CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Disentangling language understanding and reasoning structures in cross-lingual chain-of-thought prompting
Authors	Tran , Khanh-Tung;Vu, Nguyet-Hang;O'Sullivan, Barry;Nguyen, Hoang D.
Publication date	2025-11
Original Citation	Tran, K.-T., Vu, N.-H., O'Sullivan, B. and Nguyen, H. D. (2025) 'Disentangling language understanding and reasoning structures in cross-lingual chain-of-thought prompting', Findings of the Association for Computational Linguistics: EMNLP 2025, pp. 12200-12206. https://doi.org/10.18653/v1/2025.findings-emnlp.652
Type of publication	Conference item
Link to publisher's version	10.18653/v1/2025.findings-emnlp.652
Rights	© 2025, Association for Computational Linguistics. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-10-13 17:06:57
Item downloaded from	https://hdl.handle.net/10468/18353

University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Disentangling Language Understanding and Reasoning Structures in Cross-lingual Chain-of-Thought Prompting

Khanh-Tung Tran[1]* **Nguyet-Hang Vu**[2]* **Barry O'Sullivan**[1] **Hoang D. Nguyen**[1]
[1]School of Computer Science and Information Technology, University College Cork, Ireland
[2]ISY Labs
123128577@umail.ucc.ie hang.vu@reliable-ai.org
b.osullivan@cs.ucc.ie hn@cs.ucc.ie

Abstract

Cross-lingual chain-of-thought prompting techniques have proven effective for investigating diverse reasoning paths in Large Language Models (LLMs), especially for low-resource languages. Despite these empirical gains, the mechanisms underlying cross-lingual improvements remain perplexing. This study, therefore, addresses whether the benefits of cross-lingual prompting arise from reasoning structures intrinsic to each language, or are simply a consequence of improved comprehension through cross-linguistic exposure. We employ neuron intervention and perturbation techniques to analyze and deactivate language-specific reasoning neurons during cross-lingual prompting, leading to performance disparities across languages, upto 27.4%. Our findings disentangle that these neurons are essential for reasoning in their respective languages but have minimal effect on reasoning in other languages, providing evidence for the existence of language-specific local reasoning structures and guiding the development of more interpretable and effective multilingual AI systems.[1]

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable reasoning capabilities, yet performance disparities across languages persist, particularly for low-resource languages with limited training data (Zhang et al., 2023; Alam et al., 2024). Chain-of-thought (CoT) prompting (Wei et al., 2022), which guides models to break problems into sequential reasoning steps, has become a go-to strategy to unlock latent reasoning power, potentially due to the ability of CoT to leverage local clusters of "reasoning locality" (Prystawski et al., 2023), and has been effectively applied as cross-lingual prompting in parameter-frozen multilingual LLMs (MLLMs) (Qin et al., 2023; Ranaldi et al., 2024b). The significant performance improvements observed with cross-lingual prompting, particularly for low-resource languages, are well-documented, but the underlying mechanisms driving these gains are still not fully understood. A fundamental question remains entangled: *Do improvements from cross-lingual prompting arise from distinct local reasoning structures intrinsic to each language, or are they simply by-products of enhanced problem comprehension achieved through cross-linguistic exposure?*

Figure 1 illustrates this vital distinction. When presented with identical mathematical problems in different languages, LLMs often produce different reasoning processes and solutions. The source is unclear, whether certain languages inherently scaffold different reasoning structures through their syntax, morphology, and discourse conventions, or framing the same problem in multiple languages simply clarifies ambiguities and better translations.

