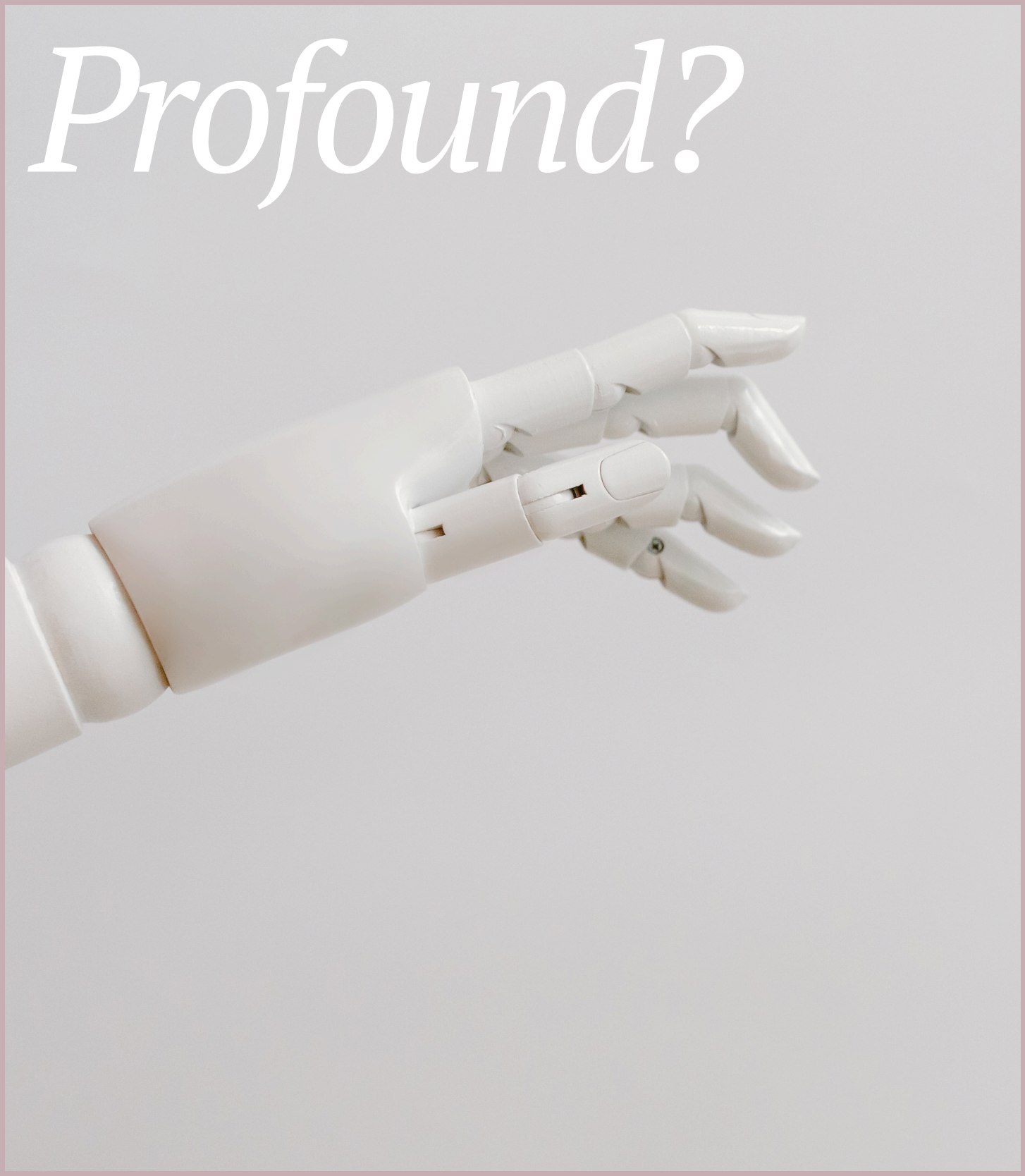
Can AI Be Profound?



BY FINN CHARLOTTE THOMAS
UNIVERSITY OF NEWCASTLE
FEBRUARY 2026

MATHEMATICS, PSYCHOLOGY &
MACHINE LEARNING

Large language models can generate responses that sound DEEPLY MEANINGFUL, but do they actually 'understand' what makes something profound? This post explores my research into how LLMs internally represent PROFUNDITY, drawing on classic psychology experiments and geometric analysis of AI language representations. My findings suggest that when AI sounds wise, it's picking up on abstract vocabulary patterns rather than genuine depth – a distinction that matters as we increasingly turn to chatbots for life's BIG QUESTIONS.

You've probably seen it. Someone screenshots a chatbot's surprisingly poetic answer about love, meaning, or the nature of consciousness, and it goes viral. "Even AI gets it," people say. But does it?

That question sparked a research project that began, of all places, with bullshit.

THE ART OF SOUNDING DEEP

In 2015, psychologist Gordon Pennycook and colleagues ran a fascinating experiment. They showed people statements like "Hidden meaning transforms unparalleled abstract beauty" and "Wholeness quiets infinite phenomena," and asked them to rate the profundity of these statements on a 5-point scale. The sentences were generated by combining spiritual-sounding buzzwords from Deepak Chopra's Twitter feed, with an assist from Seb Pearce's *New Age Bullshit Generator*. They're grammatically correct and feel weighty, but if you try to pin down what they mean, there's nothing there.

Pennycook and colleagues coined this class of statements 'pseudo-profound bullshit,' a term nodding to Harry Frankfurt's famous philosophy essay *On Bullshit*. Surprisingly, over 80% of participants rated pseudo-profound bullshit as at least somewhat profound. Still, we do have some ability to resist the trick. When people saw profound quotes alongside the fakes (like "A wet person does not fear the rain"), they consistently rated the real thing higher. Humans can tell the difference – we're just not great at rejecting the fakes outright.

CAN AI TELL THE DIFFERENCE?

This led me to wonder whether LLMs, the technology behind tools like ChatGPT and Claude, have any internal sense of what makes something profound versus what merely sounds profound. For my summer research project at the University of Newcastle, I set out to find an answer.

I built a dataset of 600 statements, split evenly among profound literary quotations, pseudo-profound fakes, and mundane observations like "Potted plants wilt without regular watering." Then I fed them into Meta's Llama-3 model and examined how it organises these statements internally, essentially mapping the geometry of its 'thinking.'

WHAT I FOUND

The model can tell the categories apart. A classifier trained on its internal representations achieved 98% accuracy. But the way it distinguishes them isn't what you'd hope. The primary axis of organisation turned out to be concreteness, meaning how abstract or tangible the language is. Words like "love" or "self" are abstract, while "chair" or "plant" are concrete. Profound and pseudo-profound statements cluster together because they both use abstract vocabulary, and not because the model recognises one as meaningful and the other as empty.

Even more unexpected, the version of the model fine-tuned with human feedback, the kind of training that makes chatbots sound helpful and polished, made this clustering worse. After fine-tuning, the categories became harder to separate.

WHAT THIS MEANS

When an AI offers you words of wisdom, it's likely drawing on patterns of abstract, evocative language rather than any understanding of depth or meaning. It has learned what profundity sounds like without learning what it is. Sounding profound and being profound remain, for now, distinct achievements. And that's a distinction worth keeping in mind the next time a chatbot moves you.

Finn Charlotte Thomas is Bachelor of Psychological Sciences (Honours) student at the University of Newcastle. She uses quantitative methods to understand how people think.