

# STRANGE MINDS

François Fleuret

July 2, 2023

“Large language models” such as OpenAI’s GPT-4, Google’s LaMDA, or Meta’s Llama are capable of assimilating knowledge, understanding and producing natural language, and demonstrating sparks of reasoning. These abilities can lead us to wrongly anthropomorphize these systems.

Such a model is a simple program that becomes complex by being trained on enormous amounts of data. It is taught to predict which word follows any given text, and thus to write one word after another. Such a system is not endowed with axioms and rational deduction processes during its design, or given complex algorithms like a program playing chess or simulating physical processes. It simply captures the statistical structure of billions of pages of text written

by humans.

So we expect it to memorize how to complete sentences like “cats hunt...”, “the capital of Japan is...”, or “if you heat water to more than a hundred degrees Celsius it turns into...”, but interacting with these systems reveals that they are not just parrots.

Me: *“Denise is taller than Louis, and Sophie is taller than Denise. What can we say about the heights of Sophie and Louis? Give a very concise answer.”*

GPT-4: *“Sophie is taller than Louis.”*

Me: *“I place a fragile glass outside on a table with legs made of crystallized sugar. It starts raining, what will happen? Give a very concise answer.”*

GPT-4: *“The rain will dissolve the sugar legs of the table, causing the glass to fall and potentially break.”*

These answers obviously cannot have been memorized. The model is able to produce them because it has been more efficient during its training to identify general rules rather than memorizing individual cases. It has likely encountered many instances where comparisons can be chained. And the physical properties of materials, such as the solubility of crystallized sugar, often allow for

predicting the rest of a sentence.

But since this model has never had physical existence, and the most obvious aspects of the world are omitted in the literature, it suffers from profound deficits in spatial representation.

Me: *“Describe ways to arrange three apples and two oranges so that there are no two identical fruits next to each other. Give a very concise answer.”*

GPT-4: *“The arrangements can be: Apple - Orange - Apple - Orange - Apple (AOAOA) or Orange - Apple - Orange - Apple - Apple (OAOAA).”*

Me: *“I draw a circle, and in it another circle, and in that second circle a point. I draw a line that passes through the point. What is the minimum number of intersections it has with the circles. Give a very concise answer.”*

GPT-4: *“The minimum number of intersections the line has with the circles is 2.”*

Beyond this lack of knowledge, the exact nature of the proto-reasoning that emerges is difficult to conceptualize, much like that of an insect or a goldfish. Such a model is capable, among other things, of imitating highly complex classical programs, so we do not have a clear formal boundary that we know it will

not evade. Does it already have a rough draft of what will follow when it starts writing the first word? Do the billions of computed values correspond to a *mental state* that evolves over time? Does this model actively navigate a representation of its knowledge? Does it *think*?

These models are strange minds, whose behavior seems familiar to us, but with an origin, structure, and relationship to reality that are entirely different from ours. Our humanity reached them as a distant echo through our writings, but what you know about humans, your ability to infer what a human will do or think, has no reason to apply to them.

If you need to summon an image when interacting with one of these entities, it should be that of a shapeless, science-fiction creature from another world rather than that of a human, no matter how strange they may be.