Research Paper

Improving LLM Accuracy in API Orchestration

Executive Summary

Research findings on semantic layers, declarative orchestration, and token efficiency

orbital

Enterprises are increasingly using LLMs to orchestrate APIs - selecting endpoints, resolving identifiers, sequencing calls. This is the foundation of AI agents, copilots, and automated workflows.

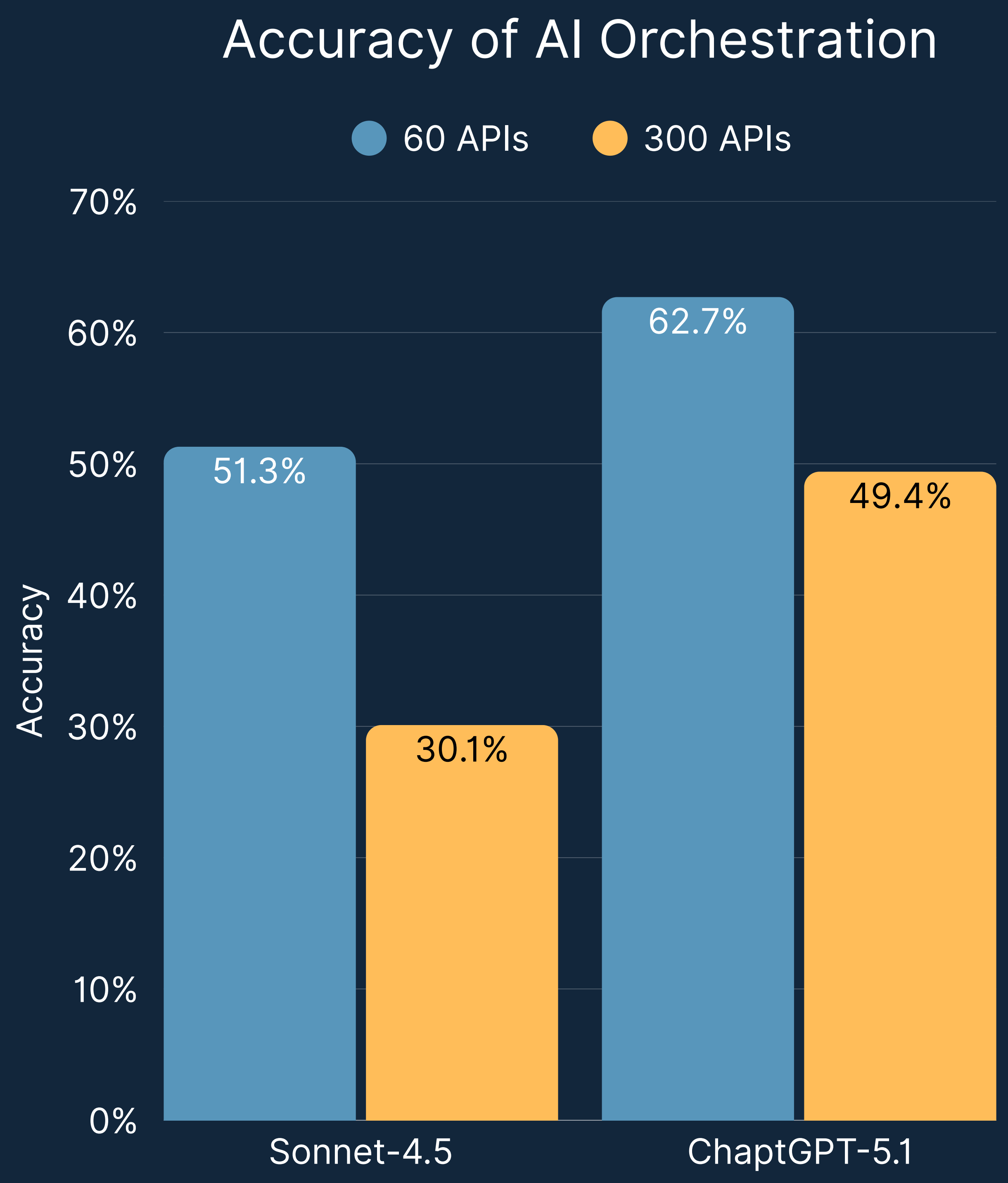
We ran a detailed study to explore:

How good are LLM's at building a plan requiring orchestrating multiple APIs together?

At even modest real-world scale, AI agents fail to plan reliably.

With 300 API endpoints, Anthropic's **Sonnet 4.5 scored only 30%** on our accuracy tests.

A typical enterprise API estate exceeds 600 endpoints.[1]



