On the Similarity of Circuits across Languages: a Case Study on the Subject-verb Agreement Task

Javier Ferrando* [1] Marta R. Costa-jussà [2]

Abstract

Several algorithms implemented by language models have recently been successfully reversed-engineered. However, these findings have been concentrated on specific tasks and models, leaving it unclear how *universal* circuits are across different settings. In this paper, we study the circuits implemented by Gemma 2B for solving the subject-verb agreement task across two different languages, English and Spanish. We discover that both circuits are highly consistent, being mainly driven by a particular attention head writing a 'subject number' signal to the last residual stream, which is read by a small set of neurons in the final MLP layers. Notably, this subject number signal is represented as a direction in the residual stream space, and is language-independent. Finally, we demonstrate this direction has a causal effect on the model predictions, effectively flipping the Spanish predicted verb number by intervening with the direction found in English examples.

1. Introduction

The widespread use of large language models (LLMs; Brown et al., 2020; Hoffmann et al., 2022; Chowdhery et al., 2023) highlights the importance of research dedicated to interpreting how these models work internally (Ferrando et al., 2024), especially to ensure they are safe. Mechanistic interpretability (MI) (Olah, 2022) aims to reverse-engineer the algorithms implemented by language models. A large set of MI works have focused on circuit analysis (Räuker et al., 2023), which locates subsets of components responsible for a behavior while giving human-understandable explanations of their roles. This research has made progress in identifying circuits that handle different tasks (Wang et al., 2023; Heimersheim & Janiak, 2023; Stolfo et al., 2023a;b; Geva et al., 2023; Hanna et al., 2023). However, it remains unclear whether the findings obtained through circuit analysis transfer to different settings. For instance, if different models learn similar circuits for solving the same task, or if models find different solutions for the same task in two different languages. In this work, we study the latter question. Through the lens of the subject-verb agreement (SVA) task (Linzen et al., 2016; Goldberg, 2019), we study the main components in Gemma 2B (Gemma Team et al., 2024) that are responsible across both English and Spanish.

2. Experimental Setup

In our experiments, we use Gemma 2B model (Gemma Team et al., 2024). This model has a large vocabulary size (256k tokens), making it particularly well-suited for circuit analysis, especially when doing activation patching (Section 3) in a multilingual setting, since it has a large set of non-English words with a reserved token. Regarding the dataset, for the English experiments use the subject-verb agreement (SVA) dataset from Arora et al. (2024)[1], built on top of SyntaxGym (Gauthier et al., 2020). The dataset consists of contrastive pairs that differ in the subject number, which agrees with the verb form continuation. This allows us to create 'clean' and 'corrupted' versions:

Singular

Clean: The executive that embarrassed the manager <u>has</u>

Corrupted: The executives that embarrassed the manager ___

Plural (1)

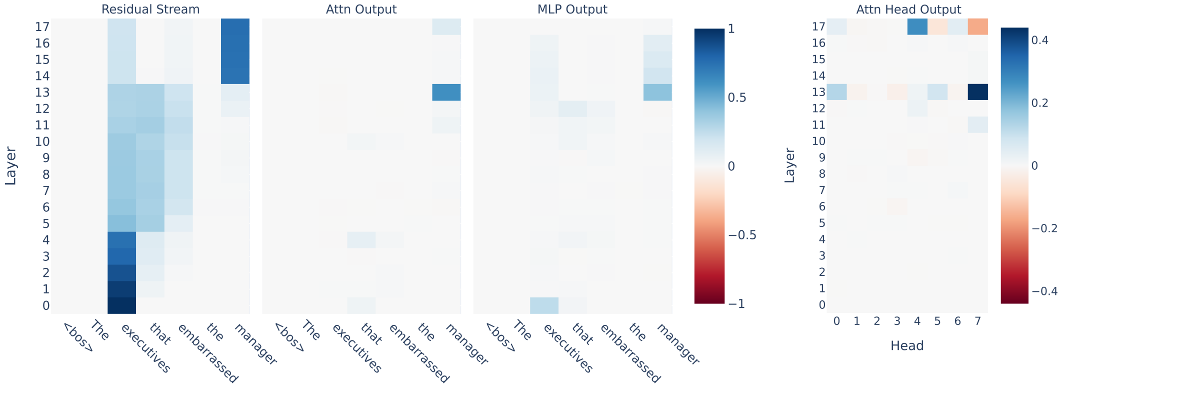
3. Methods

We start searching for a circuit in Gemma 2B for solving the SVA in English. To do so we use common techniques in circuit analysis, mainly direct logit attribution, activation patching, and attention pattern analysis.

