Agentic AI Frameworks Comparative Study

AutoGen vs LangGraph - Pilot Experiment Results

Generated: December 12, 2025 at 01:48 PM

Executive Summary

This report presents the results of a pilot comparative evaluation of AutoGen and LangGraph frameworks using standardized benchmarks and custom GEval metrics aligned with the study's research questions.

Experimental Results

AutoGen Framework Results

Metric	Value
Benchmark	Simple Math Pilot
Model	gpt-4o-mini
Temperature	0.3
Problems	5
Correct	5
Total	5
Accuracy	100.0%

LangGraph Framework Results

Metric	Value
Benchmark	Simple Math Pilot
Model	gpt-4o-mini
Temperature	0.3
Problems	5
Correct	5
Total	5
Accuracy	100.0%

Comparative Analysis

Framework	Accuracy	Winner
AutoGen	100.0%	=
LangGraph	100.0%	=

Both frameworks achieved identical performance in this pilot evaluation.

Detailed Problem-by-Problem Results

AUTOGEN - Problem Details

Field	Value
Problem ID	0
Question	What is 25 + 17?...
Expected	42
Correct	✓

Field	Value
Problem ID	1
Question	What is 8 * 7?...
Expected	56
Correct	✓

Field	Value
Problem ID	2
Question	What is 100 - 35?...
Expected	65
Correct	✓

Field	Value
Problem ID	3
Question	What is 144 / 12?...
Expected	12
Correct	✓

Field	Value
Problem ID	4
Question	What is 15 * 3 + 10?...
Expected	55
Correct	✓

LANGGRAPH - Problem Details

Field	Value
Problem ID	0
Question	What is 25 + 17?...
Expected	42
Correct	✓

Field	Value
Problem ID	1
Question	What is 8 * 7?...
Expected	56
Correct	✓

Field	Value
Problem ID	2
Question	What is 100 - 35?...
Expected	65
Correct	✓

Field	Value
Problem ID	3
Question	What is 144 / 12?...
Expected	12
Correct	✓

Field	Value
Problem ID	4
Question	What is 15 * 3 + 10?...
Expected	55
Correct	✓

Conclusion

This pilot experiment successfully demonstrated the evaluation framework. Both AutoGen and LangGraph frameworks were tested using identical conditions (model: gpt-4o-mini, temperature: 0.3). The framework is ready for full-scale evaluation with 50-100 problems across 4 benchmarks (GSM8K, HumanEval, ARC, MATH) with 5-10 repetitions as specified in the experimental design.