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

Javier Ferrando[1] **Marta R. Costa-jussà**[2]
[1]Universitat Politècnica de Catalunya
[2]FAIR, Meta
jferrandomonsonis@gmail.com

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 MLPs. 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.[1]

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 and 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)[2], 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 has

Corrupted: The executives that embarrassed the manager ___ (1)

Plural

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.

[1]Code will be released upon acceptance.

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

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 10115–10125
November 12-16, 2024

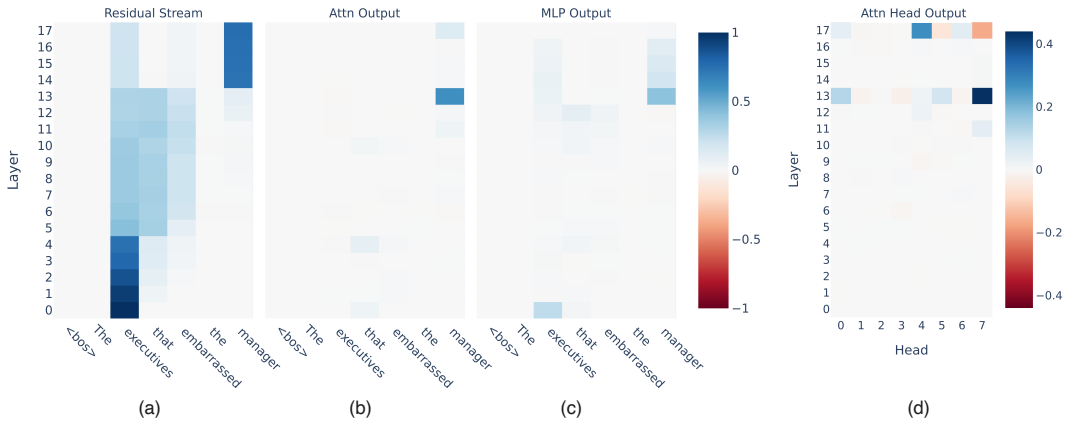

Figure 1: English dataset activation patching results on the logit difference metric on (a) the residual streams (b) attention blocks outputs, (c) MLP outputs, and (d) on attention heads at the last position.

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 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 and Neubig, 2022; Wang et al., 2023) of the two possible verb continuations (g and b):

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

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:

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

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.

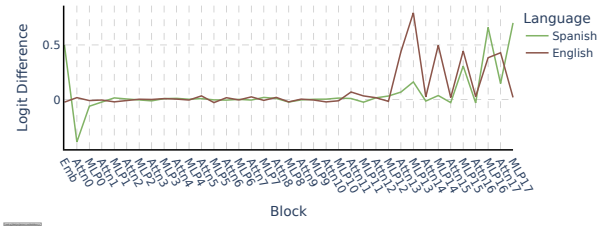

Figure 2: Average contribution to the logit difference by each model component.

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[3] 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 (b)), 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

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

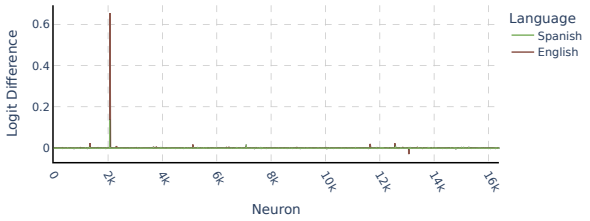

Figure 3: Average contribution to the logit difference by each neuron in MLP13.

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.

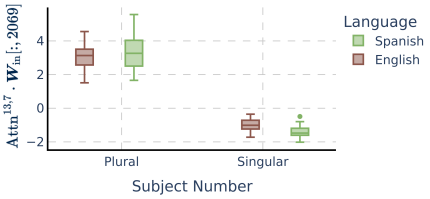

Figure 4: Dot product of the output of attention head L13H7 and the input weights of neuron 2069 in MLP13.

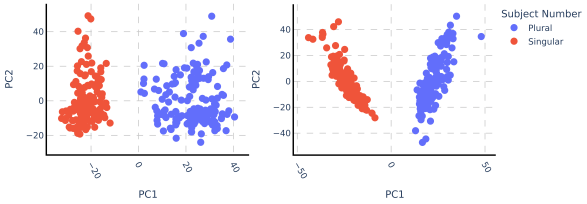

Figure 5: Projections of L13H7 outputs onto the top 2 PCs on English (left) and Spanish (right) dataset.