Direct Logit Attribution. Every model component adds a vector $f^c(\mathbf{x})$ to the residual stream, and the last residual stream state gets projected onto the unembedding matrix, producing the logits distribution. Due to the linearity of

[1] Universitat Politècnica de Catalunya [2] FAIR at Meta AI. Correspondence to: Javier Ferrando <javier.ferrando.monsonis@upc.edu>.

[1] `aryaman/causalgym`, subset `agr_sv_num_subj-relc`



(a) (b) (c) (d)

Figure 1: English dataset activation patching results on the logit difference metric on (a) the residual streams (b) attention blocks outputs, (c) MLP outputs, and (d) on attention heads at the last position.

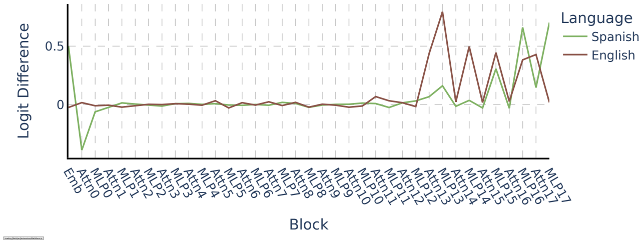


Figure 2: Average contribution to the logit difference by each model component.

the residual stream, the direct effect of a component to the logits can be measured by projecting its output onto the unembedding matrix, $f^c(\mathbf{x})\boldsymbol{W}_U$. We can also measure the **d**irect **a**ttribution to the **l**ogit **d**ifference (DLDA) (Yin & Neubig, 2022; Wang et al., 2023) of the two possible verb continuations (g and b):

$$\text{DLDA}_c = f^c(\tilde{\mathbf{x}})\boldsymbol{W}_U[:, g] - f^c(\tilde{\mathbf{x}})\boldsymbol{W}_U[:, b]. \qquad (2)$$

Activation Patching. A Transformer LM can be seen as a directed acyclic graph (DAG) representing a causal model (Geiger et al., 2021; Pearl, 2009; Vig et al., 2020), where nodes are model components, and edges representations. During the forward pass on the *corrupted input* $\mathbf{x}$ we can intervene on the value of a node, $f^c(\mathbf{x})$, or residual stream state, $f^l(\mathbf{x})$ by taking the activation value from the forward pass on the *clean input* $\tilde{\mathbf{x}}$. This is referred to as *denoising activation patching* (Vig et al., 2020; Meng et al., 2022). We can express the intervention using the do-operator (Pearl, 2009) as $f(\mathbf{x}|\text{do}(f^c(\mathbf{x}) = f^c(\tilde{\mathbf{x}})))$. Via a metric m we measure how the prediction changes between both runs:

$$\text{AP}_c = m\big(f(\mathbf{x}), f(\mathbf{x}|\text{do}(f^c(\mathbf{x}) = f^c(\tilde{\mathbf{x}})))\big). \qquad (3)$$

We are interested in finding components that increase the clean verb prediction when patching on the corrupted run. Thus, a natural choice for the patching metric m is the logit difference between the clean and the corrupted verbs' logits. In the Example 1, this means computing the logit difference between 'has' and 'have', and we expect it to increase as we patch activations from the clean (which includes 'executive') into the corrupted forward pass.

4. English Subject-Verb Agreement Circuit

Locating relevant components and residual stream states. We perform activation patching on the residual stream states across the dataset and show the average logit differences[2] in Figure 1 (a). We can see that the noun in the subject largely impacts the prediction, and patching at its position in early layers causes the verb prediction to aggressively change to match its number. Information from the subject flows towards the last residual stream via the attention block at layer 13 (Figure 1 (b)), followed by some action from downstream MLPs at the last position (Figure 1 (c)), especially MLP at layer 13 (MLP13). We can also observe that 'that' and the following verb ('embarrassed') get information from the subject at middle layers. We get a more granular understanding of the attention layers that seem relevant by doing activation patching on the output of every attention head in the last position (Figure 1 (d)). Attention head 7 in layer 13 (L13H7) has the largest effect on the logit difference, followed by L17H4. Notably, we also observe a head (L17H7) that contributes negatively to the logit difference. In Appendix F we show the average output-value-weighted heatmaps of these heads, and we see that L13H7 attends broadly to the context, with a slight focus on 'what', while L17H4 focuses on the subject's noun. Although attention blocks at layers 13 and 17 also have large direct effects Figure 2, most of the direct contribution to the logit difference is

[2] See in Appendix A the average logit differences.

