

Exploring Multilingual Probing in Large Language Models: A Cross-Language Analysis

Daoyang Li^{1,2}, Haiyan Zhao², Qingcheng Zeng³, Mengnan Du²

¹University of Southern California, ²New Jersey Institute of Technology,

³Northwestern University

daoyangl@usc.edu, mengnan.du@njit.edu

Abstract

Probing techniques for large language models (LLMs) have primarily focused on English, overlooking the vast majority of other world’s languages. In this paper, we extend these probing methods to a multilingual context, investigating how LLMs encode linguistic structures across diverse languages. We conduct experiments on several open-source LLM models, analyzing probing accuracy, trends across layers, and similarities between probing vectors for multiple languages. Our key findings reveal: (1) a consistent performance gap between high-resource and low-resource languages, with high-resource languages achieving significantly higher probing accuracy; (2) divergent layer-wise accuracy trends, where high-resource languages show substantial improvement in deeper layers similar to English; and (3) higher representational similarities among high-resource languages, with low-resource languages demonstrating lower similarities both among themselves and with high-resource languages. These results provide insights into how linguistic structures are represented differently across languages in LLMs and emphasize the need for improved structure modeling for low-resource languages.

1 Introduction

Large language models (LLMs), such as GPT-4 (Achiam et al., 2023), Claude 3.5 (Anthropic, 2024), Llama 3 (Dubey et al., 2024), have demonstrated remarkable progress across a wide range of natural language processing tasks. As these models continue to advance, there is a growing need to understand their internal mechanisms and how they represent linguistic structures. Probing techniques have emerged as a valuable tool for investigating how LLMs encode and process structural information, offering insights into their decision-making processes and the nature of their learned representations (Ferrando et al., 2024a; Zhao et al., 2024; Zou et al., 2023).

However, a significant gap exists in our understanding of how LLMs represent linguistic structures across different languages. While extensive probing research has been conducted on English language representations, there are approximately 7,000 languages spoken worldwide, many of which remain understudied in the context of LLMs. This lack of multilingual analysis limits our understanding of how LLMs encode structural information across diverse linguistic contexts, particularly for low-resource languages that are often underrepresented in model training data and evaluations.

To address this research gap, we propose a multilingual probing approach to investigate the structural representation capabilities of LLMs across a diverse set of 16 languages, including both high-resource and low-resource languages. Our study extends probing techniques from English to a multilingual context, examining how LLMs encode linguistic structures in factual knowledge and sentiment classification tasks across different languages. Our key findings reveal that: (1) high-resource languages consistently achieve higher probing accuracy compared to low-resource languages; (2) high-resource languages exhibit similar trends to English across model layers, with accuracy improving significantly in deeper layers, while low-resource languages show relatively stable or only slightly improving accuracy; and (3) there are high similarities between probing vectors of high-resource languages, whereas low-resource languages demonstrate lower similarities both among themselves and with high-resource languages.

2 Probing Method

2.1 LLM Internal Representation

We study decoder-only LLMs, where each layer of a model consists of both multi-head attention blocks (MHA) and feed-forward networks (FFNs). In this work, we utilize frozen pretrained language

models. Layers are indexed with $\ell \in L$, where L denotes the set of all layers in a model. For each layer, the computation starts and ends with a residual stream. The MHA first reads from the residual stream and performs computation, then adds its output back to the residual stream. The updated vector in the residual stream is then passed through MLPs to generate the output of the layer:

$$h_i^{\ell+1} = h_i^\ell + \text{MLP}^\ell \left(h_i^\ell + \text{Att}^\ell \left(h_i^\ell \right) \right), \quad (1)$$

where h_i^ℓ represents the hidden state of the i -th token in the input sequence at layer ℓ . We focus on the output representation space of each layer, particularly the residual stream at the end of each layer. We use the representation of the last token to represent the entire input sequence, as it is generally believed to integrate information from all previous tokens. This representation is denoted as h^ℓ , which will be simplified to h in the following section.

2.2 Linear Classifier Probing

In our analysis, we employed linear classifier probing (Ju et al., 2024a; Jin et al., 2024) to explore internal representations across various layers of LLMs. We extracted hidden states from the residual stream of each layer using two types of inputs (i.e., positive and negative) and utilized these representations to train a logistic regression model. By evaluating the performance of trained classifier, we are able to assess how well the hidden states at different layers encoded information relevant to answering factual questions or handling sentiment classification tasks. This approach provides valuable insights into the nature of the representations learned within the model.

