— AI CODING

The best books for building AI agents in 2026

By **Flowi Editorial** · *May 9, 2026 · 6 min read*

An honest list of the books worth reading if you're shipping AI agents to production this year. What each covers, where each falls short, and the gap none of them fills.

The AI agent literature in 2026 is in an awkward place. The frontier moves quarterly. Most published books are 18-24 months out of date by the time they hit Kindle. The good content lives in scattered blog posts, conference talks, and a small set of recent books that have aged surprisingly well.

This is the honest list. Six books worth your reading time if you're shipping agents to production right now. What each is good for, where each is dated, and the gap none of them fills.

The criteria

A book qualifies for this list if it answers at least one of:

1. **What's the architecture pattern** for an agent that actually ships?
2. **What's the failure mode** for that pattern, and how to engineer around it?
3. **What's the right mental model** for the underlying transformer / agent / tool-use system?

Books that are general "AI for business" overviews don't qualify — they're outdated the day they print. Books that are pure research surveys don't qualify — they're useful but they don't help you ship.

The six books

1. **"Designing Machine Learning Systems"** — Chip Huyen, O'Reilly, 2022. Still the strongest single book on production ML systems generally. Agents are a subset.
2. **"Building LLM Applications for Production"** — Suhas Pai, Manning, 2024. The most current general overview; the production patterns chapter is the best part.
3. **"AI Engineering"** — Chip Huyen, O'Reilly, 2025. The follow-up to Designing ML Systems, written for the LLM era. Less generally-applicable than its predecessor but more relevant to agents specifically.
4. **"Building Multimodal AI Applications"** — Tarun Sharma, Packt, 2025. The only book that genuinely covers the multimodal patterns (vision-language-tool agents) honestly. Some chapters are thinner than others.
5. **"Hands-On Large Language Models"** — Jay Alammar & Maarten Grootendorst, O'Reilly, 2024. The best mental-model book — explains what's happening inside the transformer well enough that you stop being confused by edge cases.
6. **"Agent Memory: The 5 Patterns That Ship in Production"** — Flowi Editorial, 2026. Disclosure: we wrote this. The first book focused specifically on the memory layer that breaks at message four.

"Designing Machine Learning Systems" — Chip Huyen

The case for: Even three years out, the framework for thinking about production ML is unmatched. The chapters on data engineering, feature stores, and monitoring apply directly to agent systems with minimal translation.

The case against: Pre-LLM era. The model-training emphasis is overweighted relative to where agent work actually lives now.

Best for: Engineers transitioning from traditional ML into LLM/agent work. The mental models transfer well; this book is the bridge.

"Building LLM Applications for Production" — Suhas Pai

The case for: Solid coverage of RAG, tool-use, and basic agent patterns. The chapter on evaluation is unusually honest about how hard it is.

The case against: Tries to cover too much. The depth on any specific pattern is shallow. By the time you're past the basics, you've outgrown the book.

Best for: Engineers shipping their first production LLM application. Skip if you've already done two or three.

"AI Engineering" — Chip Huyen

The case for: The most current treatment of the production AI stack. Excellent treatment of agentic workflows, fine-tuning decisions, and the LLMOps tooling landscape.

The case against: The agent-specific patterns chapter is good but not deep. You'll need to supplement with focused books or papers for any specific pattern.

Best for: Mid-level engineers building real systems, who want a textbook-level treatment of the stack.

"Building Multimodal AI Applications" — Tarun Sharma

The case for: The vision-language sections are the best published treatment of the topic. Real code, real failure modes, honest about what's hard.

The case against: Chapters on speech and embodied agents are thinner. The book wants to be a complete multimodal reference but only really delivers on vision-language.

Best for: Engineers shipping agents that need to reason over images or screenshots. Skip the back half.

"Hands-On Large Language Models" — Jay Alammar & Maarten Grootendorst

The case for: The best book for *understanding what's actually happening*. Alammar's visualization approach (he wrote "The Illustrated Transformer") translates to book-length and clarifies a lot of confusion.

The case against: Mental-model book, not a shipping book. You'll close it understanding the transformer better but not knowing more about production patterns.

Best for: Engineers who keep getting confused by edge cases ("why did the model do that?") and want the deeper intuition.

"Agent Memory: The 5 Patterns That Ship in Production"

Disclosure: we wrote this. Take the recommendation with that in mind.

The case for: The only book focused entirely on the memory layer. The 5 patterns (Conversation Memory, Summary Memory, Vector Memory, Knowledge Graph Memory, Hybrid Memory) are the patterns you actually need; the failure modes named for production systems are the failure modes you'll hit. Short — 4,500 words, one weekend to read, code that runs.

The case against: Narrow scope by design. If you want a book that covers tools, planning, evaluation, *and* memory, this isn't it (yet — we're writing those).

Best for: Engineers whose agents work in demos but fail at message four. Specifically the *memory* failure mode. If your agent is failing at tool use or planning, this book won't help directly.

The gap none of these fills

No book in the category covers **how to evaluate AI agents in production before they fail**. Every book has an "evaluation" chapter that's mostly theoretical — generic benchmarks, golden datasets, the classic ML stuff. None of them treat evaluation as the *operational discipline* it actually is — pre-deployment behavioral tests, ongoing regression suites, what to do when your eval breaks before your agent does.

This is where the next major book in the space needs to be written. We're working on one (book №04 in the Flowi roadmap). Until then, the gap lives in scattered blog posts, mostly from the Anthropic and OpenAI engineering teams, plus a few Substack writers (Eugene Yan, Hamel Husain, Simon Willison).

How to pick

The honest decision tree by what you most care about:

- **Bridging from traditional ML** → Designing Machine Learning Systems
- **First-time LLM application** → Building LLM Applications for Production
- **Modern stack overview** → AI Engineering
- **Vision-language agents** → Building Multimodal AI Applications
- **Understanding the transformer** → Hands-On Large Language Models
- **Agent memory specifically** → Agent Memory: The 5 Patterns That Ship in Production

A recommendation: **don't read more than two books at once.** The category moves fast enough that depth matters more than breadth. Pick the one that addresses your current production bottleneck. Read it. Ship the thing. Pick the next one.

Who should care

- **Engineers shipping their first agent:** start with Building LLM Applications for Production. Move to AI Engineering when you're past the basics.
- **Engineers whose agent works in demo but fails in production:** the issue is almost always memory, tools, or evaluation. Memory is what our book covers.
- **Solo and small-team builders:** budget book money. $19-30 per book is rounding error for the time you save not figuring out what others have already figured out.

The AI agent book market is full of titles that don't survive the next quarter. The six above are the ones that have aged or will age — pick by your specific gap.

If your agent forgets the user at message four — which is the most consistent failure mode across every "AI agent demo" — that's the territory Agent Memory: The 5 Patterns That Ship in Production was built for. The decision tree, the code, and the failure modes nobody warns you about. $19, 4,500 words, one weekend, ships on Monday.

Originally published on useflowi.app/blog/best-books-for-building-ai-agents-in-2026.
Flowi — the editorial intelligence layer for AI builders.
Daily brief at useflowi.app/blog · Monthly Dispatch at useflowi.app/dispatch.