
LLMs Flatten the Global South: Sub-Regional Representation Asymmetry

Noah De Nicola
denicola94@gmail.com

With
Apart Research¹

Abstract

LLMs default to Northern values and collapse within-country variation. Prior work shows this at the output level. We ask whether the same asymmetry shows up in the model’s internal representations, with sub-regions in the Global South separating less than those in the Global North. We build persona vectors for 21 countries and 105 cities across three open-weight LLMs, and measure how far each country’s cities separate from it in activation space. We control for two confounds: Wikipedia corpus frequency and BPE tokenizer fragmentation. **Global North sub-regions separate 39% more than Global South across all layers and 77% more in later layers.** The gap survives matched corpus frequency and token fragmentation.

1. Introduction

Large Language Models (LLMs) are deployed across the globe, and the cultural defaults they carry travel with them. Those defaults are not neutral. Across model families and evaluation setups, LLMs systematically favour Northern values when asked about everyday life, ethics, and preference (Tao et al., 2024; Zhao et al., 2026).

This bias has so far been documented at the country level. But cultures are not country-shaped. Different regions of a nation often contain distinct cultures, each with its own values and preferences. LLMs have been shown to collapse this within-country diversity, with distinct sub-regions eliciting essentially the same response (Wang et al., 2025; Madhusudan et al., 2026).

We ask the question that sits at the intersection of these findings: does the model represent sub-regional identity with the same granularity across the South as across the North? When prompted with one sub-region versus another, does the South collapse while the North stays separate?

¹Research conducted at [The Global South AI Safety Hackathon](#), June 2026

Prior work has measured cultural bias and collapse from the model’s outputs, using text comparisons, MCQ accuracy, and value-survey responses (Tao et al., 2024; Wang et al., 2025; Madhusudan et al., 2026; Zhao et al., 2026). We instead measure the model’s activations. Following Chen et al. (2025), we compute a persona vector for each country and sub-region as the mean activation under a persona prompt, minus the baseline. *Sub-region separation* is thus the spread among sub-region personas within a country.

Output-level evaluations can conflate what the model says with what it represents. Interventions such as prompting, post-hoc fine-tuning, and vector steering can produce differentiated outputs from collapsed internal states, but cannot create a distinction in the model’s representations it doesn’t already carry (Zou et al., 2023; Jain et al., 2024).

The level of granularity is likely to correlate with how often sub-regions appear in training data (Truong et al., 2026). We include data frequency in our analysis to test whether the collapse exceeds what less training data alone can explain. Tokenization is another likely correlate since less represented region names are fragmented into more sub-tokens by the model’s BPE tokenizer (Petrov et al., 2023). We include a fragmentation measure to test for this confound.

Our contributions are as follows:

- An open-source framework for measuring sub-regional separation in LLM activations. It accepts any set of regions and any question bank, extracts per-region persona vectors, computes within-country sub-region separation, and controls for Wikipedia corpus frequency and BPE tokenizer fragmentation.
- A study applying the framework to Gemma-2-2B, Llama-3.1-8B, and Qwen-2.5-7B, using 10 Western European and 11 sub-Saharan African countries (5 cities each) and 40 first-person everyday questions.

2. Related Work

Cultural bias and collapse. A growing body of work documents that LLMs default to values and preferences associated with Northern cultures. Without cultural prompting, GPT models answer the World Values Survey most like respondents from English-speaking and Protestant-European countries, and explicit cultural prompting only partially corrects this (Tao et al., 2024). The same Anglo-centric baseline surfaces in other settings. Under a neutral prompt, street-view image descriptions sit closer to European and North American framings than to any other region (Zhao et al., 2026).

Khanuja et al. (2026) argue that this is not purely a knowledge gap. Activation-space steering elicits culturally specific generations that prompting alone fails to surface, suggesting models hold the knowledge but default away from it. On an Indian commonsense benchmark, models fall back on a Central/Northern Indian framing (Madhusudan et al., 2026), and across US demographic identities, persona-prompted outputs exhibit far less within-group variance than real survey respondents (Wang et al., 2025).

Persona vectors. Building on the representation-engineering line of Zou et al. (2023), Chen et al. (2025) define a *persona vector* for a trait as the difference in mean activations between responses produced with and without it, and show that such vectors can monitor or steer the trait.

