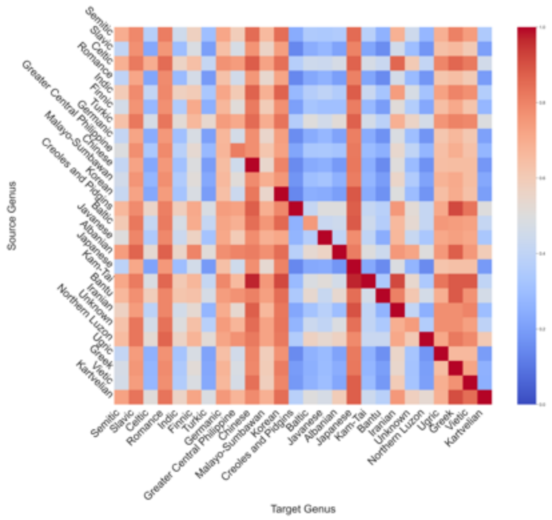


(b) Switchscores of model Llama-13b, threshold 50.

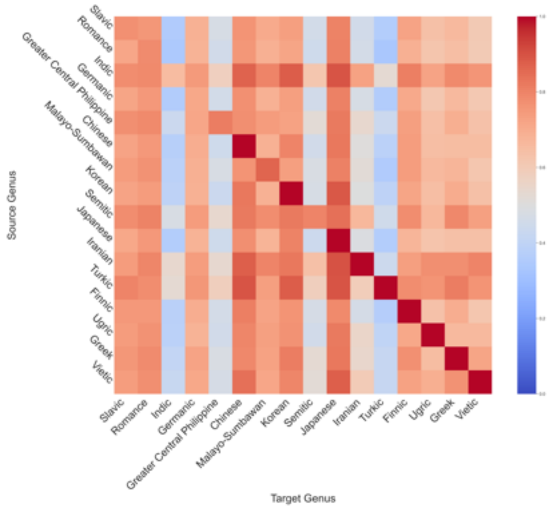


(c) Switchscores of model Llama-13b, threshold 100.

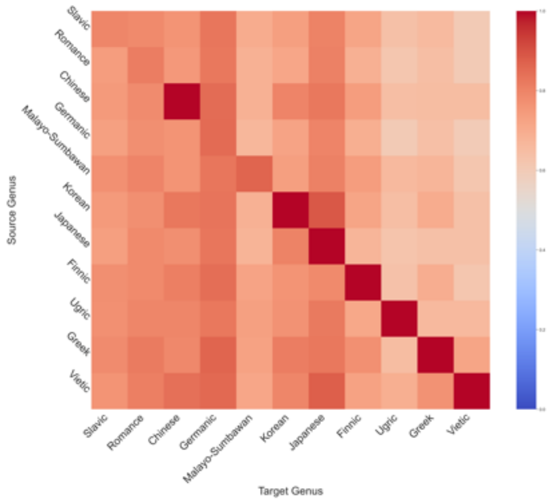
Figure 8: Switchscores Llama-13b.



(a) Switchscores of model Llama-70b, threshold 20.

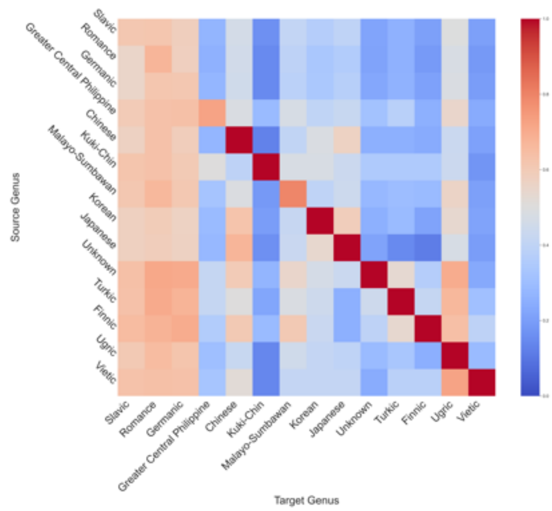


(b) Switchscores of model Llama-70b, threshold 50.

