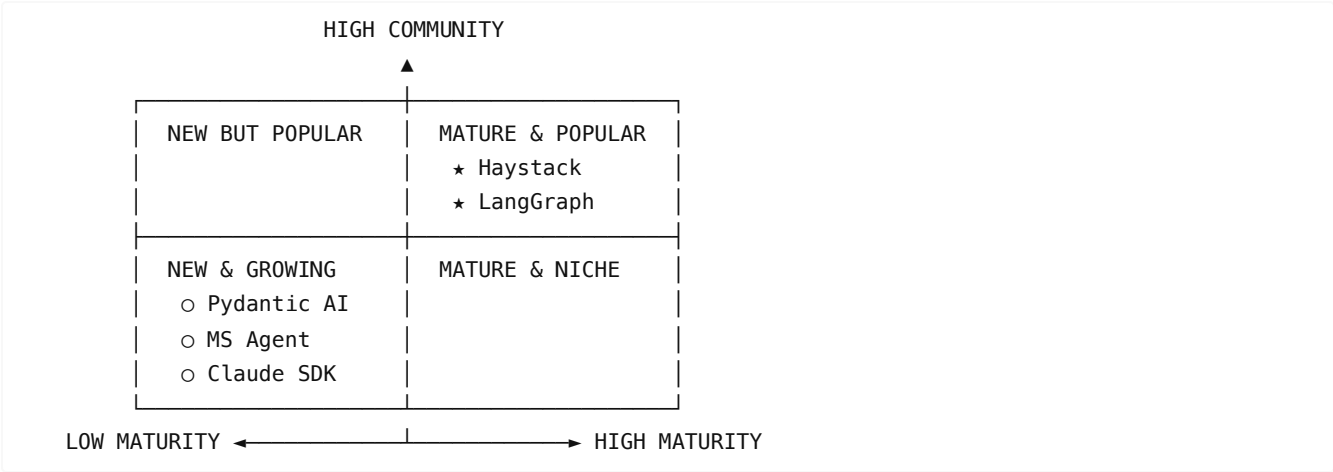


# AI Agent Framework Comparison

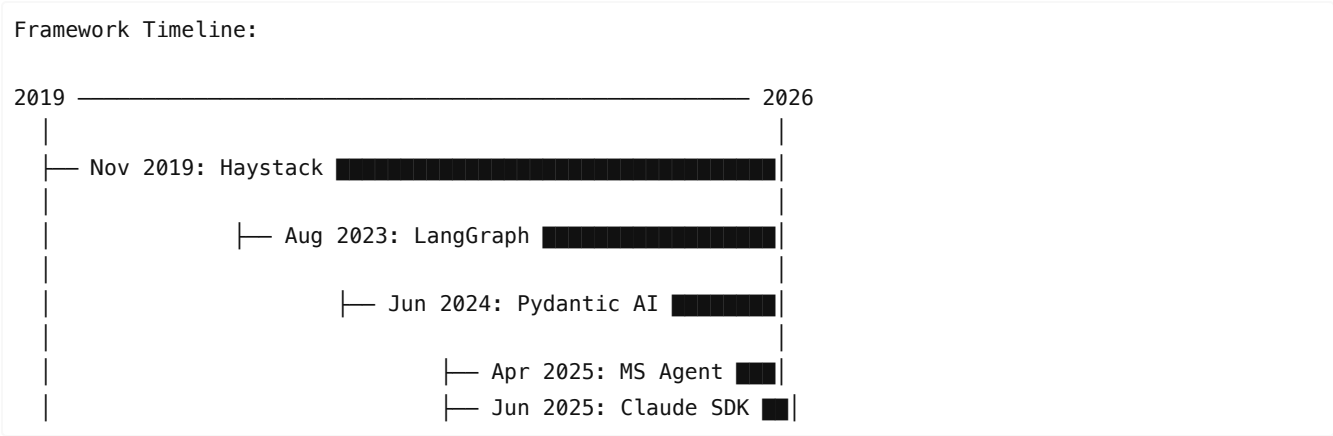
A high-level comparison of five AI agent frameworks for building multi-agent systems.

## Visual Overview

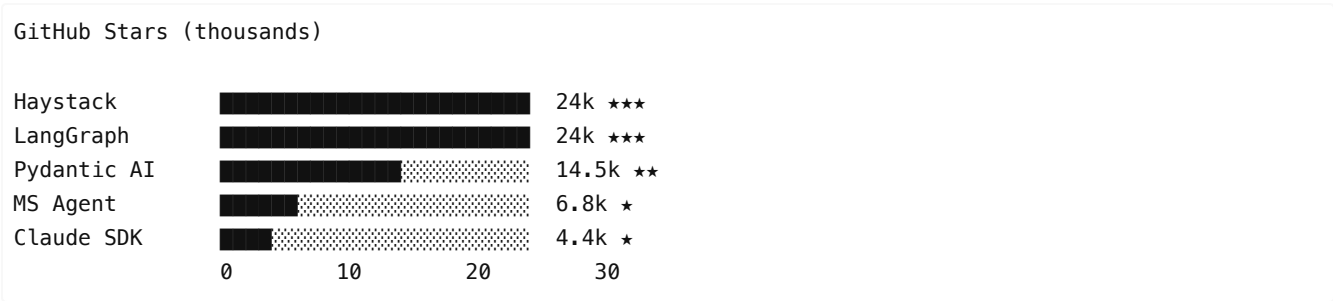
### Maturity vs Community Size



### Framework Timeline



### GitHub Stars Comparison



## Quick Reference

Framework	First Release	Stars	Primary Language	Maintainer
<a href="#">Haystack</a>	Nov 2019	24k	Python	deepset
<a href="#">LangGraph</a>	Aug 2023	24k	Python	LangChain Inc
<a href="#">Pydantic AI</a>	Jun 2024	14.5k	Python	Pydantic
<a href="#">Microsoft Agent Framework</a>	Apr 2025	6.8k	Python/C#	Microsoft
<a href="#">Claude Agent SDK</a>	Jun 2025	4.4k	Python	Anthropic

## Comprehensive Comparison Table

A single reference table covering all evaluation dimensions.

### Core Framework Attributes

Dimension	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
First Release	Nov 2019	Aug 2023	Jun 2024	Apr 2025	Jun 2025
Stable API Since*	Dec 2023 (v2.0)	Jan 2024 (v0.1)	Jun 2024 (v0.0.1)	Apr 2025 (pre-release)	Jun 2025
Maturity	6 years	2.5 years	1.5 years	9 months	7 months
GitHub Stars	24k	24k	14.5k	6.8k	4.4k
License	Apache 2.0	MIT	MIT	MIT	MIT
Primary Language	Python	Python	Python	Python/C#/TS	Python
Python Version	3.9+	3.9+	3.9+	3.10+	3.10+
Maintainer	deepset	LangChain Inc	Pydantic	Microsoft	Anthropic
Install Command	<code>pip install haystack-ai</code>	<code>pip install -U langgraph</code>	<code>pip install pydantic-ai</code>	<code>pip install agent-framework --pre</code>	<code>pip install claude-agent-sdk</code>

\*Stable API Since: Relevant for LLM learnability benchmark. DeepSeek V3's training cutoff is July 2024, so frameworks stable before this date may benefit from training data contamination (LLM has "seen" code examples). Pydantic AI (Jun 2024) is borderline; MS Agent Framework and Claude Agent SDK are completely unseen.

### Code Verbosity (Lines of Code)

**Test Task:** Australian insurance catastrophic (CAT) event verification using a two-agent pipeline:

- Weather Agent** (with tools): Geocodes a location (city, state, postcode) via Nominatim API → fetches historical weather observations from Bureau of Meteorology (BOM) API
- Eligibility Agent** (pure LLM reasoning): Applies business rules to determine claim eligibility (APPROVED if both thunderstorms AND strong wind observed, REVIEW if one, DENIED if neither)

This task tests: tool definition, API integration, multi-agent handoff, structured output, and error handling.

Metric	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Demo File LoC	203	211	186	248	363
Core Logic LoC*	~110	~120	~90	~120	~180
Verbosity Rating	Medium	Medium	Low	Medium	High

\*Core logic excludes docstrings, error handling, and fallback code. Claude Agent SDK includes MCP server setup and fallback mode, inflating LoC.

