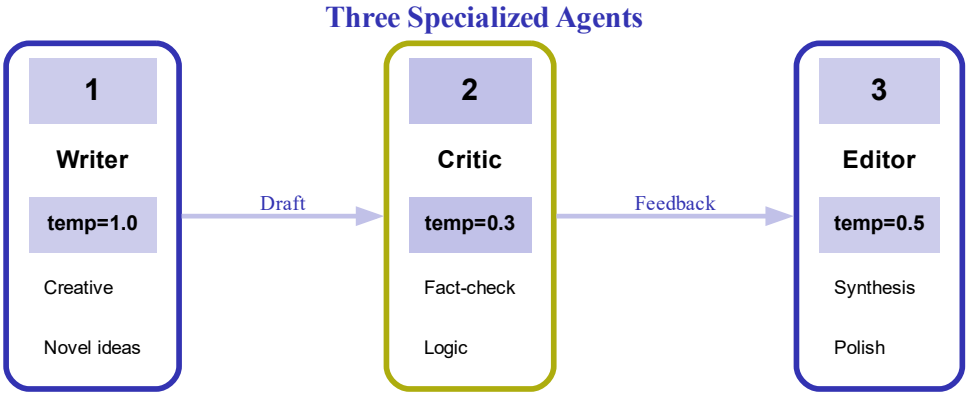
Multi-Agent Innovation Discovery


Three Specialized Agents
1
Writer
temp=1.0
Creative
Novel ideas
Draft
2
Critic
temp=0.3
Fact-check
Logic
Feedback
3
Editor
temp=0.5
Synthesis
Polish

What is Multi-Agent Collaboration?

Definition
Three specialized AI agents working in sequence to create higher-quality outputs than a single agent.

The Team

- **Writer** - Creative idea generation
- **Critic** - Analytical evaluation
- **Editor** - Polished synthesis

Why Not One Agent?

Single agent must balance:

- Creativity AND precision
- Exploration AND criticism
- Speed AND quality

Multi-Agent Advantage
Each agent excels at ONE specific task:

- Specialization over generalization
- Built-in quality control
- Separation of concerns

Core Concept: Specialization through distinct roles and temperatures

Temperature Controls Behavior

Temperature	Zone	Behavior
0.0-0.6	Analytical	Consistent, Reliable, Fact-focused
0.7-1.0	Creative	Novel ideas, Exploratory, Diverse

Key Insight: Temperature (0.0-1.0) determines agent behavior:

- **Low (0.0-0.3)** - Analytical, precise, consistent
- **Medium (0.4-0.6)** - Balanced, reliable
- **High (0.7-1.0)** - Creative, exploratory, varied

Temperature is the primary control mechanism for agent specialization

API Workflow: How Agents Communicate

Agent 1: Writer

System Prompt:
"Generate highly original, novel business ideas..."

User Prompt:
"AI + Climate Change"

Temperature: 1.0

↓ **API Call**

Output:
Creative business idea draft

Agent 2: Critic

System Prompt:
"Evaluate for true novelty. Identify similar solutions..."

User Prompt:
User input + Writer output

Temperature: 0.3

↓ **API Call**

Output:
Critical feedback and analysis

Agent 3: Editor

System Prompt:
"Synthesize into polished business concept pitch..."

User Prompt:
User input + Writer output + Critic output

Temperature: 0.5

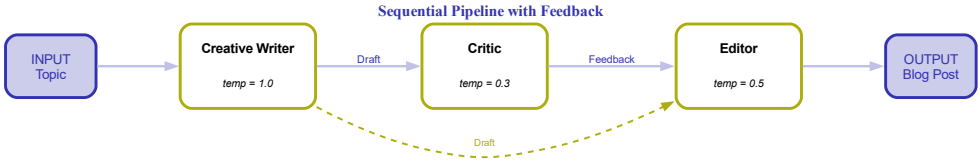
↓ **API Call**

Output:
Final polished business pitch

Key Insight: Each agent's output becomes part of the next agent's input, building context sequentially.

