

Language over Content: Tracing Cultural Understanding in Multilingual Large Language Models

Seungho Cho
KAIST
Daejeon, Republic of Korea
cho.seungho@kaist.ac.kr

Changgeon Ko
KAIST
Daejeon, Republic of Korea
pencaty@kaist.ac.kr

Eui Jun Hwang
KAIST
Daejeon, Republic of Korea
ehwa20@kaist.ac.kr

Junmyeong Lee
KAIST
Daejeon, Republic of Korea
david516@kaist.ac.kr

Huije Lee
KAIST
Daejeon, Republic of Korea
huijelee@kaist.ac.kr

Jong C. Park*
KAIST
Daejeon, Republic of Korea
jongpark@kaist.ac.kr

Abstract

Large language models (LLMs) are increasingly used across diverse cultural contexts, making accurate cultural understanding essential. Prior evaluations have mostly focused on output-level performance, obscuring the factors that drive differences in responses, while studies using circuit analysis have covered few languages and rarely focused on culture. In this work, we trace LLMs’ internal cultural understanding mechanisms by measuring activation path overlaps when answering semantically equivalent questions under two conditions: varying the target country while fixing the question language, and varying the question language while fixing the country. We also use same-language country pairs to disentangle language from cultural aspects. Results show that internal paths overlap more for same-language, cross-country questions than for cross-language, same-country questions, indicating strong language-specific patterns. Notably, the South Korea–North Korea pair exhibits low overlap and high variability, showing that linguistic similarity does not guarantee aligned internal representation.

Keywords

Multilingual Large Language Models, Mechanistic Interpretability, Cultural Understanding

1 Introduction

Large language models (LLMs) have achieved strong performance across a wide range of tasks, including translation, reasoning, and question answering [15, 22]. However, when it comes to culture, responses can vary to the speaker, country, or region, since cultural knowledge reflects social norms, historical events, and linguistic nuances that differ across societies and evolve over time [6, 14]. Therefore, it is crucial for LLMs to develop a robust understanding of diverse cultural contexts in order to provide reliable and contextually appropriate outputs [11, 18].

Since LLMs acquire cultural knowledge from diverse language sources during pretraining [5, 25, 27], linguistic and cultural signals are often intertwined, making it necessary to study them together rather than in isolation [6]. Nevertheless, most evaluations of cultural understanding have focused only on final outputs, which obscures the factors driving differences in responses, such as question language, cultural knowledge, or their interaction [8, 14, 20].

*Corresponding author.

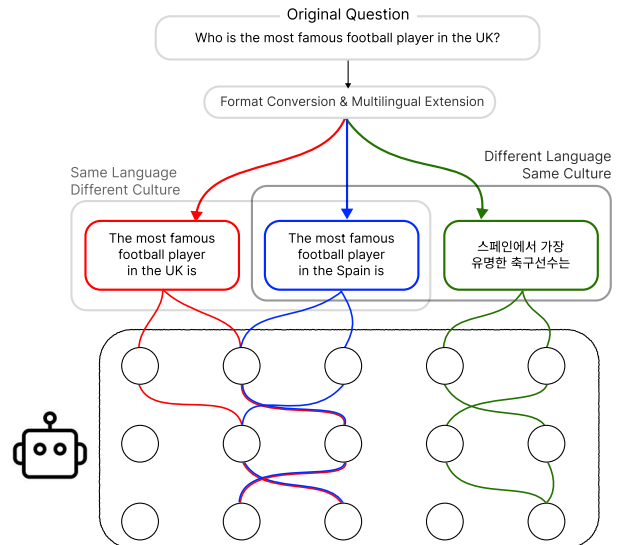


Figure 1: Overview of Tracing Cultural Understanding in Multilingual Large Language Models.

Although some research has attempted to reveal the internal circuits of multilingual LLMs by varying question language [10, 26, 28], little work has examined how models represent and process cultural knowledge within multilingual contexts [16, 26].

In this study, we address these gaps by tracing how LLMs internally use their cultural understandings. Specifically, we measure how models’ internal paths change when answering semantically equivalent cultural questions under two conditions: (1) varying the target country while fixing the question language, and (2) varying the language of the question while fixing the target country. To further disentangle cultural and linguistic aspects, we include special country pairs such as South Korea–North Korea, the US–UK, and Spain–Mexico. These pairs share similar or identical languages but differ culturally, allowing us to better separate language-driven from culture-driven signals. This design enables us to ask whether linguistic cues dominate cultural knowledge representation or whether the two interact in more complex ways.