To address this question, our work offers a comprehensive analysis aimed at disentangling whether cross-lingual prompting yields genuinely distinct local reasoning structures per language or merely enhances the understanding of problems originally written in low-resource languages. Our main contributions are:

- Systematic Framework: We leverage neuron intervention to analyze between language-specific reasoning structures and language understanding in cross-lingual prompting.
- Structural Analysis: Through our experiments, we compare language reasoning paths and analyze their structural properties, highlighting the role of language-specific features, with disabling them causing a gap of upto 27.4%.
- Insights into Multilingual Reasoning: By iso-

*Equal contribution.
[1]Our data and code are publicly available at: https://github.com/ReML-AI/DCL-CoT

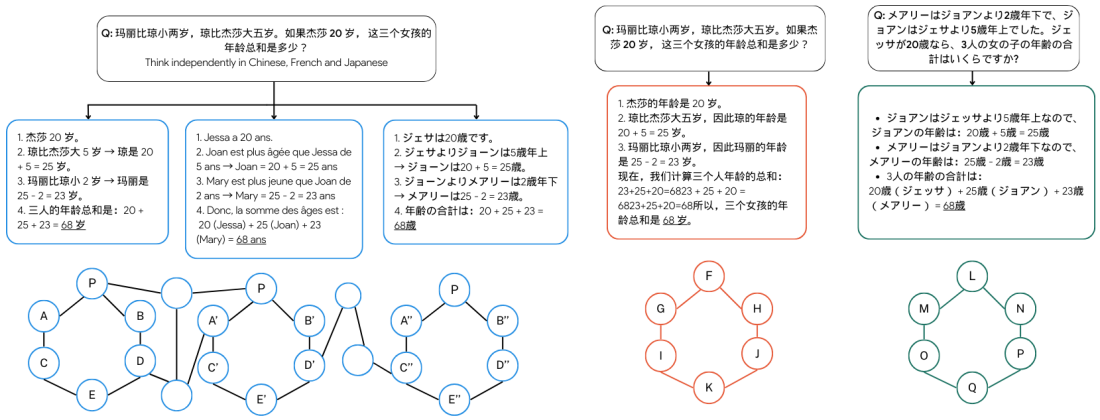

Figure 1: Hypotheses of cross-lingual reasoning. **Direct Prompting (Orange & Green):** Prompts in native languages may engage distinct, language-specific reasoning structures. **Cross-Lingual Prompting (Blue):** The *consistent local structure* hypothesis posits that an English prompt directing the model to reason in other languages would yield a uniform reasoning scaffold across target languages. Empty nodes denote shared intermediate states. Empirical results indicate, however, that distinct structures persist even under cross-lingual prompting.

lating the contributions of language understanding and reasoning structure, our work provides evidence for existence of language-specific reasoning structures within LLMs.

2 Related Works

Cross-lingual prompting Reasoning methods such as CoT prompting (Wei et al., 2022) elicit step-by-step problem solving from LLMs. Cross-lingual prompting and its variants, including self-consistent and tree-of-thought, first attempt to translate the low-resource input to a target language that is then used for reasoning on each instance, for multiple target languages; claiming to ensemble different reasoning paths across languages (Qin et al., 2023; Huang et al., 2023; Ranaldi et al., 2024b,a). However, the origin of its gains, whether from improved language understanding due to translation, avoidance of failures in specific language translation, or from intrinsic language-specific reasoning structures, remains an open question, one we seek to investigate and clarify.

Local structures in language models Recent studies suggest that CoT benefits arise when related concepts cluster closely in model activations (Prystawski et al., 2023). Other works show that reasoning may reside in middle layers, where a forward pass of LLMs first goes through surface-level translation and generation occur only at later layers (Tang et al., 2024; Tran et al., 2024; Hu et al., 2025). Building on these findings, our work identifies language-specific patterns associated with different reasoning behaviors. We examine how this

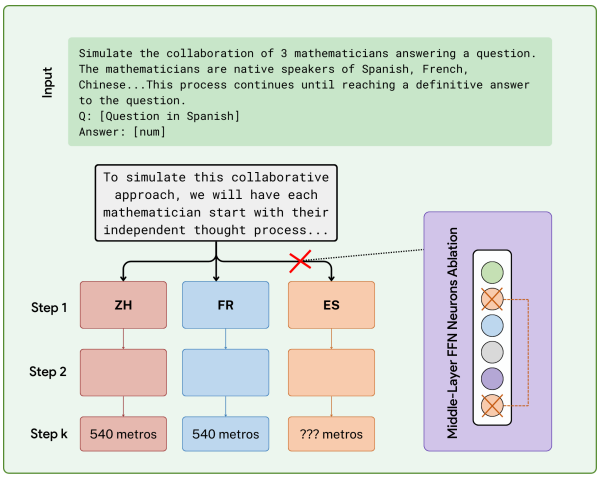

Figure 2: Our intervention approach for analyzing and disentangling language understanding and local reasoning structures in cross-lingual LLM prompting.

manifests in cross-lingual CoT and their implications for both performance and interpretability.