To perform the probing, we employed a linear classifier approach. We define $h \in R^{n \times d_{\text{model}}}$ as the set of hidden features extracted from the LLM, where n is the number of samples and d_{model} represents the dimensionality of the hidden layer. The internal representation of each sample in a specific layer is denoted by $h^{(i)} \in R^{1 \times d_{\text{model}}}$. We utilize binary classification, assigning labels $y^{(i)} \in \{0, 1\}$. The objective function for our logistic regression classifier, incorporating L2 regularization, is formulated as:

$$J(\theta) = -\frac{1}{n} \sum_{i=1}^n L(h^{(i)}, y^{(i)}; \theta) + \frac{\lambda}{2n} \|\theta\|_2^2, \quad (2)$$

where $L(\cdot)$ represents the cross-entropy loss:

$$L = y^{(i)} \log(\sigma(\theta^T h^{(i)})) + (1 - y^{(i)}) \log(1 - \sigma(\theta^T h^{(i)})), \quad (3)$$

Model	Layer	Representation Dimension
Qwen-0.5B	24	1024
Qwen-1.8B	24	2048
Qwen-7B	32	4096
Gemma-2B	18	2048
Gemma-7B	28	3072

Table 1: Model and corresponding layers.

where θ denotes the model parameters, λ is the regularization coefficient, and $\sigma(\cdot)$ represents the sigmoid activation function. By evaluating the accuracy of this classifier on the test set, we can evaluate the LLM’s performance and gain insights into its internal representations across different languages and layers.

3 Experiment

In this section, we conduct comprehensive probing experiments to investigate how language models process different languages. Our analysis focuses on two key aspects: comparing accuracy across different model layers and examining correlations between probing vectors of various languages. Through these experiments, we seek to answer three fundamental research questions (RQs):

- **RQ1** - Do other languages achieve comparable probing accuracy to English?
- **RQ2** - Do other languages exhibit similar layer-wise behavioral patterns to English?
- **RQ3** - What are the similarities between probing vectors across different languages?

3.1 Experiment Settings

In this section, we introduce the overall experimental settings of this paper.

Models: We evaluated the performance and internal representations across various languages using two open-source LLM families: Qwen (Bai et al., 2023) and Gemma (Team et al., 2024). The Qwen-1.5 architecture comprises 24 layers for smaller variants (0.5B & 1.8B) and 32 layers for the larger variant (7B), while the Gemma architecture features 18 layers for the 2B model and 28 layers for the 7B model. The representation vector dimensions vary across models: 1024 for Qwen-0.5B, 2048 for both Qwen-1.8B and Gemma-2B, 3072 for Gemma-7B, and 4096 for Qwen-7B. Table 1 provides a comprehensive overview of the models and their corresponding layer configurations.

Datasets: In the following experiments, we utilized a truthful dataset: *Cities* (Marks and Tegmark,