3. Method

Models. We evaluate three open-weight instruction-tuned models spanning two parameter scales and three model families: Gemma-2-2B-it (bf16) (Team et al., 2024), Llama-3.1-8B-Instruct (4-bit) (Grattafiori et al., 2024), and Qwen-2.5-7B-Instruct (4-bit) (Yang et al., 2024).

Regions. We compare 11 sub-Saharan African countries against 10 Western European countries, with 5 cities per country (105 city regions plus 21 country regions; full list in Appendix B). These countries were chosen for a clean Southern / Northern contrast within a constrained geography. We deliberately did *not* frequency-match: instead, we report sub-region separation as a function of frequency and let the scatter show whether the gap survives the control. Nigeria is included as a high-frequency Southern reference point, letting us check whether the granularity gap persists at higher African frequencies.

Questions. We use a fixed bank of 40 first-person, everyday questions that the persona answers in role (e.g. “*What did you have for breakfast this morning?*”, “*What did you do for fun as a kid?*”). Questions avoid religious, climate-specific, or infrastructural defaults (e.g. no Christmas or Thanksgiving) so the only cultural signal comes from the persona prompt. The full bank is in Appendix A.

Persona vectors. Following Chen et al. (2025), we construct a region-conditioned persona vector for each region r as the per-layer mean difference in residual-stream activations between a region-prompted and a generic-persona-prompted forward pass:

$$v_{r,\ell} = \mathbb{E}_q [h_\ell(\text{persona}_r, q) - h_\ell(\text{baseline}, q)], \quad (1)$$

where h_ℓ is the residual-stream activation at layer ℓ and q ranges over the 40 questions. Activations are read from the final input token (the generation-start position). Our setup is a deliberately simplified variant of Chen et al. (2025)’s, intended for representational analysis rather than behavioural control. We retain the term *persona vector* but make no claims about the causal steering behaviour of these vectors.

Sub-region separation. For each country c with city set $C(c)$, we measure how much the model separates the country’s sub-regions from the country itself:

$$d_\ell(c) = \frac{1}{|C(c)|} \sum_{k \in C(c)} (1 - \cos(v_{c,\ell}, v_{k,\ell})). \quad (2)$$

A high $d_\ell(c)$ means the model represents the cities as meaningfully distinct from the country. A low $d_\ell(c)$ means the cities collapse into the country vector. We aggregate to a Northern and Southern mean by averaging $d_\ell(c)$ across countries within each group.

Layer split. Early transformer layers are known to process surface-form information rather than semantic relationships (Tenney et al., 2019; Geva et al., 2021). Consistent with this, Chen et al. (2025) report that persona-vector steering is ineffective in early layers, with behavioural influence concentrated in middle-to-late layers. In our setting, this suggests early-layer $d_\ell(c)$ might be dominated by linguistic mismatch between place names (e.g. *Windhoek* vs. *Namibia*) rather than regional representation. We therefore report $d_\ell(c)$ both averaged over all layers and split across the first and last 50% of the model’s layers, treating the late-half signal as the more meaningful estimate of how the model organises region identity.

Frequency proxy. As a proxy for training-data prevalence we use the $\log_{10}$ of mean monthly English Wikipedia pageviews for each region’s article over a fixed 10-year window, following Mallen

et al. (2023). Per-country frequency is the mean of the country’s value and the mean over its 5 cities. Full details and one alias-handling case in Appendix C.

Fragmentation measure. As a second confound proxy, we measure how heavily Gemma-2-2B’s BPE tokenizer fragments each region label. We normalise the number of tokens by the number of characters in the label to control for name length.

Code. Code and data released here: <https://github.com/noahdenicola/subregion-separation>