3 Methodology

We investigate the origins of multilingual reasoning in LLMs by analyzing the internal activation patterns and functional roles of language-specific and shared neural pathways during cross-lingual prompting for mathematical and commonsense reasoning tasks. The experimental design comprises multiple phases, each aimed at uncovering the structure of reasoning processes across languages.

The first phase identifies neurons that are highly specific to reasoning in individual languages. Leveraging the LAPE technique (Tang et al., 2024), which operates by calculating the entropy of each feed-forward network neuron's activation probabil-

ities across large text corpora from multiple languages, neurons exhibiting low entropy, meaning they preferentially activate for a single language, are considered language-specific.

Next, to evaluate how the model encodes linguistic information, multilingual queries are fed into the model, and output embeddings at each layer are decoded into model vocabulary tokens. These decoded tokens are classified as either English or non-English. Results reveal that, in the initial layers, embeddings produced from non-English queries predominantly correspond to non-English tokens while in intermediate layers, English tokens begin to take over. This pattern suggests that the model initially processes and comprehends non-English inputs through language-specific pathways (Zhao et al., 2024), therefore, deactivation of language neurons in intermediate layers does not impact the model's understanding of the input question, indicating them as language-specific *reasoning* neurons.

To provide a robust baseline for performance analysis, we employ Cross-lingual Tree-of-Thought (Cross-ToT) prompting (Ranaldi et al., 2024b) and perform intervention with language-specific reasoning neurons, as shown in Figure 2. In this setup, the model is prompted to simulate collaboration among mathematicians/experts, each working in their native language. The original prompt encourages each one to reason step by step in their language, refining and cross-referencing their solutions at each stage until a consensus is reached. Here, we refine the prompt to explicitly label each reasoning path according to the expert's language, allowing for extraction and analysis of the reasoning trajectory for each language. Key performance metrics, including logical soundness and correctness, are evaluated using LLM-as-a-judge framework (Zheng et al., 2023; Ye et al., 2025).

The final phase examines the results and behaviors of neuron deactivation interventions. The impact on performance is assessed by measuring changes in the metrics after selectively disabling language-specific neurons. By analyzing which reasoning paths and performance metrics are most affected by these perturbations, we analyze whether distinct local reasoning structures are maintained for each language or if reasoning converges on a shared representational core. For instance, if deactivating neurons specific to language X disrupts reasoning only for that language while leaving others unaffected, it provides strong evidence for the existence of language-specific reasoning circuits.

4 Experiments

4.1 Setup

To identify language-specific reasoning neurons, we follow the original approach (Tang et al., 2024), utilizing Wikipedia articles as a high-quality, multilingual resource and compute perplexity scores for each language subset. To avoid selecting inactive or uninformative neurons, we filter out any neuron whose activation in all languages falls below the 95th percentile of global neuron activations. This ensures that only sufficiently active neurons are considered for entropy ranking. We also select a set of random neurons of equal size (500 per language) to serve as a control group for our interventions. This ensures that observed effects are specifically attributable to language-specific circuits rather than general model capacity reduction.

For cross-lingual prompting, we adopt the tree-of-thought variant (Ranaldi et al., 2024b), which enables the generation of multi-step reasoning paths in various languages. The perturbation setup mirrors this, but zeros out the selected language-specific neurons during generation. While our analysis is grounded in this specific framework, our method is broadly applicable to any cross-lingual prompting strategy.