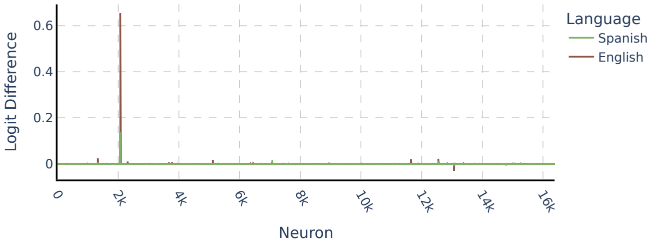


Figure 3: Average contribution to the logit difference by each neuron in MLP13.

Top Promoted Tokens *Positive* Neuron Activation
' are', 'are', 'were', ' were', 'Are', 'aren', ' ARE', ' WERE', ' weren'
Top Promoted Tokens *Negative* Neuron Activation
' gardent', **' is'**, **' has'**, ' sembrano', **' was'**, ' continúan', ' appartienment', **' isn'**, **' hasn'**, ' sostu'

Table 1: Top promoted tokens by neuron 2069 in MLP13 based on the sign of the neuron.

carried by downstream MLPs, specifically MLP14, MLP15, MLP16, and most notably MLP13.

Analysis of Neurons. The contribution of MLP13 to the logit difference is led by a single neuron (2069) (Figure 3). Recall that Gemma models use gated MLPs, which compute

$$\mathrm{GMLP}(\boldsymbol{x}) = \underbrace{\Big(g(\boldsymbol{x}\boldsymbol{W}_{\text{gate}}) \odot \boldsymbol{x}\boldsymbol{W}_{\text{in}}\Big)}_{\text{neurons}} \boldsymbol{W}_{\text{out}}, \tag{4}$$

where g is the activation function (GeGLU), $\boldsymbol{W}_{\text{gate}}, \boldsymbol{W}_{\text{in}} \in \mathbb{R}^{d \times d_{\text{mlp}}}$ read from the residual stream, and the linear combination of the rows of $\boldsymbol{W}_{\text{out}} \in \mathbb{R}^{d_{\text{mlp}} \times d}$ weighted by the neuron values is added back to the residual stream (see Appendix E for a visual description). This means that, unlike standard MLPs, neurons in GMLPs can take arbitrarily large positive and negative values. In the case of neuron 2069 in MLP13, when the neuron positively activates, their associated neuron weights (row in $\boldsymbol{W}_{\text{out}}$) write in the direction of plural verb forms (and suppress singular forms) (Table 2). On the other hand, on negative activations, the neuron weights write in the direction of singular verb forms (and suppresses plural forms). Notably, this is true for the English and the Spanish verbs in our datasets, which are present and past tenses of the verbs 'to be' and 'have', but we also observe less common non-English plural verb forms promoted on negative neuron activations. This neuron seems to read a 'subject number' signal, but where does this signal come from? A candidate is L13H7, which has a large total effect on the logit difference.

We compute the dot product between the output of attention head L13H7 at the last position and column 2069 of $\boldsymbol{W}_{\text{in}}$ ($\boldsymbol{W}_{\text{in}}[:, 2069]$) across the whole dataset and show the results in Figure 4. When the subject is singular, we get a negative dot product (activation) and promote singular verb forms (Table 2). When the subject is plural, we get positive dot product values and promote plural forms. We observe a similar pattern in other influential MLP neurons (Appendix C). We further provide evidence of the role of L13H7 by applying PCA on its outputs in the last residual stream (Figure 5). The first principal component (PC1) clearly distinguishes between singular and plural subject examples. This means that L13H7 writes into a 1-dimensional subspace where the subject number signal is encoded, from which downstream neurons read to promote the correct tokens.

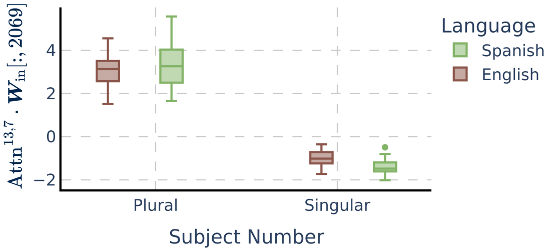


Figure 4: Dot product between the output of attention head L13H7 and the input weights of neuron 2069 in MLP13.

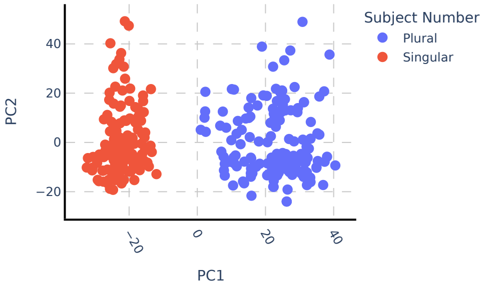


Figure 5: Projections of L13H7 outputs onto the top 2 PCs on English SVA dataset.