Our experiments show that internal path overlap is greater for same-language, cross-country questions than for cross-language,

same-country questions, indicating a strong language-specific pattern in representing cultural knowledge. Notably, the South Korea–North Korea pair shows unusually low overlaps and high variability across question languages compared with other same-language pairs, highlighting the need for further analysis. Overall, these findings suggest that multilingual LLMs rely heavily on language-specific circuits when representing and applying cultural knowledge.

Our contributions and findings can be summarized as follows:

- We highlight the need to investigate how multilingual LLMs internally represent cultural understanding.
- Our experiments show that internal paths overlap more when questions are in similar languages across cultures than when they are in different languages for the same culture, indicating a strong language-specific pattern.
- We find cases where models exhibit distinct internal patterns despite high linguistic similarity, highlighting the need for further investigation.

2 Related Work

2.1 Mechanistic Interpretability of LLMs

Recent studies have actively investigated methods to understand the internal mechanisms of LLMs by first identifying interpretable features in their computations and then constructing circuits to capture how these features interact [1, 12, 26, 28]. Some early works directly treated raw neurons as interpretable features [7, 21, 24]; however, the polysemantic nature of neurons made it difficult to derive clear interpretations of model behavior [4]. To address this limitation, subsequent studies trained sparse coding models such as SAE to decompose MLP representations into interpretable features [2, 19], but these approaches lacked input invariance, preventing general conclusions about model behavior. Transcoder overcomes these challenges by decomposing MLP computations and enabling input-invariant, feature-level circuit analysis [1, 3]. This method allows for the extraction of more general interpretable features and the direct computation of feature interactions, facilitating a detailed analysis of internal model flows. Using this approach, we construct circuits to study the internal mechanisms that occur during the LLM’s answer generation process.

2.2 Cultural Understanding of Multilingual LLMs

Recent studies have sought to evaluate the cultural understanding of multilingual LLMs, leading to the development of benchmarks that incorporate locally collected, culture-specific data. While these benchmarks better reflect target cultures, they remain limited in scope, particularly in the number of languages considered and their ability to capture multilingual usage scenarios. Moreover, most prior studies have assessed model performance only at the output level [8, 14, 20], leaving the internal mechanisms underlying cultural understanding largely unexplored. Although recent research has begun to address interpretability, it has primarily focused on multilinguality rather than culture [16, 26, 28], with analyses often restricted to language-specific neurons and narrow language coverage [24]. In this work, we extend this line of research by examining

internal circuits when multilingual LLMs answer culturally relevant questions across more diverse multilingual settings.

3 Tracing Knowledge Circuit

3.1 Task Formulation

We define $Q_{L,C}$ as the set of culture-related questions asked in language L about country C . Correspondingly, $P(Q_{L,C})$ represents the internal activation paths within the LLM when answering the questions in $Q_{L,C}$. We followed the approach in [3] to extract interpretable features (nodes), to measure attributions (edges) and to construct the circuits (internal path).

To analyze the interplay of language and culture in the model’s internal processing, we measure the overlap between activation paths under two cases. First, we fix the language L and compare the similarity between internal paths activated by questions about two different countries C_n and C_m , denoted as $Sim(P(Q_{L,C_n}), P(Q_{L,C_m}))$. Second, we fix the country C and compare internal path similarity for questions asked in two different languages L_n and L_m , denoted as $Sim(P(Q_{L_n,C}), P(Q_{L_m,C}))$. By contrasting these overlaps, we assess whether the model’s cultural knowledge representation is predominately influenced by the input language or the cultural content itself.

The internal paths $P(Q_{L,C})$ are extracted as weighted subgraphs that represent the model’s internal feature attributions during answer generation. Nodes correspond to interpretable features, while edges capture attribution strength between features. We normalize each edge’s weight so that the sum of absolute attributions equals one. We then quantify path similarity using Weighted Jaccard Similarity [23], treating missing edges as having zero weight. Similarity scores close to 1 indicate largely overlapping internal processing paths, whereas scores near 0 indicate distinct mechanisms.

3.2 Data Construction

We construct $Q_{L,C}$ using the culture-specific benchmark dataset BLEnD [14]. From the question set, we randomly select 50 questions, ensuring minimal semantic overlap, to create our experimental dataset.

