Token Usage Breakdown by Component

2.1 Orchestrator Agent (Magentic Manager)

Estimated total: ~5,500–6,000 tokens

a) Task Ledger – Facts (Initial Planning)

- **Prompt:** ~1,500–2,000 tokens
    - User request
    - Agent descriptions
    - System instructions
- **Completion:** ~200–300 tokens
- **Total:** ~1,700–2,300 tokens

---

b) Task Ledger – Plan (Planning Phase)

- **Prompt:** ~2,000–2,500 tokens
    - Report schema (large)
    - Team descriptions
    - Scenario instructions
    - Previous facts
- **Completion:** ~300–400 tokens
- **Total:** ~2,300–2,900 tokens

---

c) Progress Ledger Checks (Multiple Iterations)

- **Observed from logs (lines 468–469):**
    - Prompt: 2,632 tokens
    - Completion: 235 tokens
- **Total:** 2,867 tokens
- **Note:** Called **multiple times** during orchestration

---

d) Final Answer Generation

- **Observed from logs (lines 528–529):**
    - Prompt: 2,113 tokens
    - Completion: 128 tokens
- **Total:** 2,241 tokens

📌 Orchestrator Subtotal

≈ 5,500–6,000 tokens

2.2 SQL Agent

Estimated total: ~2,000–2,500 tokens

- **Prompt:** ~1,500–2,000 tokens
  - Full database schema (major contributor)
  - User request
  - SQL generation rules
  - Prior context
- **Completion:** ~200–300 tokens
  - Generated SQL
  - Execution status
- **Invocation:** `calculate_tokens()` → `agents.py:845`

2.3 Analysis Agent

Estimated total: ~1,500–2,000 tokens

- **Prompt:** ~1,200–1,500 tokens
  - Analysis instructions
  - SQL output data
  - User request context
- **Completion:** ~433 tokens (from logs)
- **Invocation:** `calculate_tokens()` → `agents.py:983`

2.4 Follow-up Suggestion Agent

Estimated total: ~500–800 tokens

- **Prompt:** ~400–600 tokens
  - Summarized dataframe context
  - Suggestion instructions

- **Completion:** ~100–200 tokens
  - 3–4 follow-up questions

2.5 Message Passing & Coordination Overhead

Estimated total: ~500–1,000 tokens

- Agent-to-agent messages
- Context forwarding
- Chat history accumulation

3. Overall Token Consumption (Per Request)

Component	Estimated Tokens
Orchestrator Agent	5,500 – 6,000
SQL Agent	2,000 – 2,500
Analysis Agent	1,500 – 2,000
Follow-up Agent	500 – 800
Message Overhead	500 – 1,000
Total	**~12,000 – 14,000**

4. Root Causes of High Token Usage

1. Large prompts per agent
2. Database schema (~1,000–1,500 tokens) sent **every SQL call**
3. Report templates (~500–800 tokens)
4. Agent instruction blocks (~500–1,000 tokens each)
5. Separate **facts + plan** orchestration phases
6. Multiple **progress ledger checks**
7. Agent coordination messages
8. Final answer generation as a full LLM call
9. Chat history accumulation
10. Each agent response adds to future context
11. Full conversation history passed between agents
12. ChatHistoryTruncationReducer applies **only to main chat**
13. Agent plugin threads do **not** truncate history
14. Magentic pattern mandates multiple LLM calls
15. Each progress check is a separate API call

16. Task ledger updates repeated verbatim