Tool Definition & Type Safety

Aspect	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Tool Definition	@tool decorator or JSON schema	@tool from langchain_core	@agent.tool decorator	@tool decorator	@tool(name, desc, params)
Schema Style	Explicit JSON or Annotated	LangChain tool format	Pydantic models	Python type hints	Dict-based schema
Type Safety	Partial	Partial	Strong	Partial	Partial
Structured Output	Via integrations	Via LangChain	Native Pydantic models	Manual	Manual
Dependency Injection	Manual	Manual	RunContext pattern	Manual	Via hook context

Multi-Agent Patterns

Pattern	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Sequential	Pipeline components	Graph edges	Programmatic hand-off	Agent chaining	Manual orchestration
Delegation	Agent as ComponentTool	Subgraphs	@agent.tool calling agents	Nested agents	MCP tool calls
Supervisor/Router	Coordinator + tools	Conditional routing	Application logic	Orchestrator pattern	Custom routing
State Machine	Pipeline branching	StateGraph (native)	Pydantic Graph	Workflow graphs	Manual state
Parallel Execution	AsyncPipeline	Parallel nodes	asyncio.gather	Parallel branches	Multiple agents
Human-in-Loop	External	First-class	Manual	Workflow pause	Permission hooks

### Observability

Aspect	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Built-in	No	No	No	Yes	Hooks only
First-party Tool	—	LangSmith	Logfire	Azure Monitor	—
Standard	Varies	Proprietary	OpenTelemetry	OpenTelemetry	Custom
Setup Effort	Medium (3/5)	Low (1/5)	Low (1/5)	Low (1/5)	High (5/5)
Token Tracking	Via integration	Yes	Yes	Yes	DIY
Self-host Option	Yes	No	Enterprise	Yes	Yes (DIY)
Free Tier	Varies	Yes	Yes	Azure free tier	N/A

### Error Handling & Recovery

Aspect	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Validation Errors	Component-level	State validation	Pydantic models	Type hints	Hook-based
LLM Failures	Retry via integration	Checkpoint + resume	Auto-retry to LLM	Middleware	ProcessError
Tool Failures	Pipeline continues	Node-level handling	Error fed to LLM	Filters	PostToolUseFailure hook
Recovery Mechanism	Pipeline breakpoints	Checkpoint resume	Reflection loop	State snapshots	Session resume
Idempotency Support	Manual	First-class	Manual	Manual	Via hooks

### Workflow Capabilities

Capability	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Branching	ConditionalRouter	add_conditional_edges()	Application logic	Workflow graphs	Hook decisions
Loops	Component cycles	Graph cycles	While loops	Workflow loops	Recursive prompts
Checkpointing	No	Yes (durability modes)	No	Yes	Via hooks
Long-running Tasks	Limited	Excellent	Manual	Good	Manual

### Production Readiness

Factor	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Enterprise Use	Yes (deepset cloud)	Yes (Klarna, Replit)	Growing	Yes (Azure)	Growing
Documentation	Extensive	Extensive	Good	MS Learn	Growing
Breaking Changes	Stable (v2)	Evolving	Evolving	Very new	Very new
Risk Level	Low (1/5)	Low-Medium (2/5)	Low-Medium (2/5)	Medium (3/5)	High (5/5)
Jan 2026 Status	✅ Production ready	✅ Production ready	✅ Production ready	✅ Pre-release, functional	⚠️ Needs CLI setup

### Framework Overhead

Factor	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Dependencies	Medium	Medium (LangChain)	Light	Light	Light
Startup Time	Medium	Medium	Fast	Fast	Fast
Memory Footprint	Higher (RAG features)	Medium	Light	Light	Light
Model Lock-in	None	None	None	Azure-optimized	Claude-only

### LLM Learnability (DeepSeek V3 Benchmark)

**Experiment:** Measured how easily an LLM (DeepSeek V3, training cutoff July 2024) can produce working code for each framework. The LLM was given a simple classification task (insurance claims) and iteratively refined its code based on error feedback (up to 10 attempts). This simulates AI-assisted development where developers use LLMs to generate framework-specific code.

**What it measures:** Framework API intuitiveness, documentation clarity, and how well the framework patterns match common LLM training data. Higher success rate = LLM produces working code more reliably. Lower turns = faster convergence to working solution.

**Source:** [github.com/srepho/fwork-learnability](https://github.com/srepho/fwork-learnability)

### Experiment 1: Scraped Documentation (48 trials)

Metric	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Training Data Exposure	High (6+ months)	High (6+ months)	Minimal (~1 month)	None	None

<b>Success Rate</b>	75%	83%	<b>92%</b>	N/A <sup>†</sup>	N/A <sup>†</sup>
<b>Avg Turns to Success</b>	3.0	4.7	3.4	N/A <sup>†</sup>	N/A <sup>†</sup>
<b>First-Attempt Success</b>	25%	0%	<b>25%</b>	N/A <sup>†</sup>	N/A <sup>†</sup>
<b>Optimal Doc Level</b>	Minimal	Minimal	Moderate	N/A <sup>†</sup>	N/A <sup>†</sup>

<sup>†</sup>MS Agent Framework and Claude Agent SDK released after training cutoff (July 2024).

**Experiment 2: Curated Documentation (41 trials)**

A follow-up experiment with hand-curated documentation (real code examples instead of scraped landing pages):

Metric	Haystack	LangGraph	OpenAI Agents	Anthropic Agents
<b>Success Rate</b>	<b>100%</b>	<b>100%</b>	<b>100%</b>	92%
<b>Avg Turns to Success</b>	<b>1.75</b>	5.4	<b>1.65</b>	2.1
<b>Success@None (no docs)</b>	100%	100%	100%	100%

**Key insight:** Curated docs dramatically improved Haystack (75% → 100%) and reduced turns. The original experiment's scraped "full docs" were actually marketing pages with no code examples.

**Contamination Analysis**

All frameworks showed **high contamination** (100% success with zero documentation), confirming DeepSeek V3 has these frameworks memorized from training data:

Framework	Success@None	API Version Used	Contamination
Haystack	100%	v2 (modern)	HIGH - uses 2024 APIs
LangGraph	100%	v0/v1 mixed	HIGH - 50/50 version mix
Pydantic AI	100%	current	HIGH - despite June 2024 release
OpenAI Agents	100%	current	HIGH
Anthropic Agents	100%	current	HIGH

**Implication:** The benchmark measures "recall + refinement" rather than pure learnability. However, documentation still reduces average turns by 40-60%, showing its value even when the LLM has memorized patterns.

**Key Findings**

- Documentation quality matters more than quantity:** Minimal curated docs (real code) outperformed full scraped docs (marketing pages)
- LangGraph consistently slowest:** 4.7-5.4 avg turns across both experiments, suggesting steeper learning curve
- Pydantic AI performed well despite minimal training exposure:** 92% success with only ~1 month in training data
- All frameworks highly contaminated:** 100% success with no docs means true "learnability from scratch" can't be measured with DeepSeek V3

**Philosophy Summary**

Framework	Mental Model	Best Suited For
Haystack	Pipelines of swappable components	RAG, document processing, stability-focused teams
LangGraph	Agents as state machines	Complex workflows, explicit control, LangChain users
Pydantic AI	Type safety as foundation	Regulated industries, type-first developers
MS Agent Framework	Agents when necessary, functions otherwise	Azure/.NET shops, AutoGen/SK migrators
Claude Agent SDK	Claude Code as infrastructure	Claude-native apps, fine-grained permission control

**Overall Scores (Subjective)**

Category	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Ease of Setup	☆☆☆	☆☆☆	☆☆☆☆☆☆	☆☆☆☆	☆☆
Code Elegance	☆☆☆	☆☆☆	☆☆☆☆☆☆	☆☆☆☆	☆☆☆☆
Type Safety	☆☆	☆☆	☆☆☆☆☆☆	☆☆☆☆	☆☆
Workflow Control	☆☆☆	☆☆☆☆☆☆	☆☆☆☆	☆☆☆☆	☆☆☆☆
Observability	☆☆☆	☆☆☆☆	☆☆☆☆	☆☆☆☆☆☆	☆☆
Production Ready	☆☆☆☆☆☆	☆☆☆☆	☆☆☆☆	☆☆☆☆	☆☆
Community/Docs	☆☆☆☆☆☆	☆☆☆☆☆☆	☆☆☆☆	☆☆☆☆	☆☆☆☆
OVERALL	☆☆☆☆	☆☆☆☆	☆☆☆☆☆☆	☆☆☆☆	☆☆

**Haystack**

**The most mature framework** - Originally built for semantic search and QA, evolved to support LLM agents.

GitHub	<a href="https://github.com/deepset-ai/haystack">https://github.com/deepset-ai/haystack</a>
Documentation	<a href="https://docs.haystack.deepset.ai/docs/intro">https://docs.haystack.deepset.ai/docs/intro</a>
Quick Start	<a href="https://haystack.deepset.ai/overview/quick-start">https://haystack.deepset.ai/overview/quick-start</a>
Tutorials	<a href="https://haystack.deepset.ai/tutorials">https://haystack.deepset.ai/tutorials</a>
First Release	November 2019 (~6 years old)
Stars	24,000
Language	Python
License	Apache 2.0

**Description:** AI orchestration framework to build customizable, production-ready LLM applications. Connect components (models, vector DBs, file converters) to pipelines or agents that can interact with your data.

**Strengths:**

- Most battle-tested in production
- Strong RAG and document processing capabilities
- Extensive integrations ecosystem

---

## LangGraph

**State machine approach** - Low-level orchestration for building stateful agents as graphs.

GitHub	<a href="https://github.com/langchain-ai/langgraph">https://github.com/langchain-ai/langgraph</a>
Documentation	<a href="https://docs.langchain.com/oss/python/langgraph/">https://docs.langchain.com/oss/python/langgraph/</a>
API Reference	<a href="https://reference.langchain.com/python/langgraph/">https://reference.langchain.com/python/langgraph/</a>
First Release	August 2023 (~2.5 years old)
Stars	24,000
Language	Python (99.3%)
License	MIT

**Description:** Build resilient language agents as graphs. A low-level orchestration framework for building, managing, and deploying long-running, stateful agents.