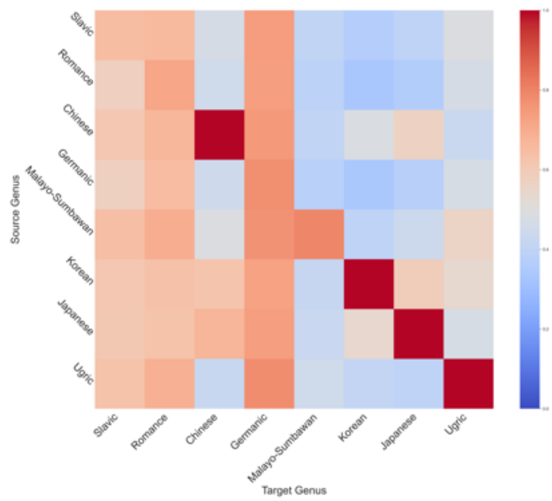


(c) Switchscores of model Llama-70b, threshold 100.

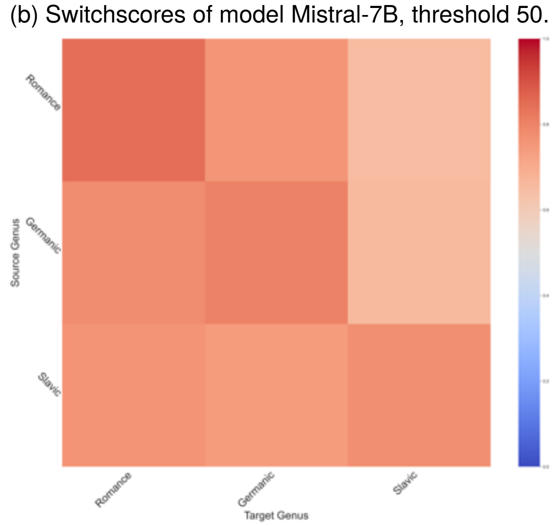
Figure 9: Switchscores Llama-70b.



(a) Switchscores of model Mistral-7B, threshold 20.

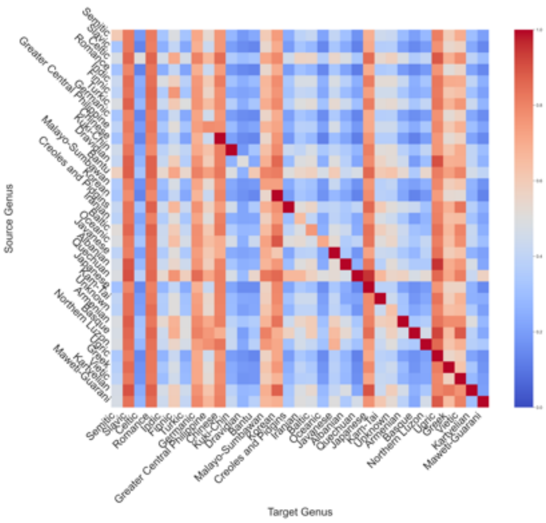


(b) Switchscores of model Mistral-7B, threshold 50.

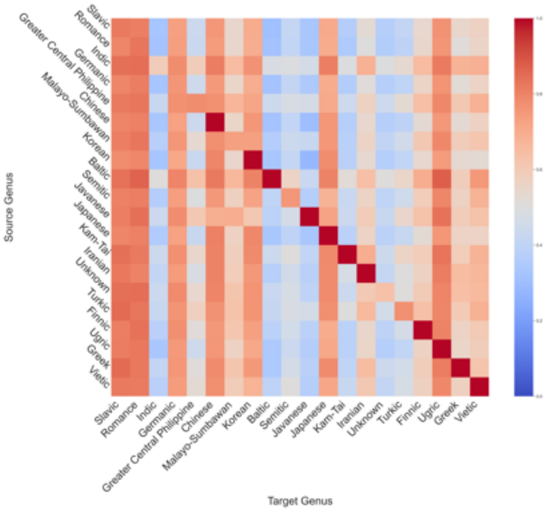


(c) Switchscores of model Mistral-7B, threshold 100.