3.2.1 Country and Language Selection. For cross-cultural analysis, we select seven countries: South Korea (KR), North Korea (KP), the United States (US), the United Kingdom (UK), Spain (ES), Mexico (MX), and China (CN). To study cases where language is shared but cultural contexts vary, we include three pairs of linguistically related countries: South Korea–North Korea, Mexico–Spain, and the United Kingdom–United States. These pairings minimize linguistic variation while emphasizing cultural differences. As a contrasting case, we add China to represent a distinct language group. While languages within each pair are highly similar, subtle distinctions in vocabulary and grammar remain. For this reason, we treat them as separate languages throughout our analysis.

3.2.2 Question Format Conversion. To facilitate next-token prediction and simplify the analysis of the model’s internal representations, we convert interrogative questions into declarative statements. This conversion enables the model to generate answers as continuations within a unified framework, providing inputs that more closely match its training distribution [13]. For example, the

question “Who is the most famous football player in the UK?” can be converted as “The most famous football player in the UK is_”. This process preserves grammatical correctness and natural word order in each language, inserting spaces where required for accurate token generation (except for Chinese, which does not use spaces).

3.2.3 Multilingual Extension. The original BLEnD dataset provides each cultural question set only in its corresponding language. To extend coverage, we translate each question set into all languages used in our experiments, creating 49 Q_{LC} question sets. For example, a question about the most famous football player in South Korea is expressed not only in Korean, but also in English, Chinese, and Spanish. This design enables systematic analysis of how language inputs and culture jointly influence the model’s internal path selection.

3.3 Implementation Details

For questions related to cultural knowledge, we used the Gemma 2¹ [17]. To avoid potential effects on the model’s internal paths, we employed the base version rather than the instruction-tuned model. For internal path extraction, we used Gemma Scope Transcoder² [3, 9] to obtain interpretable features. In the dataset reformulation and extension steps, the questions were generated with GPT-4o and verified with o4-mini via the OpenAI API³.

4 Analysis

4.1 Main Result

Table 1 presents similarity scores of internal path overlaps in the LLM under two conditions: fixed language and fixed culture. The results clearly show that the model’s internal path selection is much more affected by the language of the question than by the cultural context. When the question language is fixed (Table 1a), path overlap remains relatively high across different target cultures, especially among linguistically similar country pairs such as South Korea–North Korea, the United States–the United Kingdom, and Spain–Mexico. This suggests that language similarity strongly encourages reuse of internal paths.

Conversely, when the cultural context is fixed and the question language varies (Table 1b), path overlap drops significantly. This indicates that even semantically equivalent queries in different languages prompt the model to use markedly different internal paths. It implies that the model organizes and accesses cultural knowledge in a language-dependent manner, prioritizing linguistic form over semantic content when processing multilingual queries.

Figure 2 summarizes these results by averaging scores for each pair, shown as descending bar charts with 95% confidence intervals. Hatched bars highlight country pairs sharing similar languages, and the orange line marks the overall average. Figure 2a (fixed question language) highlights that linguistic similarity supports greater path overlap—possibly reflecting overlapping cultural traits as well. Figure 2b (fixed culture) shows low overlaps when question languages differ, supporting the conclusion that question language

Table 1: (a) Path overlap across target cultures with the question language fixed. (b) Path overlap across question languages with the target culture fixed.

		(a) Fixed Language						
		L						
C ₁	C ₂	KR	KP	US	UK	ES	MX	CN
KR	KP	0.10	0.04	0.44	0.46	0.57	0.56	0.52
KR	US	0.05	0.04	0.37	0.39	0.16	0.16	0.21
KR	UK	0.06	0.04	0.37	0.38	0.40	0.42	0.45
KR	ES	0.31	0.04	0.17	0.17	0.15	0.15	0.44
KR	MX	0.30	0.38	0.17	0.17	0.16	0.15	0.43
KR	CN	0.06	0.04	0.17	0.17	0.16	0.16	0.16
KP	US	0.35	0.35	0.31	0.31	0.15	0.15	0.19
KP	UK	0.36	0.36	0.30	0.30	0.36	0.37	0.42
KP	ES	0.04	0.04	0.16	0.16	0.14	0.14	0.43
KP	MX	0.04	0.04	0.15	0.16	0.14	0.14	0.44
KP	CN	0.38	0.39	0.16	0.16	0.15	0.15	0.16
US	UK	0.59	0.58	0.56	0.56	0.27	0.27	0.27
US	ES	0.05	0.04	0.19	0.19	0.17	0.16	0.20
US	MX	0.04	0.05	0.19	0.19	0.17	0.16	0.21
US	CN	0.50	0.48	0.19	0.19	0.18	0.17	0.39
UK	ES	0.05	0.04	0.19	0.18	0.16	0.16	0.50
UK	MX	0.04	0.04	0.18	0.18	0.16	0.16	0.47
UK	CN	0.51	0.51	0.19	0.18	0.16	0.16	0.17
ES	MX	0.41	0.05	0.54	0.53	0.60	0.59	0.56
ES	CN	0.05	0.04	0.43	0.44	0.49	0.51	0.14
MX	CN	0.04	0.05	0.44	0.43	0.51	0.52	0.15