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 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

$$\text{GMLP}(\boldsymbol{x}) = \underbrace{\left(g(\boldsymbol{x}\boldsymbol{W}_{\text{gate}}) \odot \boldsymbol{x}\boldsymbol{W}_{\text{in}}\right)}_{\text{neurons}} \boldsymbol{W}_{\text{out}}, \quad (4)$$

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 1). 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 1). 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.

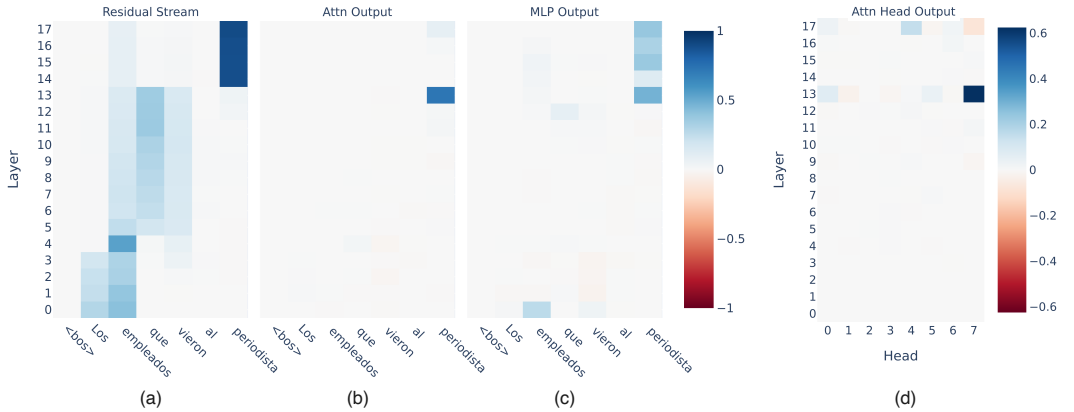

Figure 6: Spanish dataset activation patching results on the logit difference metric on (a) the residual streams (b) attention blocks outputs, (c) MLP outputs, and (d) on attention heads at the last position.

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 ___

Plural

(5)

Spanish circuit is consistent with the English circuit. With activation patching we see a similar pattern to that 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 (Figure 5). Here, we study whether this direction, found in 50 English examples ($\text{PC1}_{\text{English}}$) 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)

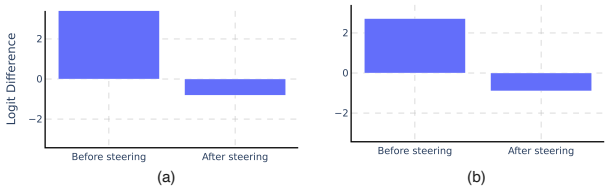

Figure 7: Spanish average logit difference in (a) singular subject and (b) plural subject examples, before and after steering the prediction with $\text{PC1}_{\text{English}}$.

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

where the coefficient α scales the unit norm $\text{PC1}_{\text{English}}$ vector to match $\text{Attn}_n^{13,7}$ norm. Results show that adding $\text{PC1}_{\text{English}}$ successfully flips the Spanish verb number prediction to plural (Figure 7 (a)) on examples with singular subject, and that subtracting $\text{PC1}_{\text{English}}$ 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 Related Work

Our work builds upon and extends several key studies in the field of probing neural language models for syntactic knowledge, particularly focusing on agreement mechanisms in multilingual contexts.

Causal probing studies have provided evidence for the existence of specific syntactic agreement neurons in language models (Finlayson et al., 2021; De Cao et al., 2022; Lakretz et al., 2019). Furthermore, research has demonstrated that models like BERT (Devlin et al., 2019) rely on a linear encoding of grammatical number to solve the number agreement task (Lasri et al., 2022). However, these studies have primarily focused on monolingual models, leaving a gap in our understanding of multilingual contexts.

Moving beyond monolingual studies, Chi et al. (2020) investigated universal grammatical relations in multilingual BERT. They developed a structural probe (Hewitt and Manning, 2019), learning a linear mapping to a syntactic subspace where syntactic features overlap between languages. This study demonstrated the potential for identifying cross-lingual syntactic similarities in multilingual models. Mueller et al. (2022) conducted a causal analysis of syntactic agreement neurons in multilingual language models. Via counterfactual interventions, they discovered significant overlap between languages in terms of neurons that causally influence syntactic agreement.