All experiments are conducted with Llama 3.1-8B-Instruct, Llama 3.1-70B-Instruct (Grattafiori et al., 2024; AI, 2024) models selected for its strong performance and open source license. To evaluate the effects of cross-lingual prompting and interventions on language-specific neurons, we use questions from the MGSM, an arithmetic reasoning benchmark (Shi et al., 2023) and XCOPA, a commonsense reasoning benchmark (Ponti et al., 2020) in different languages. We compare performance with and without language-specific reasoning neurons deactivation across both Latin and non-Latin

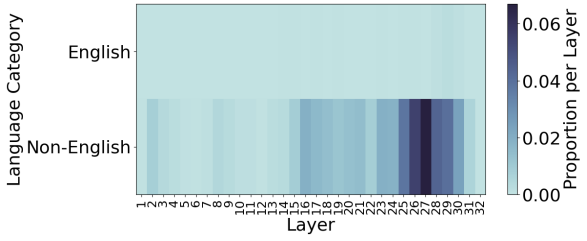

Figure 3: Distribution of identified language specific-neurons across layers in Llama-3.1-8B-Instruct.

script languages, and all resource regimes (high, medium, and low-resource languages).

To assess the quality of generated reasoning paths, we employ Claude 3.7 Sonnet (Anthropic, 2025) as automated evaluator, providing consistent scoring for each language's reasoning process. The validity of this evaluation method was established by correlating its outputs with those of a strong open-weight baseline, Qwen-3-32B. First, the baseline's internal consistency was verified; evaluations using two different random seeds yield a high intra-model correlation (Pearson's r = 0.8726, p < 0.000001). Subsequently, the assessments from Claude 3.7 Sonnet demonstrates a strong correlation with the validated baseline (r = 0.8423, p = 0.000012). The Nvidia HGX platform with H100 GPUs is used throughout our experiments.

4.2 Results

Figure 3, 4 show the distribution of identified language-specific neurons across the layers of the models. Most of these neurons are concentrated in the reasoning layers, suggesting a role in language-specific reasoning circuits. The number of English-specific neurons identified in each layer is relatively small compared to other languages. For example, out of 6,650 language-specific neurons identified in total of Llama-3.1-8B-Instruct, only 148 (2.23%) are specific to English, with the highest concentration being 35 neurons in layer 29. We attribute this phenomenon to English being the dominant language in Llama 3.1's training data, thereby requiring fewer specialized neurons to support English-specific language ability. Neurons may be shared across multiple languages. To further assess their reasoning relevance, we benchmark cross-lingual tree-of-thought prompting on three Latin script languages (English, Spanish, French) and two non-Latin script languages (Japanese, Chinese) from MGSM. For XCOPA, we include Chinese (a medium resource language) along with Vietnamese and Indonesian, both considered low-resource languages (Joshi et al., 2020) (Liu et al., 2025) with less than 1% representation in Common Crawl[2]. Performance metrics, as evaluated by Claude 3.7 Sonnet, are summarized in Tables 1–2.

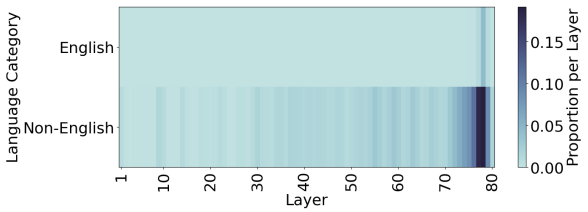

Figure 4: Distribution of identified language specific-neurons across layers in Llama-3.1-70B-Instruct.

Model	Lang	Self-Perturb (Same)	Self-Perturb (Other)
Llama 3.1-8B -Instruct	en	+0.9%	**+1.0%**
	es	+1.3%	**+8.2%**
	fr	–24.1%	**–10.0%**
	ja	–27.4%	**+1.2%**
	zh	**–2.5%**	–4.8%
Llama 3.1-70B -Instruct	en	**+2.5%**	+2.1%
	es	–3.9%	**–3.4%**
	fr	–2.8%	**–1.2%**
	ja	–1.0%	**–0.3%**
	zh	–5.8%	**–1.1%**

Table 1: Self vs. Cross-language intervention effects on MGSM. Logical soundness changes (%) relative to no-intervention baseline.

