

The background is a dark blue gradient with an abstract digital theme. It features several glowing, semi-transparent cubes and rectangular blocks arranged in a complex, interconnected pattern. These blocks are covered in a fine grid of small, glowing blue dots, resembling a data matrix or a digital cityscape. Bright blue and red light beams or rays emanate from various points on the blocks, creating a sense of dynamic energy and data flow. The overall aesthetic is futuristic and high-tech.

Smart Workflows with LLMs: Parallelization & Routing

How to Build Efficient, Specialized AI Systems

Overview

- What are LLM workflows?
- Why structure matters in building AI systems
- Today's focus: Parallelization and Routing

What is Parallelization?

- LLMs work on multiple tasks simultaneously
- Used to speed up processes, enhance diversity, or increase confidence
- Outputs are combined programmatically

LLM calls



Aggregator



Out

In







Variants of Parallelization

- 1. Sectioning
 - - Divide task into independent subtasks
 - - Each handled by a different LLM call
- 2. Voting
 - - Run the same task multiple times
 - - Aggregate diverse results to find best/most common





When to Use Parallelization

- Tasks are independent and can be done in parallel
- You want multiple takes (for creativity or confidence)
- Task has multiple aspects needing focused handling

Benefits of Parallelization

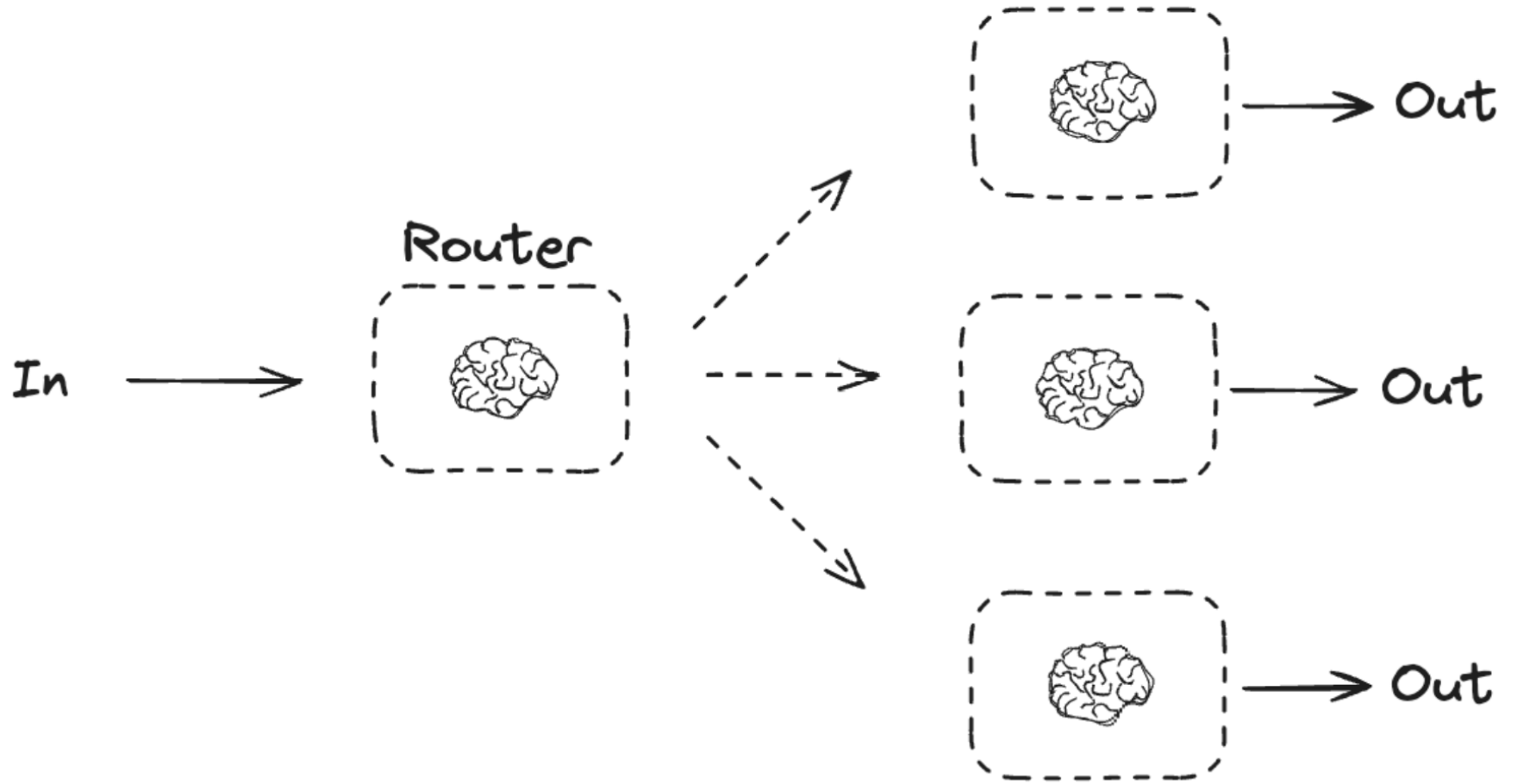
-  Faster results
-  Better focus per subtask
-  More creative options
-  Improved reliability through consensus

Examples of Parallelization

-  Content writing: Sections handled independently
-  Legal review: Each clause analyzed separately
-  Idea generation: Multiple runs for diverse insights
-  Research synthesis: Each paper summarized in parallel

What is Routing?

- Classify the input and route it to the right follow-up task
- Enables specialized handling of each type of input
- Prevents performance drops from 'one-size-fits-all' prompts







Why Routing Matters

- Separates concerns — each task has its own optimized logic
- Allows tailored prompts for better performance
- Ensures robustness across a range of input types

When to Use Routing

- Your task has clearly defined categories
- Different types of input require different handling strategies
- You can accurately classify input using LLMs or ML classifiers

Examples of Routing

-  Chatbot: Route billing vs. tech support queries
-  Content engine: Route news vs. opinion pieces
-  Research assistant: Route to summarizer vs. translator
-  Education app: Route based on grade level or subject

Summary

-

Workflow	Best For	Benefits
Parallelization	Independent or multi-angle tasks	Speed, focus, confidence
Routing	Categorically distinct inputs/tasks	Specialization, prompt tuning