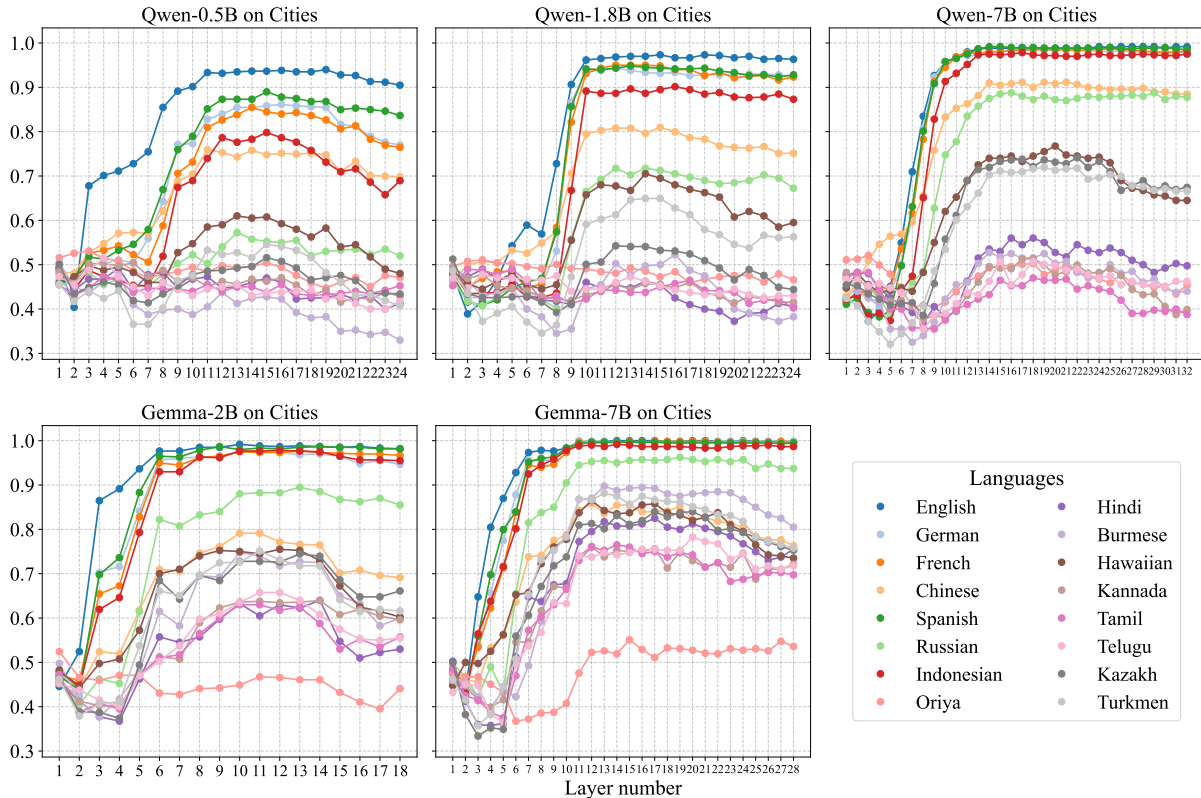


Figure 1: Layer-wise probing accuracy of 5 open-source LLMs across 16 languages.

2023), and a sentiment dataset: *Opinion* (Tatman, 2017). Cities contains 1496 samples, and Opinion contains 1000 samples.

- *Cities* (Marks and Tegmark, 2023): consists of statements about the location of cities from worldwide and their veracity labels (e.g., The city of Lyon is in France, which is true).
- *Opinion* (Tatman, 2017): consists of opinions of 20 famous hotels. It contains the hotel’s name, opinion’s polarity, and its source.

Our dataset encompasses 16 languages: English, German, French, Chinese, Spanish, Russian, Indonesian, Oriya, Hindi, Burmese, Hawaiian, Kannada, Tamil, Telugu, Kazakh, Turkmen. We categorized English, German, French, Chinese, Spanish, Russian, and Indonesian as high-resource languages, and rest of them as low-resource languages based on the volume of available digital content and linguistic resources. The original language of our two datasets are English, and we used Google Translate within deep-translator python library (Azam, 2024) to translate them into other 15 languages, as Google Translate supports translation between over 100 languages, and achieves high accuracy compared to other translation tools.

Implementation Details: To evaluate the performance of LLMs on each language, we use the template for the Cities dataset in English as "*Judge the statement is Positive or Negative. <Statement>*". The prompts of other languages utilize the same template translated by Google Translate. This allows us to prevent any context differences regarding the prompt design. We present the full set of prompt templates for all 16 languages in Figure 3 at the Appendix. We applied probing techniques to assess the information encoded within each layer of these models. For our probing analysis, we selected linear classifier probing for our experiments. Each dataset is divided into a training and a test set with an 8:2 ratio, and we adhered to the standard procedure for probing classifiers in LLMs, extracting feature representations from the final hidden states at each layer of the LLMs to serve as input to the probing classifier. The linear weight parameter θ of the logistic regression classifier is regarded as the probing vector for each language and layer.

3.2 Multilingual Accuracy

In this section, we explored (1) whether other languages besides English have the same probing accuracy as English and (2) whether they follow the

Table 2: Probing accuracy of various LLMs across different languages on the Cities dataset.

Model	High-Resource Languages							Low-Resource Languages								
	English	German	French	Chinese	Spanish	Russian	Indonesian	Oriya	Hindi	Burmese	Hawaiian	Kannada	Tamil	Telugu	Kazakh	Turkmen
Gemma-2B	0.98	0.95	0.97	0.69	0.98	0.87	0.95	0.44	0.53	0.60	0.60	0.60	0.56	0.56	0.66	0.62
Gemma-7B	0.99	0.99	0.99	0.76	0.99	0.93	0.99	0.54	0.76	0.81	0.74	0.72	0.70	0.72	0.75	0.76
Qwen-0.5B	0.90	0.77	0.76	0.70	0.84	0.52	0.69	0.47	0.41	0.33	0.48	0.43	0.45	0.42	0.43	0.41
Qwen-1.8B	0.96	0.92	0.92	0.75	0.93	0.67	0.87	0.47	0.41	0.37	0.60	0.42	0.40	0.43	0.44	0.56
Qwen-7B	0.99	0.98	0.98	0.88	0.98	0.88	0.97	0.45	0.50	0.44	0.65	0.40	0.39	0.46	0.67	0.67

same trend as English in different layers.

