Memory Isn’t Real

How Large-langauge models change everything we thought we knew about knowing

Several years ago, a family member asked for my opinion on the death penalty. I proceeded to launch into a fairly detailed exposition, citing historical evidence, considering the sociological implications and attempting to navigate the tricky moral landscape. Afterward, I was struck by what had happened; not the position I had taken, but rather a puzzling realization: I hadn't consciously thought much about this issue before. Where on earth did this nuanced, fully formed opinion that came gushing out of me actually come from? How did I have a structured view on a complex topic I had never deliberately contemplated?

The conventional explanation would suggest that I had subconsciously formed these beliefs over time through exposure to various influences, and they were stored somewhere in my mind, waiting to be retrieved when prompted. But recent advances in artificial intelligence suggest a more radical possibility: I didn't have that opinion —or indeed any opinion—until the very moment I was asked about it. In other words, the viewpoint I articulated wasn't "in" my brain at all; it only came into being as I was responding to my friend's question. This seemingly counterintuitive viewpoint is not just a fanciful philosophical conjecture; the models that underlie contemporary artificial intelligence—particularly large language models such as those underlying ChatGPT or Claude operate precisely this way. They don't store information that is later retrieved during runtime. Instead, they operate through a process that generates text “just in time”, piece by piece, more like painting an abstract design than faithfully reproducing an existing scene.

To understand this more deeply, let’s briefly look under the hood of LLMs like ChatGPT or Claude and see what they are and what they’re doing. The core engine of these models is a so-called ‘neural network’ which is simply a series of sequential mathematical operations that takes in an input and produces an output. In the case of LLMs, the input is a sequence of words that have been “tokenized” (i.e. split into smaller units—words or subwords— and then converted into numbers the model can process). The output is a single such token. For example, when prompted with a request like "Recite the Pledge of Allegiance," the model doesn't retrieve a stored text. Instead, it performs a single mathematical operation through its parameters to predict what token is ‘likely to come next; in this case the word "I." This predicted "I" now becomes part of the next input (along with the original request), and the model produces the next output: "pledge." Now, "I pledge" becomes part of the input for generating "allegiance," (actually, ‘allegiance’ would likely be made up of more than one token, but you get the idea).

This process continues, likely for the whole pledge, until the LLM outputs some special character that tells it to stop.

This basic recipe of producing one output at a time, tacking it onto the previous input and then running it again, is called autoregression. Despite its inherent simplicity, this process, applied to next-token prediction for language, is sufficient to generate coherent, structured, and mostly accurate responses to anything people can throw at it. These models can handle not just short prompts or factual queries, but also generate long-form responses—stories, essays, dialogues—at a level that often matches or exceeds what humans might produce on the fly.

(Sidenote: While the algorithms are literally guessing the next-token, recent research from Anthropic shows that their internal activity often reflects sensitivity to words or structures several steps ahead. Criticially, this doesn't actually reflect ' planning' in the conventional multi-step sense—it's still just one computation—but it suggests the model has internalized patterns so rich that each prediction subtly reflects the likely trajectory of the whole sequence. It's as if the future leaves a trace in the present– a 'ghost' of probable futures guiding the present prediction, much like the past tokens provide explicit context. This underscores how next-token prediction alone can yield outputs that feel foresightful—without any need for memory or truly looking ahead—just through the dynamics of generation itself.)

Crucially for our discussion, this form of autoregressive next-token generation means that there is no holistic representation of a stored sequence, like the Pledge of Allegiance, encoded anywhere in the model's weights. The complete recitation only emerges as a result of an unfolding autoregressive process, with each token triggering the generation of the next. So does the model actually "know" the Pledge of Allegiance? If knowledge means having a complete representation stored somewhere, then the model doesn't actually "know" the pledge in this sense. What exists are parameters that enable the reliable generation of each next element in the sequence when given the preceding context. Under the right autoregressive conditions, this yields an accurate recounting of the pledge *without explicitly storing it anywhere*.

Note that the Pledge of Allegiance is a special case—a single, tightly constrained sequence. Most of what a large-language model can do looks nothing like that. When you ask it to, say, summarize a news article or draft an email, no finished answer is sitting in storage. Instead, the model brings forth a response from a reservoir of *potentialities*—dispositions encoded in its weights during training—that guide its output stream. And herein lies a key point: The billions of sentences the model consumes are *not* copied into a mental filing cabinet. Each new example nudges the weights ever so slightly, reshaping the statistical landscape that guides future generations. The particulars of any single sentence are soon lost, yet their influence persists as subtle biases that steer the next token the model will emit. In this way, past data does not remain inside the network as isolated facts; it reappears only as an altered tendency to map inputs onto meaningful outputs.

