How AI chatbots like ChatGPT or Bard work – visual explainer

Large language models (LLMs) are the fundamental architecture behind chatbots like ChatGPT or Bard. A question typed in to ChatGPT, such as "What is the capital of France", has to be processed by an LLM in order to produce an answer like "The capital of France is Paris". Here's a visual walk-through of how this type of artificial intelligence works. That reweighting step is what the LLM technicians call a "transformer", and the principle of re-evaluating the weights based on the salience given to previous bits of the text is what they call "attention". The LLM applies these steps to every part of a given conversation. So that if you ask "What is the capital of France?" it can re-evaluate capital to probably mean "city" not "financial resources" when it gets the added input of "France". And when you subsequently ask "How many people live there?" it has already assigned enough salience to the idea of "Paris (city)" that it can conclude that "there" is standing in for "Paris". Attention is widely considered a breakthrough development in natural language AI, but it doesn't make for the successful models on its own. Each of those models then goes through extensive training, partly to master the question-and-response format, and often to weed out unacceptable responses – sometimes sexist or racist – that would arise from an uncritical adoption of the material in the training corpus. Notes Most of the visualisations are illustrative but informed by conversations with industry experts, to whom thanks, and by interaction with publicly available LLMs. The vector for happy is from the BERT language model using the transformers Python package.