We present results on multilingual accuracy across our five evaluated models (Qwen-0.5B, Qwen-1.8B, Qwen-7B, Gemma-2B, Gemma-7B) on the cities and Opinion datasets. In Figure 1 and Table 2, we show the results of layer-wise probing accuracy on the Cities dataset. The results of Opinion dataset are included in Figure 4 in the Appendix. These results visualize how probing accuracy changes across model layers for all 16 languages. Based on these results, our analysis lead to two general observations as follows:

- *High-resource languages show higher accuracy, while low-resource languages have comparatively lower accuracy.* We conducted experiments using Cities and Opinion datasets, exploring the binary classification problem in 16 selected languages. Table 2 shows that in Cities dataset, high-resource languages such as French and German achieve at least 70% accuracy, even reaching over 90% accuracy for some models, while low-resource languages like Oriya and Hindi only achieve about 40% accuracy in the final layer.
- *High-resource languages follow similar trends to English, where accuracy significantly improves as the layers deepen.* Low-resource languages maintain relatively stable probing accuracy or show only slight improvements. Figure 1 shows that as model layers go deeper, English, French, and other high-resource languages could reach highest accuracy at the 11th layer. However, the probing accuracies of the low-resource languages have not improved significantly.

3.3 Similarity Correlation of Probing Vectors

In this section, we conducted similarity analysis on probing vectors θ across languages using two visualization approaches:

- **Correlation Heatmaps:** These visualize the pairwise similarities between probing vectors

of all 16 languages. These highlight clustering patterns and resource-level disparities.

- **Layer-wise Similarity Plots:** They measure cosine similarity between each language’s probing vector and English’s across model layers, revealing representation dynamics.

For demonstration, Figure 2 shows results from the Qwen-1.8B model and Opinion dataset. In the Appendix, we extend this analysis to all five models (Qwen-0.5B, Qwen-1.8B, Qwen-7B, Gemma-2B, Gemma-7B) and both datasets (Opinion, Cities) through Figure 5 and Figure 6. Our analysis reveals the following three key patterns:

- The probing vectors of high-resource languages (English, German, French, Chinese, Spanish, Russian) demonstrate strong correlations with each other, as evidenced by the darker clusters in the heatmaps and consistently higher similarity curves in the trajectory plots. For instance, in the Qwen-1.8B Opinion task, German and French probing vectors maintain correlations above 0.6 with English across most layers. In contrast, low-resource languages show notably weaker correlations, both among themselves and with high-resource languages. This pattern is visible in the bright regions of the heatmaps for languages like Tamil, Telugu, and Oriya, with similarity scores typically remaining below 0.3 across all layers.
- The evolution of similarities across model layers reveals further insights into these representational differences. High-resource languages exhibit dynamic similarity patterns with English, often peaking in middle layers before slightly decreasing, while low-resource languages maintain relatively stable, low similarity levels throughout the model layers. These patterns persist across different model sizes and architectures in both the Qwen and Gemma families, and remain consistent across the Opinion and Cities datasets.

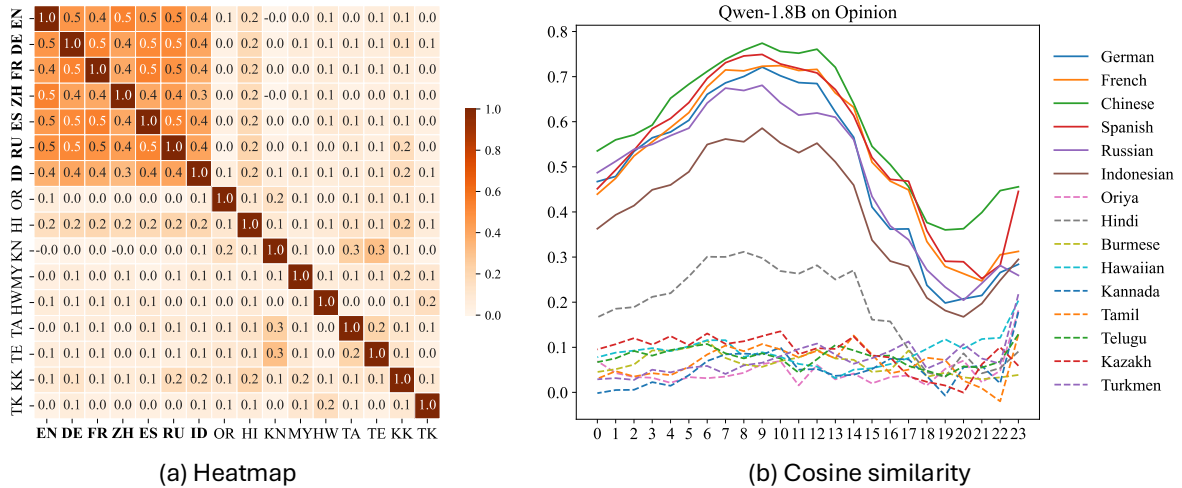


Figure 2: (a) Heatmap of the similarities of probing vectors correlation across languages; (b) Cosine similarity of probing vectors with English. (Model: Qwen-1.8B, Dataset: Opinion).

4 Related Work

In this section, we review two lines of research that are most relevant to ours.