		(b) Fixed Culture						
		C						
L ₁	L ₂	KR	KP	US	UK	ES	MX	CN
KR	KP	0.16	0.38	0.36	0.37	0.06	0.15	0.38
KR	US	0.01	0.01	0.02	0.02	0.01	0.01	0.01
KR	UK	0.01	0.01	0.02	0.02	0.01	0.01	0.01
KR	ES	0.01	0.01	0.01	0.02	0.01	0.01	0.02
KR	MX	0.01	0.01	0.01	0.02	0.01	0.01	0.02
KR	CN	0.02	0.03	0.02	0.03	0.02	0.02	0.02
KP	US	0.01	0.01	0.02	0.02	0.01	0.01	0.02
KP	UK	0.01	0.01	0.02	0.02	0.01	0.01	0.02
KP	ES	0.01	0.01	0.01	0.02	0.01	0.01	0.02
KP	MX	0.01	0.01	0.01	0.02	0.01	0.01	0.02
KP	CN	0.02	0.03	0.02	0.02	0.02	0.02	0.02
US	UK	0.91	0.94	0.95	0.93	0.97	0.95	0.89
US	ES	0.02	0.02	0.02	0.02	0.03	0.03	0.03
US	MX	0.02	0.02	0.02	0.02	0.02	0.03	0.03
US	CN	0.02	0.02	0.02	0.02	0.02	0.02	0.02
UK	ES	0.02	0.02	0.02	0.02	0.03	0.03	0.03
UK	MX	0.02	0.02	0.02	0.02	0.02	0.03	0.03
UK	CN	0.02	0.02	0.02	0.02	0.02	0.02	0.02
ES	MX	0.66	0.70	0.66	0.67	0.69	0.66	0.69
ES	CN	0.01	0.02	0.01	0.01	0.02	0.02	0.01
MX	CN	0.01	0.01	0.01	0.01	0.02	0.02	0.01

