and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Lera Boroditsky, Lauren A. Schmidt, and Webb Phillips. 2003. Sex, syntax, and semantics. In Dedre Gentner and Susan Goldin-Meadow, editors, *Language in Mind: Advances in the Study of Language and Thought*, pages 61–79. MIT Press, Cambridge, MA.

Nicola Cancedda. 2024. Spectral filters, dark signals, and attention sinks. *arXiv preprint arXiv:2402.09221*.

Steven Cao, Nikita Kitaev, and Dan Klein. 2020. Multilingual alignment of contextual word representations.

Yihong Chen, Kelly Marchisio, Roberta Raileanu, David Ifeoluwa Adelani, Pontus Stenetorp, Sebastian Riedel, and Mikel Artetxe. 2023. Improving language plasticity via pretraining with active forgetting.

Rochelle Choenni and Ekaterina Shutova. 2020. What does it mean to be language-agnostic? probing multilingual sentence encoders for typological properties. *arXiv preprint arXiv:2009.12862*.

Arthur Conmy, Augustine N Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *arXiv preprint arXiv:2304.14997*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020a. Unsupervised cross-lingual representation learning at scale.

Alexis Conneau, Shijie Wu, Haoran Li, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. Emerging cross-lingual structure in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6022–6034.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. LLM.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Anna Di Natale, Max Pellert, and David Garcia. 2021. Colexification networks encode affective meaning. *Affective Science*, 2(2):99–111.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2023. Do multilingual language models think better in english?

Mor Geva, Avi Caciularu, Kevin Ro Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space.

Charles Goddard. 2023. Llama-polyglot-13b. `https://huggingface.co/chargoddard/llama-polyglot-13b`. Accessed: 2024-01-22.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *arXiv preprint arXiv:2305.01610*.

Bofeng Huang. 2023. vigogne-2-13b-instruct. `https://huggingface.co/bofenghuang/vigogne-2-13b-instruct`. Accessed: 2024-01-22.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Wayne 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.

Jaavid Aktar Husain, Raj Dabre, Aswanth Kumar, Ratish Puduppully, and Anoop Kunchukuttan. 2024. Romansetu: Efficiently unlocking multilingual capabilities of large language models models via romanization.

Daekeun Kim. 2023. Llama-2-ko-dpo-13b. `https://huggingface.co/daekeun-ml/Llama-2-ko-DPO-13B`. Accessed: 2024-01-22.

Kenneth Li, Aspen K Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Jindřich Libovický, Rudolf Rosa, and Alexander Fraser. 2020. On the language neutrality of pre-trained multilingual representations. *arXiv preprint arXiv:2004.05160*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Samuel Marks and Max Tegmark. 2023. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Giovanni Monea, Maxime Peyrard, Martin Josifoski, Vishrav Chaudhary, Jason Eisner, Emre Kıcıman, Hamid Palangi, Barun Patra, and Robert West. 2023. A glitch in the matrix? locating and detecting language model grounding with fakepedia. *arXiv preprint arXiv:2312.02073*.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021. First align, then predict: Understanding the cross-lingual ability of multilingual bert. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*.

Nostalgebraist. 2020. Interpreting gpt: The logit lens. LessWrong.

Rafael E. Núñez and Eve Sweetser. 2006. With the future behind them: Convergent evidence from aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science*, 30(3):401–450.

OpenAI. 2023. Gpt-4 technical report.

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2022. Multilingual bert has an accent: Evaluating english influences on fluency in multilingual models.

Rait Piir. 2023. Finland's chatgpt equivalent begins to think in estonian as well. ERR News.

Björn Plüster. 2023. LeoLM: Ein Impuls für Deutschsprachige LLM-Forschung. `https://laion.ai/blog-de/leo-lm/`. Accessed: 2024-01-22.

Lucia Quirke, Lovis Heindrich, Wes Gurnee, and Neel Nanda. 2023. Training dynamics of contextual n-grams in language models. *arXiv preprint arXiv:2311.00863*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nina Rimsky. 2023. Decoding intermediate activations in Llama-2-7b. LessWrong.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1599–1613, Minneapolis, Minnesota. 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. 2022. Language models are multilingual chain-of-thought reasoners.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27.

Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.

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.

Wenhao Zhu, Shujian Huang, Fei Yuan, Shuaijie She, Jiajun Chen, and Alexandra Birch. 2024. Question translation training for better multilingual reasoning.

Wenhao Zhu, Yunzhe Lv, Qingxiu Dong, Fei Yuan, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Extrapolating large language models to non-english by aligning languages.

A Additional methodological details

A.1 Word translation

A detail that we omitted in the main paper for brevity is how we translate the English words resulting from the procedure outlined in Sec. 3.3 to French, German, and Russian. During these translations we translated both the individual words alongside their cloze sentences using DeepL.[5] For each word translation, we include the context of the cloze task to disambiguate homonyms. We then filter the translations to remove words that have the same prefix token across English and the

[5] `https://www.deepl.com/translator`

target language. For example, the French translation of the word "photograph", "photographier", shares the "photo" prefix token. Additionally, we parse through the translations and filter any cloze translations where the target word doesn't align with the expected word from the individual word translation, which was due to failures in the DeepL translation. These filterings result in a different number of final words across the different languages.