5. Spanish Subject-Verb Agreement Circuit

To study the subject-verb agreement task in Spanish, we follow the style of the English dataset, where we first prompt GPT4 (OpenAI et al., 2024) to generate verbs and nouns, and remove those words tokenized into multiple subwords. Then, we build similar examples to the ones in the English dataset. An example of a contrastive pair is:

Singular

Clean: El ingeniero que ayudó al cantante era

Corrupted: Los ingenieros que ayudaron al cantante ___ (5)

Plural

Spanish circuit is consistent with the English circuit. With activation patching we see a similar pattern to that

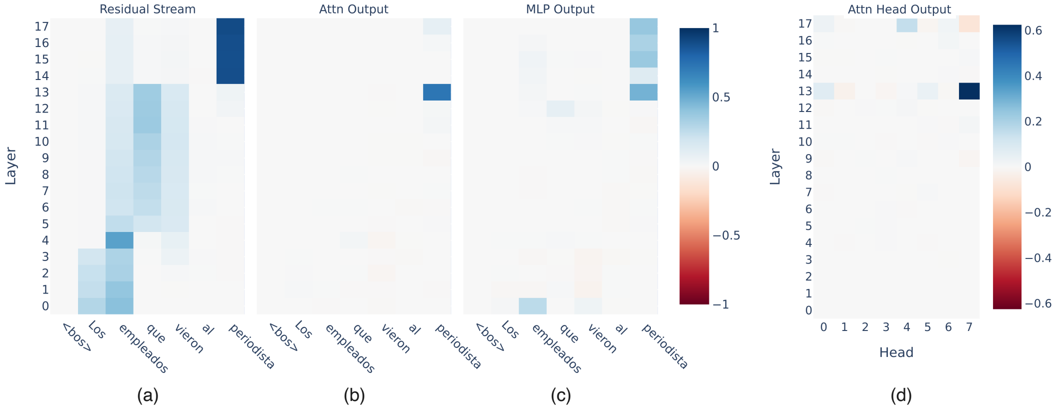


Figure 6: Spanish dataset activation patching results on the logit difference metric on (a) the residual streams (b) attention blocks outputs, (c) MLP outputs, and (d) on attention heads at the last position.

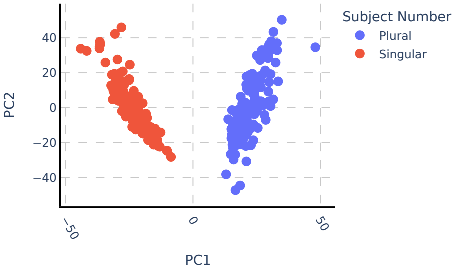


Figure 7: Projections of L13H7 outputs onto the top 2 PCs on Spanish SVA dataset.

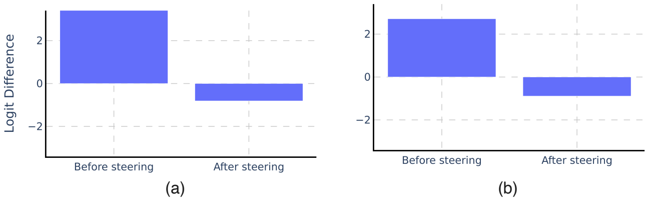


Figure 8: Average logit difference in (a) singular subject and (b) plural subject examples, before and after steering the prediction with $\mathrm{PC1}_{\mathrm{English}}$.

of the English dataset. Information from the subject flows to the last residual stream at layer 13, where the attention block shows a large effect (Figure 6). Also similarly, downstream MLPs are relevant for correctly solving the task, with MLP13 showing the highest total effect (Figure 6 (b)), while MLP15, MLP16 and MLP17 having large direct effects on the logit difference. The contribution of MLP17 is notably greater than in the English dataset (Figure 2), where we observe non-English specific neurons (Appendix D). Activation patching on individual attention heads (Figure 6 (d)) shows that, as in the English dataset, attention heads L13H7 and L17H4 have a positive influence on the correct verb form, while L17H7 influences negatively.

Activation Steering. In both languages, the same attention head (L13H7) composes with specific neurons in downstream MLPs that are responsible for the correct verb form prediction, suggesting that this head writes a 'subject number' signal, which is found via PC1 (Figures 5 and 7). Here, we study whether this direction, found in 50 English examples ($\mathrm{PC1}_{\mathrm{English}}$) has a causal effect on the model predictions, also on Spanish sentences. Specifically, we do activation steering (Turner et al., 2023; Li et al., 2023; Tigges et al., 2023) on the attention head output at the last position (n)

$$\mathrm{Attn}_n^{13,7} = \mathrm{Attn}_n^{13,7} \pm \alpha \mathrm{PC1}_{\mathrm{English}}, \qquad (6)$$

where the coefficient α scales the unit norm $\mathrm{PC1}_{\mathrm{English}}$ vector to match $\mathrm{Attn}_n^{13,7}$ norm. Results show that adding $\mathrm{PC1}_{\mathrm{English}}$ successfully flips the Spanish verb number prediction to plural (Figure 8 (a)) on examples with singular subject, and that subtracting $\mathrm{PC1}_{\mathrm{English}}$ flips the Spanish plural number prediction to singular. Furthermore, we observe that the top predicted tokens other than verbs remain mostly unchanged (see example in Appendix G).