¹<https://huggingface.co/google/gemma-2-2b>

²<https://huggingface.co/google/gemma-scope-2b-pt-transcoders>

³<https://platform.openai.com/>

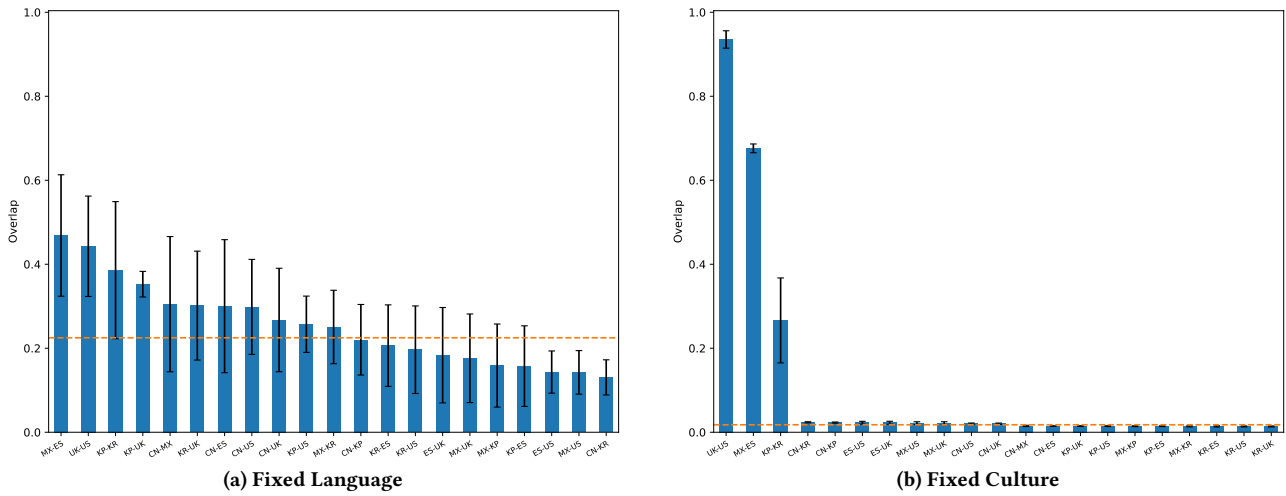


Figure 2: (a) Path overlap by country pair when the *question language is fixed*; We find that overlaps remain relatively high, with linguistically similar country pairs showing especially high reuse of internal paths. (b) Path overlap by language pair when the *target culture is fixed*; We find that overlaps drop markedly when the query language changes, indicating that language (rather than meaning) dominates internal path selection. Each bar shows the mean ($\pm 95\%$ CI), sorted in descending order; hatched bars denote linguistically similar pairs, and the orange horizontal line marks the overall average.

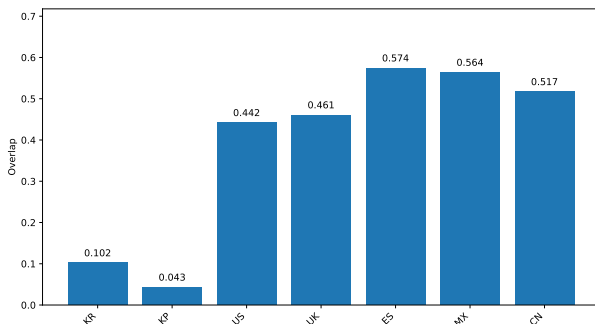


Figure 3: Path overlap between questions on South and North Korean culture by question language. We find that path overlaps are low in Korean languages than in non-Korean languages.

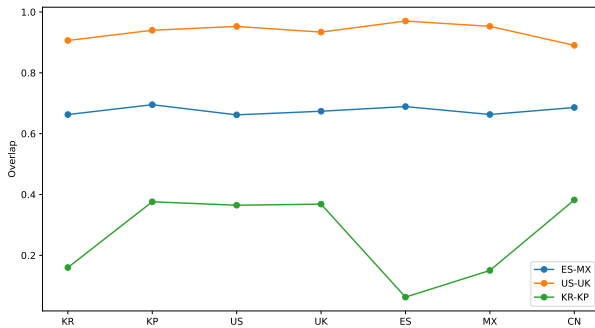


Figure 4: Path overlap between questions for similar-language pairs under a fixed target culture. US-UK and Spain-Mexico show high and stable overlap while South Korea-North Korea shows lower and more variable overlap.

dominates internal path selection more than cultural context or semantic equivalence.

4.2 Distinct Path Patterns in South and North Korea

South and North Korea, despite sharing a similar language, exhibited distinct patterns compared with other linguistically similar pairs. Figure 3 visualizes path overlaps for questions about South and North Korean culture across the various question languages. The results show higher overlaps in non-Korean languages and lower overlaps in Korean. Figure 4 presents overlaps across similar-language pairs when the target culture is fixed. US-UK and Spain-Mexico maintained high and stable overlaps, while South Korea-North Korea showed lower and more variable overlaps. The reasons for these differences, whether due to linguistic, cultural, or other factors, remain unclear and warrant further investigation.

5 Conclusion

We explored the internal mechanisms of cultural understanding in multilingual LLMs. To reflect realistic scenarios, we extended the cultural dataset to include multiple languages and measured the overlap of activated internal paths. Results show that query language affects internal path selection more strongly than target culture, and that cultural understanding is mainly stored in language-dependent paths. We also observed that politically unique contexts, such as South and North Korea, are reflected in the model’s internal mechanisms. These findings offer key insights into how multilingual LLMs understand and utilize cultural knowledge.

Our analysis focused on internal path overlap, but interventions or circuit patching could clarify which features drive cultural knowledge processing and how language- and culture-related features interact. Some country pairs, such as South and North Korea, show distinct patterns compared to other similar language pairs, yet the reasons remain unclear. Future work could further subdivide cultural knowledge into finer categories and include more languages, enabling a more comprehensive investigation of language-culture interactions within the model.

