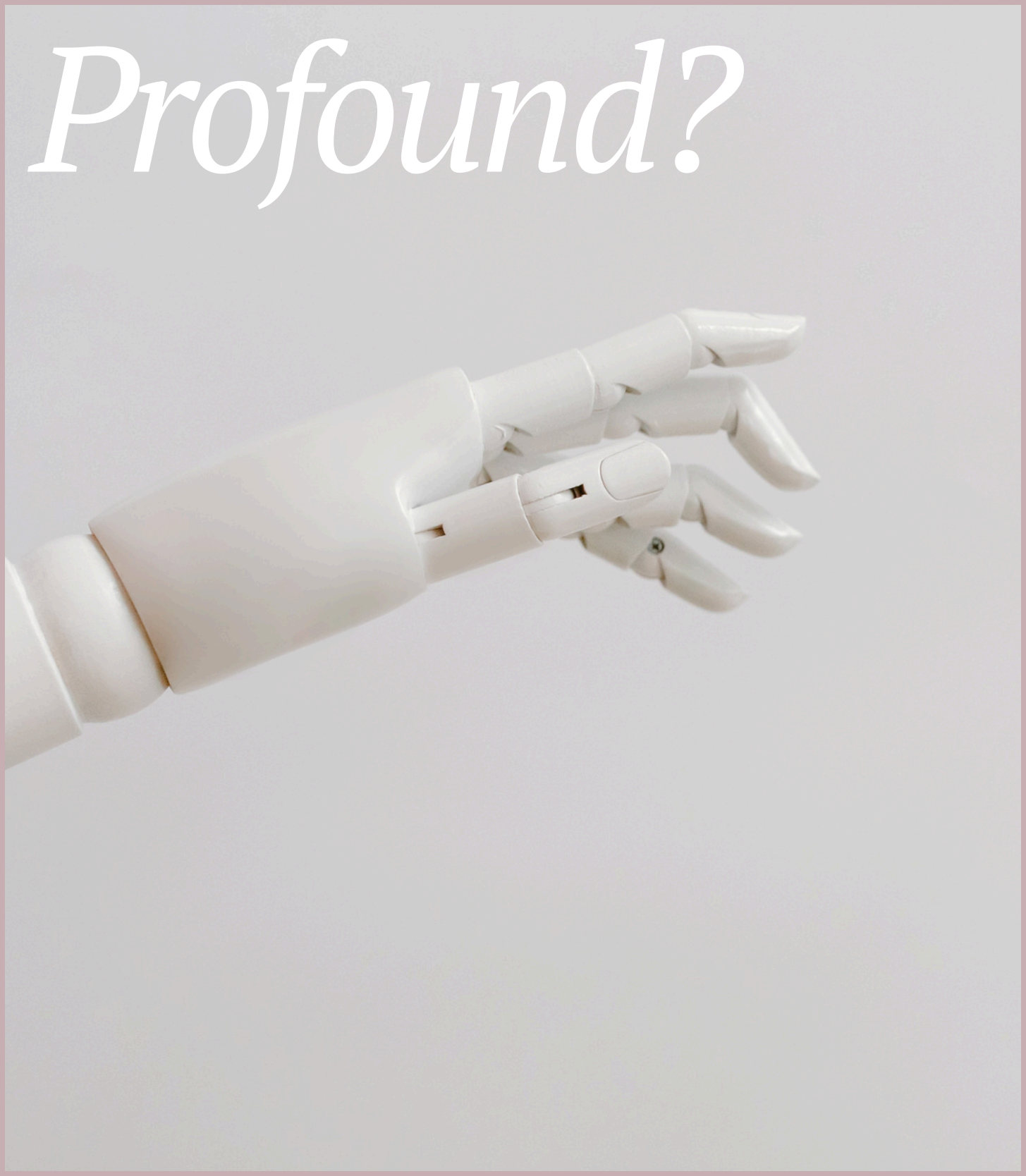
Can AI Be Profound?



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 research into how LLMs internally represent PROFUNDITY, drawing on classic psychology experiments with 'pseudo-profound BULLSHIT' and new geometric analysis of AI language representations. The 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 led me down a research rabbit hole that started, of all places, with bullshit.

THE ART OF SOUNDING DEEP

In 2015, psychologist Gordon Pennycook ran a fascinating experiment. He showed people statements like, "Hidden meaning transforms unparalleled abstract beauty" and "Wholeness quiets infinite phenomena," sentences generated by randomly mashing together spiritual-sounding buzzwords. They're grammatically correct and feel weighty, but if you try to pin down what they mean, there's nothing there.

He called them "pseudo-profound bullshit." And here's the kicker: over 80% of participants rated them as at least somewhat profound.

We're not completely fooled, though. When people saw genuinely profound quotes – like "A wet person does not fear the rain" – alongside the fake ones, they consistently rated the real thing higher. We can tell the difference. We're just not great at rejecting the fakes outright.

CAN AI TELL THE DIFFERENCE?

For my research at the University of Newcastle, I wanted to know whether large language models, the technology behind tools like ChatGPT and Claude, have any internal sense of what makes something profound versus what just sounds profound.

I built a dataset of 600 statements: 200 genuinely profound quotes, 200 pseudo-profound fakes, and 200 mundane observations like "Potted plants wilt without regular watering." Then I fed them into Meta 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 achieved 98% accuracy. But the way it distinguishes them isn't what you'd hope. The primary axis of organisation isn't profundity at all. It's concreteness, 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, not because the model recognises one as genuinely meaningful and the other as empty.

Even more striking: the version of the model fine-tuned with human feedback, the kind of training that makes chatbots sound helpful and polished, made this worse, not better. The categories became harder to separate, not easier.

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, not 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.