Limitations

- **Main limitation:** We presented an extended study of *how* two pre-trained language models process human and machine-generated inputs, but we did not provide an account of *why* we are observing the processing differences we are seeing. We noticed, for example, that *M-cont* prompts activate units associated to punctuation marks and special characters. We do not know, however, in which way these units contribute to retrieving the correct answer in the target semantic tasks, nor how the optimization procedure chances upon them. This is our priority for future work.

- Our work is limited to the OPT family of models trained on the English language, to the LAMA semantic tasks and to the AutoPrompt and OptiPrompt prompt extraction methods. A straightforward direction for future work is to extend our analysis to more models (including instruction-tuned models, as instruction tuning might have a significant impact on how models respond to unnatural input), languages, data-sets and prompt extraction algorithms.

Ethics Statement

The advent of publicly accessible LM interfaces such as ChatGPT has heated up the debate around the broader impact of LMs. While there is a variety of possible societal issues to consider (Weidinger et al., 2022), we believe that a better understanding of how LMs process information is a crucial part of bias and harm containment. If we do not understand the models, we cannot control their behaviour, and we are exposed to intentional adversarial attacks and other forms of unintentional model misuse. The very existence of completely opaque but empirically effective machine-generated prompts is proof of how counterintuitive the behaviour of LMs can be, and of how little we understand them. We thus believe that our investigations of "unnatural language processing" fit well into the broader program of improving our scientific understanding of LMs, in order to make them more predictable, controllable and, ultimately, safer.

Acknowledgements

We thank Emmanuel Chemla, Emily Cheng, Nathanaël Carraz Rakotonirina, Xavier Suau, the members of the UPF COLT lab, the members of the Barcelona Apple Machine Learning Research group and the participants in the EviL seminar for helpful feedback and suggestions. Our work was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 101019291). This paper reflects the authors' view only, and the ERC is not responsible for any use that may be made of the information it contains.

References

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill*. Https://distill.pub/2020/circuits.

Damai Dai, Li Dong, Yaru Hao, Zhifang Sui, Baobao Chang, and Furu Wei. 2022. Knowledge neurons in pretrained transformers. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8493–8502.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. 2022. RLPrompt: Optimizing discrete text prompts with reinforcement learning. In *Proceedings of EMNLP*, pages 3369–3391, Abu Dhabi, United Arab Emirates.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Hady ElSahar, Pavlos Vougiouklis, Arslen Remaci, Christophe Gravier, Jonathon S. Hare, Frédérique Laforest, and Elena Paslaru Bontas Simperl. 2018. T-rex: A large scale alignment of natural language with knowledge base triples. In *International Conference on Language Resources and Evaluation*.

Stephanie Evert. 2005. *The Statistics of Word Cooccurrences*. Ph.D dissertation, Stuttgart University.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of EMNLP*, pages 5484–5495, Online and Punta Cana, Dominican Republic.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. In *Proceedings of the EMNLP BlackboxNLP Workshop*, pages 240–248, Brussels, Belgium.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2022. Demystifying prompts in language models via perplexity estimation. https://arxiv.org/abs/2212.04037.

Ashim Gupta, Giorgi Kvernadze, and Vivek Srikumar. 2021. BERT and family eat word salad: experiments with text understanding. In *Proceedings of AAAI*, pages 12946–12954, Online.

Yoichi Ishibashi, Danushka Bollegala, Katsuhito Sudoh, and Satoshi Nakamura. 2023. Evaluating the robustness of discrete prompts. In *Proceedings of EACL*, pages 2373–2384, Dubrovnik, Croatia.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Daniel Khashabi, Xinxi Lyu, Sewon Min, Lianhui Qin, Kyle Richardson, Sean Welleck, Hannaneh Hajishirzi, Tushar Khot, Ashish Sabharwal, Sameer Singh, and Yejin Choi. 2022. Prompt waywardness: The curious case of discretized interpretation of continuous prompts. In *Proceedings of NAACL*, pages 3631–3643, Seattle, WA.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of EMNLP*, pages 3045–3059, Punta Cana, Dominican Republic.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher Manning, Christopher Ré, Diana Acosta-Navas, Drew Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic evaluation of language models. https://arxiv.org/abs/2211.09110.

Zachary C Lipton. 2018. The mythos of model interpretability. *Communications of the ACM*, 61(10):36–43.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

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.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Joon-Sung Park, Joseph O'Brien, Carrie Cai, Meredith Ringel Morris, Percy Liang, and Michael Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. https://arxiv.org/abs/2304.03442.

Fabian Pedregosa, Gaël Varoquaux, , Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings EMNLP*, pages 2463–2473, Hong Kong, China.

Thang Pham, Trung Bui, Long Mai, and Anh Nguyen. 2021. Out of Order: How important is the sequential order of words in a sentence in natural language understanding tasks? In *Findings of ACL*, pages 1145–1160, Online.

Nathanaël Rakotonirina, Roberto Dessì, Fabio Petroni, Sebastian Riedel, and Marco Baroni. 2023. Can discrete information extraction prompts generalize across language models? In *Proceedings of ICLR*, Kigali, Rwanda. Published online: https://openreview.net/group?id=ICLR.cc/2023/Conference.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, et al. 2021. Hopfield networks is all you need. In *International Conference on Learning Representations 2021*.

Taylor Shin, Yasaman Razeghi, Robert Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting knowledge from language models with automatically generated prompts. In *Proceedings of EMNLP*, pages 4222–4235, Online.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. 2021a. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. In *Proceedings of EMNLP*, pages 2888–2913, Punta Cana, Dominican Republic.

Koustuv Sinha, Prasanna Parthasarathi, Joelle Pineau, and Adina Williams. 2021b. UnNatural Language Inference. In *Proceedings of ACL*, pages 7329–7346, Online.

Chelsea Voss, Nick Cammarata, Gabriel Goh, Michael Petrov, Ludwig Schubert, Ben Egan, Swee Kiat Lim, and Chris Olah. 2021. Visualizing weights. *Distill*. Https://distill.pub/2020/circuits/visualizing-weights.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing NLP. In *Proceedings of EMNLP*, pages 2153–2162, Hong Kong, China.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of NAACL*, pages 2300–2344, Seattle, WA.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sasha Brown, Zac Kenton, Will Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of risks posed by language models. In *Proceedings of FAccT*, pages 214–229, Seoul, Korea.

Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. 2023. Socratic models: Composing zero-shot multimodal reasoning with language. In *Proceedings of ICLR*, Kigali, Rwanda. Published online: https://openreview.net/group?id=ICLR.cc/2023/Conference.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open pre-trained transformer language models. https://arxiv.org/abs/2205.01068.

Zexuan Zhong, Dan Friedman, and Danqi Chen. 2021. Factual probing is [MASK]: Learning vs. learning to recall. In *Proceedings of NAACL*, pages 5017–5033, Online.

Andy Zou, Zifan Wang, Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. http://arxiv.org/abs/2307.15043.