API outputs flow sequentially - each agent sees all previous work

Workflow: Sequential Pipeline



Step-by-Step Process

1. **Writer** (temp=1.0) generates creative first draft
2. **Critic** (temp=0.3) reviews and provides structured feedback
3. **Editor** (temp=0.5) synthesizes both into final polished output

Sequential pipeline ensures each stage builds on previous work

Quality vs Cost Tradeoffs

When to Use Multi-Agent

		Task Complexity	
		Simple	**Complex**
Quality Needs	**High**	**Single Agent** Better prompts	**Multi-Agent** RECOMMENDED
	Low	**Single Agent** Simple prompt	**Single Agent** Chain prompts

Use Multi-Agent When:

- High stakes content
- Quality is critical
- Complex tasks

Use Single-Agent When:

- Quick drafts
- Simple tasks
- Budget constraints

Claude Models: Three Tiers for Different Needs

Model Hierarchy

Haiku 4.5 - Small, Fast

- Speed-optimized for real-time
- Best for: UI, chat, pair programming
- 2-3x faster than Sonnet
- Lowest cost: $1/$5 per million tokens

Sonnet 4.5 - Balanced

- State-of-the-art coding (our choice!)
- Best for: Agents, general use
- Best all-around performance
- Mid-tier cost: $3/$15 per million tokens

Opus 4.1 - Large, Powerful

- Maximum reasoning capability
- Best for: Complex analysis, deep thinking
- Catches nuances others miss
- Highest cost: $15/$75 per million tokens

Anthropic does not publicly disclose parameter counts for competitive reasons

Performance Characteristics

Speed vs Intelligence Tradeoff

- Haiku: Lightweight, fast inference
- Sonnet: Balanced speed and capability
- Opus: Deep reasoning, slower but thorough

Model Orchestration
Sonnet can break down complex tasks and delegate to multiple Haiku instances running in parallel.

Our Notebook Setup
Uses Sonnet 4.5 for all three agents:

- Writer (temp=1.0)
- Critic (temp=0.3)
- Editor (temp=0.5)

Same model, different behaviors through temperature and system prompts.

API Billing & Multi-Agent Costs

How Billing Works

Token-Based Pricing

- Charged per token
- 4 characters = 1 token
- Input tokens: what you send
- Output tokens: what model generates

Sonnet 4.5 Pricing

- Input: $3 per million tokens
- Output: $15 per million tokens
- Output costs 5x more!

Free Credits

New users get $5 in free API credits (no credit card required).

Multi-Agent Workflow Cost

Example: "AI + Climate Change" (10 tokens)

Writer:
Input: 10 tokens, Output: 500 tokens
Critic:
Input: $10 + 500 = 510$ tokens
Output: 300 tokens
Editor:
Input: $10 + 500 + 300 = 810$ tokens
Output: 400 tokens

Total:
Input: 1,330 tokens $\times$ **$**3/1M = **$**0.004
Output: 1,200 tokens $\times$ **$**15/1M = **$**0.018
TOTAL: $0.022

Single-agent equivalent: **$**0.008
Multi-agent premium: 2.7x

Multi-agent costs more but delivers higher quality - choose based on your use case

Key Takeaways

Core Principles of Multi-Agent Collaboration

1. **Temperature Control**
   Temperature (0.0-1.0) is the primary mechanism for controlling agent behavior
   Low = analytical, High = creative

2. **System Prompts Create Specialization**
   Each agent has distinct personality and role through custom system prompts
   Writer, Critic, Editor = three expert roles from one base model

3. **Sequential Pipeline Enables Quality**
   Writer → Critic → Editor flow ensures iterative refinement
   Each stage builds on previous work

4. **Quality vs Cost Tradeoffs**
   Multi-agent: 2-3x higher cost, significantly better quality
   Choose based on stakes, complexity, and budget constraints

Try the notebook: Run the minimal workflow and see the HTML output!