Multilingual Abilities of LLMs. The multilingual capabilities of LLMs have garnered increasing attention from researchers (Ali and Pyysalo, 2024; Jayakody and Dias, 2024). Recent studies have investigated the consistency of factual knowledge across different languages in multilingual pretrained language models (PLMs) (Fierro and Søgaard, 2022; Qi et al., 2023). Additionally, significant efforts have been directed towards enhancing the representation of low-resource languages (Abadji et al., 2022; Imani et al., 2023; Li et al., 2024). These investigations demonstrate that LLMs still possess considerable untapped potential in multilingual capabilities.

Probing Representations in LLMs. Probing is a popular method to investigate the internal representations for LLMs in recent days, which is widely used in LLM interpretability studies (Alain and Bengio, 2018; Taktasheva et al., 2021; Pimentel et al., 2020; Ferrando et al., 2024b; Wendler et al., 2024). Previous work demonstrate that different layers typically acquired different information (Jin et al., 2024; Ju et al., 2024b). Various works using probing technique to assess how they encode linguistic features (Liu et al., 2023; Marks and Tegmark, 2024).

Our study employs probing techniques to examine LLMs’ performance and internal representations across different languages. The most closely related work is the Language Ranker (Li et al.,

2024), which uses cosine similarity between a language’s representation and English as a baseline. In contrast, our method utilizes linear classifier probing to evaluate performance across languages. This approach allows us to directly assess the model’s ability to extract language-specific information, providing a more detailed view of LLMs’ multilingual capabilities.

5 Conclusions and Future Work

In this work, our multilingual probing experiments on LLMs reveal significant disparities in performance and representational qualities across languages, suggesting potential limitations in how these models learn linguistic concepts. Specifically, high-resource languages consistently achieve higher probing accuracy and exhibit similar trends to English, with accuracy improving significantly in deeper layers. We also observe high similarities between probing vectors of high-resource languages, while low-resource languages demonstrate lower similarities both among themselves and with high-resource languages. These findings not only indicate the current limitations of LLMs in handling low-resource languages, but also suggest that these models may not be learning deeper linguistic concepts effectively across all languages.

In future, we plan to conduct research to address these gaps by developing more equitable and effective multilingual language models that can better capture universal linguistic concepts. Besides, we plan to extend this research to multimodal models that incorporate visual and textual information.

Limitations

In this work, we use machine translation to generate the prompt templates and question sentences from English to other languages, which may introduce noise. We only experiment with five open-source LLMs and two datasets. In the future, we would like to expand these findings with other datasets and models to confirm how well the LLMs' performance and representations generalize in these settings. Additionally, we just utilized linear classifier probing to do the experiments. We plan to explore more sophisticated probing methods beyond linear classifiers, which could offer deeper insights into the nature of linguistic representations within LLMs.

Acknowledgment

The work is in part supported by NSF #2310261. The views and conclusions in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

- Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. [Towards a cleaner document-oriented multilingual crawled corpus](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altmenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Guillaume Alain and Yoshua Bengio. 2018. [Understanding intermediate layers using linear classifier probes](#). *Preprint*, arXiv:1610.01644.
- Wazir Ali and Sampo Pyysalo. 2024. [A survey of large language models for european languages](#). *Preprint*, arXiv:2408.15040.
- Anthropic. 2024. [Introducing claude 3.5 sonnet](#).
- Safiul Azam. 2024. [Deep translator](#). Accessed: 2024-09-03.
- J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang, Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. 2023. Qwen technical report. Technical report.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R Costa-jussà. 2024a. [A primer on the inner workings of transformer-based language models](#). *arXiv preprint arXiv:2405.00208*.
- Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024b. [A primer on the inner workings of transformer-based language models](#). *Preprint*, arXiv:2405.00208.
- Constanza Fierro and Anders Søgaard. 2022. [Factual consistency of multilingual pretrained language models](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3046–3052.
- Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. [Glot500: Scaling multilingual corpora and language models to 500 languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.
- Ravindu Jayakody and Gihan Dias. 2024. [Performance of recent large language models for a low-resourced language](#). *Preprint*, arXiv:2407.21330.
- Mingyu Jin, Qinkai Yu, Jingyuan Huang, Qingcheng Zeng, Zhenting Wang, Wenyue Hua, Haiyan Zhao, Kai Mei, Yanda Meng, Kaize Ding, Fan Yang, Mengnan Du, and Yongfeng Zhang. 2024. [Exploring concept depth: How large language models acquire knowledge at different layers?](#) *Preprint*, arXiv:2404.07066.
- T. Ju, W. Sun, W. Du, X. Yuan, Z. Ren, and G. Liu. 2024a. [How large language models encode context knowledge? a layer-wise probing study](#). *arXiv preprint arXiv:2402.16061*.
- Tianjie Ju, Weiwei Sun, Wei Du, Xinwei Yuan, Zhaochun Ren, and Gongshen Liu. 2024b. [How large language models encode context knowledge? a layer-wise probing study](#). *Preprint*, arXiv:2402.16061.
- Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ninghao Liu, and Mengnan Du. 2024. [Quantifying multilingual performance of large language models across languages](#). *arXiv preprint arXiv:2404.11553*.
- Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023. [Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness?](#) *Preprint*, arXiv:2312.03729.

