

Token Usage Analysis Report – Magentic Multi-Agent Orchestration

Use Case: Get Top 10 PO details

Total Tokens Used: 10,546

Prompt Tokens: 9,671 (91.7%)

Completion Tokens: 875 (8.3%)

1. Executive Summary

This report analyzes token usage across agents during a single Magentic orchestration run. The execution is prompt-heavy due to large schemas, repeated orchestration steps, and lack of truncation in agent-level threads.

2. Overall Token Breakdown

Total Tokens: 10,546

Prompt Tokens: 9,671

Completion Tokens: 875

3. Orchestrator Agent Breakdown

- Task Ledger – Facts: ~1,700–2,300 tokens
 - Task Ledger – Plan: ~2,300–2,900 tokens
 - Progress Ledger Checks: 2,867 tokens
 - Final Answer Generation: 2,241 tokens
- Subtotal:** ~5,500–6,000 tokens

4. SQL Agent

Estimated Usage: ~2,000–2,500 tokens

Major contributor is repeated inclusion of full database schema.

5. Analysis Agent

Estimated Usage: ~1,500–2,000 tokens

Includes SQL output, analysis instructions, and result explanation.

6. Follow-up Suggestion Agent

Estimated Usage: ~500–800 tokens

Used for generating suggested follow-up questions.

7. Message Passing Overhead

Estimated Usage: ~500–1,000 tokens

Includes agent-to-agent communication and accumulated chat history.

8. Root Causes of High Token Usage

- Large static prompts (schemas, templates)
- Multiple orchestration phases
- No truncation in agent threads
- Repeated agent descriptions

9. Recommended Optimizations

- Apply chat history truncation to agent threads
- Reduce database schema size
- Cache agent instructions
- Reduce progress ledger frequency

10. Conclusion

Token usage is expected for Magentic orchestration but can be reduced by 40–60% with prompt optimization and truncation strategies.