7 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.

8 Limitations

We recognize two main limitations in our study. First, we focused solely on Gemma 2B model. This choice is motivated by its large vocabulary size, which aids in studying multilingual settings. Results obtained on Gemma 2B do not guarantee they generalize to other models. Second, our study is limited to two languages: English and Spanish. Although we identified a language-agnostic subject number direction in the model's representation space, demonstrating its generality across these two languages, we cannot conclude that the same applies to all other languages, particularly those that are more linguistically distant.

Acknowledgments

The authors acknowledge the anonymous reviewers for the for their valuable insights and constructive feedback. We also thank Neel Nanda for the positive feedback and encouragement to convert this project into a publication. Javier Ferrando is supported by the Spanish Ministerio de Ciencia e Innovación through the project PID2019-107579RB-I00 / AEI / 10.13039/501100011033.

References

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *Preprint*, arXiv:2402.12560.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira,

Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. 2022. Sparse interventions in language models with differentiable masking. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 16–27, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A primer on the inner workings of transformer-based language models.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc.

Google Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. 2024. Gemma: Open models based on gemini research and technology. *ArXiv*.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *ArXiv*.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Advances in Neural Information Processing Systems*, volume 36, pages 76033–76060. Curran Associates, Inc.

Stefan Heimersheim and Jett Janiak. 2023. A circuit for python docstrings in a 4-layer attention-only transformer. *AI Alignment Forum*.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero,

Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. 2022. An empirical analysis of compute-optimal large language model training. In *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc.

Yair Lakretz, German Kruszewski, Theo Desbordes, Dieuwke Hupkes, Stanislas Dehaene, and Marco Baroni. 2019. The emergence of number and syntax units in LSTM language models. 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 11–20, Minneapolis, Minnesota. Association for Computational Linguistics.

Karim Lasri, Tiago Pimentel, Alessandro Lenci, Thierry Poibeau, and Ryan Cotterell. 2022. Probing for the usage of grammatical number. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8818–8831, Dublin, Ireland. Association for Computational Linguistics.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2023. Inference-time intervention: Eliciting truthful answers from a language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, pages 17359–17372. Curran Associates, Inc.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Chris Olah. 2022. Mechanistic interpretability, variables, and the importance of interpretable bases. *Transformer Circuits Thread*.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Fe-

lipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *ArXiv*.

Judea Pearl. 2009. *Causality*, 2 edition. Cambridge University Press.

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. 2023. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. *Arxiv*.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023a. A mechanistic interpretation of arithmetic reasoning in language models using causal mediation analysis. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7035–7052, Singapore. Association for Computational Linguistics.

Alessandro Stolfo, Yonatan Belinkov, and Mrinmaya Sachan. 2023b. Understanding arithmetic reasoning in language models using causal mediation analysis. *Arxiv*.

Curt Tigges, Oskar John Hollinsworth, Atticus Geiger, and Neel Nanda. 2023. Linear representations of sentiment in large language models. *Arxiv*.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. Activation addition: Steering language models without optimization. *Preprint*, arXiv:2308.10248.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

A Logit Differences Clean and Corrupted prompts

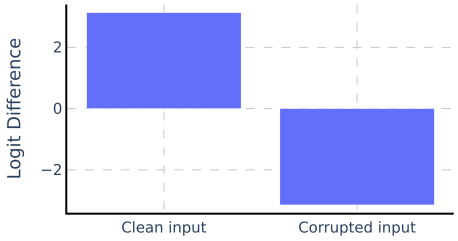

Figure 8: Logit Difference on clean and corrupted inputs. English dataset.

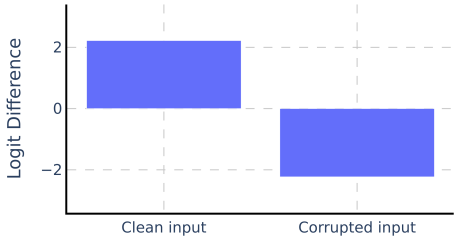

Figure 9: Logit Difference on clean and corrupted inputs. Spanish dataset.