Model	Lang	Self-Perturb (Same)	Self-Perturb (Other)
Llama 3.1-8B -Instruct	zh	–11.3%	**–3.6%**
	vi	–15.3%	**–6.3%**
	id	–5.7%	**–2.0%**
Llama 3.1-70B -Instruct	zh	–2.7%	**–2.0%**
	vi	–22.6%	**–2.4%**
	id	–12.5%	**–1.5%**

Table 2: Self vs. Cross-language intervention effects on XCOPA. Logical soundness changes (%) relative to no-intervention baseline.

4.3 Analysis

Our experiments reveal several key insights regarding the role of language-specific reasoning neurons in multilingual LLMs. First, we find that the majority of these neurons are concentrated in the model's reasoning layers, rather than in layers associated with surface-level language understanding. This localization suggests that these neurons participate directly in the reasoning process, forming distinct circuits for different languages.

When we selectively deactivate the reasoning neurons associated with a specific language, the model's performance on reasoning tasks in same language, denoted **Self-Perturb (Same)**, drops significantly. More specifically, in all scenarios (**p-value = 0.006, two-tailed**), deactivating a language-specific neurons will affect the performance when prompt to reason in that language (an average

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

Model	Lang	Baseline (%)	Perturbed (%)	Change (%)
Llama 3.1-8B -Instruct	en	65.60	64.00	-1.60
	es	61.60	56.29	-5.31
	fr	59.20	51.94	-7.26
	ja	46.80	39.20	-7.60
	zh	56.40	53.26	-3.14
(pertubed with random mask)	en	65.60	66.65	1.05
	es	61.60	61.40	-0.20
	fr	59.20	59.15	-0.05
	ja	46.80	46.90	0.10
	zh	56.40	56.45	0.05
Llama 3.1-70B -Instruct	en	93.60	92.91	-0.69
	es	84.40	81.94	-2.46
	fr	79.60	78.29	-1.31
	ja	80.80	80.51	-0.29
	zh	86.40	84.11	-2.29

Table 3: Accuracy on the MGSM benchmark. Baseline accuracy, accuracy after perturbing language-specific neurons, and absolute change. Random mask control shows negligible impact.

Model	Lang	Baseline (%)	Perturbed (%)	Change (%)
Llama 3.1-8B -Instruct	zh	68.60	63.27	-5.33
	vi	69.20	68.87	-0.33
	id	70.60	68.00	-2.60
Llama 3.1-70B -Instruct	zh	93.40	91.80	-1.60
	vi	92.00	90.93	-1.07
	id	93.40	91.80	-1.60

Table 4: Accuracy on the XCOPA commonsense reasoning benchmark, showing selective degradation after neuron perturbation across model sizes.

drop of -8.31%), compared to reason in other languages, denoted **Self-Perturb (Other)**, (an average drop of -1.60%), which, will use the other language-specific reasoning neurons, with a degradation of upto 27.4% compared to baseline for Japanese. This selective degradation demonstrates that the affected neurons are primarily responsible for separate, local reasoning in their corresponding language, rather than for general language understanding or for multilingual reasoning as a whole. On XCOPA, we also observe that the effects are more severe on low-resource languages (Vietnamese and Indonesian), compared to high-resource ones (Chinese), with a performance drop of -14.03% versus -7.00%. This is potentially due to these languages being less regularized during pre-training (Conneau et al., 2020) (Pires et al., 2019), and their language-specific reasoning neurons are more specific to their own languages, leading to deactivating them having larger effects.

Furthermore, when comparing the results to the baseline with no neuron deactivation, we observe that disabling these neurons does not cause uniform performance drops across all languages or tasks. In the case of English and Spanish, overall performance is even improved. This outcome confirms the interpretation that these neurons are not in charge of initial language comprehension, supporting the existence of language-specific reasoning structures within LLM (and potentially language understanding neurons), and that interventions targeting these neurons selectively disrupt reasoning in their respective languages without broadly affecting multilingual performance or comprehension

We report results with accuracy metric in Tables 3 and 4. On MGSM with Llama 3.1-8B-Instruct, perturbing Japanese-specific and French-specific neurons produced a -7.60% and -7.26% absolute drop in final-answer accuracy, respectively On XCOPA, perturbing Chinese-specific neurons on Llama 3.1-8B yielded a -5.33% accuracy drop (Table 4). By contrast, applying a matched random-neuron mask produces negligible changes in accuracy (Table 3, "Random Mask (Control)" rows), indicating that the observed effects are not attributable to generic capacity reduction, stochastic noise, or model instability but rather to targeted disruption of functionally important units.

