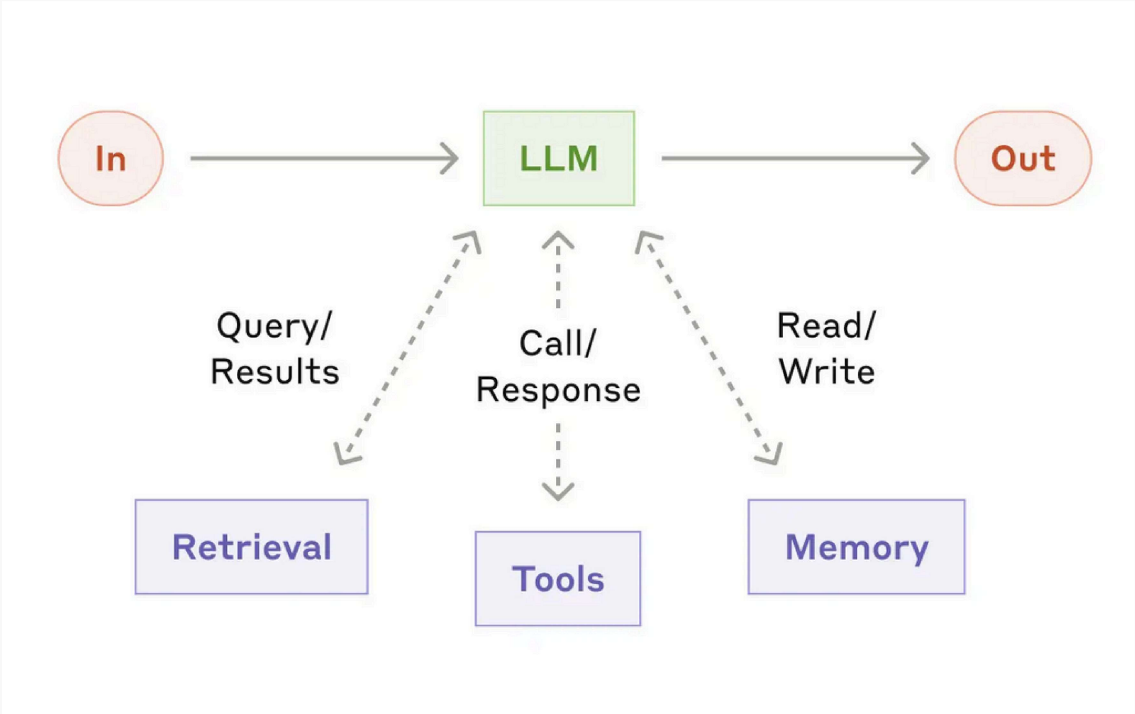
Building block

The augmented LLM

LLM enhanced with augmentations such as retrieval, tools, and memory



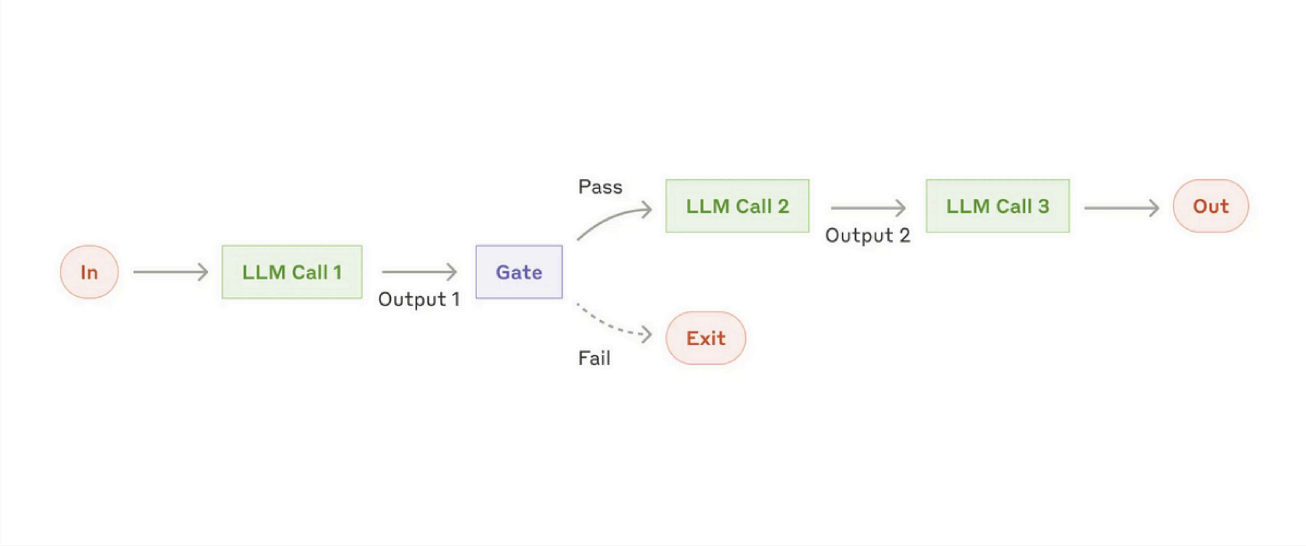
Based on Anthropic's "Building Effective Agents" guide
anthropic.com/engineering/building-effective-agents

linkedin.com/in/neehanthreddy github.com/neehanthreddym

Workflow

Prompt chaining

Breaks down a task into a series of steps, with each LLM call processing the preceding one's output.



