Accuracy by Composition Level

Legend: L0 (Single), L1 (Chain), L2 (Parallel), L3 (DAG)

Model	L0 (Single)	L1 (Chain)	L2 (Parallel)	L3 (DAG)
Mistral Small	92%	61%	89%	42%
Mistral Large	88%	60%	89%	42%
Mistral Medium	90%	58%	89%	38%
Gemini 2.0 Flash (OpenRouter)	77%	55%	88%	38%
Llama 3.1 8B (Groq)	81%	53%	70%	38%
Llama 4 Scout 17B (Groq)	91%	53%	28%	
Llama 3.3 70B (Groq)			16%	