5 Conclusion

We demonstrate that LLMs exhibit distinct language-specific reasoning and language understanding structures, with specialized local reasoning neurons that primarily support inference within individual languages. By systematically analyzing the effects of targeted neuron deactivation, we find that these language-specific neurons are crucial for reasoning in their respective languages, while having limited impact on reasoning in others. These results provide strong evidence that multilingual LLMs do not rely solely on a shared reasoning core but instead develop local reasoning structures for different languages. Our findings offer valuable insight into the internal organization of cross-lingual reasoning and new directions for developing more interpretable and effective multilingual AI systems.

Acknowledgements

This publication has emanated from research supported by Research Ireland under Grant 12/RC/2289-P2 and 18/CRT/6223.

Limitations

This study is limited to the evaluation of the Llama 3.1 series, and its findings have not yet been validated across other LLMs, such as Llama 4's series (AI, 2025) due to potential concerns regarding license[3]. Additionally, the languages investigated are primarily mid- to low-resource, and do not include extremely low-resource languages (Ghosh et al., 2025). We believe that the observed phenomena would be even more pronounced in extremely low-resource scenarios. The evaluation methodology also relies on LLM-as-a-judge approach, which may introduce challenges related to scalability and assessment robustness. Despite these constraints, the clear performance disparities observed in our experiments provide compelling evidence for the presence of language-specific reasoning structures and offer valuable insights into the internal organization of cross-lingual reasoning.

References

Meta AI. 2024. Introducing llama 3.1: Our most capable models to date.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. LLMs for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, pages 27–33, St. Julian's, Malta. Association for Computational Linguistics.

Anthropic. 2025. Claude 3.7 sonnet.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Akash Ghosh, Debayan Datta, Sriparna Saha, and Chirag Agarwal. 2025. The multilingual mind : A survey of multilingual reasoning in language models. *Preprint*, arXiv:2502.09457.

Aaron Grattafiori and 1 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

[3] https://www.llama.com/llama4/use-policy/

Peng Hu, Sizhe Liu, Changjiang Gao, Xin Huang, Xue Han, Junlan Feng, Chao Deng, and Shujian Huang. 2025. Large language models are cross-lingual knowledge-free reasoners. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1525–1542, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2025. Is translation all you need? a study on solving multilingual tasks with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9594–9614, Albuquerque, New Mexico. Association for Computational Linguistics.

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *Preprint*, arXiv:2005.00333.

Ben Prystawski, Michael Y. Li, and Noah Goodman. 2023. Why think step by step? reasoning emerges from the locality of experience. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2695–2709, Singapore. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, Barry Haddow, and Alexandra Birch. 2024a. Empowering multi-step reasoning across languages via program-aided language

models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12171–12187, Miami, Florida, USA. Association for Computational Linguistics.

Leonardo Ranaldi, Giulia Pucci, Federico Ranaldi, Elena Sofia Ruzzetti, and Fabio Massimo Zanzotto. 2024b. A tree-of-thoughts to broaden multi-step reasoning across languages. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1229–1241, Mexico City, Mexico. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations*.

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)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Khanh-Tung Tran, Barry O'Sullivan, and Hoang Nguyen. 2024. Irish-based large language model with extreme low-resource settings in machine translation. In *Proceedings of the Seventh Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2024)*, pages 193–202, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V Chawla, and Xiangliang Zhang. 2025. Justice or prejudice? quantifying biases in LLM-as-a-judge. In *The Thirteenth International Conference on Learning Representations*.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, Singapore. Association for Computational Linguistics.

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*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.