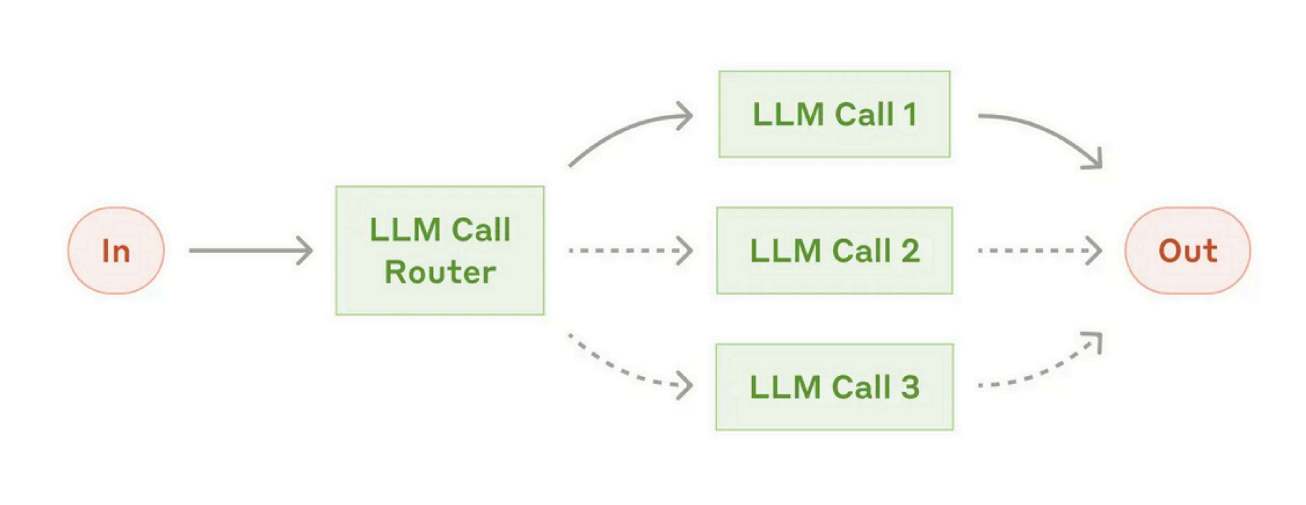
When to use

When a task can be simply and neatly divided into fixed subtasks, this process is perfect. The primary objective is to trade off latency for increased accuracy by simplifying each LLM call.

Workflow

Routing

Identifies an input and routes it to a certain follow-up task.



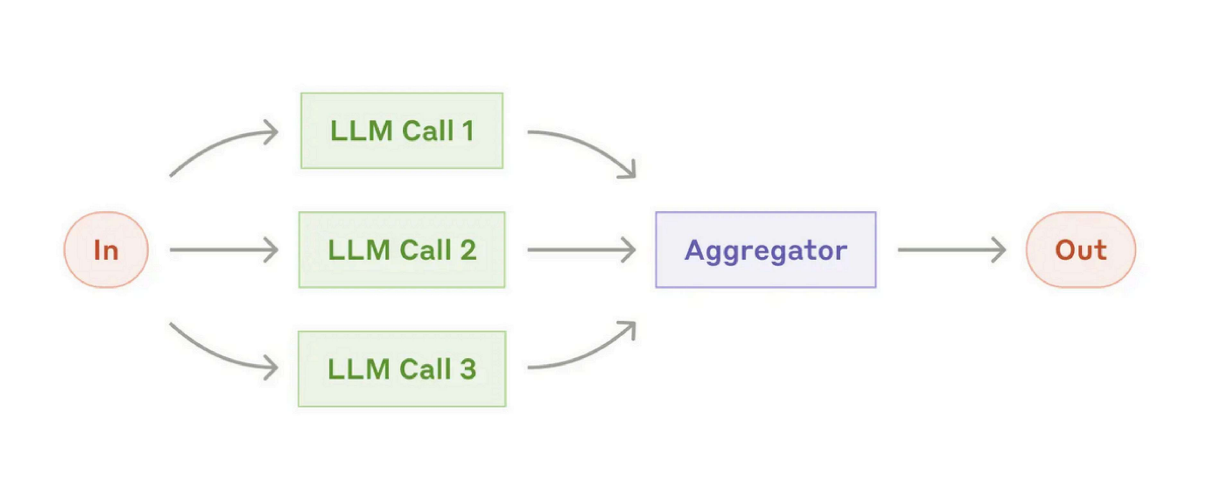
When to use

Routing is a good way to manage complicated activities that have clear categories that are better off being handled individually. It can also be done accurately by an LLM or a more traditional classification model or algorithm.