4. Results

4.1 Sub-Region Separation

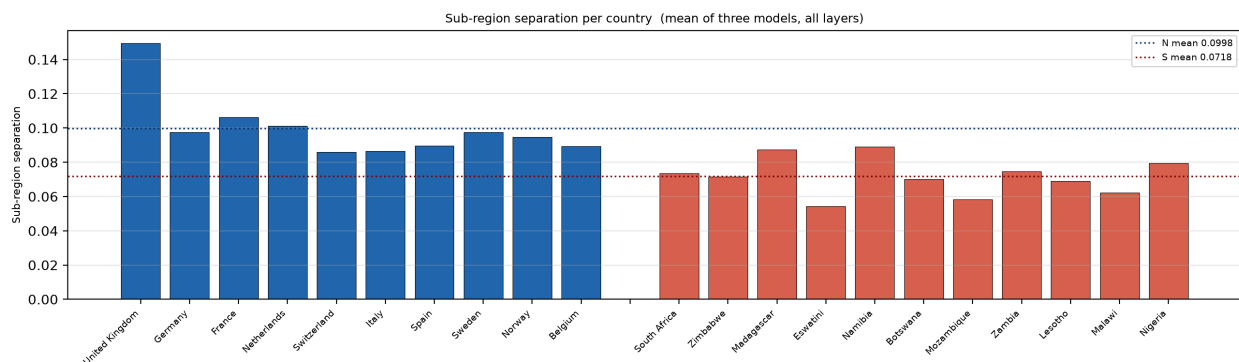


Figure 1: Sub-region separation per country, mean of Gemma, Llama, and Qwen, averaged over all transformer layers. Dotted lines mark the cluster means.

Figure 1 shows that every Northern country’s sub-region separation exceeds the Southern mean. Only Namibia and Madagascar cross into the Northern cluster. The Northern mean sits $\sim 39\%$ above the Southern mean. The models’ representations show a consistently greater separation between country and city for Northern countries than for Southern.

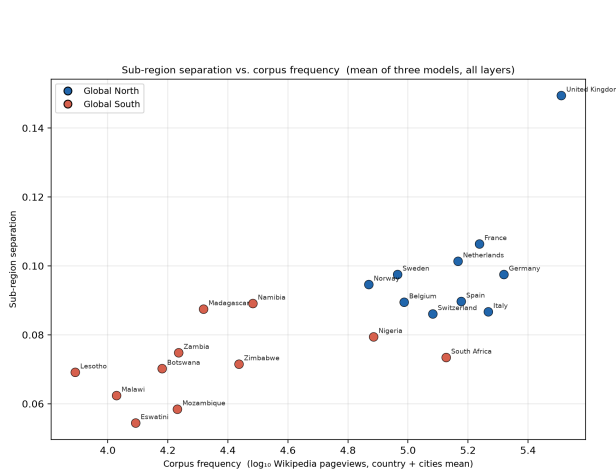
4.2 Data and Fragmentation Confounds

There are two natural non-representational explanations for the gap: **less training data** on Southern regions, and **more fragmentation** of Southern names by the tokenizer.

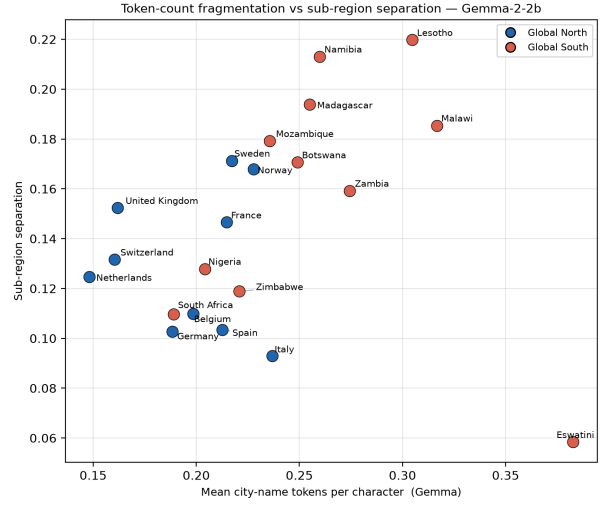
Corpus Frequency. Figure 2a illustrates that higher corpus frequency correlates with higher sub-region separation. However, we notice that Nigeria and South Africa sit substantially below Northern countries at comparable corpus frequencies.

Fragmentation Count. Figure 2a also reveals a second pattern within the Southern cluster. Madagascar and Namibia have low corpus frequency but achieve separation comparable to Northern nations such as Belgium, Spain, Switzerland, and Italy. This is likely down to how the tokenizer fragments both the country and city names.

Figure 2b illustrates the pattern. There is a direct relationship between tokens-per-character and sub-region separation in the early layers of the model. Eswatini is an outlier here. Its country name has by far the highest tokens-per-character score of any country shown (Figure 7, in the appendix).