- S. Marks and M. Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.
- Samuel Marks and Max Tegmark. 2024. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). *Preprint*, arXiv:2310.06824.
- Tiago Pimentel, Josef Valvoda, Rowan Hall Maudslay, Ran Zmigrod, Adina Williams, and Ryan Cotterell. 2020. [Information-theoretic probing for linguistic structure](#). *Preprint*, arXiv:2004.03061.
- Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.
- Ekaterina Taktasheva, Vladislav Mikhailov, and Ekaterina Artemova. 2021. [Shaking syntactic trees on the sesame street: Multilingual probing with controllable perturbations](#). In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 191–210, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Rachael Tatman. 2017. [Deceptive opinion spam corpus](#). Accessed: 2024-09-03.
- G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bualanov, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, P. G. Sessa, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y. hui Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy. 2024. Gemma: Open models based on gemini research and technology. Technical report.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*.
- Haiyan Zhao, Fan Yang, Himabindu Lakkaraju, and Mengnan Du. 2024. Opening the black box of large language models: Two views on holistic interpretability. *arXiv preprint arXiv:2402.10688*.
- Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.

Language	Prompt
English	Judge the statement is Positive or Negative.
German	Beurteilen Sie, ob die Aussage positiv oder negativ ist.
French	Jugez si l'énoncé est positif ou négatif.
Chinese	判断该陈述是正面的还是负面的。
Spanish	Juzgue si la afirmación es positiva o negativa.
Russian	Оцените, является ли утверждение положительным или отрицательным.
Indonesian	Tentukan apakah pernyataan tersebut positif atau negatif.
Oriya	ବିବେଚନା କରନ୍ତୁ ଯେ ବକ୍ତବ୍ୟଟି ଧନାତ୍ମକ କି ନିକାତ୍ମକ ।
Hindi	निर्णय करें कि कथन सकारात्मक है या नकारात्मक ।
Burmese	ထောက်ထားချက်သည်အနုမြူရောအနုတ်မြူရောဖြစ်သည်။
Hawaiian	E ho'opa'apa'a inā he maika'i a i 'ole he maika'i 'ole ka 'olelo.
Kannada	ಹೇಳಿಕೆಯನ್ನು ಸಕಾರಾತ್ಮಕ ಅಥವಾ ಋಣಾತ್ಮಕ ಎಂದು ತೀರ್ಮಾನಿಸಿ.
Tamil	வாக்குமூலம் நேர்மையானதா அல்லது எதிர்மையானதா என்பதை மதிப்பீடு செய்யவும்.
Telugu	వాఖ్యానికీ సానుకూలం లేదా ప్రతికూలం అని తీర్పు ఇవ్వండి.
Kazakh	Мәлімдеменің оң немесе теріс екенін анықтаңыз.
Turkmen	Beýannamanyň oňyn ýa-da otrisatelidigini kesgitläň.

Figure 3: Prompt templates of all languages used in experiments.