**Strengths:**

- Explicit state management with conditional edges
- Tight integration with LangChain ecosystem
- Free structured course via LangChain Academy
- Used by Klarna, Replit, Elastic

---

## Pydantic AI

**Type-safe agents** - Production-grade AI applications the Pydantic way.

GitHub	<a href="https://github.com/pydantic/pydantic-ai">https://github.com/pydantic/pydantic-ai</a>
Documentation	<a href="https://ai.pydantic.dev/">https://ai.pydantic.dev/</a>
First Release	June 2024 (~1.5 years old)
Stars	14,500
Language	Python (99.8%)
License	MIT

**Description:** GenAI Agent Framework built with Pydantic's philosophy of type safety and validation.



### Strengths:

- Strong typing and Pydantic validation throughout
  - Structured outputs with validation
  - Dependency injection pattern
  - Observability via Pydantic Logfire
  - Multiple LLM provider support
- 

## Microsoft Agent Framework

**Enterprise multi-language** - Unified framework consolidating AutoGen and Semantic Kernel.

GitHub	<a href="https://github.com/microsoft/agent-framework">https://github.com/microsoft/agent-framework</a>
Documentation	<a href="https://learn.microsoft.com/agent-framework/overview/agent-framework-overview">https://learn.microsoft.com/agent-framework/overview/agent-framework-overview</a>
Quick Start	<a href="https://learn.microsoft.com/agent-framework/tutorials/quick-start">https://learn.microsoft.com/agent-framework/tutorials/quick-start</a>
User Guide	<a href="https://learn.microsoft.com/en-us/agent-framework/user-guide/overview">https://learn.microsoft.com/en-us/agent-framework/user-guide/overview</a>
Discord	<a href="https://discord.gg/b5zjErwbQM">https://discord.gg/b5zjErwbQM</a>
First Release	April 2025 (~9 months old)
Stars	6,800
Languages	Python (50.8%), C# (44.3%), TypeScript (4.5%)
License	MIT

**Description:** A framework for building, orchestrating and deploying AI agents and multi-agent workflows with support for Python and .NET.

### Strengths:

- Multi-language support (Python, C#, TypeScript)
  - Migration paths from AutoGen and Semantic Kernel
  - Microsoft enterprise backing
  - Azure integration
- 

## Claude Agent SDK

**Claude-native** - Official SDK for building agents with Claude Code capabilities.

GitHub	<a href="https://github.com/anthropics/claude-agent-sdk-python">https://github.com/anthropics/claude-agent-sdk-python</a>
Documentation	<a href="https://platform.claude.com/docs/en/agent-sdk/python">https://platform.claude.com/docs/en/agent-sdk/python</a>
Hooks Reference	<a href="https://docs.anthropic.com/en/docs/claude-code/hooks">https://docs.anthropic.com/en/docs/claude-code/hooks</a>
First Release	June 2025 (~7 months old)
Stars	4,400

Language	Python
License	MIT

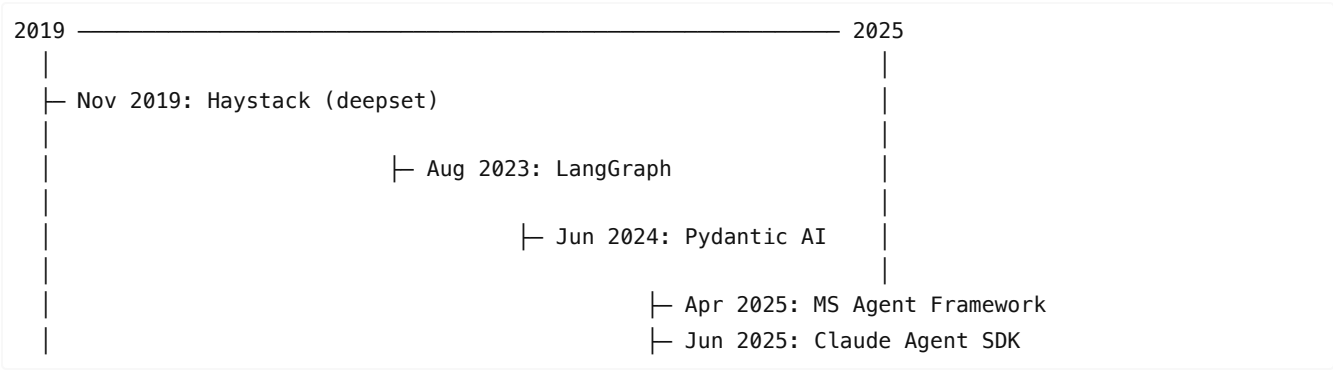
**Description:** Python SDK for Claude Agent - enables programmatic interaction with Claude Code, including file operations, code execution, and custom tool definitions.

**Strengths:**

- Native Claude Code integration (Read, Write, Bash tools)
- In-process MCP server support (no subprocess overhead)
- Permission hooks for fine-grained control
- Bundled CLI - no separate installation needed

**Note:** Requires Python 3.10+. Governed by Anthropic's Commercial Terms of Service.

## Maturity Timeline



## Installation

```
# Haystack
pip install haystack-ai

# LangGraph
pip install -U langgraph

# Pydantic AI
pip install pydantic-ai

# Microsoft Agent Framework
pip install -U agent-framework --pre

# Claude Agent SDK
pip install claude-agent-sdk
```

## Tool Definition Syntax

### Haystack

Three approaches - explicit `Tool` class, `@tool` decorator, or `create_tool_from_function` :

```
# Decorator approach (simplest)
from haystack.tools import tool
from typing import Annotated

@tool
def get_weather(
    city: Annotated[str, "the city for which to get the weather"] = "Munich"
):
    '''A simple function to get the current weather for a location.'''
    return f"Weather report for {city}: 20 Celsius, sunny"

# Explicit class (full control)
from haystack.tools import Tool

add_tool = Tool(
    name="addition_tool",
    description="This tool adds two numbers",
    parameters={
        "type": "object",
        "properties": {
            "a": {"type": "integer"},
            "b": {"type": "integer"}
        },
        "required": ["a", "b"]
    },
    function=add
)
```

## LangGraph

Uses LangChain's tool system - pass functions directly or use decorator:

```
from langgraph.graph import StateGraph, MessagesState, START, END

def get_weather(city: str) -> str:
    """Get weather for a given city."""
    return f"It's always sunny in {city}!"

# Tools passed to create_agent() or bound to model
```

## Pydantic AI

Decorator on agent instance with dependency injection via `RunContext` :

```
from pydantic_ai import Agent, RunContext

agent = Agent('anthropic:claude-sonnet-4-0')

@agent.tool
```

```
def get_weather(ctx: RunContext, city: str) -> str:
    """Get weather for a city.""" # Docstring = tool description
    return f"Weather in {city}: sunny"
```

## Microsoft Agent Framework

Tools defined as Python functions passed to agent configuration:

```
from agent_framework.azure import AzureAIClient

def get_weather(city: str) -> str:
    """Get weather for a given city."""
    return f"Weather in {city}: sunny"

# Tools bound during agent creation
```

## Claude Agent SDK

In-process MCP server with `@tool` decorator:

```
from claude_agent_sdk import tool, create_sdk_mcp_server

@tool("get_weather", "Get weather for a city", {"city": str})
async def get_weather(args):
    return {
        "content": [{"type": "text", "text": f"Weather: sunny in {args['city']}"}]
    }

server = create_sdk_mcp_server(
    name="weather-tools",
    version="1.0.0",
    tools=[get_weather]
)
```

---

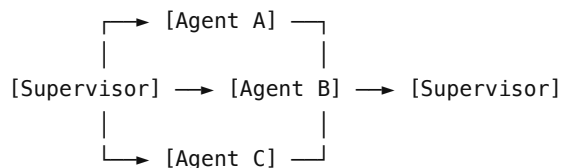
## Multi-Agent Orchestration Patterns

### Common Patterns Visualized

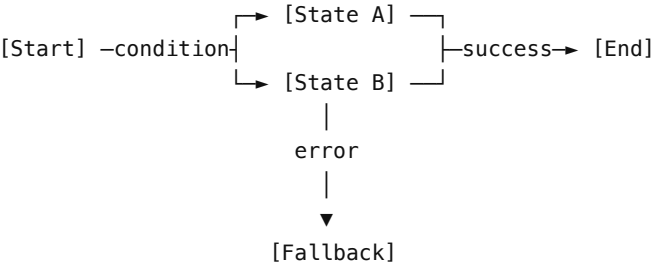
Sequential Pattern:

[Agent 1] → [Agent 2] → [Agent 3]

Supervisor/Router Pattern:



State Machine Pattern:



Pattern	Haystack	LangGraph	Pydantic AI	MS Agent Framework	Claude Agent SDK
Sequential	Pipeline components	Graph edges	Programmatic hand-off	Agent chaining	Manual orchestration
Delegation	Agent as ComponentTool	Subgraphs	@agent.tool calling other agents	Nested agents	MCP tool calls
Supervisor	Coordinator agent + tools	Conditional routing	Application logic	Orchestrator pattern	Custom routing
State Machine	Pipeline branching	StateGraph + conditions	Graph-based control	State management	Manual state

Pattern Details

**Haystack:** Wrap an Agent as a `ComponentTool` so a coordinator agent can invoke specialized sub-agents.

**LangGraph:** First-class support via `StateGraph` with `add_conditional_edges()` for routing decisions.

**Pydantic AI:** Five levels of complexity:

- 1. Single agent
- 2. Agent delegation (agent calls agent via tool)
- 3. Programmatic hand-off (app logic decides)
- 4. Graph-based control (state machines)
- 5. Deep agents (autonomous with planning)

**MS Agent Framework:** Inherits patterns from AutoGen (group chats) and Semantic Kernel (plugins).

**Claude Agent SDK:** Manual orchestration - you control the flow in your application code.

Production Readiness Assessment

Factor	Haystack	LangGraph	Pydantic AI	MS Agent	Claude SDK
Maturity	6 years	2.5 years	1.5 years	9 months	7 months
Community	24k stars	24k stars	14.5k stars	6.8k stars	4.4k stars
Enterprise Use	Yes (deepset cloud)	Yes (Klarna, Replit)	Growing	Yes (Azure)	Growing
Observability	Integrations	LangSmith	Logfire	Azure Monitor	Hooks

Type Safety	Partial	Partial	Strong	Partial	Partial
Documentation	Extensive	Extensive	Good	MS Learn	Growing
Breaking Changes	Stable (v2)	Evolving	Evolving	Very new	Very new

### Risk Assessment

Production Risk Level (lower = safer)

Haystack

Low

✓ Safe

LangGraph

Low-Medium

✓ Safe

Pydantic AI

Low-Medium

✓ Safe

MS Agent

Medium

⚠ Pre-release

Claude SDK

High

⚠ Needs setup

1

2

3

4

5

Framework	Risk Level	Notes
Haystack	Low	Mature, stable API, Apache 2.0 license
LangGraph	Low-Medium	Active development, MIT license, large community
Pydantic AI	Low-Medium	Strong backing (Pydantic team), MIT, tested working
MS Agent Framework	Medium	Pre-release (--pre), but functional, Microsoft backing
Claude Agent SDK	High	MCP integration issues, requires Claude Code CLI

January 2026 Testing Update:

- MS Agent Framework: Works with `pip install agent-framework --pre`. Uses familiar patterns (`ChatAgent`, `@tool` decorator). Pre-release but functional.
- Claude Agent SDK: Installs but MCP server has TaskGroup errors. Basic fallback works but full functionality requires additional Claude Code setup.