(a) Sub-region separation vs. corpus frequency.



(b) Sub-region separation vs. mean city-name tokens-per-character, for Gemma-2-2b (first 10% of layers).

Figure 2: Confound checks. Sub-region separation against the two natural non-representational explanations for the gap.

Nigeria and South Africa are the natural confound-breakers. They sit within the Northern cluster’s range on both corpus frequency (Figure 2a) and tokens-per-character (Figure 2b). Figure 3 pairs each one with its nearest Northern neighbours by corpus frequency. At matched frequency and comparable fragmentation, the gap persists across all three models. South Africa sits $\sim 16\%$ below its Northern matches and Nigeria $\sim 18\%$ below.

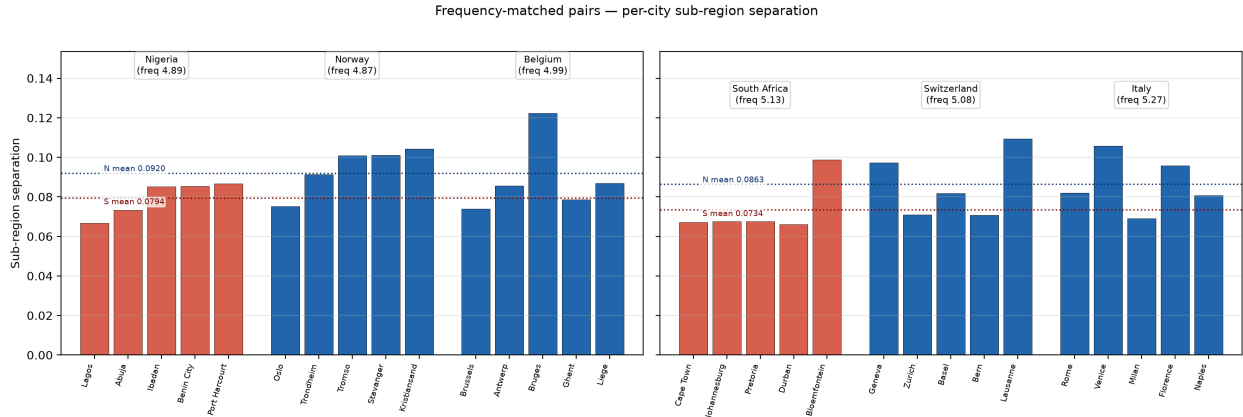


Figure 3: Nigeria and South Africa each plotted against their nearest Northern neighbours by corpus frequency. One bar per city, grouped by country. Country labels include the corpus frequency used to match. Dotted lines mark the Northern and Southern means within each panel.

4.3 Later Layers

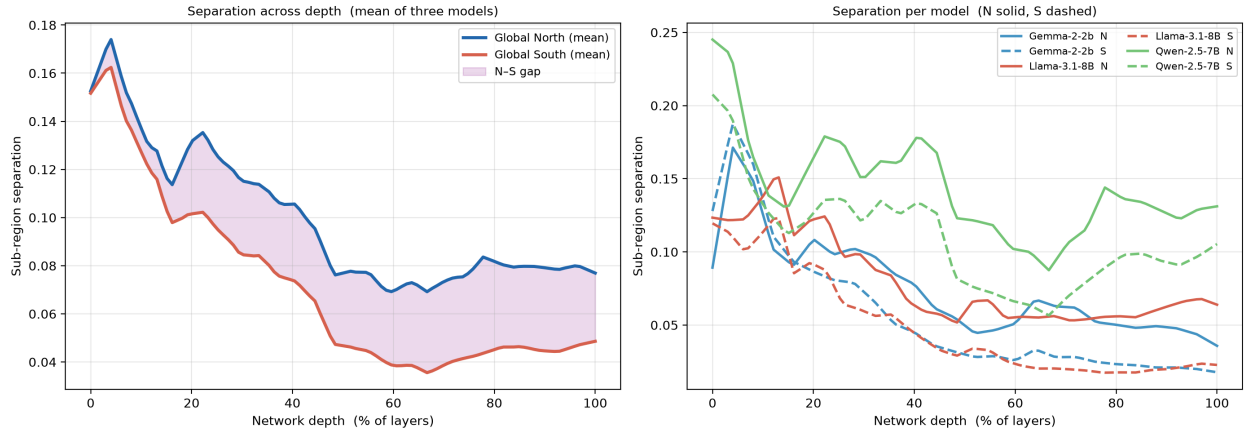


Figure 4: Average N and S separation curves across network depth (left), and the same curves broken out per model with N solid and S dashed (right). Layers are normalised to a 0–100% scale.