B Logit Difference by Neurons in MLPs

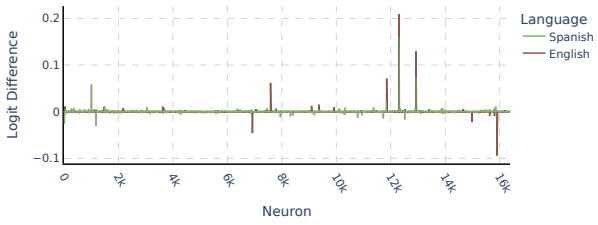

Figure 10: Average contribution to the logit difference by each neuron in MLP15.

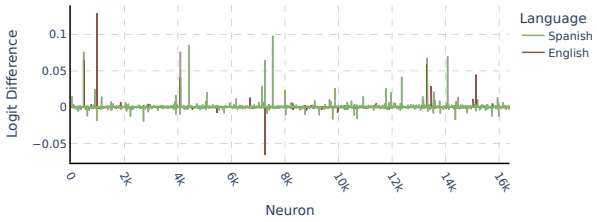

Figure 11: Average contribution to the logit difference by each neuron in MLP16.

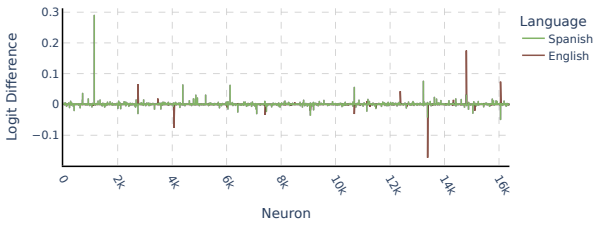

Figure 12: Average contribution to the logit difference by each neuron in MLP17.

C Attention Head L13H7 Composition with Downstream Neurons

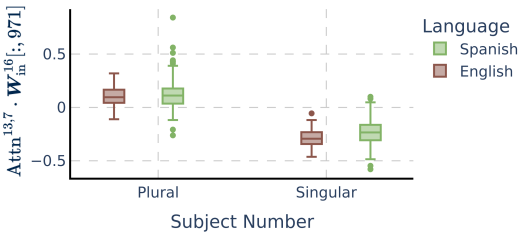

Figure 13: Values of the dot product between the output of attention head L13H7 and the input weights of neuron 971 in MLP16.

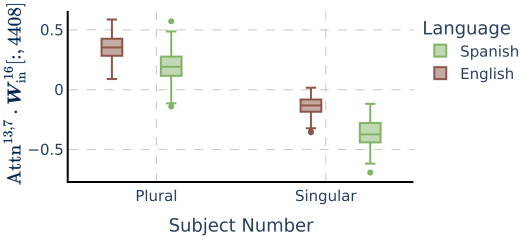

Figure 14: Values of the dot product between the output of attention head L13H7 and the input weights of neuron 4408 in MLP16.

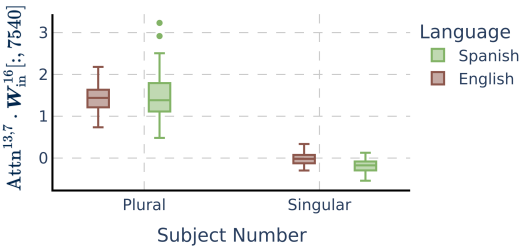

Figure 15: Values of the dot product between the output of attention head L13H7 and the input weights of neuron 7540 in MLP16.

D Neuron 1138 in MLP17

Neuron 1138 in MLP17 only activates on sentences with plural subjects. This can be seen in Figure 17, the dot-product of $\boldsymbol{W}_{\text{gate}}[:, 1138]$ with L13H7 output is negative for singular subjects, meaning that it doesn't activate. In contrast, on plural subjects the dot product of $\boldsymbol{W}_{\text{gate}}[:, 1138]$ and L13H7 output is positive, and $\boldsymbol{W}_{\text{in}}[:, 1138]$ 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.

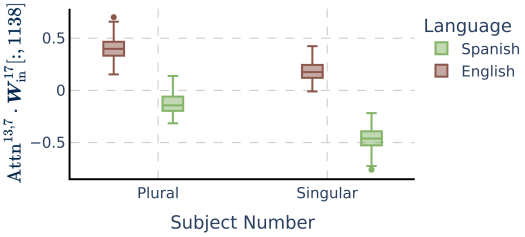

Figure 16: Values of the dot product between the output of attention head L13H7 and the input weights $\boldsymbol{W}_{\text{in}}$ of neuron 1138 in MLP17.

