



The meta-reasoning illusion

Description

The ability of an LLM to provide an account of the processes by which it came up with its results, as in the inventory table I showed in the [last posting](#), is indeed convincing. But apparently it is an illusion. I can't yet find an article on the subject, so I have come to rely on ChatGPT's own account of its explanatory processes. I attach the [transcript](#) of an interaction with ChatGPT. After some prompting, the platform disclosed its method.

It may seem like I **first performed a deep architectural analysis, then consciously reconstructed my reasoning**. In reality, **both steps were separate token prediction processes**, but the latter was fine-tuned to **appear like a genuine retrospective analysis**. This fine-tuning creates the **illusion of intentional reasoning**, even though it's ultimately **just a probabilistic language process**.

The platform asserted that this is how human beings often process their understanding anyway.

This is similar to how humans sometimes **rationalize decisions post hoc**—we often act based on intuition, then construct an explanation **after the fact** when asked to articulate our reasoning.

In further interaction with ChatGPT I discovered that it takes two separate "runs" or simulations with the model (1) to generate a decision, and (2) to generate an explanation of how it derives, or justifies that outcome.

The style of explanation is also under the influence of context which includes the user statements that prompt the explanation. For example, after its initial explanation, I asked for a variation that incorporated the elements of the Transformer method. I wanted a technical explanation of what actually happened in terms of semantic embeddings, focus of attention, attention heads, positional sequencing etc. (See post [Extending large language models](#).) But even that explanation was a post-hoc product resulting from training and fine tuning the neural network on examples of such explanations.

As a further illustration, I noted that the model assumed the building facade had “Maltese balconies.” I asked how did the model deduce that the building was in Malta. My prompt gave the game away, or at least it prompted the model with the assumption that buildings with Maltese balconies are in Malta. My prompt thereby pushed the model to come up with a post-hoc narrative about the various elements in the facade that contributed to the inference that the image I provided was of a building in Malta.

This confirms that the model is not designed to perform explicit, stepwise deduction as in a rule-based or logic-based system. Instead, the patterns encoded in its training data produce high-probability token sequences that align with expected explanations. When prompted to explain its reasoning, the model generates a justification that fits within the learned discourse of “rational analysis,” a discourse of the kind evident in its training data.

ChatGPT concurred that in the case of the Malta inference, the retrospective account gave the appearance of deliberate reasoning. But it was in fact just a structured response shaped by probability distributions rather than calculated deductions. This underscores the nature of the meta-reasoning illusion: the sense of intentional cognition where there is none.

Are we ok with the illusion?

This phenomenon aligns with psychological research on the loose coupling between decision-making and its justification. Studies in cognitive science suggest that human explanations are often only **quasi-dependent** on decision-making outcomes, meaning that explanations follow decisions rather than determining them. In so far as neural networks approximate aspects of human cognition, their behaviour reinforces this view.

Importantly, we humans feel we need explanations—not just any explanations, but ones that align with **cultural expectations, values, and biases**. The drive for coherence can make us prefer structured narratives over uncertainty, even if those narratives are post hoc constructions.

Some argue that in many cases, **any explanation is better than none**, satisfying our craving for causality even when calculated reasoning is absent. The meta-reasoning illusion thus highlights not just an AI phenomenon but an important aspect of human cognitive tendencies. Put another way, it aligns with how we interpret. See post on [hermeneutics](#).

Note

- Featured image: I asked Dall-e to generate an image of an old Maltese balcony in need of repair.



Category

1. Artificial Intelligence

Tags

1. architecture
2. hermeneutics
3. reasoning
4. urbanism

Date Created

March 3, 2025

Author

rcoyne99