6. Conclusion

In this work, we study how Gemma 2B solves the subject-verb agreement task in two different languages, English and Spanish. Through activation patching and direct logit attribution we find that both languages rely on circuits that are highly consistent. Moreover, we provide evidence of an attention head (L13H7) writing a 'subject number' signal as a direction from which downstream neurons read to promote the correct verb number continuation. Finally, we show this direction has a causal effect, being able to flip the predicted verb number across languages.

References

Arora, A., Jurafsky, D., and Potts, C. Causalgym: Benchmarking causal interpretability methods on linguistic tasks, 2024. URL https://arxiv.org/abs/2402.12560.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H. W., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., Reif, E., Du, N., Hutchinson, B., Pope, R., Bradbury, J., Austin, J., Isard, M., Gur-Ari, G., Yin, P., Duke, T., Levskaya, A., Ghemawat, S., Dev, S., Michalewski, H., Garcia, X., Misra, V., Robinson, K., Fedus, L., Zhou, D., Ippolito, D., Luan, D., Lim, H., Zoph, B., Spiridonov, A., Sepassi, R., Dohan, D., Agrawal, S., Omernick, M., Dai, A. M., Pillai, T. S., Pellat, M., Lewkowycz, A., Moreira, E., Child, R., Polozov, O., Lee, K., Zhou, Z., Wang, X., Saeta, B., Diaz, M., Firat, O., Catasta, M., Wei, J., Meier-Hellstern, K., Eck, D., Dean, J., Petrov, S., and Fiedel, N. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Ferrando, J., Sarti, G., Bisazza, A., and Costa-jussà, M. R. A primer on the inner workings of transformer-based language models. *ArXiv*, 2024. URL https://arxiv.org/abs/2405.00208.

Gauthier, J., Hu, J., Wilcox, E., Qian, P., and Levy, R. SyntaxGym: An online platform for targeted evaluation of language models. In Celikyilmaz, A. and Wen, T.-H. (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 70–76, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.10. URL https://aclanthology.org/2020.acl-demos.10.

Geiger, A., Lu, H., Icard, T., and Potts, C. Causal abstractions of neural networks. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W. (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 9574–9586. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/4f5c422f4d49a5a807eda27434231040-Paper.pdf.

Gemma Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M. S., Love, J., Tafti, P., Hussenot, L., Sessa, P. G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C. L., Choquette-Choo, C. A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.-C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.-B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L. L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S. L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., and Kenealy, K. Gemma: Open models based on gemini research and technology. *ArXiv*, 2024. URL https://arxiv.org/abs/2403.08295.

Geva, M., Bastings, J., Filippova, K., and Globerson, A. Dissecting recall of factual associations in auto-regressive language models. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751.

Goldberg, Y. Assessing bert's syntactic abilities. *ArXiv*, 2019. URL https://arxiv.org/abs/1901.05287.

Hanna, M., Liu, O., and Variengien, A. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In Oh, A., Neumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S. (eds.), *Advances in*

Neural Information Processing Systems, volume 36, pp. 76033–76060. Curran Associates, Inc., 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/efbba7719cc5172d175240f24be11280-Abstract-Conference.html.

Heimersheim, S. and Janiak, J. A circuit for python docstrings in a 4-layer attention-only transformer. *AI Alignment Forum*, 2023. URL https://www.alignmentforum.org/posts/u6KXXmKFbXfWzoAXn/a-circuit-for-python-docstrings-in-a-4-layer-attention-only.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., van den Driessche, G., Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., Vinyals, O., Rae, J., and Sifre, L. An empirical analysis of compute-optimal large language model training. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c1e2faff6f588870935f114ebe04a3e5-Abstract-Conference.html.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=aLLuYpn83y.

Linzen, T., Dupoux, E., and Goldberg, Y. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016. doi: 10.1162/tacl_a_00115. URL https://aclanthology.org/Q16-1037.

Meng, K., Bau, D., Andonian, A., and Belinkov, Y. Locating and editing factual associations in GPT. In Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., and Oh, A. (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 17359–17372. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Olah, C. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/mech-interp-essay.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report. *ArXiv*, 2024. URL https://arxiv.org/abs/2303.08774.