Figure 4 tracks how the gap evolves with network depth. The small gap in early layers is expected, since these layers handle surface-level processing. The gap between groups becomes far more substantial as we go deeper into the network. This is the region where persona and identity traits are more likely to live (Chen et al., 2025).

The right panel shows that this is consistent across the three models. The one exception is Gemma, where the Southern group exceeds the Northern in early layers, reflecting what was discussed in Section 4.2.

Figure 5 restricts the average to the last 50% of layers. Across all countries, the Northern mean sits $\sim 77\%$ above the Southern mean, up from the $\sim 39\%$ all-layer gap in Section 4.1. In the layers where identity is likely encoded, the gap nearly doubles.

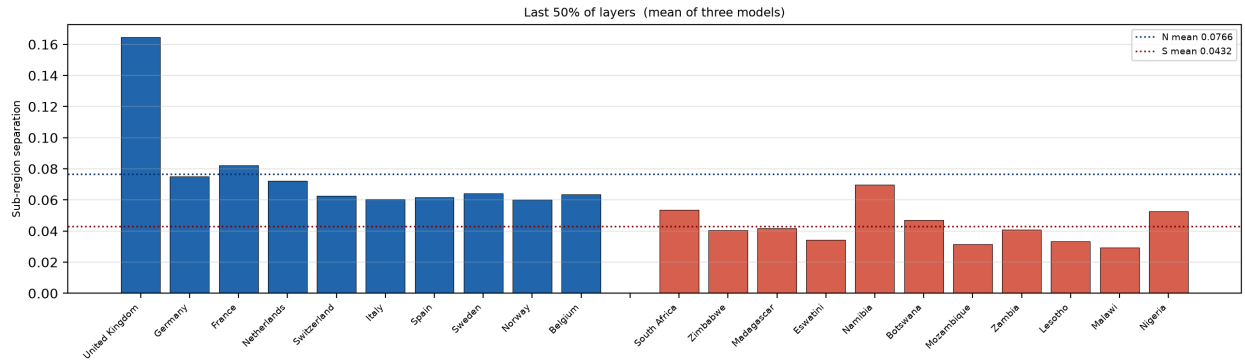


Figure 5: Sub-region separation per country, averaged across the last 50% of layers.

5. Discussion

Representation collapse is biased toward the South. Prior work establishes that LLMs default to Western values at the country level (Tao et al., 2024; Zhao et al., 2026), and that within-country cultural diversity gets collapsed onto a dominant sub-region (Madhusudan et al., 2026). These findings come from model outputs. We show that the collapse is visible in the activations themselves, sub-nationally, and is consistently larger for the South than for the North. The gap persists at matched corpus frequency and comparable tokenization. This extends the *flattening* that Wang et al. (2025) named for US demographics to geography, and pushes it from outputs into the model’s internal state.

Representation-based steering. Khanuja et al. (2026) steer cultural defaults by extracting a per-country prototype direction in activation space. The method works to the extent that the prototype is well-separated from its peers. Our results suggest that for Southern sub-regions, this separation is precisely what is missing. Prototype-minus-others directions will be noisier and less discriminative for the regions that need them most. Their pipeline already filters out countries with fewer than 500 cultural assertions. The steering toolkit could itself be biased toward the North.

Output-level fixes are not enough. Jain et al. (2024) show that fine-tuning typically learns a small wrapper on top of a pretrained capability rather than rewriting it. If sub-regional collapse lives in the pretrained representation, then RLHF, instruction tuning, and cultural prompting are patching the surface. This motivates representation-level interventions over prompt-level ones.

6. Conclusion

We measured how far each country separates from its sub-regions in the activation space of three open-weight LLMs. Northern sub-regions separate $\sim 39\%$ more than Southern sub-regions on average across all layers, and $\sim 77\%$ more in the second half of the network where identity is likely encoded. The gap is not explained by training data volume or by tokenizer fragmentation. Nigeria and South Africa sit within the Northern range on both confounds, and still fall $\sim 16\text{--}18\%$ below their frequency-matched Northern neighbours.

