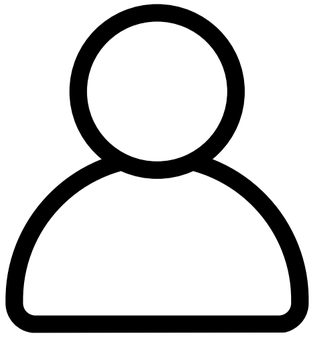
RAG:
RETRIVAL AUGMENTED GENERATION

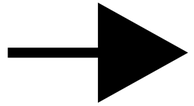
QUESTION:

if you want to teach an LLM your domain-specific knowledge

SIMPLE LLM APPLICATION



Query + prompt



LLM

PROBLEMS W SIMPLE APPROACH?

PROBLEMS W SIMPLE APPROACH?

HALLUCINATION

FINE-TUNING?

FINE-TUNING?

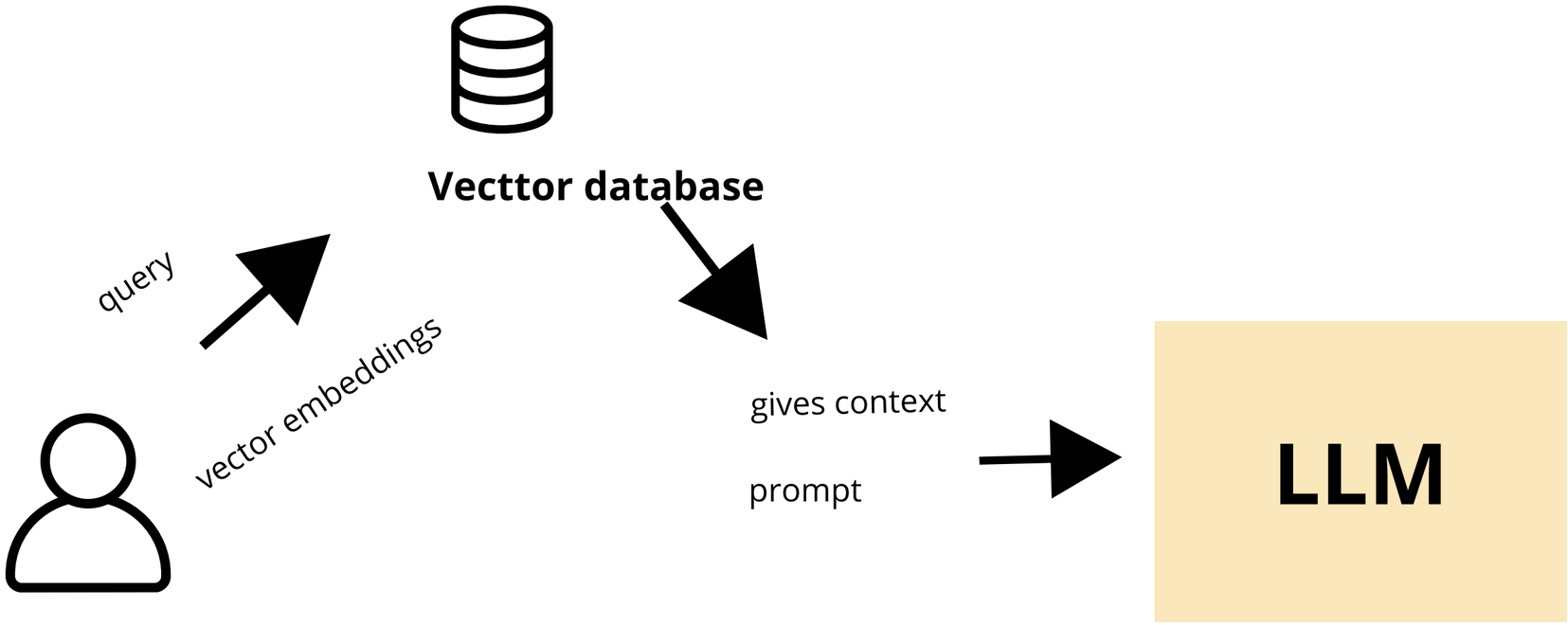
Expensive — training cost + time
Data-hungry — needs thousands of high-quality labeled examples
Static — can’t update facts easily (model must be retrained)
Output fusion — mixing multiple sources → model guesses

RAG

- The model doesn’t try to remember everything
- Instead, it searches your knowledge base during inference

- “open-book exam” instead of “memorize the whole book”

RAG PIPELINE



RAG PIPELINE