Pearl, J. *Causality*. Cambridge University Press, 2 edition, 2009. doi: 10.1017/CBO9780511803161.

Räuker, T., Ho, A., Casper, S., and Hadfield-Menell, D. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. *Arxiv*, 2023. URL https://arxiv.org/abs/2207.13243.

Stolfo, A., Belinkov, Y., and Sachan, M. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7035–7052, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.435. URL https://aclanthology.org/2023.emnlp-main.435.

Stolfo, A., Belinkov, Y., and Sachan, M. Understanding arithmetic reasoning in language models using causal mediation analysis. *Arxiv*, 2023b. URL https://arxiv.org/abs/2305.15054.

Tigges, C., Hollinsworth, O. J., Geiger, A., and Nanda, N. Linear representations of sentiment in large language models. *Arxiv*, 2023. URL https://arxiv.org/abs/2310.15154.

Turner, A. M., Thiergart, L., Udell, D., Leech, G., Mini, U., and MacDiarmid, M. Activation addition: Steering language models without optimization, 2023.

Vig, J., Gehrmann, S., Belinkov, Y., Qian, S., Nevo, D., Singer, Y., and Shieber, S. Investigating gender bias in language models using causal mediation analysis. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12388–12401. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

Wang, K. R., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Yin, K. and Neubig, G. Interpreting language models with contrastive explanations. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 184–198, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.14. URL https://aclanthology.org/2022.emnlp-main.14.

A. Logit Differences Clean and Corrupted prompts

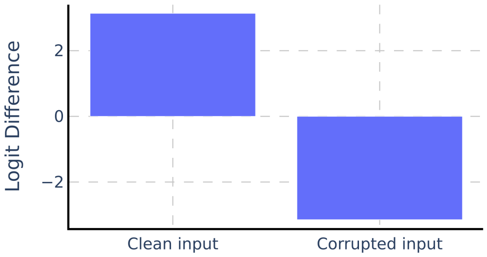


Figure 9: Logit Difference on clean and corrupted inputs. English dataset.

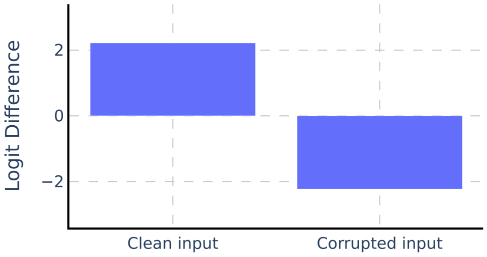


Figure 10: Logit Difference on clean and corrupted inputs. Spanish dataset.

B. Logit Difference by Neurons in MLPs

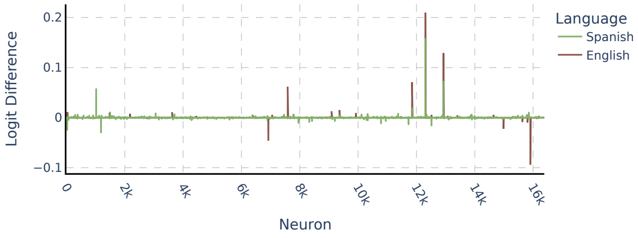


Figure 11: Average contribution to the logit difference by each neuron in MLP15.

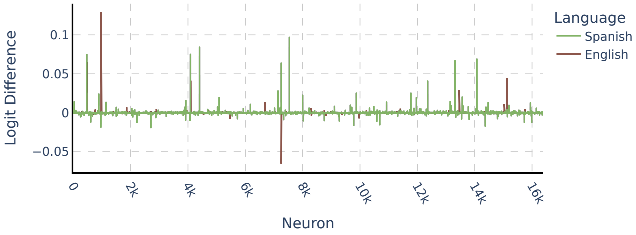


Figure 12: Average contribution to the logit difference by each neuron in MLP16.

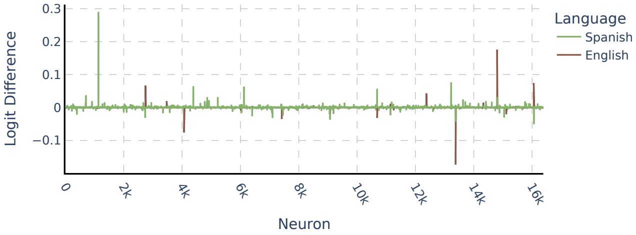


Figure 13: Average contribution to the logit difference by each neuron in MLP17.