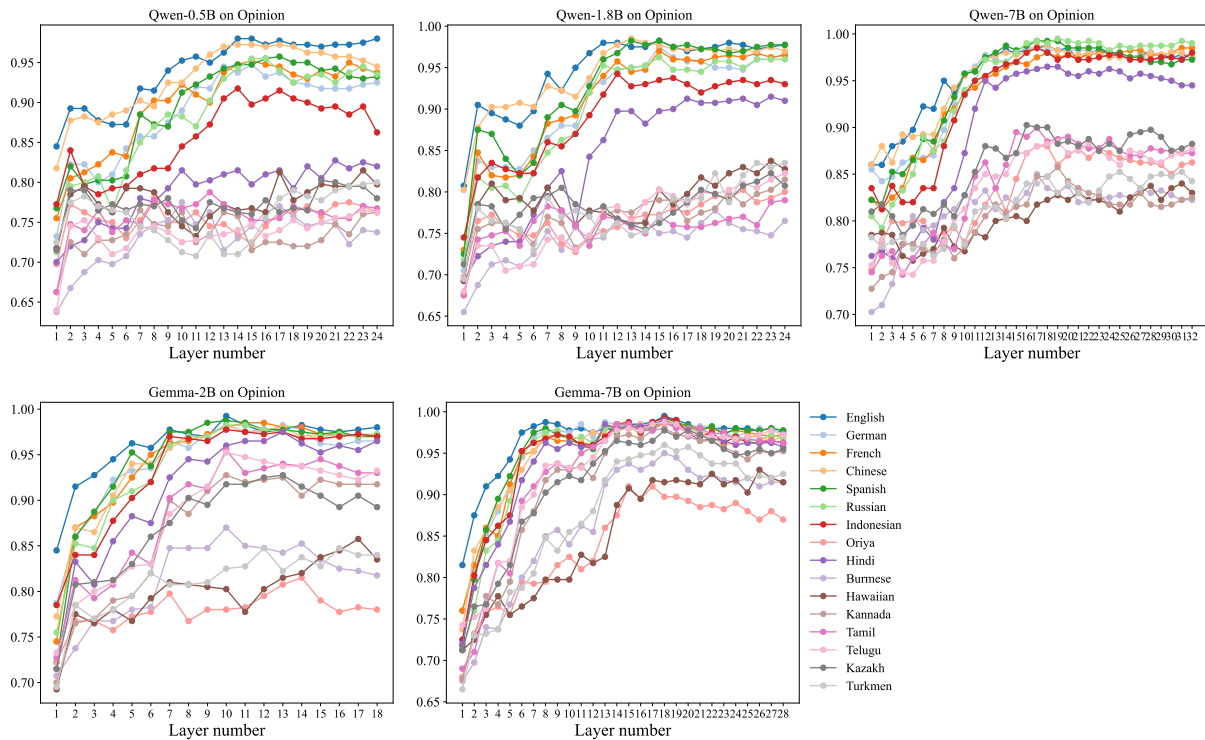


Figure 4: Additional results for multilingual accuracy of Qwen and Gemma Series Model on the Opinion Dataset

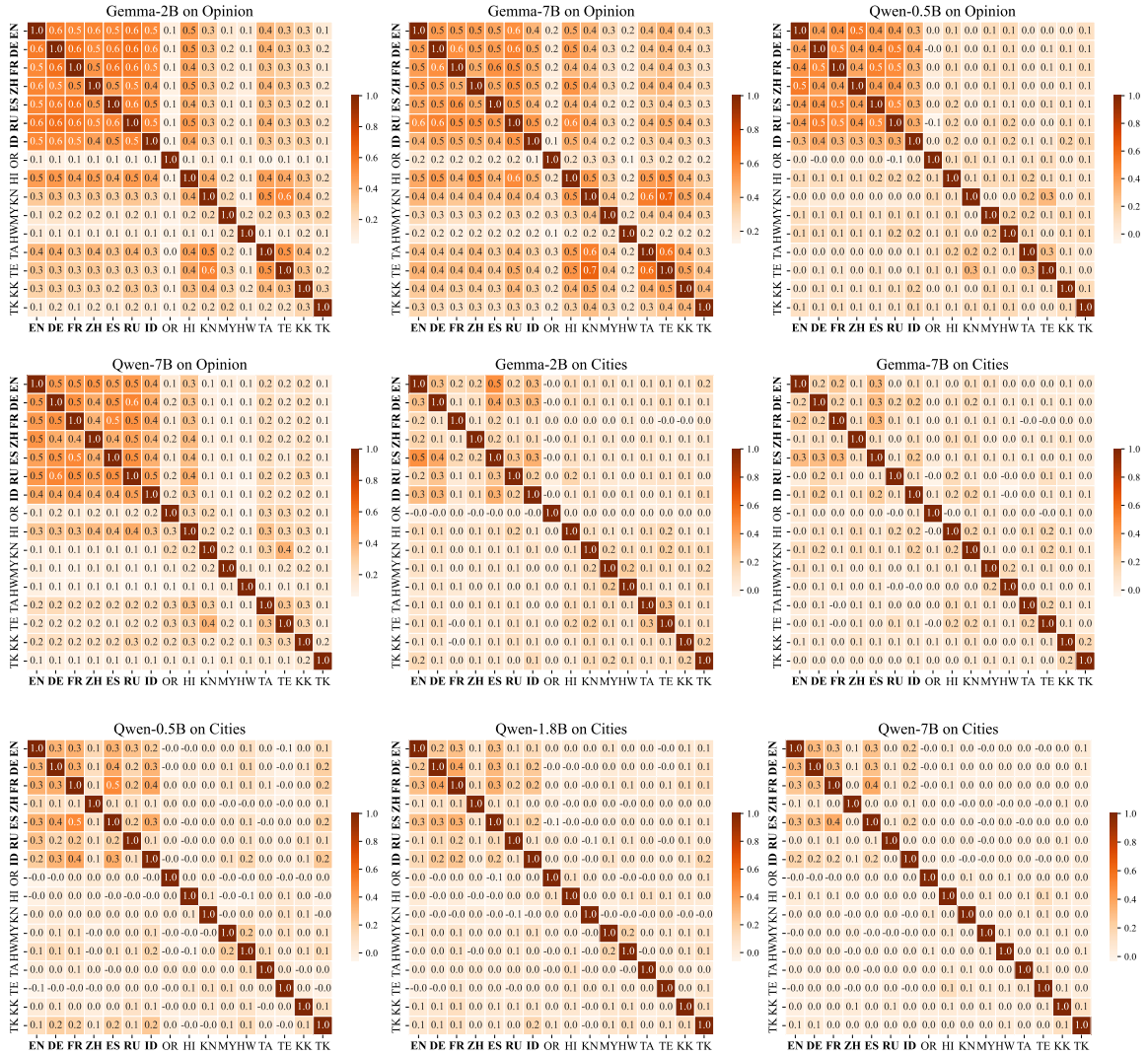


Figure 5: Heatmap of the similarities of probing vectors correlation across languages.

