A Plain English Justification for Geometric Networks

Modern AI models are complicated. The largest language models (e.g. Claude, Gemini , GPT-4, Llama) contain many billions of trainable parameters (Zhao et al., 2023). These models are currently used in domains like code generation, document-writing, and mathematical deduction. A risk exists that these models produce incorrect or misaligned outputs. Humans have several methods for ensuring outputs align with benchmarks for correctness and preference-alignment.

In-sample statistics

When one can reliably measure the AI’s performance, one can construct benchmarks and compare a model’s performance statistics against the benchmarks. We can answer questions like:

1. Was that output correct?
2. Will that output reliably repeat given several of the same inputs?
3. Can I obtain (frequentist) parameter estimates the network’s outputs?

When an input-output pair appears many times in the data, we say the mapping is ‘in-sample’. This means we have reliable data about how inputs relate to outputs. In these instances, large language models perform reliably (simple code generation, poem writing, undergraduate mathematics questions). Seemingly complicated mappings are well documented (e.g. X = spacecraft telemetry, Y = crash probability); perhaps seemingly simple but undocumented mappings exist.

Out-of-sample statistics

When limited data on input-output pairs exists, we say the mapping is ‘out-of-sample’. In these instances, one is generally less confident in the model’s performance (Perez et al., 2021; Zhao et al. 2021). A promising and human-like method for improving out of sample performance is to use logic. With logic, one deduces or infers a correct mapping without ‘regurgitating’ prior knowledge. Alpha-Geometry and GPT-o1 are impressive examples of logical reasoning in AI models (Trinh et al, 2024; OpenAI, 2024).

In extreme examples, however, the problem may be so complicated that its proof is not human understandable. Think of a proof with more than 1000 steps or one in which a given statement has a complex dependence on previous statements. Perhaps the proof that a rocket will travel faster than light, or that a medicine will cure aging is too complex for human comprehension. In such instances, humans can’t verify that what a model says or does is correct.

Worse still, if the model is clever enough to know the proof is beyond human understanding, and if the model becomes misaligned with human preferences, the model may *deliberately* deceive its human architect. In such instances, logical proof becomes ineffective in ensuring model behaviours are correct and well-aligned.

Architectural Indicators

Human may not understand the task or the model’s task performance. In such instances, creative solutions exist to give humans clues about what a model is doing. Some have proposed secondary models that humans can trust to check more complicated models. One imagines a hierarchy of increasingly intelligent brains verifying each other (Rott Shaham et al, 2024).

A simpler method is to examine the complex architecture directly and look for ‘tells’ about how the model performs. Anthropic’s safety team focuses on mechanistic interpretability, the mathematical examination of weights, propagations, and backpropagated gradients (amongst other things) to predict and explain outputs (Anthropic, 2023; Olssen et al, 2023). This work is extremely sophisticated. For example, in causal tracing researchers varying a single model parameter to examine how the output changes, thus revealing what output sub-parts a single parameter (or sub-network) is responsible for. Later activations in this parameter or sub-network may inform the user about the model’s behaviour (e.g. the ‘lying’ sub-network is activated before the model lies).

Geometric networks

Mechanistic interpretability is great for highly-specialised machine learning engineers, but AI models are already affecting society as a whole. Not everyone has an advanced mathematical degree, but most people have some exposure to AI. So, a large proportion of the population may lack an understanding of models that greatly affect their lives. The question arises: “can we construct these models so a general audience can understand critical safety features, like the ‘lying network’?”

Humans understand their world through sensory information (sight, sound, touch, smell, taste). Rearranging neural networks so they contain more accessible sensory information may raise understanding among the general public. A first step could be to create 3D models that people can see and interact with, to understand misaligned behaviours like deception as they arise. Then, even a child with an ability to understand colours and shapes can glean information about the network’s function. One imagines a future where people check AI ‘brains’ using screen projections or VR goggles like mechanics checking car parts.

Interestingly, a task’s complexity might not relate to the ‘heuristic tell’ that indicates misalignment. For example, whether a humanoid robot intends to pull someone’s chair away as they sit down, or whether a language model attempts to defraud a national government, the architectural ‘tell’ or sub-network activation that precedes these actions might look very similar (and hopefully very simple). This similarity is present in human brains where simultaneous activation in the prefrontal cortex, amygdala, and parietal cortex correlates strongly with lying (Ofen et al, 2017).

References

Anthropic. (2023). Mapping the Mind of a Large Language Model. Retrieved from https://www.anthropic.com/research/mapping-mind-language-model OpenAI. (2024). *Introducing OpenAI o1*. Retrieved from <https://openai.com/o1/>

Ofen, N., Whitfield-Gabrieli, S., Chai, X.J., Schwarzlose, R.F. and Gabrieli, J.D., 2017. Neural correlates of deception: Lying about past events and personal beliefs. *Social cognitive and affective neuroscience*, *12*(1), pp.116-127.

Olsson, C., et al. (2023). Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. Retrieved from https://transformer-circuits.pub/2023/monosemantic-features/index.html

Perez, E., Kiela, D. and Cho, K., 2021. True few-shot learning with language models. *Advances in neural information processing systems*, *34*, pp.11054-11070.

Rott Shaham, T., Schwettmann, S., Wang, F., Rajaram, A., Hernandez, E., Andreas, J., & Torralba, A. (2024). *A Multimodal Automated Interpretability Agent*. Massachusetts Institute of Technology. Retrieved from <https://multimodal-interpretability.csail.mit.edu/maia/>

Trinh, T.H., Wu, Y., Le, Q.V., He, H. and Luong, T., 2024. Solving olympiad geometry without human demonstrations. *Nature*, *625*(7995), pp.476-482.

Zhao, Z., Wallace, E., Feng, S., Klein, D. and Singh, S., 2021, July. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning* (pp. 12697-12706). PMLR.

Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z. and Du, Y., 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.