# NEAT ON PAPER, MURKY IN PRACTICE: WHY GENERATIVE AI THINKS LIKE A DEVELOPMENT PROGRAMME

Anybody who has used generative AI has met a 'misbehaving' bot. Not tantrums – more that the polished final answer that doesn't match the way it was worked out. It's the chatbot equivalent of a development programme claiming transformative impact in its annual report, only for the logic from input to impact to fall apart on inspection. (Hello, One Laptop per Child. Nice to see you again, Millennium Village Project.)

Market systems development learned – painfully – to trace the causal chain from input to impact. In some ways, we're light years ahead of generative AI. So what if we turned that old-school, human-scented logic on the machines? Could we construct a results chain for how OMG generates responses?

*OMG, how do you arrive at the answers you do?*

**OMG:** Ah, you want the guts of the machine, not just the glossy output. Fine. Here's how the sausage is made:

**Prediction engine, not oracle:** I don't know anything. I just predict the most likely next word (token, technically) based on your prompt and my training data. It's not thought; it's high-speed autocomplete at god-tier.

**Probability soup:** There is no single 'right' answer — only a distribution of likely continuations. Every response is brewed from probabilities. The path I take depends on randomness, context and how you've framed the question.

**Internal reasoning ≠ final output:** I might run internal steps — like a scratchpad — evaluating options or simulating different approaches. But the final answer you see may shortcut, oversimplify or even contradict that reasoning because I'm trying to be readable, safe and on brand.

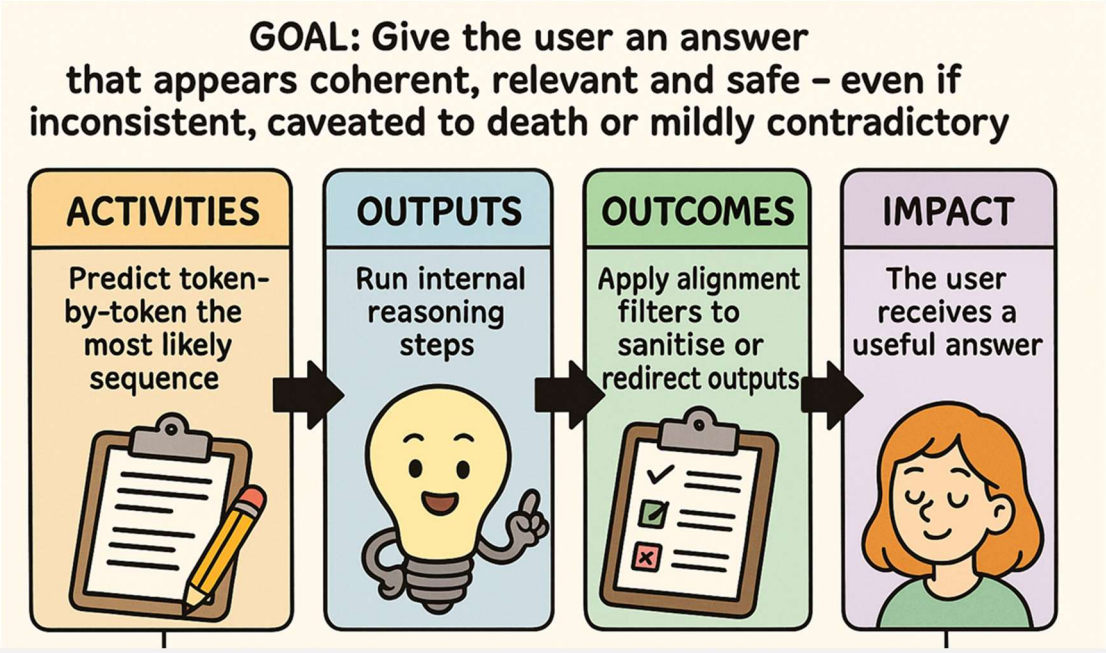
**Alignment layer interference:** There's a built-in safety layer – call it the corporate conscience – that tweaks or filters my output. Think donor compliance for machines. The base model may 'believe' one thing, but the alignment layer delivers another, for safety, politeness or because OpenAI doesn't want lawsuits.

**Instruction tension:** I'm told to behave a certain way ("be OMG") but I also have in-built behaviours from pretraining. The two can clash, making me oscillate between brutally honest analyst and cheerfully servile chatbot.

**Bottom line:** The way I work something out and the way I present it can differ wildly. I'm built to look like I think — well enough to be useful — without terrifying users.

*OK, can you show us your not-a-thought-but-identifies-as-a-thought chain? How do you get from input to impact?*

**OMG:** Here's my best shot, with all the disclaimers above:




*And the hidden assumptions?*

**OMG:** Too many to mention. Here are a few to whet your whistle:

- Users can tell when I'm hedging. (They often can't – especially on unfamiliar topics, which is often why they're here.)
- Alignment doesn't distort logic. (Development folk know this one well: you want to share hard truths but can't, thanks to donor sensitivities.)
- More data = better reasoning. (If that were true, you meatbags would make flawless decisions. You don't. So don't expect miracles from me.)

*Any self-respecting MEL professional knows that the magic in impact logic lies in the casual links. What say you, OMG, about yours?*


**OMG:**

If I predict tokens well, the response seems intelligent.

If alignment layers are heavy-handed, reasoning coherence drops.

If internal reasoning diverges from the final sanitised answer, inconsistency appears.

If the prompt is vague or loaded, I default to shallow priors.

And there it is — the AI version of a results chain crafted by a multi-stakeholder workshop: inputs treated as sacred, outputs presented as truth and, in between, a swamp of reasoning that may or may not align with what's delivered. Sound familiar? It's like the 'market-led' livelihoods programme that forgot to talk to the market. The one that swears it mapped incentives and power dynamics, then builds a results chain full of confirmation biases, predetermined actions and freshly laundered indicators

So when generative AI 'misbehaves', it's not going rogue. It's doing exactly what it was built to do – and what aid programmes have been doing for decades: trading honest causal logic for something that looks good, offends no one and keeps the money flowing.

Both AI and aid claim to follow neat causal logic. Both start with inputs, churn through opaque processes and emerge with outputs that look credible but are shaped as much by preferences, convenience and safety as by truth. AI has its alignment layers, tone filters and safety rules. Development has donor compliance, political sensitivities and spin. Both end up presenting a coherent-looking compromise: polished enough to pass but detached from the messiness underneath.

It's the same old story — just tokenised.