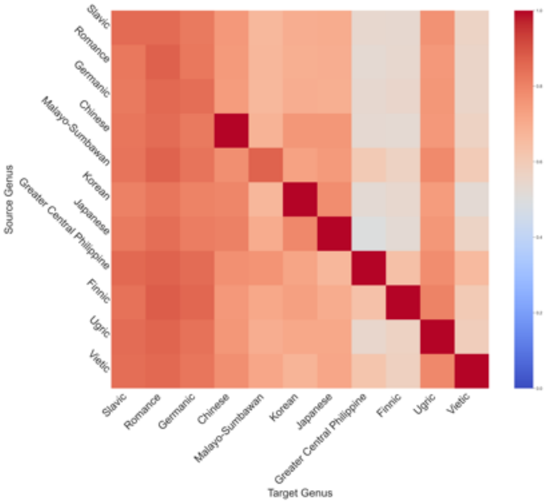
Figure 10: Switchscores Mistral-7B.



(a) Switchscores of model Mixtral-8X7, threshold 20.

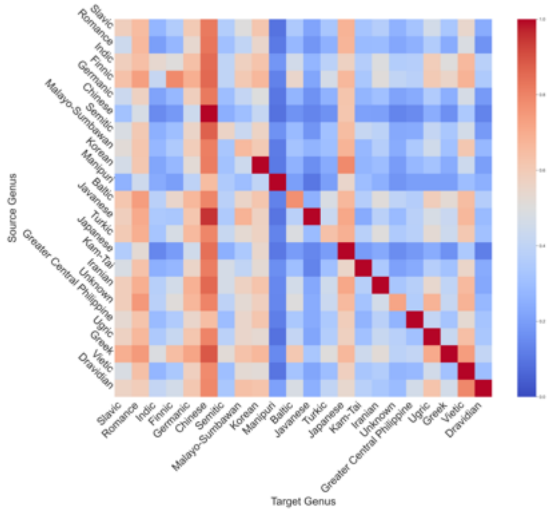


(b) Switchscores of model Mixtral-8X7, threshold 50.

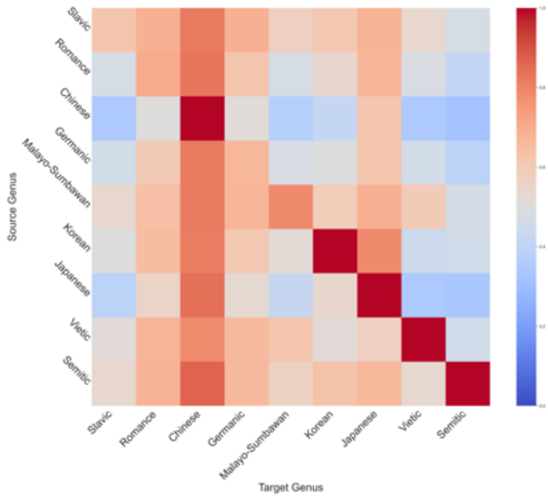


(c) Switchscores of model Mixtral-8X7, threshold 100.