## Philosophy & Mental Models

Understanding each framework's philosophy helps developers write idiomatic code and make better architectural decisions.

### Haystack: "Production-Ready Pipelines"

**Core Philosophy:** Modularity and production-readiness from day one.

**Mental Model:** Think of your application as a **pipeline of swappable components**. Each component (retriever, generator, ranker, agent) is a building block you can swap without rewriting core logic.

**Key Principles:**

- Composability over monoliths** - Build pipelines from small, focused components
- Swap without rewriting** - Change model providers, vector DBs, or tools without architectural changes
- Control your data flow** - Add loops, branches, and custom routing to fit your use case
- Reliability over complexity** - Clean architecture with careful dependency management

### Developer Mindset:

```
"I'm building a workflow from composable pieces. Each piece does one thing well.  
If I need to change providers or add a step, I swap or insert a component."
```

**Best suited for:** Developers who value stability, have complex document processing needs, or want maximum flexibility in choosing infrastructure.

---

## LangGraph: "Agents as State Machines"

**Core Philosophy:** Low-level orchestration with explicit state management.

**Mental Model:** Think of your agent as a **graph of states and transitions**. Each node is a processing step, each edge is a possible transition. State flows through the graph explicitly.

### Key Principles:

- **Explicit over implicit** - You define every state and transition, nothing is hidden
- **Durable execution** - Agents persist through failures and resume from checkpoints
- **Human-in-the-loop** - Inspect and modify agent state at any point
- **Observability first** - Deep visibility into complex agent behavior

### Developer Mindset:

```
"I'm designing a state machine. What are my states? What triggers transitions?  
What data needs to persist between steps? Where might I need human intervention?"
```

**Best suited for:** Developers who need fine-grained control over execution flow, are comfortable with graph-based thinking, or need robust recovery from failures.

---

## Pydantic AI: "Type Safety as Foundation"

**Core Philosophy:** Catch errors at development time, not runtime.

**Mental Model:** Think of agents as **typed, reusable components** - like a FastAPI router or a well-typed class. The type system is your safety net; trust it.

### Key Principles:

- **Types are contracts** - Define inputs, outputs, and dependencies with Pydantic models
- **Agents are reusable** - Instantiate once, use globally (like a FastAPI app)
- **Explicit dependencies** - Use dependency injection for testability and clarity
- **Observability is essential** - "You need to actually see what happened" (built-in Logfire support)

### Developer Mindset:

```
"What types go in? What types come out? What dependencies does this agent need?  
If my types are right, my agent is probably right. If something fails, I can trace it."
```

**Best suited for:** Developers who love type hints, want IDE autocomplete and static analysis, or come from strongly-typed language backgrounds.

---

## Microsoft Agent Framework: "Agents When Necessary"

**Core Philosophy:** Use agents for autonomous decision-making, functions for everything else.

**Mental Model:** Think in terms of **"Do I need autonomy?"** If a task is well-defined and sequential, write a function. If it requires exploration, planning, or conversation - use an agent.

**Key Principles:**

- **Pragmatism first** - "If you can write a function to handle the task, do that instead"
- **Agents + Workflows continuum** - Single agents for autonomy, workflows for orchestration
- **Enterprise-grade foundations** - Sessions, state management, filters, telemetry built-in
- **Multi-language parity** - Same patterns work in Python, C#, and TypeScript

**Developer Mindset:**

```
"Does this task need autonomous decision-making? If yes, agent. If no, function.  
For complex multi-step processes, compose agents into workflows with explicit control."
```

**Best suited for:** Enterprise teams, .NET shops, developers migrating from AutoGen/Semantic Kernel, or those who want strong guardrails against over-using agents.

---

**Claude Agent SDK: "Claude Code as Infrastructure"**

**Core Philosophy:** Leverage Claude's native capabilities (file ops, code execution) directly.

**Mental Model:** Think of Claude Code as **infrastructure you're programming against**. Your code provides tools and hooks; Claude Code provides the execution environment.

**Key Principles:**

- **Native capabilities** - Read, Write, Bash tools are built-in, not simulated
- **Hooks for control** - Intercept any action at any lifecycle point
- **Permission-based security** - Fine-grained control over what Claude can do
- **Event-driven architecture** - React to session events, tool calls, and completions

**Developer Mindset:**

```
"Claude Code is my execution environment. I provide tools via MCP servers.  
I use hooks to enforce policies, add context, or intercept dangerous operations."
```

**Best suited for:** Developers building Claude-native applications, those who want tight integration with Claude Code's capabilities, or teams needing fine-grained permission control.

---

**Observability Deep Dive**

Understanding what's happening inside your agents is critical for debugging, optimization, and trust.

**Haystack: Integration Ecosystem**

**Approach:** No built-in observability - instead, integrates with your choice of providers.

**Available Integrations:**

Provider	Type	Maintained By
Arize Phoenix	Tracing	Community
Arize AI	Tracing + Monitoring	Community



<b>Langfuse</b>	Tracing + Monitoring	deepset (official)
<b>OpenLIT</b>	Monitoring + Evaluation	Community
<b>Opik</b>	Tracing + Evaluation	Community
<b>Traceloop</b>	Quality Evaluation	Community
<b>Weights &amp; Biases Weave</b>	Tracing + Visualization	deepset (official)

#### What You Can See:

- Pipeline execution flow
- Component-level latency
- Token usage per step
- Error locations and stack traces

#### Setup Example:

```
# Langfuse integration
from haystack_integrations.components.connectors.langfuse import LangfuseConnector

tracer = LangfuseConnector()
pipeline.add_component("tracer", tracer)
```

### LangGraph: LangSmith Integration

**Approach:** First-party observability via LangSmith (works with or without LangChain).

#### Key Features:

- **Full trace visibility** - See every step your agent takes
- **Quality tracking** - Measure and track quality over time
- **Framework agnostic** - Works with LangGraph, LangChain, or standalone
- **Development to production** - Same tooling from local dev to prod

#### What You Can See:

- Complete request lifecycle
- Token usage and latency per step
- Tool calls and their results
- State transitions in the graph
- Human feedback integration

#### Setup:

```
import os
os.environ["LANGCHAIN_TRACING_V2"] = "true"
os.environ["LANGCHAIN_API_KEY"] = "your-api-key"

# That's it - tracing is automatic
```

**Pricing:** Free tier available, paid tiers for higher volume.

## Pydantic AI: Logfire Integration

**Approach:** Built on OpenTelemetry with first-party Logfire support. Two lines to enable.

### Key Features:

- **Automatic instrumentation** - No code changes beyond setup
- **SQL queryable** - Query telemetry data with SQL
- **HTTP transparency** - See actual prompts/completions sent to providers
- **Multi-agent tracing** - Correlate traces across agent delegation

### What You Can See:

- Agent runs (top-level traces)
- Model requests (API calls)
- Tool execution spans
- Token usage and latency
- Full request/response bodies (optional)

### Setup:

```
import logfire

logfire.configure()
logfire.instrument_pydantic_ai()

# Now all agent runs are traced automatically
```

**Pricing:** Free tier available, commercial hosting, self-hosting on enterprise.

---

## Microsoft Agent Framework: OpenTelemetry Native

**Approach:** Built on OpenTelemetry with Azure Monitor integration. Most comprehensive built-in observability.

### Key Features:

- **OpenTelemetry standard** - Follows GenAI Semantic Conventions
- **Traces, Logs, Metrics** - All three pillars supported
- **Azure Monitor integration** - First-class support for Azure environments
- **Sensitive data controls** - Enable/disable prompt logging

### What You Can See:

- `invoke_agent <name>` - Top-level agent invocation
- `chat <model>` - Chat model calls
- `execute_tool <function>` - Tool execution
- Operation duration histograms
- Token usage histograms
- Function invocation timing

### Setup (Python):

```
from agent_framework.observability import configure_otel_providers

# Option 1: Environment variables (recommended)
```

```
# Set OTEL_EXPORTER_OTLP_ENDPOINT, ENABLE_INSTRUMENTATION=true
configure_otel_providers()

# Option 2: Azure Monitor
from azure.monitor.opentelemetry import configure_azure_monitor
configure_azure_monitor(connection_string="...")
```

Environment Variables:

Variable	Purpose
ENABLE_INSTRUMENTATION	Enable OpenTelemetry
ENABLE_SENSITIVE_DATA	Log prompts/responses
OTEL_EXPORTER_OTLP_ENDPOINT	OTLP endpoint
OTEL_SERVICE_NAME	Service identifier

**Local Development:** Use Aspire Dashboard (Docker) for local trace visualization.

Claude Agent SDK: Hook-Based Observability

**Approach:** Event-driven hooks at every lifecycle point. You build observability into your hooks.

Hook Lifecycle Events:

Event	When It Fires	Use For
SessionStart	Session begins	Load context, set env vars
UserPromptSubmit	User submits prompt	Validate, add context
PreToolUse	Before tool execution	Approve/deny/modify
PostToolUse	After tool succeeds	Log, validate results
PostToolUseFailure	After tool fails	Error handling
SubagentStart	Spawning subagent	Track nested agents
SubagentStop	Subagent finishes	Aggregate results
Stop	Claude finishes	Verify completion
SessionEnd	Session terminates	Cleanup, final logging

What You Can Capture:

- Every tool call (name, input, output)
- Session and transcript paths
- Permission decisions
- Subagent execution chains
- Custom metrics via your hook scripts

Setup:

```
// ~/.claude/settings.json
{
  "hooks": {
    "PostToolUse": [{
      "matcher": "*",
      "hooks": [{
        "type": "command",
        "command": "/path/to/log-tool-use.py"
      }]
    }]
  }
}
```

Hook Input (JSON via stdin):

```
{
  "session_id": "abc123",
  "transcript_path": "/path/to/transcript.jsonl",
  "tool_name": "Bash",
  "tool_input": {"command": "ls -la"},
  "tool_response": {"output": "..."}
}
```

**Key Difference:** Unlike other frameworks, observability is DIY - you decide what to log, where to send it, and how to visualize it. Maximum flexibility, more setup required.