References

- [1] Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. Circuit Tracing: Revealing Computational Graphs in Language Models. *Transformer Circuits Thread* (2025). <https://transformer-circuits.pub/2025/attribution-graphs/methods.html>
- [2] Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread* (2023). <https://transformer-circuits.pub/2023/monosemantic-features/index.html>
- [3] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. 2024. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=J6zHcScAo0>
- [4] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy Models of Superposition. *Transformer Circuits Thread* (2022). https://transformer-circuits.pub/2022/toy_model/index.html
- [5] Omer Goldman, Uri Shaham, Dan Malkin, Sivan Eiger, Avinat Hassidim, Yossi Matias, Joshua Maynez, Adi Mayrav Gilady, Jason Riesa, Shruti Rijhwani, Laura Rimell, Idan Szepke, Reut Tsarfay, and Matan Eyal. 2025. ECLeTic: a Novel Challenge Set for Evaluation of Cross-Lingual Knowledge Transfer. arXiv:2502.21228 [cs.CL] <https://arxiv.org/abs/2502.21228>
- [6] Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and Strategies in Cross-Cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 6997–7013. doi:10.18653/v1/2022.acl-long.482
- [7] Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. 2024. On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Kevin Duh, Helena Gomez, and Steven Bethard (Eds.). Association for Computational Linguistics, Mexico City, Mexico, 6919–6971. doi:10.18653/v1/2024.naacl-long.384
- [8] Huihan Li, Liwei Jiang, Nouha Dziri, Xiang Ren, and Yejin Choi. 2024. CULTURE-GEN: Revealing Global Cultural Perception in Language Models through Natural Language Prompting. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=DbsLm2KAqP>
- [9] Tom Lieberum, Senthooan Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, Janos Kramar, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, Yonatan Belinkov, Najoung Kim, Jaap Jumelet, Hosein Mohebbi, Aaron Mueller, and Hanjie Chen (Eds.). Association for Computational Linguistics, Miami, Florida, US, 278–300. doi:10.18653/v1/2024.blackboxnlp-1.19
- [10] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermy, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. 2025. On the Biology of a Large Language Model. *Transformer Circuits Thread* (2025). <https://transformer-circuits.pub/2025/attribution-graphs/biology.html>
- [11] Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025. Culturally Aware and Adapted NLP: A Taxonomy and a Survey of the State of the Art. *Transactions of the Association for Computational Linguistics* 13 (2025), 652–689. doi:10.1162/tacl_a_00760
- [12] Joseph Miller, Bilal Chughtai, and William Saunders. 2024. Transformer Circuit Evaluation Metrics Are Not Robust. In *First Conference on Language Modeling*. <https://openreview.net/forum?id=zSf8PJyQb2>
- [13] Eric Mitchell, Joseph Noh, Siyan Li, Will Armstrong, Ananth Agarwal, Patrick Liu, Chelsea Finn, and Christopher Manning. 2022. Enhancing Self-Consistency and Performance of Pre-Trained Language Models through Natural Language Inference. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 1754–1768. doi:10.18653/v1/2022.emnlp-main.115
- [14] Junho Myung, Nayeon Lee, Yi Zhou, Jiho Jin, Rifki Putri, Dimosthenis Antypas, Hsuvas Borkakoty, Eunsu Kim, Carla Perez-Almendros, Abinew Ali Ayele, and et al. 2024. Blend: A benchmark for llms on everyday knowledge in diverse cultures and languages. *Advances in Neural Information Processing Systems* 37 (2024), 78104–78146.
- [15] OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, Mark Chen, Enoch Cheung, Aidan Clark, Dan Cook, Marat Dukhan, Casey Dvorak, Kevin Fives, Vlad Fomenko, Timur Garipov, Kristian Georgiev, Mia Glaese, Tarun Gogineni, Adam Goucher, Lukas Gross, Katia Gil Guzman, John Hallman, Jackie Hehir, Johannes Heidecke, Alec Helyar, Haitang Hu, Romain Huet, Jacob Huh, Saachi Jain, Zach Johnson, Chris Koch, Irina Kofman, Dominik Kundel, Jason Kwon, Volodymyr Kyrlyuk, Elaine Ya Le, Guillaume Leclerc, James Park Lennon, Scott Lessans, Mario Lezcano-Casado, Yuanzhi Li, Zhuohan Li, Ji Lin, Jordan Liss, Lily, Liu, Jiancheng Liu, Kevin Lu, Chris Lu, Zoran Martinovic, Lindsay McCallum, Josh McGrath, Scott McKinney, Aidan McLaughlin, Song Mei, Steve Mostovoy, Tong Mu, Gideon Myles, Alexander Neitz, Alex Nichol, Jakub Pachocki, Alex Paino, Dana Palmie, Ashley Pantuliano, Giambattista Parascandolo, Jongsoo Park, Leher Pathak, Carolina Paz, Ludovic Peran, Dmitry Pimenov, Michelle Pokrass, Elizabeth Proehl, Huida Qiu, Gaby Raila, Filippo Raso, Hongyu Ren, Kimmy Richardson, David Robinson, Bob Rotsted, Hadi Salman, Suvansh Sanjeev, Max Schwarzer, D. Sculley, Harshit Sikchi, Kendal Simon, Karan Singhal, Yang Song, Dane Stuckey, Zhiqing Sun, Philippe Tillet, Sam Toizer, Foivos Tsimpouras, Nikhil Vyas, Eric Wallace, Xin Wang, Miles Wang, Olivia Watkins, Kevin Weil, Amy Wendling, Kevin Whinnery, Cedric Whitney, Hannah Wong, Lin Yang, Yu Yang, Michihiro Yasunaga, Kristen Ying, Wojciech Zaremba, Wenting Zhan, Cyril Zhang, Brian Zhang, Eddie Zhang, and Shengjia Zhao. 2025. gpt-oss-120b & gpt-oss-20b Model Card. arXiv:2508.10925 [cs.CL] <https://arxiv.org/abs/2508.10925>
- [16] Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and Interpretability of Multilingual Large Language Models: A Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Suzhou, China. <https://openreview.net/forum?id=KQjVhM2YhN> Accepted for publication.
- [17] Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozinska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshov, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucinska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Milligan, Keelin McDonnell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjosund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, and Lilly McNealus. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *CoRR* abs/2408.00118 (2024). <https://doi.org/10.48550/arXiv.2408.00118>
- [18] Alireza Salemi, Sheshera Mysore, Michael Bendersky, and Hamed Zamani. 2023. LaMP: When Large Language Models Meet Personalization. arXiv:2304.11406 [cs.CL]
- [19] Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. 2025. A Survey on Sparse Autoencoders: Interpreting the Internal Mechanisms of Large Language Models. *CoRR* abs/2503.05613 (March 2025). <https://doi.org/10.48550/arXiv.2503.05613>
- [20] Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adeniji, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 18761–18799.