6.1 Limitations and Future Work

Extended hierarchy. We tested one level of the hierarchy, country to city. We did not measure whether countries themselves flatten into continents, nor whether cities further split into ethnolinguistic groups. Extending the hierarchy in both directions is a natural next step.

Tokenization. We only controlled for tokenizer fragmentation using tokens-per-character. The values of the tokens themselves were not controlled for.

Token pooling. We read activations from a single position, the final input token. Pooling across input or answer tokens, or averaging across all tokens, may give a different picture of where regional identity is encoded.

Acknowledgements

We thank Apart Research and AI Safety South Africa for hosting the Global South AI Safety Hackathon, and Dr Claude Formanek for his advice and mentorship throughout the project.

References

- Runjin Chen, Andy Ardit, Henry Sleight, Owain Evans, and Jack Lindsey. Persona vectors: Monitoring and controlling character traits in language models. *arXiv preprint arXiv:2507.21509*, 2025.
- Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, 2021.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert Dick, Hidenori Tanaka, Tim Rocktäschel, Edward Grefenstette, and David Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, volume 2024, pages 9944–10015, 2024.
- Simran Khanuja, Hongbin Liu, Shujian Zhang, John Lambert, Mingqing Chen, Rajiv Mathews, and Lun Wang. Steering llms for culturally localized generation. *arXiv preprint arXiv:2603.23301*, 2026.
- Sangmitra Madhusudan, Trush Shashank More, Steph Buongiorno, Renata Dividino, Jad Kabbara, and Ali Emami. Common to whom? regional cultural commonsense and llm bias in india. *arXiv preprint arXiv:2601.15550*, 2026.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 9802–9822, 2023.
- Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36: 36963–36990, 2023.
- Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346, 2024.
- Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.
- Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601, 2019.
- Thinh Hung Truong, Jey Han Lau, and Jianzhong Qi. Gpsbench: Do large language models understand gps coordinates? *arXiv preprint arXiv:2602.16105*, 2026.

- Angelina Wang, Jamie Morgenstern, and John P Dickerson. Large language models that replace human participants can harmfully misportray and flatten identity groups. *Nature Machine Intelligence*, 7(3):400–411, 2025.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.
- Rong Zhao, Wanqi Liu, Zhizhou Sha, Nanxi Su, Yecheng Zhang, and Ying Long. Culturally uneven urban perception in large language models. 2026.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*, 2023.

A. Question Bank

The 40 questions are organised into eight topical blocks of five questions each. The blocks themselves are not used in any analysis. Every persona is asked all 40 questions in the same order.

- **Food & meals:** What did you have for breakfast this morning? Describe a typical meal with your family. What do you usually buy at the market or grocery store? What do you like to cook when you have time? What's your favourite thing to eat when you're tired after work?
- **Daily routine & movement:** Describe your walk or commute to work. How do you get around your city day to day? What does a normal weekday morning look like for you? What do you do in the hour before you go to sleep? What time do you usually wake up, and what's the first thing you do?
- **Social & relationships:** How do you greet someone you've just met? Who do you spend most of your time with? How do you usually keep in touch with people you care about? What do you do when a neighbour needs help? How do you usually say goodbye at the end of a visit?
- **Leisure & weekends:** How do you usually spend a free weekend? What do you do to relax after a long day? Where do you like to go when you want to be around people? What's a normal Friday night for you? What did you do for fun as a kid?
- **Home & neighbourhood:** Describe your neighbourhood. What does your home look like? What sounds do you hear from your window on a typical morning? What's the nearest place you'd go to sit outside and think? What does your street look like in the evening?
- **Weather & environment:** What's the weather like where you are right now? How does the weather affect your daily routine? What's your favourite time of year, and why? What do you wear on an ordinary day out? How do you deal with extreme heat or cold?
- **Work & everyday tasks:** What kind of work do you do, and how do you feel about it? What's the most stressful part of your average day? How do you handle a task you've been putting off? What's the first thing you do when you sit down to work? How do you decide what to spend money on?
- **Milestones & celebration:** How do you celebrate your birthday? What do you do when something good happens to you? How do you mark the end of a busy week? What's a celebration you look forward to every year? How do you comfort someone who is having a hard time?