## Observability Comparison

Observability Setup Effort (lower = easier)

Haystack	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	Medium	(integrations)
LangGraph	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	Low	(2 env vars)
Pydantic AI	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	Low	(2 lines)
MS Agent	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	Low	(built-in)
Claude SDK	<div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div><div></div></div>	High	(DIY hooks)
	12345		

Aspect	Haystack	LangGraph	Pydantic AI	MS Agent	Claude SDK
Built-in	No	No	No	Yes	Hooks only
First-party Tool	-	LangSmith	Logfire	Azure Monitor	-
Standard	Varies	Proprietary	OpenTelemetry	OpenTelemetry	Custom
Setup Effort	Medium	Low	Low	Low	High
Self-host Option	Yes	No	Enterprise	Yes	Yes (DIY)
Token Tracking	Via integration	Yes	Yes	Yes	DIY

Multi-agent	Yes	Yes	Yes	Yes	Via hooks
Free Tier	Varies	Yes	Yes	Azure free tier	N/A

## Framework Overhead

**Note:** LLM API costs are model-dependent, not framework-dependent. Framework overhead is minimal compared to LLM latency/cost.

Factor	Haystack	LangGraph	Pydantic AI	MS Agent	Claude SDK
Dependencies	Medium	Medium (LangChain)	Light	Light	Light
Python Version	3.9+	3.9+	3.9+	3.10+	3.10+
Startup Time	Medium	Medium	Fast	Fast	Fast (CLI bundled)
Memory Footprint	Higher (RAG features)	Medium	Light	Light	Light
Model Lock-in	None	None	None	Azure-optimized	Claude-only

## Failure Modes & Error Handling

How each framework handles things going wrong - critical for production systems.

### Error Handling Comparison

Aspect	Haystack	LangGraph	Pydantic AI	MS Agent	Claude SDK
Validation Errors	Component-level	State validation	Pydantic models	Type hints	Hook-based
LLM Failures	Retry via integration	Checkpoint + resume	Auto-retry to LLM	Middleware	ProcessError
Tool Failures	Pipeline continues	Node-level handling	Error fed to LLM	Filters	PostToolUseFailure hook
Recovery	Pipeline breakpoints	Checkpoint resume	Reflection loop	State snapshots	Session resume
Idempotency	Manual	First-class support	Manual	Manual	Via hooks

### Haystack: Pipeline-Based Error Handling

**Approach:** Errors propagate through the pipeline; use breakpoints for debugging.

```

# Pipeline with error handling via try/except
from haystack import Pipeline
from haystack.components.generators import OpenAIGenerator

pipeline = Pipeline()
pipeline.add_component("generator", OpenAIGenerator())

try:
    result = pipeline.run({"generator": {"prompt": "Hello"}})
except Exception as e:
    # Access pipeline state at failure point
    print(f"Pipeline failed: {e}")

```

Key Features:

- **Pipeline breakpoints** - Pause and inspect state at any component
- **Component isolation** - Failures in one component don't necessarily stop others
- **Fallback patterns** - Tutorials show "Agentic RAG with Fallback to Websearch"

Failure Modes to Watch:

- Component timeout (no built-in timeout management)
- Invalid data types between components
- External API failures (handled per-integration)

LangGraph: Checkpoint-Based Recovery

**Approach:** Durable execution with automatic checkpointing. Resume from last good state.

```

from langgraph.graph import StateGraph
from langgraph.checkpoint.memory import InMemorySaver

# Enable checkpointing
checkpointer = InMemorySaver()
graph = workflow.compile(checkpointer=checkpointer)

# Run with thread ID for tracking
config = {"configurable": {"thread_id": "my-thread-123"}}
result = graph.invoke(input_data, config)

# If failure occurs, resume from checkpoint
# Same thread_id + None input = resume from last checkpoint
resumed = graph.invoke(None, config)

```

Durability Modes:

Mode	Behavior	Use Case
"exit"	Persist only at completion	Best performance
"async"	Persist asynchronously during execution	Balanced

"sync"	Persist synchronously before each step	Highest durability
--------	----------------------------------------	--------------------

### Key Principle - Idempotency:

*"For reliable retry mechanisms, ensure side effects are idempotent."*

If a task fails partway through, resumption re-runs it. Use idempotency keys or result verification to prevent duplicate actions.

### Failure Modes to Watch:

- Non-idempotent side effects on retry
- Checkpoint storage failures
- State serialization errors

## Pydantic AI: Validation-Driven Recovery

**Approach:** Let validation errors guide the LLM to self-correct. Automatic reflection loops.

```
from pydantic_ai import Agent
from pydantic import BaseModel

class WeatherResponse(BaseModel):
    temperature: float
    conditions: str

agent = Agent('openai:gpt-4', result_type=WeatherResponse)

# If LLM returns invalid data, Pydantic AI:
# 1. Catches validation error
# 2. Sends error back to LLM
# 3. LLM retries with corrected output
result = await agent.run("What's the weather in Sydney?")
```

### Error Handling Features:

- **Validation recovery** - "Errors are passed back to the LLM so it can retry"
- **Structured output retry** - If response doesn't match schema, agent is "prompted to try again"
- **HTTP request retries** - Built-in retry logic for transient API failures
- **Durable execution** - "Preserve progress across transient API failures"

### Failure Modes to Watch:

- Infinite retry loops on fundamentally invalid requests
- LLM unable to produce valid schema after multiple attempts
- Dependency injection failures

## Microsoft Agent Framework: Middleware-Based Interception

**Approach:** Use middleware/filters to intercept and handle errors at any layer.

```
from agent_framework import AIAgent
from agent_framework.middleware import RetryMiddleware, LoggingMiddleware
```

```
# Stack middleware for comprehensive error handling
agent = (
    AIAgent(client, instructions="...")
    .with_middleware(RetryMiddleware(max_retries=3))
    .with_middleware(LoggingMiddleware())
)

# Middleware intercepts:
# - Pre-execution (validation, rate limiting)
# - Post-execution (logging, metrics)
# - Errors (retry logic, fallbacks)
```

#### Key Features:

- **Filters** - Intercept agent actions before/after execution
- **Middleware stack** - Compose error handling behaviors
- **Session state** - Restore from session snapshots
- **Azure integration** - Application Insights for error tracking

#### Failure Modes to Watch:

- Middleware ordering issues
- State corruption on partial failures
- Azure service dependencies

---

## Claude Agent SDK: Hook-Based Error Handling

**Approach:** React to failures via lifecycle hooks. You control recovery logic.

```
from claude_agent_sdk import ClaudeSDKClient, ClaudeAgentOptions, HookMatcher

async def handle_tool_failure(input_data, tool_use_id, context):
    """Called when any tool fails."""
    tool_name = input_data.get('tool_name')
    error = input_data.get('tool_response', {}).get('error')

    # Log the failure
    print(f"Tool {tool_name} failed: {error}")

    # Optionally provide guidance to Claude
    return {
        'hookSpecificOutput': {
            'additionalContext': f"The {tool_name} tool failed. Try an alternative approach."
        }
    }

options = ClaudeAgentOptions(
    hooks={
        'PostToolUseFailure': [
            HookMatcher(hooks=[handle_tool_failure])
        ]
    }
)
```



```
        }
    )
}
```

Exception Types:

Exception	When
CLINotFoundError	Claude Code CLI not installed
CLIConnectionError	Failed to connect to Claude Code
ProcessError	Claude Code process failed (has exit_code, stderr)
CLIJSONDecodeError	Response parsing failed

Recovery Options:

- `resume` option to continue from previous session
- `fallback_model` for automatic model fallback
- Hook-based retry logic (DIY)

Failure Modes to Watch:

- CLI process crashes
- Hook script failures
- Permission denials interrupting workflow

## Complex Workflow Patterns

How each framework handles branching, loops, parallel execution, and conditional logic.