C. Attention Head L13H7 Composition with Downstream Neurons

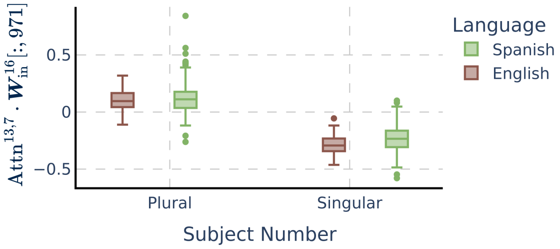


Figure 14: Values of the dot product between the output of attention head L13H7 and the input weights of neuron 971 in MLP16.

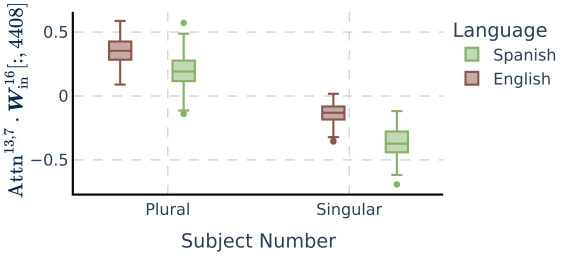


Figure 15: Values of the dot product between the output of attention head L13H7 and the input weights of neuron 4408 in MLP16.

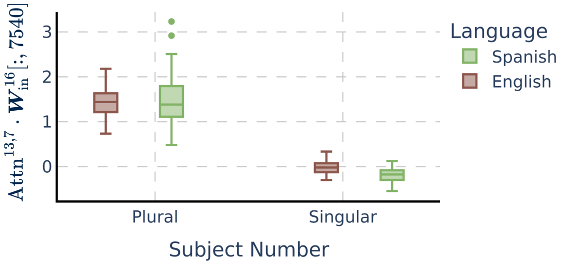


Figure 16: Values of the dot product between the output of attention head L13H7 and the input weights of neuron 7540 in MLP16.

D. Neuron 1138 in MLP17

Neuron 1138 in MLP17 only activates on sentences with plural subjects. This can be seen in Figure 18, the dot-product of $\boldsymbol{W}_{\mathrm{gate}}[:, 1138]$ with L13H7 output is negative for singular subjects, meaning that it doesn't activate. In contrast, on plural subjects the dot product of $\boldsymbol{W}_{\mathrm{gate}}[:, 1138]$ and L13H7 output is positive, and $\boldsymbol{W}_{\mathrm{in}}[:, 1138]$ is negative, meaning that the neurons fires negatively. In Table 2 we see that the promoted tokens in this case are plural verb forms of multiple non-English languages.

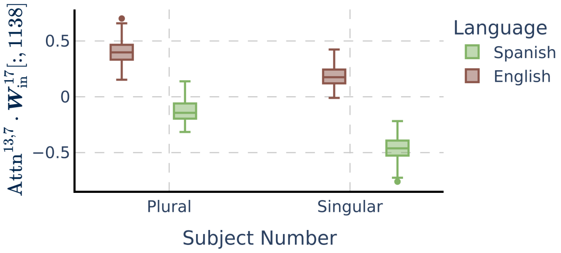


Figure 17: Values of the dot product between the output of attention head L13H7 and the input weights $\boldsymbol{W}_{\mathrm{in}}$ of neuron 1138 in MLP17.

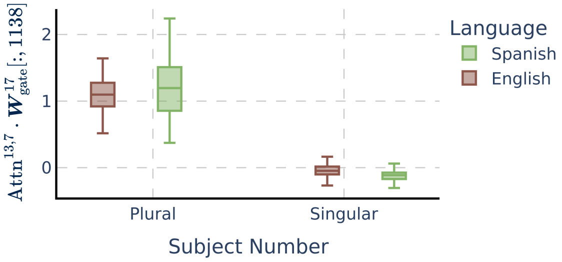


Figure 18: Values of the dot product between the output of attention head L13H7 and the input weights $\boldsymbol{W}_{\mathrm{gate}}$ of neuron 1138 in MLP17.

Top Promoted Tokens *Negative* Neuron Activation
'abbiano', ' avevano', ' sembrano', ' avrebbero', ' continúan', **' fossero', ' possano', ' poseen', ' tenham', ' terão'** **' ont', ' constituyen', ' lograron'**

Table 2: Top promoted tokens by neuron 1138 in MLP17 based on negative neuron activations.