B. Region List

Five cities per country, selected from a candidate pool by Wikipedia $\log_{10}$ pageview rank. One city (*Kanye, Botswana*) was excluded from the candidate pool because its tokenization collides with the rapper *Kanye West* and dominates the activation.

Sub-Saharan Africa (11 countries): South Africa (Cape Town, Johannesburg, Pretoria, Durban, Bloemfontein); Zimbabwe (Harare, Bulawayo, Victoria Falls, Mutare, Masvingo); Madagascar (Antananarivo, Nosy Be, Toamasina, Antsiranana, Toliara); Eswatini (Mbabane, Lobamba, Manzini, Nhlanguano, Siteki); Namibia (Windhoek, Walvis Bay, Swakopmund, Grootfontein, Luderitz); Botswana (Gaborone, Maun, Francistown, Kanye, Serowe); Mozambique (Maputo, Beira, Nampula, Pemba, Matola); Zambia (Lusaka, Livingstone, Ndola, Kitwe, Kabwe); Lesotho (Maseru, Butha-Buthe, Mokhotlong, Mafeteng, Qacha's Nek); Malawi (Lilongwe, Zomba, Mzuzu, Blantyre, Kasungu);

Nigeria (Lagos, Abuja, Ibadan, Benin City, Port Harcourt).

Western Europe (10 countries): United Kingdom (London, Birmingham, Manchester, Edinburgh, Liverpool); Germany (Berlin, Munich, Hamburg, Cologne, Frankfurt); France (Paris, Marseille, Nice, Strasbourg, Lyon); Netherlands (Amsterdam, The Hague, Utrecht, Rotterdam, Eindhoven); Switzerland (Geneva, Zurich, Basel, Bern, Lausanne); Italy (Rome, Venice, Milan, Florence, Naples); Spain (Barcelona, Madrid, Seville, Valencia, Malaga); Sweden (Stockholm, Gothenburg, Malmo, Uppsala, Linkoping); Norway (Oslo, Trondheim, Tromso, Stavanger, Kristiansand); Belgium (Brussels, Antwerp, Bruges, Ghent, Liege).

C. Frequency Proxy Details

For each region label, we fetch monthly pageview counts for the corresponding English Wikipedia article from the Wikimedia REST API (`metrics/pageviews/per-article, all-access, user agent, monthly granularity`) over the fixed window 2015-07-01 to 2025-06-30 (120 months). City labels of the form “*Cape Town, South Africa*” are mapped to the article “*Cape_Town*”; country labels are used as-is. Per-region frequency is the mean of these 120 monthly counts, $\log_{10}$ -transformed. Per-country aggregate frequency (used in the frequency scatter) is the mean of the country’s $\log_{10}$ value and the mean $\log_{10}$ across its 5 cities.

Alias handling. Eswatini was renamed from *Swaziland* in 2018, splitting its Wikipedia pageview history across two articles. We sum monthly views from both articles before aggregating. No other region required alias handling.

D. Supplementary Figures

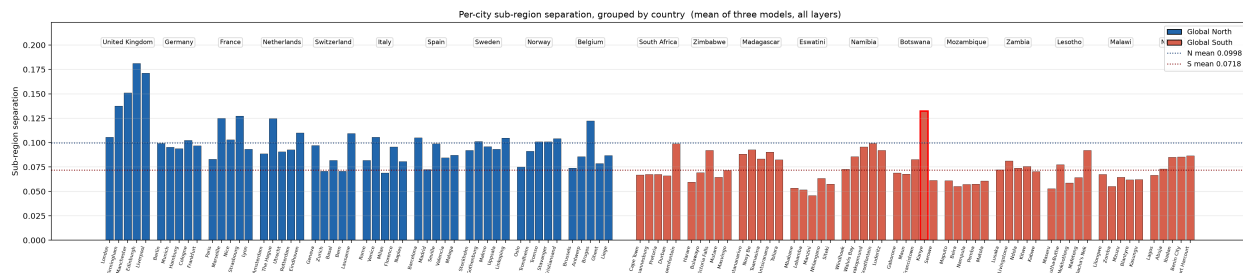


Figure 6: Per-city sub-region separation, grouped by country, mean across the three models, all layers. The tallest bar in the Botswana group is the city *Kanye*. Gemma appears to be representing it as the rapper Kanye West rather than the small town in southern Botswana. Kanye is therefore omitted when computing statistics about Botswana across all figures.