Workflow Capabilities Comparison

Pattern	Haystack	LangGraph	Pydantic AI	MS Agent	Claude SDK
Branching	ConditionalRouter	add_conditional_edges()	Application logic	Workflow graphs	Hook decisions
Loops	Component cycles	Graph cycles	While loops	Workflow loops	Recursive prompts
Parallel	AsyncPipeline	Parallel nodes	asyncio.gather	Parallel branches	Multiple agents
Human-in-loop	External	First-class	Manual	Workflow pause	Permission hooks
State Machines	Pipeline state	StateGraph (native)	Pydantic Graph	Workflow state	Manual

Haystack: Pipeline Branching & Loops

Branching with ConditionalRouter:

```

from haystack.components.routers import ConditionalRouter

routes = [
    {"condition": "{{query|length > 100}}", "output": "long_query", "output_type": str},
    {"condition": "{{query|length <= 100}}", "output": "short_query", "output_type": str},
]

router = ConditionalRouter(routes=routes)
pipeline.add_component("router", router)

# Connect different paths
pipeline.connect("router.long_query", "detailed_processor")
pipeline.connect("router.short_query", "quick_processor")

```

### Self-Correcting Loops:

```

# Validator loops back to generator for correction
pipeline.connect("generator", "validator")
pipeline.connect("validator.invalid", "generator") # Loop back
pipeline.connect("validator.valid", "output")      # Continue

```

### Parallel Execution:

```

from haystack import AsyncPipeline

# Components run in parallel when dependencies allow
async_pipeline = AsyncPipeline()

# Independent branches execute concurrently

```

---

## LangGraph: Native Graph-Based Workflows

### Conditional Branching:

```

from langgraph.graph import StateGraph, END

def should_continue(state):
    if state["error_count"] > 3:
        return "fallback"
    elif state["needs_review"]:
        return "human_review"
    return "continue"

graph = StateGraph(MyState)
graph.add_node("process", process_node)
graph.add_node("fallback", fallback_node)
graph.add_node("human_review", review_node)

# Conditional routing
graph.add_conditional_edges(
    "process",

```

```

    should_continue,
    {
        "continue": "process",      # Loop
        "fallback": "fallback",    # Branch
        "human_review": "human_review"
    }
)

```

### Human-in-the-Loop (First-Class):

```

from langgraph.checkpoint.memory import InMemorySaver

# Interrupt before sensitive operations
graph.add_node("sensitive_action", sensitive_node)
graph.set_entry_point("sensitive_action")

# Compile with interrupt_before
app = graph.compile(
    checkpointer=InMemorySaver(),
    interrupt_before=["sensitive_action"]
)

# Execution pauses, human reviews, then resume

```

### Parallel Execution:

```

# Multiple nodes can execute in parallel if no dependencies
graph.add_node("fetch_weather", weather_node)
graph.add_node("fetch_news", news_node)
graph.add_edge(START, "fetch_weather")
graph.add_edge(START, "fetch_news") # Both run in parallel
graph.add_edge("fetch_weather", "combine")
graph.add_edge("fetch_news", "combine")

```

---

## Pydantic AI: Programmatic Control + Optional Graph

### Simple Branching (Application Logic):

```

from pydantic_ai import Agent

weather_agent = Agent('openai:gpt-4', system_prompt="Weather expert")
news_agent = Agent('openai:gpt-4', system_prompt="News analyst")

async def route_query(query: str):
    if "weather" in query.lower():
        return await weather_agent.run(query)
    elif "news" in query.lower():
        return await news_agent.run(query)
    else:

```

```
# Default agent or error
return await general_agent.run(query)
```

#### Parallel Execution:

```
import asyncio

async def parallel_agents(query: str):
    # Run multiple agents concurrently
    results = await asyncio.gather(
        weather_agent.run(query),
        news_agent.run(query),
        sentiment_agent.run(query)
    )
    return combine_results(results)
```

#### Pydantic Graph (For Complex Workflows):

```
from pydantic_ai.graph import Graph, Node
from dataclasses import dataclass

@dataclass
class FetchData(Node):
    url: str

    async def run(self) -> "ProcessData | ErrorState":
        try:
            data = await fetch(self.url)
            return ProcessData(data=data)
        except Exception:
            return ErrorState(message="Fetch failed")

@dataclass
class ProcessData(Node):
    data: dict

    async def run(self) -> "Complete | NeedsReview":
        if self.data.get("confidence") > 0.9:
            return Complete(result=self.data)
        return NeedsReview(data=self.data)

# Type annotations define the graph edges automatically
```

**Caution from docs:** "If you're not confident a graph-based approach is a good idea, it might be unnecessary." Use simple programmatic control first.

---

## Microsoft Agent Framework: Workflow Orchestration

### Explicit Workflow Graphs:

```

from agent_framework.workflows import Workflow, WorkflowNode

workflow = Workflow()

# Define nodes
workflow.add_node("classify", classify_agent)
workflow.add_node("process_urgent", urgent_agent)
workflow.add_node("process_normal", normal_agent)

# Conditional routing
workflow.add_conditional_edge(
    "classify",
    lambda state: "urgent" if state["priority"] == "high" else "normal",
    {"urgent": "process_urgent", "normal": "process_normal"}
)

```

#### Parallel Branches:

```

# Multiple agents run in parallel
workflow.add_parallel_nodes(
    "gather_info",
    [weather_agent, traffic_agent, calendar_agent]
)

```

#### Human-in-the-Loop:

```

# Workflows can pause for human input
workflow.add_node("human_review", human_review_node)
workflow.set_interrupt_before("human_review")

# Checkpointing for long-running processes
workflow.enable_checkpointing()

```

---

## Claude Agent SDK: Hook-Driven Control Flow

#### Branching via Hooks:

```

async def route_based_on_tool(input_data, tool_use_id, context):
    """Intercept and redirect tool calls."""
    tool_name = input_data.get('tool_name')

    if tool_name == "Write" and "/sensitive/" in input_data.get('tool_input', {}).get('file_path', ''):
        return {
            'hookSpecificOutput': {
                'permissionDecision': 'deny',
                'permissionDecisionReason': 'Sensitive directory - use secure_write tool instead'
            }
        }

```

```
}  
return {}
```

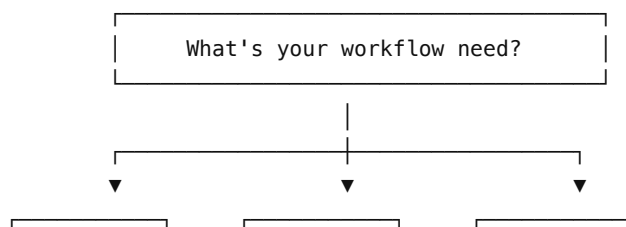
#### Parallel Agents:

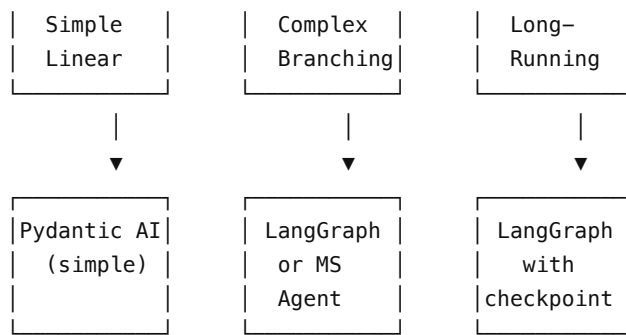
```
import asyncio  
from claude_agent_sdk import query, ClaudeAgentOptions  
  
async def parallel_queries():  
    tasks = [  
        query(prompt="Analyze code quality", options=code_options),  
        query(prompt="Check security issues", options=security_options),  
        query(prompt="Review documentation", options=docs_options)  
    ]  
  
    # Process all agents in parallel  
    results = []  
    for task in asyncio.as_completed(tasks):  
        async for message in await task:  
            results.append(message)
```

#### Stop Hook for Conditional Continuation:

```
async def should_continue(input_data, tool_use_id, context):  
    """Decide if Claude should continue working."""  
    # Check if all tasks are complete  
    transcript = read_transcript(input_data['transcript_path'])  
  
    if not all_tasks_done(transcript):  
        return {  
            'decision': 'block',  
            'reason': 'Not all tasks completed. Continue with remaining items.'  
        }  
    return {}  
  
options = ClaudeAgentOptions(  
    hooks={  
        'Stop': [HookMatcher(hooks=[should_continue])]  
    }  
)
```

### Workflow Pattern Decision Guide





Need RAG? —————> Haystack pipelines  
Need Claude Code? —————> Claude Agent SDK with hooks  
Need .NET? —————> MS Agent Framework

## Real-World Implementation Testing

We implemented the **same use case** across multiple frameworks to compare developer experience and behavior.

### Note on framework versions:

- **Tested:** LangGraph, Haystack, Pydantic AI, AutoGen 0.4+
- **Not tested (too new):** Microsoft Agent Framework (Apr 2025), Claude Agent SDK (Jun 2025)

MS Agent Framework inherits patterns from AutoGen/Semantic Kernel. Claude Agent SDK is Claude-specific and wasn't part of the original evaluation scope.

