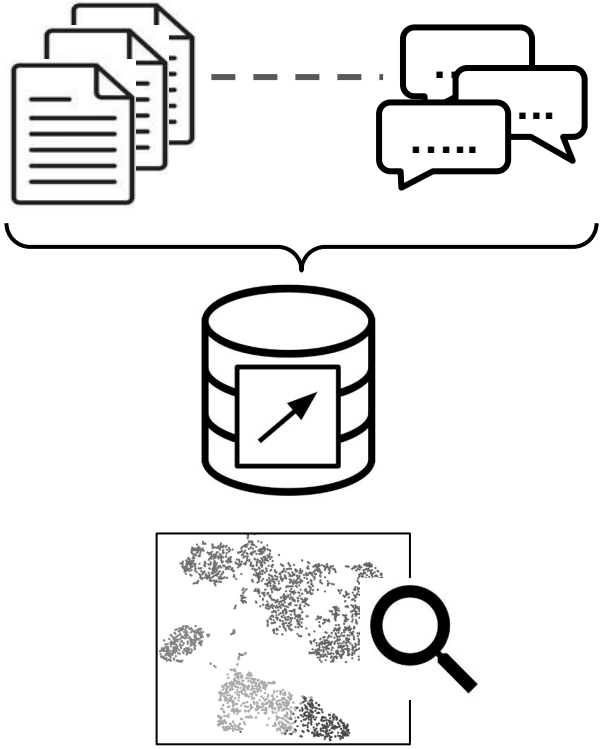
Large language models: What's real



Neil Shepherd

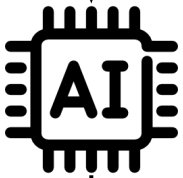
First - a primer on LLM components (simplified!)

Embedding models (aka representative models)



Generative models (e.g. Cohere, GPT3.5..)

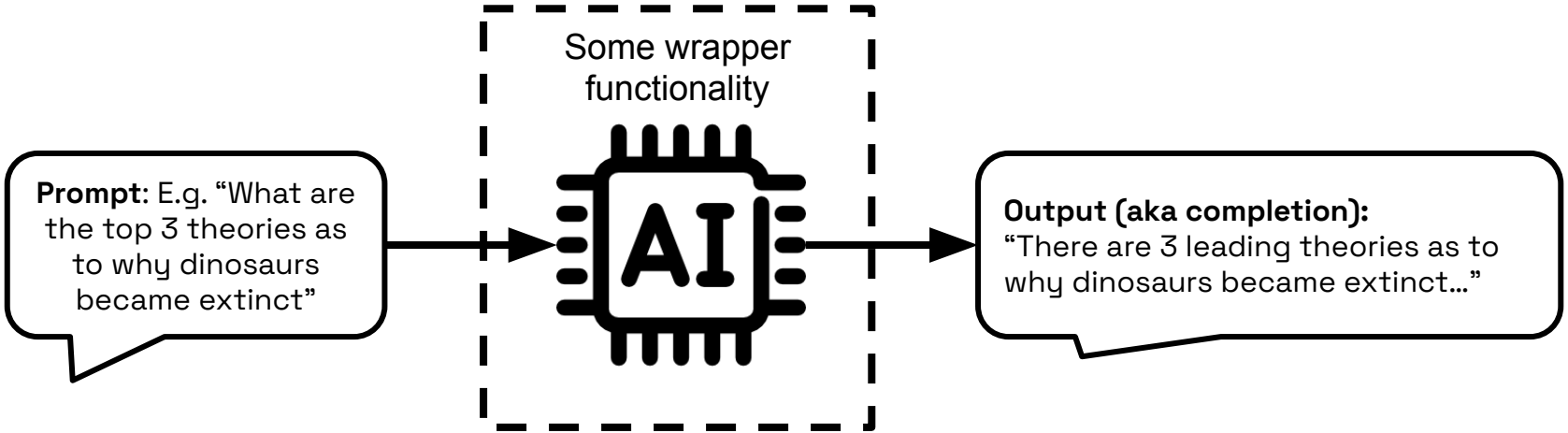
Prompt: E.g. “What are the top 3 theories as to why dinosaurs became extinct”



Output (aka completion):
“There are 3 leading theories as to why dinosaurs became extinct:

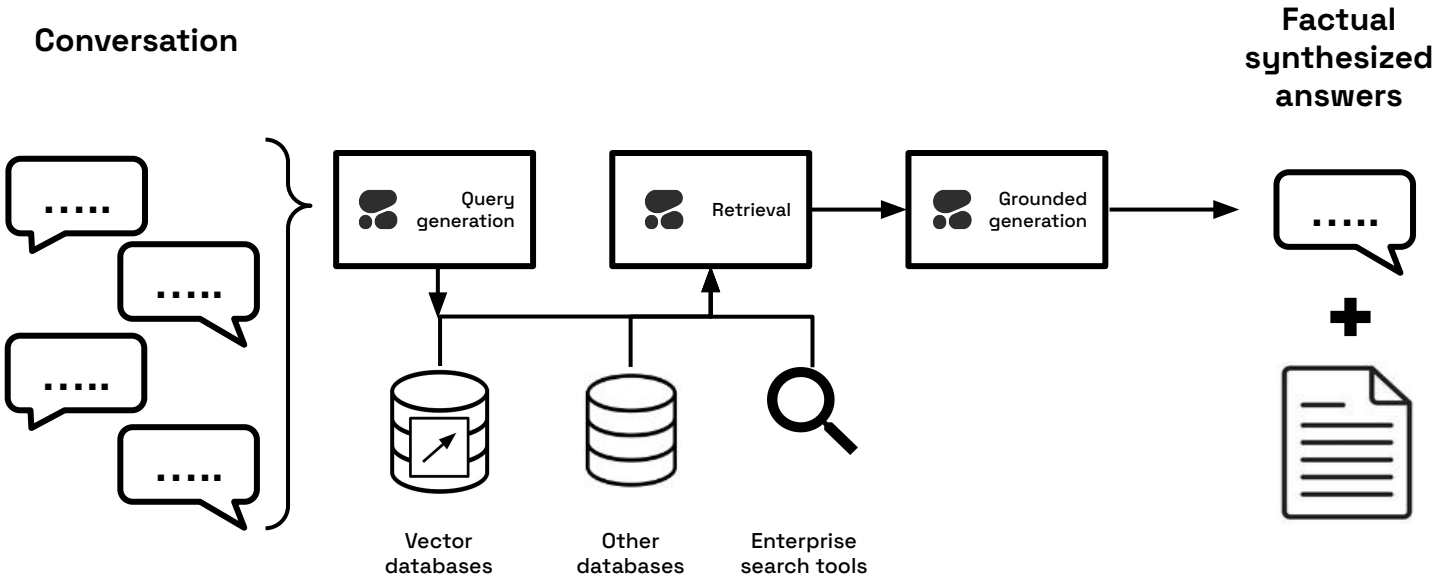
- A really big meteor
- Climate change
- COVID minus 65million

Generation only is commoditizing



Not a moat!

RAG (Retrieval Augmented Generation) is where all the value is



The list of enterprise use cases is enormous



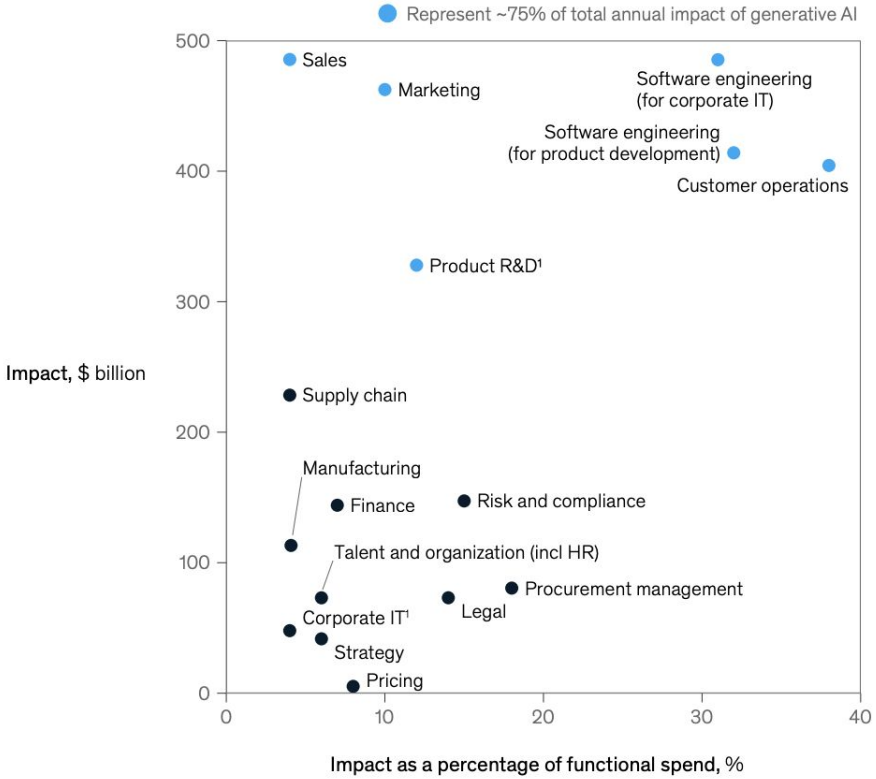
TL;DR:

~$2.8-4.4Tn impact in US

Bank on **15-20% opex savings** in white collar jobs

~80% of white collar jobs impacted

Most value appears to be in sales, marketing, software and customer operations



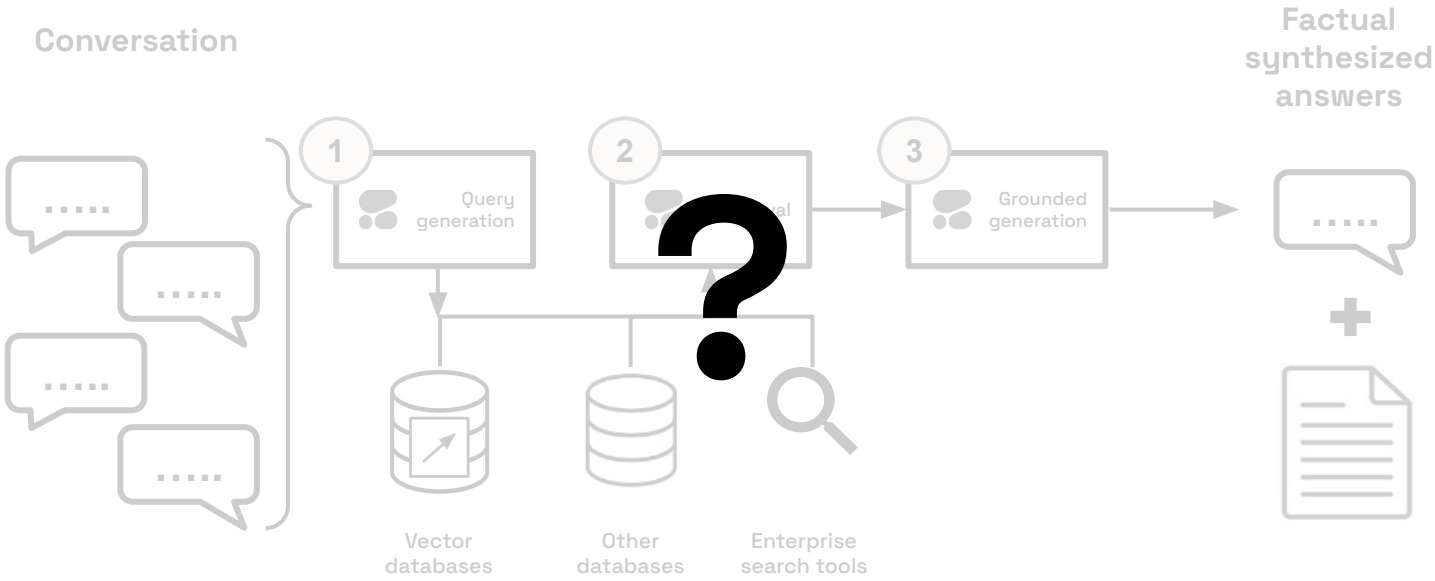
(Most) Executives are moving quickly

- **>80% of executives see Generative AI as a top 5 priority** in the next 5 years (~10% as the #1 immediate priority)
- **Generative AI POCs are working** - the majority are seeing positive impact (but it greatly helps to have a strategy)
- **Level of ambition varies** - many dipping toes but some companies are going very broad

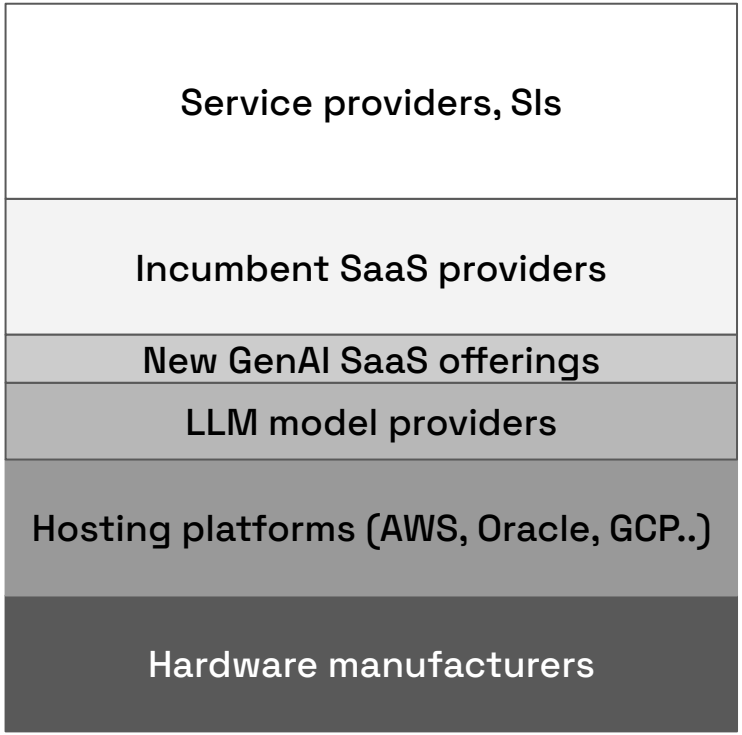


This stuff is expensive

Also, who's going to capture value delivering this?



It gets complicated, but much of the stack have excellent moats



Outlook (for 2024 anyway)



"Awesome! Let's get on this new service line ASAP"



"Quick - Add GenAI to increase value and not lose customers"



I'll explain.....



"Excellent, this is going to drive lots of compute and storage"



"This is great" (Jensen, until other HW mfrs catch up)

My take: The hype cycle will apply, just much compressed