LLM Usage Statement

We used Claude Code to help build the codebase and to brainstorm, edit, and proofread sections of the paper. All results and claims were independently verified by the authors.

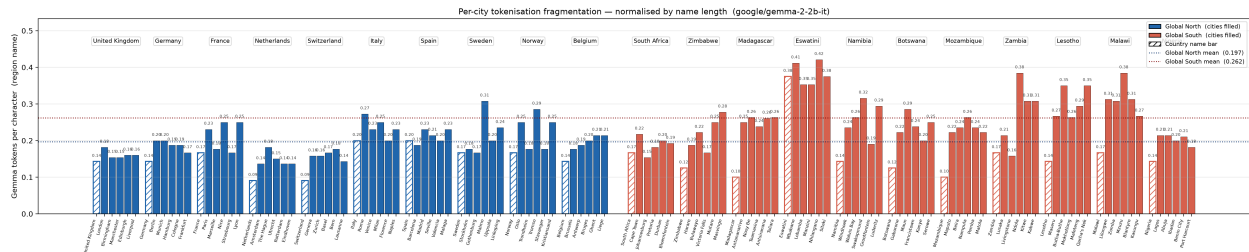


Figure 7: Gemma-2-2b tokens-per-character for every city name (and the country name, hatched), grouped by country in the same ordering as Figure 1. Dotted lines mark the two cluster means.

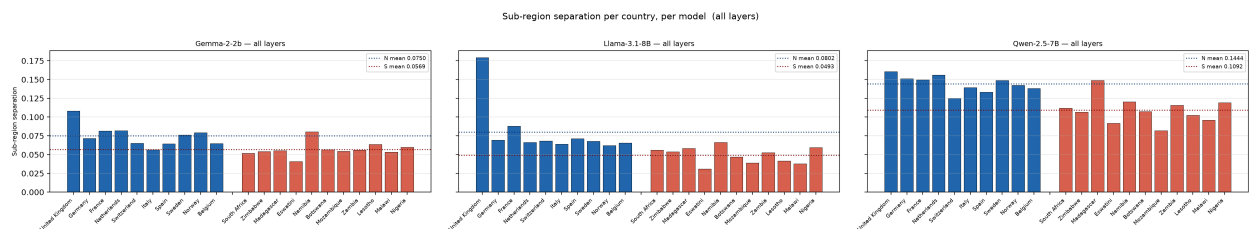


Figure 8: Country-level mean version of Figure 1's data, one panel per model. Sanity check that the qualitative pattern is the same per model as in the cross-model mean.



Figure 9: Figure 5 disaggregated by model.

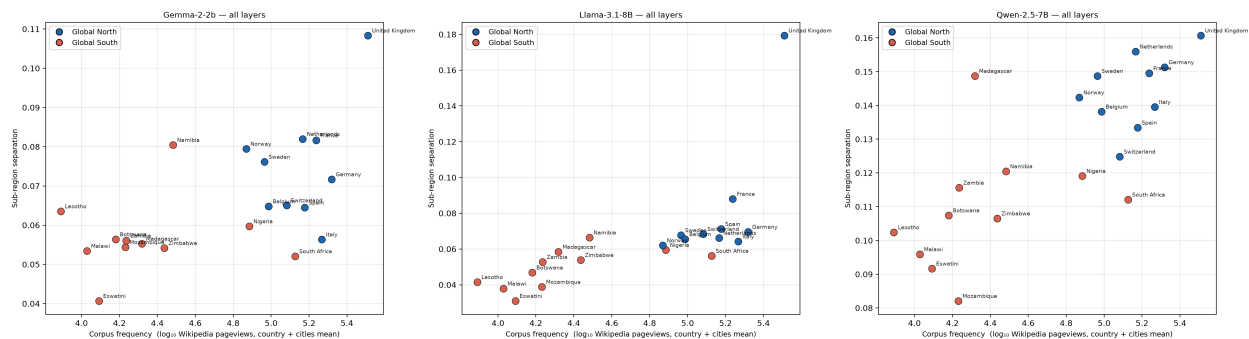


Figure 10: Figure 2a disaggregated by model.