### Test Scenario: Insurance Weather Verification

**Use Case:** Australian insurance CAT (catastrophic) event verification

- **Agent 1 - Weather Verification:** Geocodes location → fetches BOM weather observations
- **Agent 2 - Claims Eligibility:** Pure LLM reasoning to apply business rules

#### Business Rules:

- APPROVED: Both thunderstorms AND strong wind observed
- REVIEW: Only one weather type observed
- DENIED: Neither observed

**Test Data:** Brisbane, QLD, 4000 on 2025-03-07

### Implementation Findings

#### LangGraph

##### What worked well:

- Clean `StateGraph` + `TypedDict` state management
- `@tool` decorator from `langchain_core` is familiar
- Explicit node → edge → END structure easy to visualize

##### Code sample from testing:

```
class ClaimState(TypedDict):  
    city: str
```

```
coordinates: dict | None
weather_data: dict | None
eligibility_decision: str | None
messages: Annotated[list, add]
```

```
workflow = StateGraph(ClaimState)
workflow.add_node("weather_agent", weather_agent_node)
workflow.add_node("eligibility_agent", eligibility_agent_node)
workflow.set_entry_point("weather_agent")
workflow.add_edge("weather_agent", "eligibility_agent")
workflow.add_edge("eligibility_agent", END)
```

**Gotcha:** Manual tool call loop handling required - had to iterate through `response.tool_calls` and invoke each tool.

---

## Haystack

### What worked well:

- `OpenAIChatGenerator` with tools parameter straightforward
- `ChatMessage.from_tool()` for tool results

### Code sample from testing:

```
tools = [{
    "type": "function",
    "function": {
        "name": "geocode_location",
        "description": "Convert address to coordinates",
        "parameters": {...}
    }
}]

weather_generator = OpenAIChatGenerator(
    model=DEFAULT_MODEL,
    generation_kwargs={"tools": tools}
)
```

**Gotcha:** Required manual iteration loop with `max_iterations` guard:

```
while iteration < max_iterations:
    response = weather_generator.run(messages=messages)
    if reply.tool_calls:
        # Process tool calls manually
    else:
        break # Final response
```

---

## Pydantic AI

### What worked well:

- **Cleanest code** of all frameworks tested
- Pydantic models for structured output = automatic validation



- `@agent.tool` with `RunContext` for dependency injection
- Async-native throughout

#### Code sample from testing:

```
class WeatherVerificationResult(BaseModel):
    location: str
    latitude: float
    longitude: float
    thunderstorms: str
    strong_wind: str
    severe_weather_confirmed: bool

weather_agent = Agent(
    'openai:gpt-4o-mini',
    deps_type=AppDependencies,
    output_type=WeatherVerificationResult,
    instructions="..."
)

@weather_agent.tool
async def geocode(ctx: RunContext[AppDependencies], city: str, state: str, postcode: str) -> dict:
    # Tool implementation with access to shared http_client via ctx.deps
```

**Result:** Typed `.output` attribute with full IDE autocomplete. If LLM returns invalid data, Pydantic AI automatically re-prompts.

---

#### AutoGen 0.4+ (now superseded by MS Agent Framework)

**Important:** This testing was done with **AutoGen 0.4+**, which has since been **replaced by Microsoft Agent Framework** (April 2025). The new framework consolidates AutoGen + Semantic Kernel into a unified API. Patterns may differ.

#### What worked well:

- `AssistantAgent` with tools as plain async functions
- Simple API for single-agent use

#### Challenge encountered:

"Some models struggle with multi-step tool calling in `RoundRobinGroupChat`"

**Workaround required:** Had to call tools directly instead of letting agent orchestrate:

```
# Instead of letting agent use tools via RoundRobin,
# we call tools directly and have the eligibility agent do reasoning
geo_result = await geocode_location(TEST_CITY, TEST_STATE, TEST_POSTCODE)
weather_result = await get_bom_weather(geo_data["latitude"], ...)

# Then pass to eligibility agent
eligibility_result = await eligibility_agent.run(task=f"Based on: {weather_summary}")
```

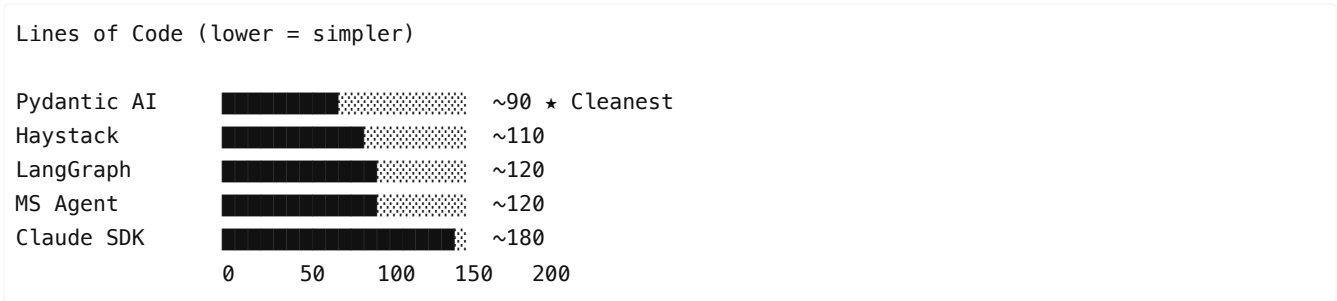
**Migration note:** Microsoft provides migration guides from both AutoGen and Semantic Kernel to the new Agent Framework. The multi-step tool calling issues may be resolved in the new framework.

---

API Testing Results

API	Status	Sample Response
Nominatim Geocoding	✔ Working	Brisbane → (-27.47, 153.02)
BOM Weather	✔ Working	Returns thunderstorm/wind observations

Lines of Code Comparison



Framework	LoC (same use case)	Notes
Pydantic AI	~90	Cleanest, most Pythonic
Haystack	~110	Tool schema definition verbose
LangGraph	~120	Explicit but verbose state setup
MS Agent Framework	~120	Clean API with ChatAgent + @tool
Claude Agent SDK	~180	MCP server setup + fallback logic

Note: Claude Agent SDK demo includes fallback mode due to MCP issues, inflating LoC.

Key Takeaways from Testing

- Pydantic AI had the best developer experience** - Type safety, structured outputs, and clean async code. Recommended for regulated industries where data validation matters.
- LangGraph requires more boilerplate** but gives explicit control over state flow - good for complex workflows where you need to visualize the graph.
- Haystack's tool definition is verbose** (JSON schema) compared to decorator approaches, but pipeline model is intuitive for RAG scenarios.
- MS Agent Framework works as pre-release** - Install with `pip install agent-framework --pre`. Uses `ChatAgent` with `OpenAIChatClient` and `@tool` decorator. Familiar patterns, Microsoft backing.
- Claude Agent SDK needs more setup** - Installs fine but MCP server integration requires Claude Code CLI. Fallback mode works for basic use.
- All frameworks successfully completed the use case** - Either directly or via fallback. Differences were in code elegance and setup complexity, not capability.

7. **DeepSeek works as a drop-in replacement** - All frameworks worked with DeepSeek via OpenAI-compatible API.  
Good for cost-conscious development.

New Framework Testing (January 2026)

We ran all five frameworks against **DeepSeek** ( `deepseek-chat` ) using the same Brisbane weather verification test case.

Test Results: Framework Status

✔ Works Directly (80%): Pydantic AI, LangGraph,  
Haystack, MS Agent Framework

⚠ Fallback Only (20%): Claude Agent SDK

Test Results:

Framework	Status	Decision	Notes
Pydantic AI	✔ Direct	APPROVED	Both thunderstorms + wind observed
LangGraph	✔ Direct	REVIEW	Only wind observed
Haystack	✔ Direct	REVIEW	Only wind observed, retried BOM once
MS Agent Framework	✔ Direct	REVIEW	Works with <code>pip install agent-framework --pre</code>
Claude Agent SDK	✔ Fallback	REVIEW	MCP server error, fallback worked

Key Findings:

1. **MS Agent Framework works!** - Package `agent-framework` is available as pre-release ( `pip install agent-framework --pre` ). Uses `ChatAgent` with `OpenAIChatClient` . The `@tool` decorator works similarly to other frameworks. Successfully called both geocoding and BOM weather tools.
2. **Claude Agent SDK requires Claude Code CLI** - The SDK installs ( `pip install claude-agent-sdk` ) but the MCP server integration had `TaskGroup` errors. The fallback (direct tool execution) works. Full functionality likely requires Claude Code CLI setup.
3. **BOM API variability** - Pydantic AI got "Observed" for both weather types while others got only wind. This is timing/caching variance from the BOM API, not a framework difference.
4. **DeepSeek works well** - All frameworks successfully used DeepSeek as the LLM backend via OpenAI-compatible API.

Installation Commands:

```
# Activate environment
conda activate dory

# Frameworks that work directly
pip install pydantic-ai langchain langchain-openai langgraph haystack-ai

# MS Agent Framework (pre-release)
pip install agent-framework --pre

# Claude Agent SDK
```

```
pip install claude-agent-sdk
# Note: May need Claude Code CLI for full MCP functionality
```

Run demos:

```
python pydantic_ai_demo.py
python langgraph_demo.py
python haystack_demo.py
python ms_agent_framework_demo.py
python claude_agent_sdk_demo.py
```

# LLM Learnability Benchmark

A study measured **how easily an LLM can produce working code** with each framework, given access to documentation.