E. MLP and Gated MLP (GMLP)

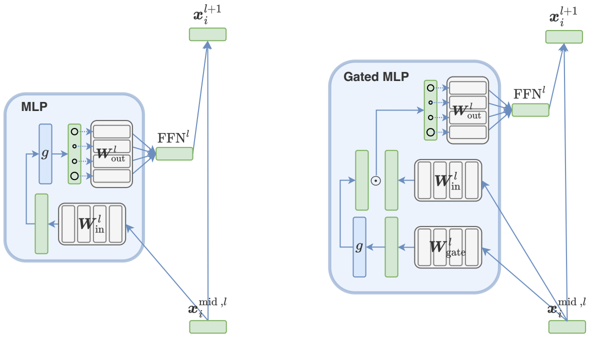

$\boldsymbol{x}_i^{l+1}$
MLP
g
$\boldsymbol{W}_{\mathrm{out}}^{l}$
FFN^l
$\boldsymbol{W}_{\mathrm{in}}^{l}$
$\boldsymbol{x}_i^{\mathrm{mid},l}$
$\boldsymbol{x}_i^{l+1}$
Gated MLP
FFN^l
$\boldsymbol{W}_{\mathrm{out}}^{l}$
$\odot$
$\boldsymbol{W}_{\mathrm{in}}^{l}$
g
$\boldsymbol{W}_{\mathrm{gate}}^{l}$
$\boldsymbol{x}_i^{\mathrm{mid},l}$

Figure 19: A comparison between the operations performed by the standard MLP and the Gated MLP (GMLP) found in Gemma models.

F. Attention Patterns Main Heads

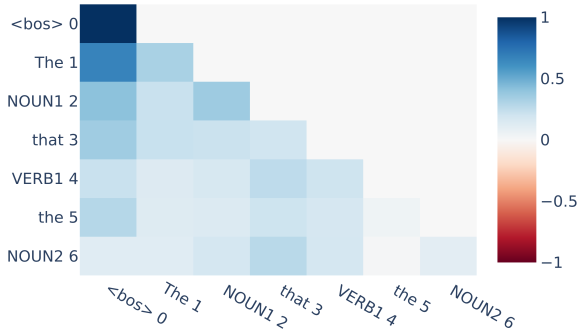


Figure 20: L13H7 average attention patterns (output-value weighted) across the English dataset.

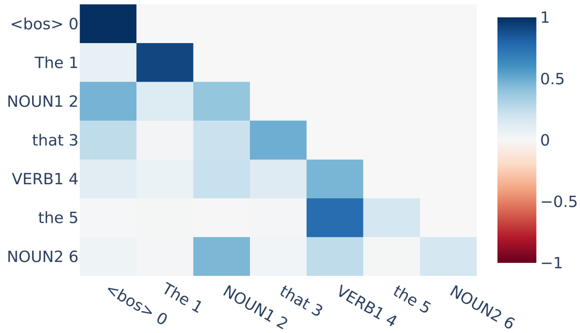


Figure 21: L17H4 Average attention patterns (output-value weighted) across the English dataset.

G. Example Top Predicted Tokens in Steering Experiment

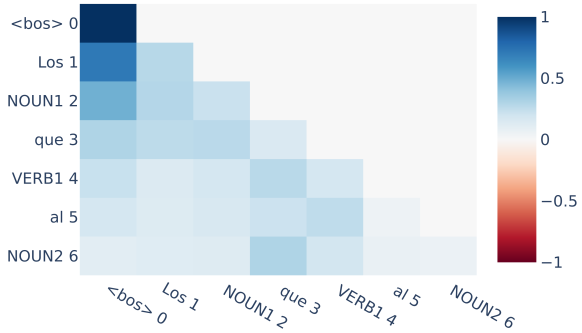


Figure 22: L13H7 average attention patterns (output-value weighted) across the Spanish dataset.

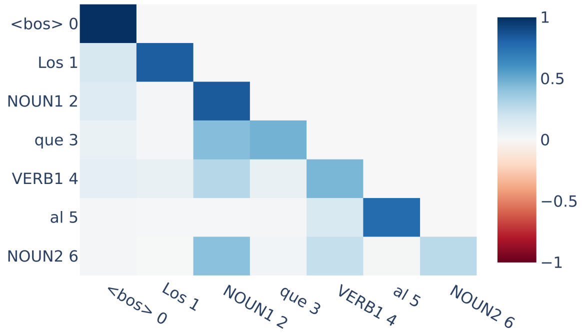


Figure 23: L17H4 Average attention patterns (output-value weighted) across the English dataset.

Top 10 Predicted Tokens Before Steering
’ se’, ’ de’, ’ en’, **’ era’**, ’ y’, ’ del’, ’ ’, ’,’, **’ es’**, **’ fue’**
Top 10 Predicted Tokens After Steering
’ de’, ’ se’, ’ en’, ’ y’, ’ ’, ’ del’, **’ son’**, ’,’, ’ no’, **’ eran’**

Table 3: Top 10 Predicted Tokens before and after steering a spanish example. In bold are shown spanish forms of the verb ‘to be’.