MINERVA | WINTER AI SCHOOL

VLLM CHALLENGE

FRANCESCA, GIUSEPPE & ALEX

THE PROBLEM

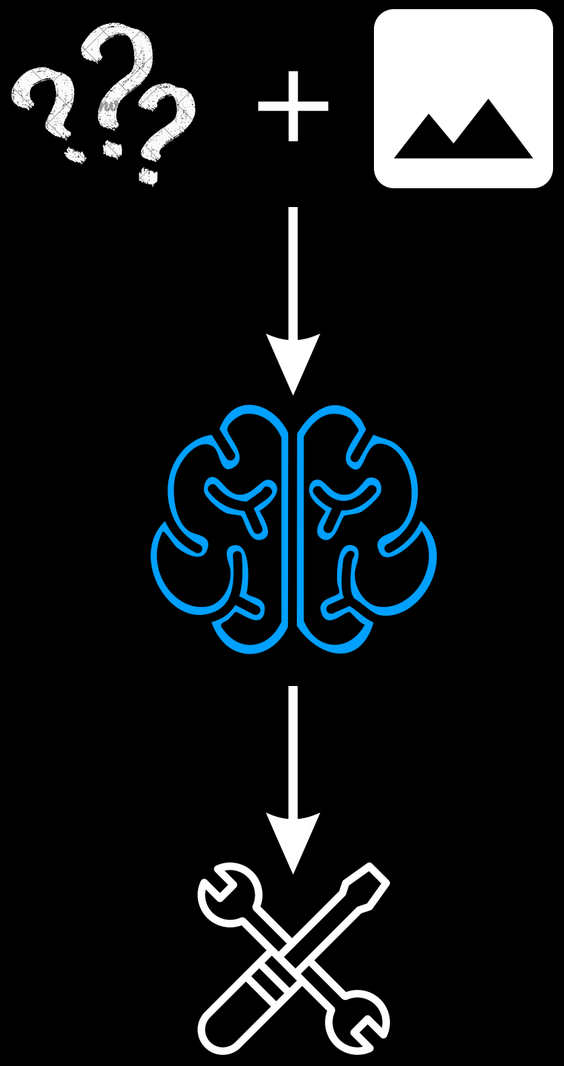
Multimodal Large Language Models (MLLMs) are not reliable on their own

Why?

1. Limited knowledge
2. Hallucinations
3. Poor document understanding

OUR VISION

Transform a standalone MLLM into an intelligent multimodal agent



FROM MODEL TO AGENT

Instead of relying on a single model, we design a system where:

1. The MLLM acts as a “Brain” (Router)
2. It understands the question + image
3. It decides which tool to use

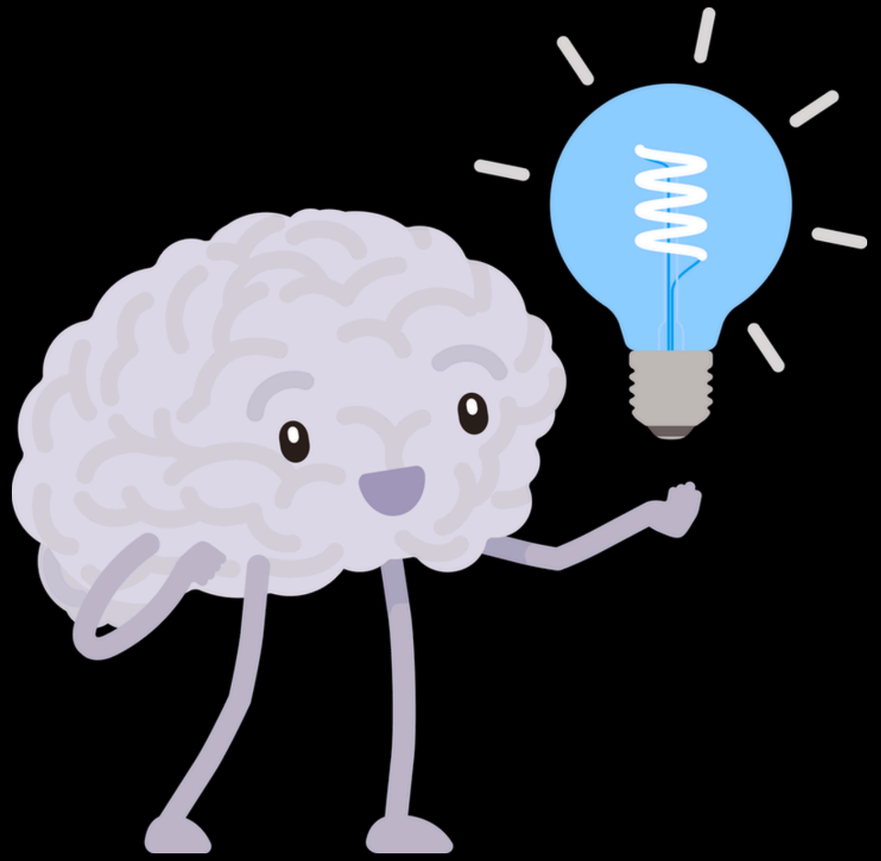
TOOL-AUGMENTED INTELLIGENCE

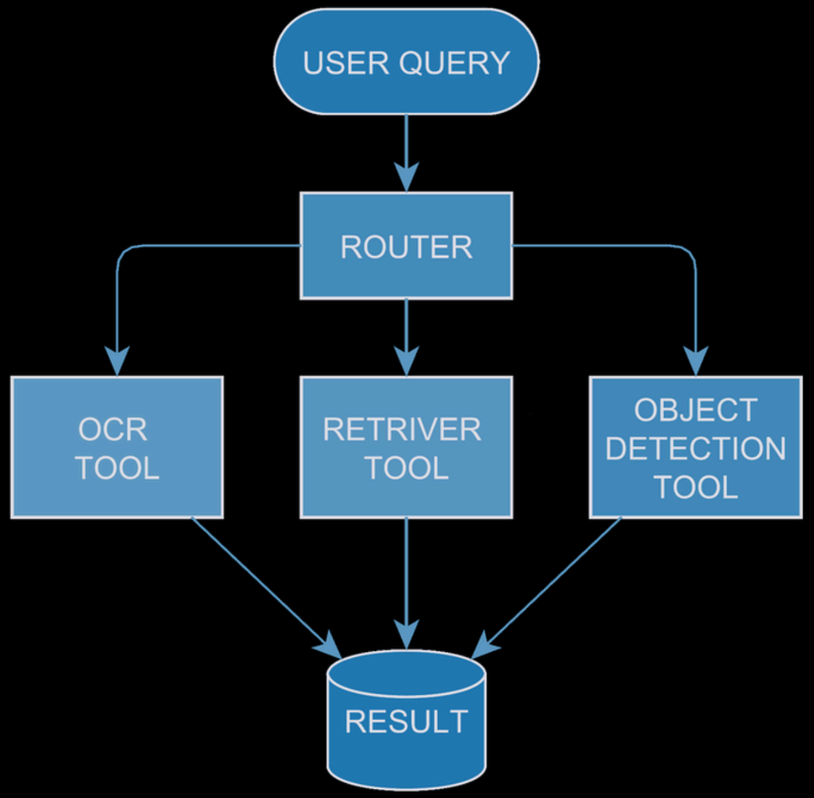
We enhance the model with specialized tools:

Retriever → adds external knowledge

Object Detector → prevents hallucinations

OCR Reader → extracts precise text





CLOSED-LOOP REASONING

The agent:

1. Analyzes the input
2. Selects the right tool
3. Uses the tool output as context
4. Generates a more accurate final answer

THANK YOU