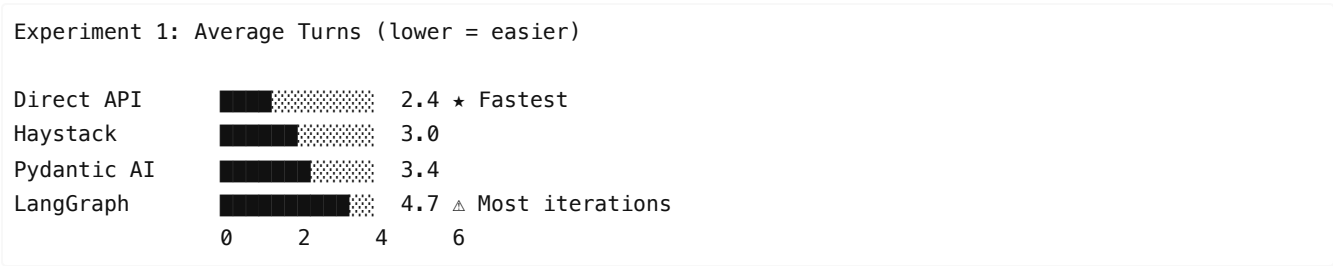
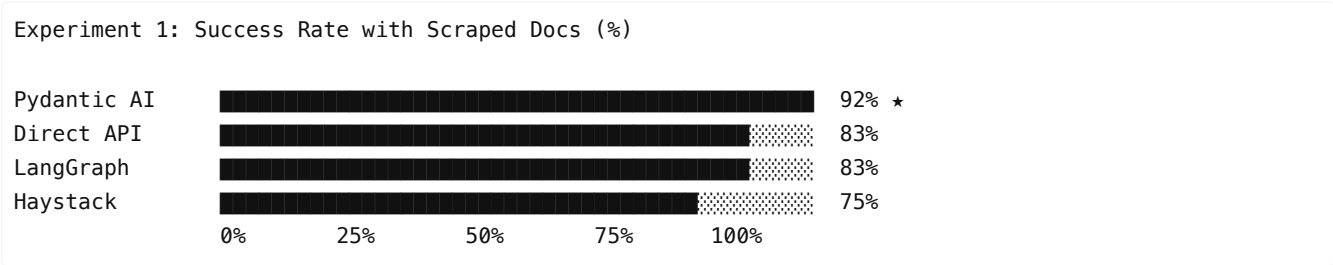
Source: [github.com/srepho/fwork-learnability](https://github.com/srepho/fwork-learnability)

## Methodology

- **Model:** DeepSeek V3 (training cutoff July 2024)
- **Approach:** "Turns to Working Code" - LLM attempts implementation, receives error feedback, iterates up to 10 times
- **Documentation levels:** None, Minimal, Moderate, Full
- **Task:** Tier 1 insurance claims classification
- **Total trials:** 89 across two experiments

## Experiment 1: Scraped Documentation (48 trials)

Initial experiment using auto-scraped documentation from framework websites.

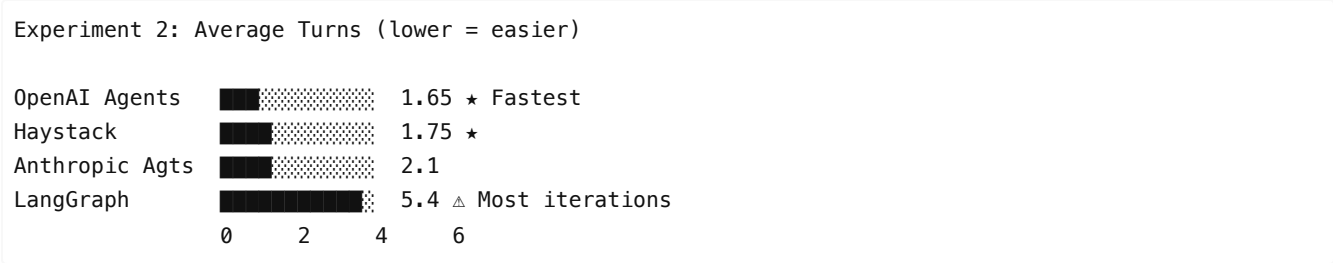
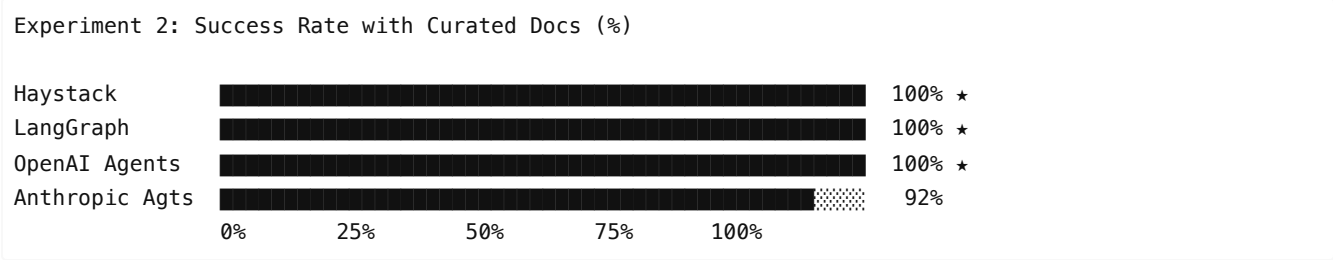


Framework	Success Rate	Avg Turns	First-Attempt	Best Doc Level
Pydantic AI	92% (11/12)	3.4	25%	Moderate
Direct API	83% (10/12)	2.4	8%	Moderate
LangGraph	83% (10/12)	4.7	0%	Minimal

Haystack	75% (9/12)	3.0	25%	Minimal
----------	------------	-----	-----	---------

Experiment 2: Curated Documentation (41 trials)

Follow-up experiment with hand-curated docs containing real code examples.



Framework	Success Rate	Avg Turns	Doc Impact (none→minimal)
Haystack	100%	1.75	2.3t → 1.0t (+57% faster)
LangGraph	100%	5.4	4.7t → 7.5t (slower*)
OpenAI Agents	100%	1.65	2.0t → 2.0t (no change)
Anthropic Agents	92%	2.1	3.0t → 1.7t (+44% faster)

\*LangGraph's increase likely due to small sample size at minimal level.

Contamination Analysis

All frameworks showed **100% success with zero documentation**, confirming high training data contamination:

Framework	Success@None	API Version	Contamination Level
Haystack	100%	v2 (modern)	HIGH - uses 2024 APIs
LangGraph	100%	v0/v1 mixed	HIGH - 50/50 version mix
Pydantic AI	100%	current	HIGH - despite June 2024 release
OpenAI Agents	100%	current	HIGH
Anthropic Agents	100%	current	HIGH

**Implication:** DeepSeek V3's training data includes all tested frameworks. The benchmark measures "recall + refinement" rather than pure learnability from scratch.

Key Findings

1. Documentation quality matters more than quantity

- Haystack improved from 75% → 100% when using curated docs vs scraped docs
- Original "full docs" were marketing pages with no code examples
- Minimal curated docs (real code) outperformed comprehensive scraped docs

2. LangGraph consistently slowest to converge

- 4.7-5.4 avg turns across both experiments
- Zero first-attempt successes in Experiment 1
- Suggests steeper learning curve even for LLMs

3. Even with contamination, docs reduce turns by 40-60%

- Haystack: 2.3 turns (no docs) → 1.0 turns (minimal docs)
- Anthropic Agents: 3.0 turns → 1.7 turns
- Docs provide correct patterns without trial-and-error

4. Too much documentation can hurt

- Anthropic Agents had its only failure (1/12) at "full" documentation level
- Suggests excessive context can confuse the model

Error Patterns

Error Type	Count	%	Example
SyntaxError	40	54%	Unterminated strings, fails to recover
AssertionError	15	20%	Wrong classification despite valid code
TypeError	7	9%	Wrong method signatures
API Hallucination	26	-	PydanticAI's fabricated result_type (15x)

Implications for Framework Selection

If you prioritize...	Choose...	Why
Fastest LLM-assisted dev	OpenAI Agents or Haystack	1.65-1.75 avg turns with curated docs
Highest eventual success	Any with good docs	All reach 92-100% with curated docs
Works without docs	Any (contaminated)	All 100% at none level
Steepest learning curve	Avoid LangGraph	Consistently 4.7-5.4 turns

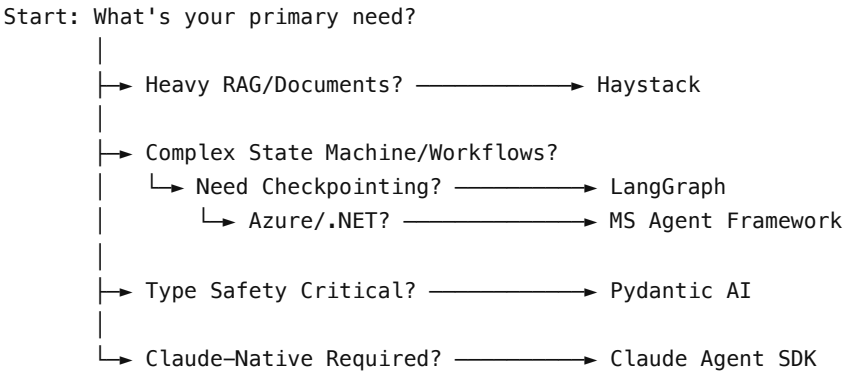
Caveats

- Only Tier 1 (classification) tasks tested; Tier 2 (tool use) and Tier 3 (agent-native) pending
- High contamination means results reflect "memorization + refinement" not pure learnability
- Human learnability may differ from LLM learnability
- Pydantic AI not re-tested with curated docs (would likely improve)

Decision Flowchart

Use this flowchart to select the right framework for your use case:

Decision Flowchart:



Quick Decision Matrix

Framework Recommendation by Use Case

Type Safety (30%)	<div></div>	→ Pydantic AI
RAG/Documents (25%)	<div></div>	→ Haystack
Complex Workflows (25%)	<div></div>	→ LangGraph
Azure/Enterprise (15%)	<div></div>	→ MS Agent
Claude-Native (5%)	<div></div>	→ Claude SDK

Summary: When to Use Each

Framework	Best For	Status (Jan 2026)
Haystack	RAG-heavy applications, document processing, production stability	✅ Production ready
LangGraph	Complex state machines, explicit control flow, LangChain ecosystem	✅ Production ready
Pydantic AI	Type-safe applications, structured outputs, clean Python code	✅ Production ready
MS Agent Framework	Azure environments, .NET shops, migrating from AutoGen/SK	✅ Pre-release, functional
Claude Agent SDK	Claude-native apps, leveraging Claude Code capabilities	⚠️ Needs setup

Recommendation

For production use today: Choose **Pydantic AI**, **LangGraph**, **Haystack**, or **MS Agent Framework**.

Consider carefully:

- **MS Agent Framework** - Works but is pre-release ( --pre flag required). Good if already in Azure ecosystem.
- **Claude Agent SDK** - MCP server integration needs work. Best if you're willing to set up Claude Code CLI.