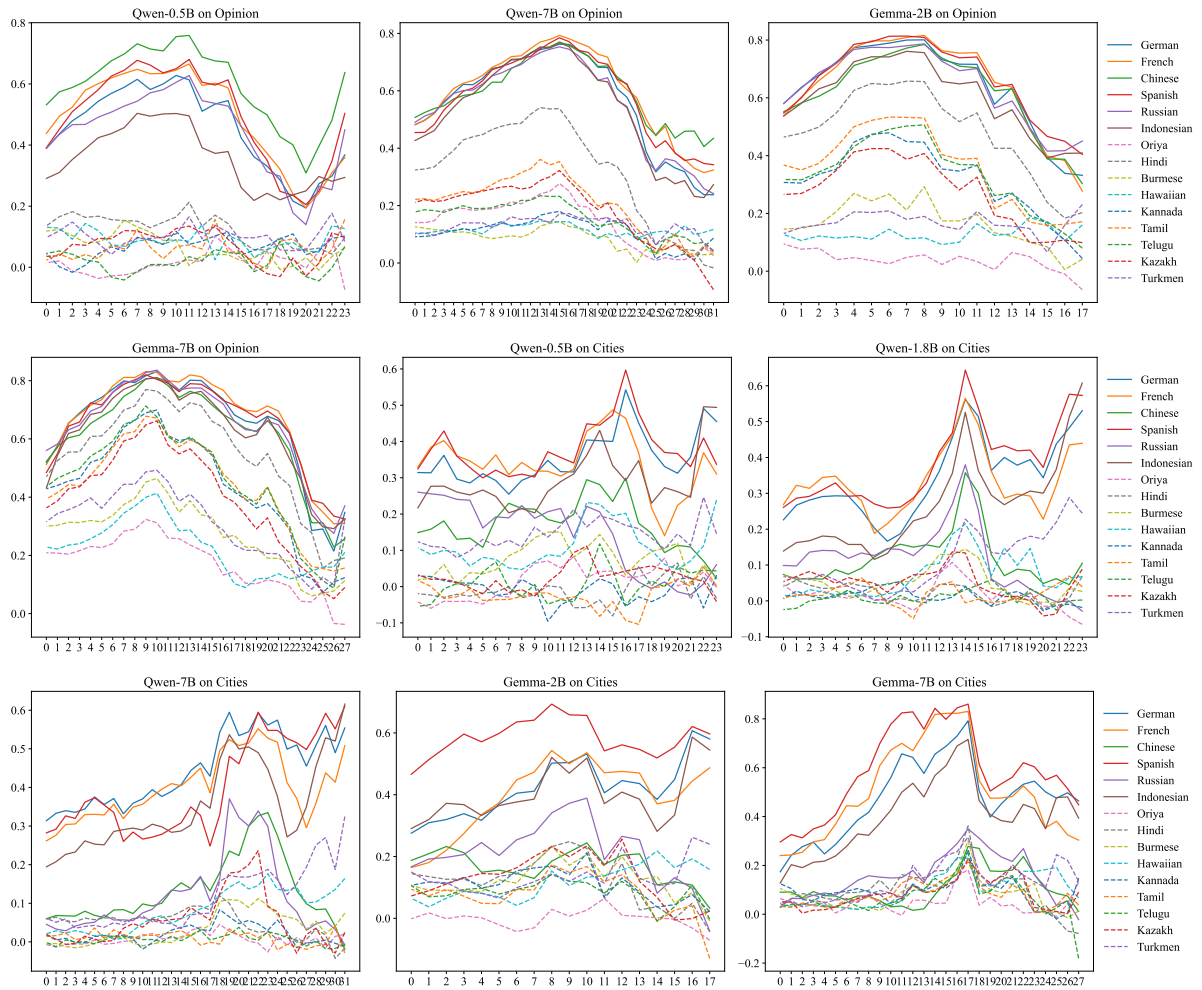


Figure 6: Cosine similarity of probing vectors with English across different language models (Qwen-0.5B, Qwen-7B, Gemma-2B, and Gemma-7B) and tasks (Opinion and Cities).