Workflow

Parallelization

LLM does several processes concurrently, either by dividing a task into separate components or by repeatedly executing the same operation in order to compare and aggregate various outcomes.



When to use

Use parallelization when you need faster execution or better results by splitting work into independent parts or comparing multiple attempts.

Additional Workflows

Orchestrator-workers

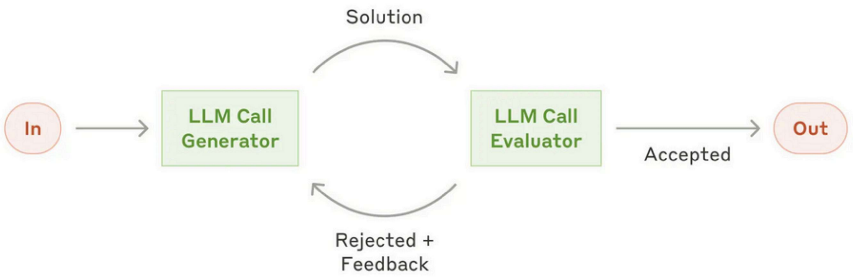
The orchestrator-workers workflow uses a central LLM to break a task into subtasks, assign them to worker LLMs, and combine their results into a final output.

Ideal for complex tasks where the necessary subtasks are unpredictable.

Evaluator-optimizer

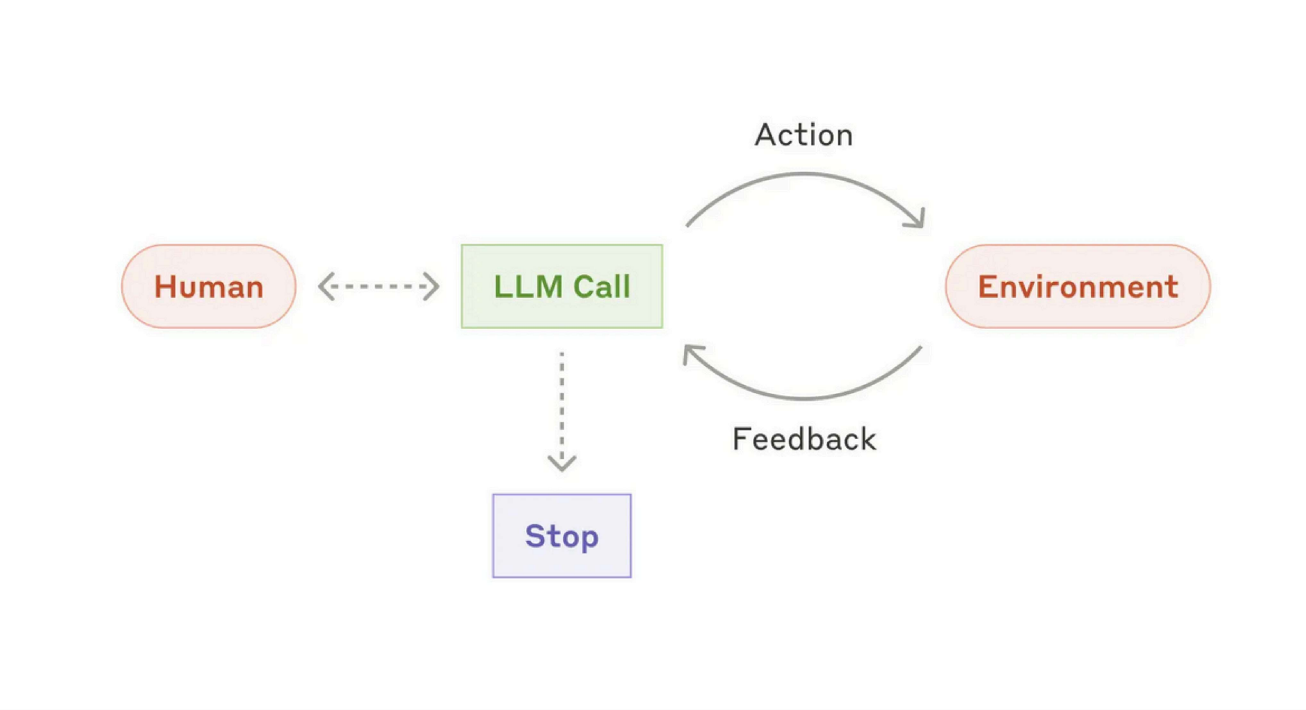
One LLM call produces a response in the evaluator-optimizer cycle, while another continuously gives evaluation and feedback.

Useful when clear evaluation criteria exist and iterative feedback can meaningfully improve results, much like refining a draft through multiple rounds of review.



Agents

Agents are LLMs that plan and act independently on complex tasks, using tools and environmental feedback in a loop, while optionally consulting humans at checkpoints or when blocked.



Use agents for open-ended tasks with unpredictable steps, where autonomous decision-making is needed, but ensure safeguards and testing to manage cost and error risk.