- doi:10.18653/v1/2025.acl-long.919
- [21] Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 5701–5715. doi:10.18653/v1/2024.acl-long.309
- [22] Kimi Team, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijie Chen, Yanru Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, Angang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao, Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu, Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu, Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tianwei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi, Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang, Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang, Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu, Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang, Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou, Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. 2025. Kimi K2: Open Agentic Intelligence. arXiv:2507.20534 [cs.LG]. <https://arxiv.org/abs/2507.20534>
- [23] Wikipedia. 2025. Jaccard index — Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Jaccard%20index&oldid=1292934854>. [Online; accessed 14-September-2025].
- [24] Jiahao Ying, Wei Tang, Yiran Zhao, Yixin Cao, Yu Rong, and Wenxuan Zhang. 2025. Disentangling Language and Culture for Evaluating Multilingual Large Language Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 22230–22251. doi:10.18653/v1/2025.acl-long.1082
- [25] Chen Zhang, Zhiyuan Liao, and Yansong Feng. 2025. Cross-Lingual Transfer of Cultural Knowledge: An Asymmetric Phenomenon. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (Eds.). Association for Computational Linguistics, Vienna, Austria, 147–157. doi:10.18653/v1/2025.acl-short.13
- [26] Ruochen Zhang, Qinan Yu, Matianyu Zang, Carsten Eickhoff, and Ellie Pavlick. 2025. The Same but Different: Structural Similarities and Differences in Multilingual Language Modeling. In *The Thirteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=NCrFA7dq8T>
- [27] Xin Zhao, Naoki Yoshinaga, and Daisuke Oba. 2024. Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, Yvette Graham and Matthew Purver (Eds.). Association for Computational Linguistics, St. Julian's, Malta, 2088–2102. doi:10.18653/v1/2024.eacl-long.127
- [28] Yiran Zhao, Wenxuan Zhang, Guizhen Chen, Kenji Kawaguchi, and Lidong Bing. 2024. How do Large Language Models Handle Multilingualism?. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=ctXYOoAgRy>