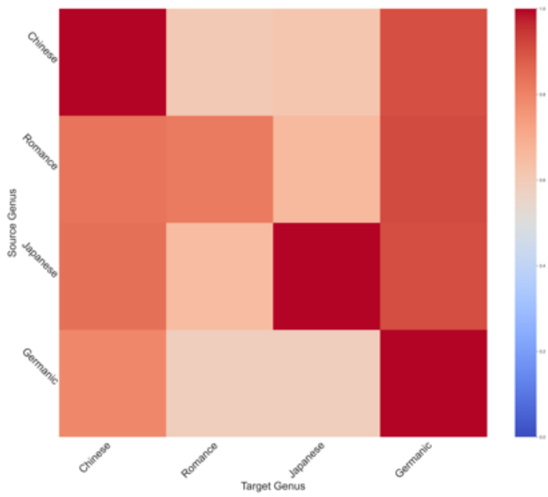
Figure 11: Switchscores Mixtral-8X7.



(a) Switchscores of model Qwen-7B, threshold 20.

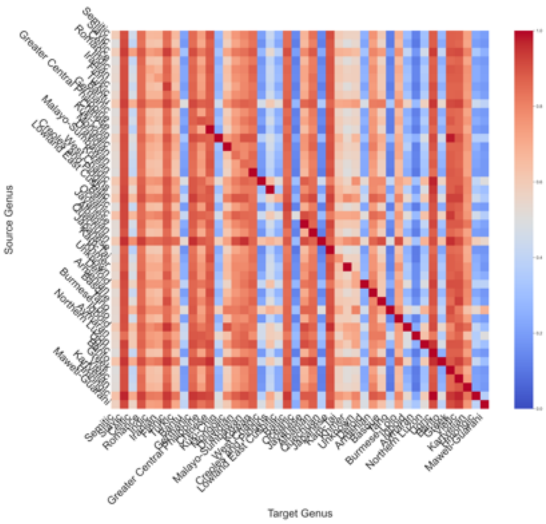


(b) Switchscores of model Qwen-7B, threshold 50.

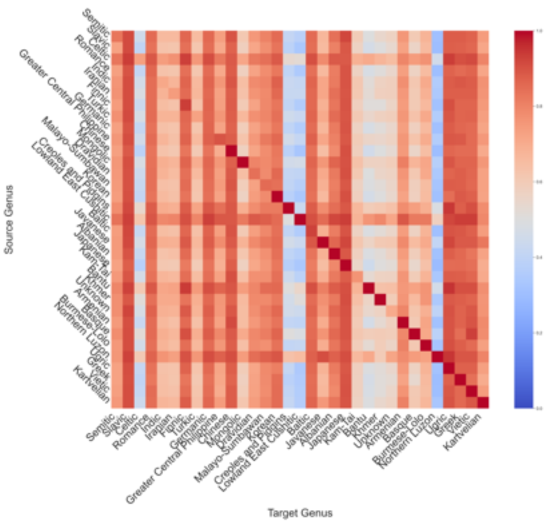


(c) Switchscores of model Qwen-7B, threshold 100.

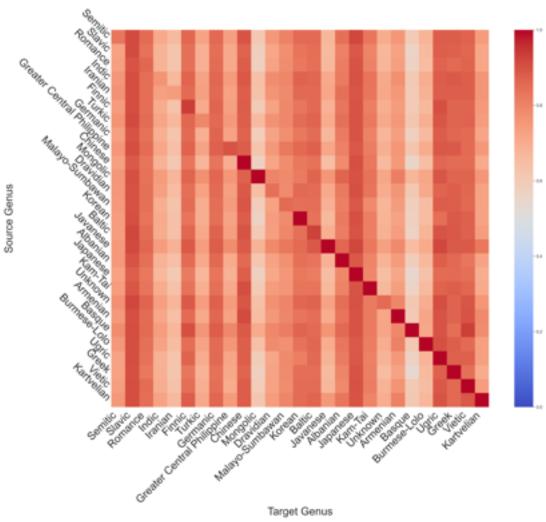
Figure 12: Switchscores Qwen-7B.



(a) Switchscores of model Apertus-7B, threshold 20.



(b) Switchscores of model Apertus-7B, threshold 50.



(c) Switchscores of model Apertus-7B, threshold 100.

Figure 13: Switchscores Apertus-7B.