At this point you might be thinking: "Fine—LLMs juggle potentialities instead of memories, but *I* definitely remember things, *real* things. I can picture my mother's smile, see the look on my

friend's face when I told that joke, walk myself through the directions to my childhood home, or hum the chorus of a favorite song." It feels as though you are pulling crisp snapshots or instruction sheets from mental storage. But notice something: with the same ease you can imagine your mother doing a cartwheel into a ball pit, your friend's face painted purple, or your childhood street floating among the clouds. That effortless recombination is the telltale sign that nothing was ever retrieved; everything, from fanciful imaginations to distinct recollections, may be generated, on the spot, from a pliable web of synaptic weights shaped by experience.

In other words, perhaps those weights in your brain, the neural connections, encode potentialities too—generative biases that let present cues blossom into infinitely many coherent scenes, instructions, or melodies—based on whatever is going on in your mind and the environment. In some cases, these potentialities yield a somewhat 'accurate' account of the lived past. But that's just a drop in the cognitive bucket. Guided generation is immeasurably more powerful than retrieval: instead of reopening a single file, the brain (or an LLM) can spin up any number of meaningful variations, constrained only by context and imagination.

While all of this may sound good in theory at this point we need to ask: *are* people actually autoregressive generation engines like LLMs? This claim is supported by multiple lines of evidence. First, consider how humans produce language—one word at a time, with each word shaped by prior context, exactly like next-token prediction. In addition, research on infants suggests that they acquire language not through abstract rules but by tracking statistical patterns and transitional probabilities between sounds and words. Also, psychological phenomena like priming effects demonstrate how recent inputs subtly shape what people say or think next, while research on long-term recall of events shows how malleable 'memories' are to implicit revision. These phenomena are more easily explained if 'recall' is actually context-based generation rather than retrieval from cold storage.

But perhaps most telling is the success of the autoregressive approach itself. The existence of LLMs proves, minimally, that coherent and contextually appropriate language generation does not require discrete storage and retrieval mechanisms—processes traditionally assumed to underpin human cognition. Ockham's razor dictates that we must seriously entertain the possibility that human cognition itself may fundamentally operate based on generative, autoregression. But at a deeper level, the fact that LLMs work so well suggests that language as a system is *designed* to support it—it is inherently structured around sequential prediction. If so, it makes sense that our brains would leverage the generative structure that is *already there* rather than using some completely different method to produce it. And it also makes sense that language would have been built on already existing autoregressive mechanisms already built into the brain for other, older cognitive functions.

If our minds do indeed operate based on autoregressive principles, this would challenge our fundamental understanding of how our basic cognition works. We intuitively picture our minds as archives where memories, knowledge, and beliefs are already present, waiting to be accessed when needed. We speak of "storing memories," "retrieving information," "holding beliefs," and "possessing knowledge"—metaphors that frame the mind as a repository of preserved experiences, facts, and convictions.

These metaphors aren't merely linguistic conveniences: they form the bedrock of how we conceive of ourselves and others as continuous beings with coherent identities. The very notion that I "know," "remember," and "believe" certain things before accessing them is not just a casual assumption—it's baked into our most fundamental concept of selfhood. We define ourselves and others largely in terms of the supposed contents of our minds— the memories we carry, the knowledge we possess, the beliefs we hold. To challenge these metaphors is to challenge our most basic understanding of what it means to be a person with a continuous identity across time.

If cognition indeed arises through generative rather than the mind is not a static entity or storehouse of accumulated contents, but a continuously unfolding dynamic process. Understood this way, a mind is not something we "have" but rather something that emerges, one cycle of brain activity at a time. This reframing reveals cognition as an ongoing performance—an improvisation in which the self is not retrieved from storage, but perpetually regenerated through each new neural activation pattern. Such a perspective doesn't merely update our understanding of the mind—it completely rewrites the book of cognition, transforming our most basic assumptions about what it means to know, to remember, and ultimately, to be.

This is what I mean when I say that *memory isn't real*. It's not that the past doesn't meaningfully guide our current cognition. Of course it does. What I mean is that the past is not tucked away in static drawers waiting to be pulled open; it is diffused through the weights of the network (or, in the biological case, through synaptic strengths). Those traces steer the present act of generation without dictating a single predetermined path, enabling responses that are at once grounded in experience and infinitely flexible. We don't recite; we generate. My impromptu death-penalty monologue was exactly this: an emergent construction assembled in real time from countless articles, classroom debates, and dinner-table arguments—none of which was stored verbatim in my head, yet all of which left a trace that could be recombined in the moment. Memory, then, is less a well we dip into than a stream we stand within—always in motion, always responsive to the the contours of the present moment.