We provide the numbers for the aggregated translation task (Table 1), repetition task (Table 2), cloze-task (Table 3), and individual translation tasks (Table 4).

	Total	Single Token
de	287	126
fr	162	88
ru	324	45
zh	353	353

Table 1: Aggregated translation task dataset sizes.

	Total	Single Token
de	104	45
en	132	132
fr	56	31
ru	115	15
zh	139	139

Table 2: Repetition task dataset sizes.

	Total	Single Token
de	104	45
en	132	132
fr	56	31
ru	115	15
zh	139	139

Table 3: Cloze task dataset sizes.

A.2 Computing language probabilities

In order to compute language probabilities, we search Llama-2's vocabulary for all tokens that could be the first token of the correct word in the respective language. In particular, we search Llama-2's vocabulary for all prefixes of the word without and with leading space.[6] For Chinese and Russian we also consider tokenizations based on the UTF-8 encodings of their unicode characters. For a language ℓ and its corresponding target word w, we define

$$P(\text{lang} = \ell) := \sum_{t_\ell \in \text{Start}(w)} P(x_{n+1} = t_\ell), \tag{3}$$

where $\text{Start}(w)$ denotes the set of starting tokens of the word w.

For example, if the correct next Chinese word is "花" ("flower"), which can be tokenized either using the single token "花" or via its UTF-8 encoding "<0xE8>·<0x8A>·<0xB1>", we have $P(\text{lang} = \text{ZH}) = P(x_{n+1} = \text{“花”}) + P(x_{n+1} = \text{“<0xE8>”})$ and $P(\text{lang} = \text{EN}) = P(x_{n+1} = \text{“f”}) + P(x_{n+1} = \text{“fl”}) + P(x_{n+1} = \text{“flow”}) + P(x_{n+1} = \text{“\_f”}) + P(x_{n+1} = \text{“\_fl”}) + P(x_{n+1} = \text{“\_flo”}) + P(x_{n+1} = \text{“\_flow”}) + P(x_{n+1} = \text{“\_flower”})$ (all the token-level prefixes of "flower" and "_flower").

[6] Represented by "_".

	de	en	fr	ru	zh
de	–	120 (120)	56 (31)	105 (15)	120 (120)
en	104 (45)	–	57 (31)	114 (15)	132 (132)
fr	93 (40)	118 (118)	–	104 (15)	118 (118)
ru	90 (41)	114 (114)	49 (26)	–	115 (115)
zh	104 (45)	132 (132)	57 (31)	115 (15)	–

Table 4: Translation statistics between languages, including total numbers and single-token translations (in brackets).

B Additional results

Here we provide the results for all languages: Chinese, English, French, German, and Russian.

Language probability. Language probability plots (with entropy heatmaps) for the aggregated translation task are in Fig. 5, for the repetition task in Fig. 7, and, for the cloze task in Fig. 9. Additionally, we provide the translation task results for individual language pairs in Fig. 11, Fig. 13, Fig. 15, Fig. 17, Fig. 19.

We observe the same pattern—noise in the early layers, English in the middle, target language in the end—across almost all languages and model sizes. The only exception is the Chinese repetition task.

Energy. Energy (Sec. 4.2) plots for the aggregated translation task are in Fig. 6, for the repetition task in Fig. 8, and, for the cloze task in Fig. 10. Additionally, we provide the translation task results for individual language pairs in Fig. 12, Fig. 14, Fig. 16, Fig. 18, Fig. 20.

Energy plots are consistent with the theory outlined in Sec. 5.

B.1 Low-resource language Estonian

We also performed our analysis with Llama-2-7B on Estonian, a low-resource language, in Fig. 21. The fact that Estonian is a low-resource language is already evident in the number of single-token words: only one out of our 99 Estonian words can be represented with a single token.

Copy task. In the copy task, Estonian behaves the most similarly to Chinese, with the Estonian probability exceeding the English probability already in the intermediate layers.

Translation task. While the success probability on the translation task after the final layer is significantly smaller than in the languages studied in the main paper, we still observe the same effect as for the other languages: the intermediate next-token distributions decoded via the logit lens concentrate their probability mass on the correct English tokens and only in the final layers transition to Estonian.

Cloze task. The Estonian cloze task seems too hard, possibly due to the extremely low resources of Estonian in the Llama-2 training data: Llama-2-7B has a 0% success probability after the last layer. Interestingly, the Estonian success probability is slightly greater than 0% in the intermediate layers, when the logit lens decodes to English. The success probability might increase if we included synonyms of the translated words or used human experts for the creation of the cloze examples instead of GPT-4.

B.2 Other models: Mistral

We also performed our analysis on Mistral-7B, a model from outside the Llama model family. The results, shown in Fig. 22, are consistent with those for Llama-2, pointing at the universality of our findings.