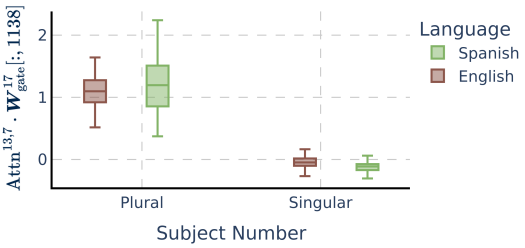

Figure 17: Values of the dot product between the output of attention head L13H7 and the input weights $\boldsymbol{W}_{\text{gate}}$ of neuron 1138 in MLP17.

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)

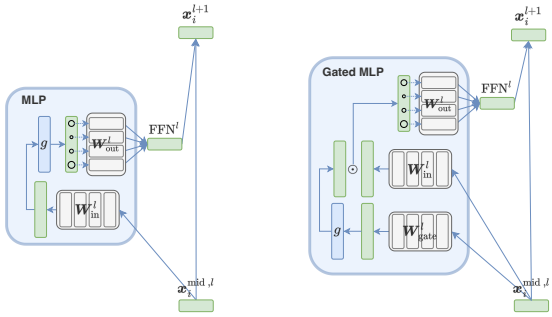

Figure 18: A comparison between the operations performed by the standard MLP and the Gated MLP (GMLP) found in Gemma models.

F Attention Patterns Main Heads

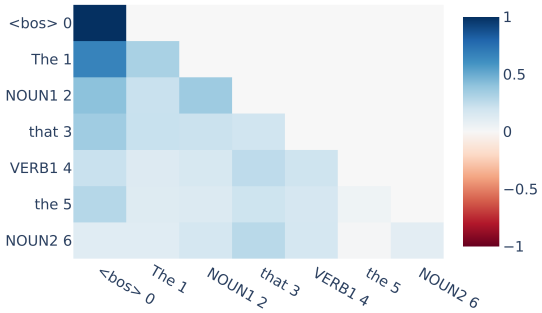

Figure 19: L13H7 average attention patterns (output-value weighted) across the English dataset.

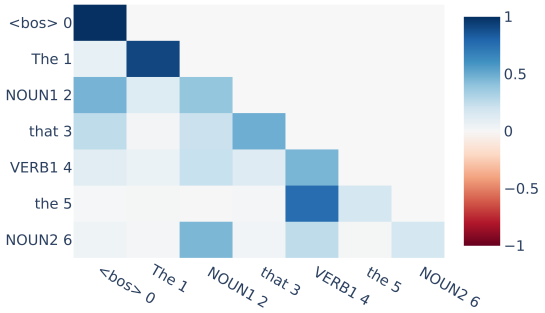

Figure 20: L17H4 Average attention patterns (output-value weighted) across the English dataset.

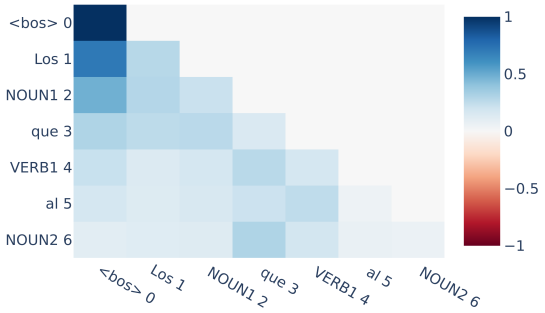

Figure 21: L13H7 average attention patterns (output-value weighted) across the Spanish dataset.

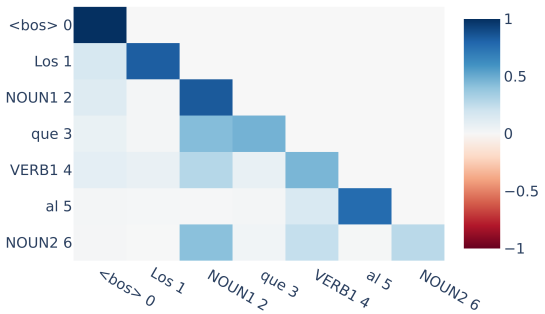

Figure 22: L17H4 Average attention patterns (output-value weighted) across the English dataset.

G Example Top Predicted Tokens in Steering Experiment

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’.