orbital

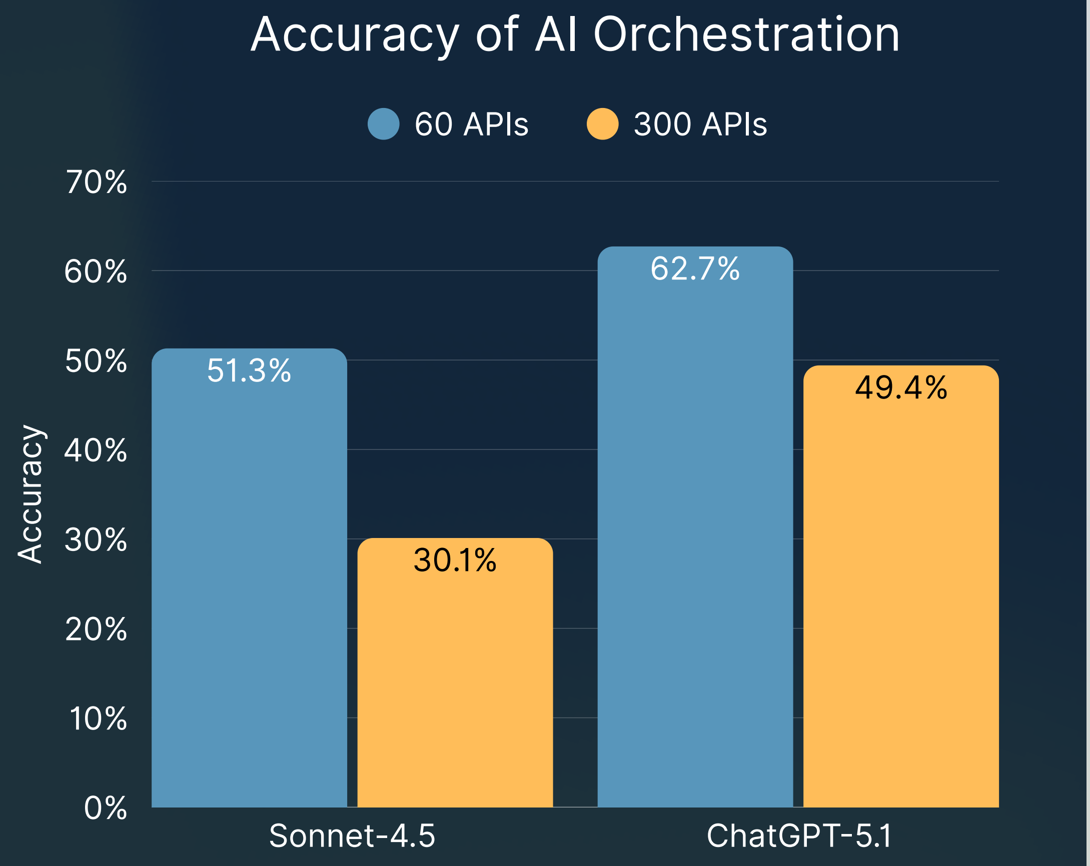
[1] IBM - https://www.ibm.com/think/insights/api-strategy

1 Planning accuracy falls to unusable levels between 60 to 300 endpoints

At 300 endpoints, flagship LLM's produced inacurate plans in ~50%–70% of test runs

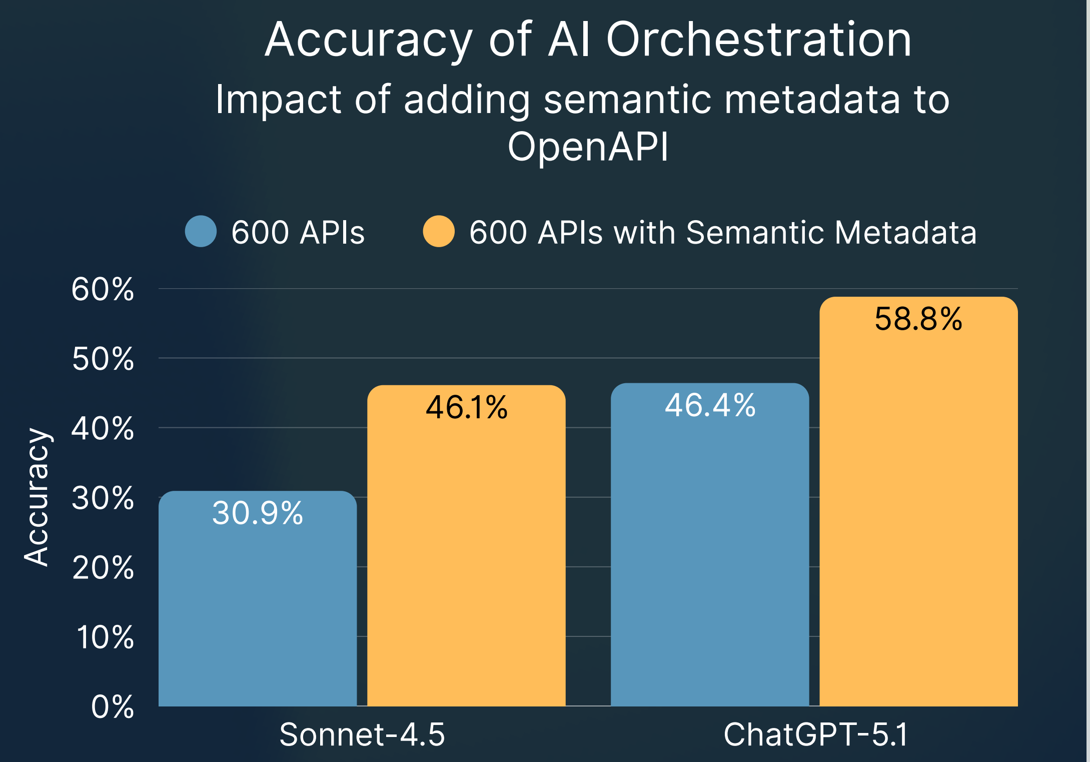
Where agents fail at scale:

- Wrong APIs selected
- Mismatched identifiers
- Incorrect fields used in calculations



2 Adding even minimal semantic metadata improves planning accuracy

Adding semantic type annotations to existing API specs is a lightweight change teams can adopt incrementally - and materially reduces the ambiguity LLMs struggle with at scale

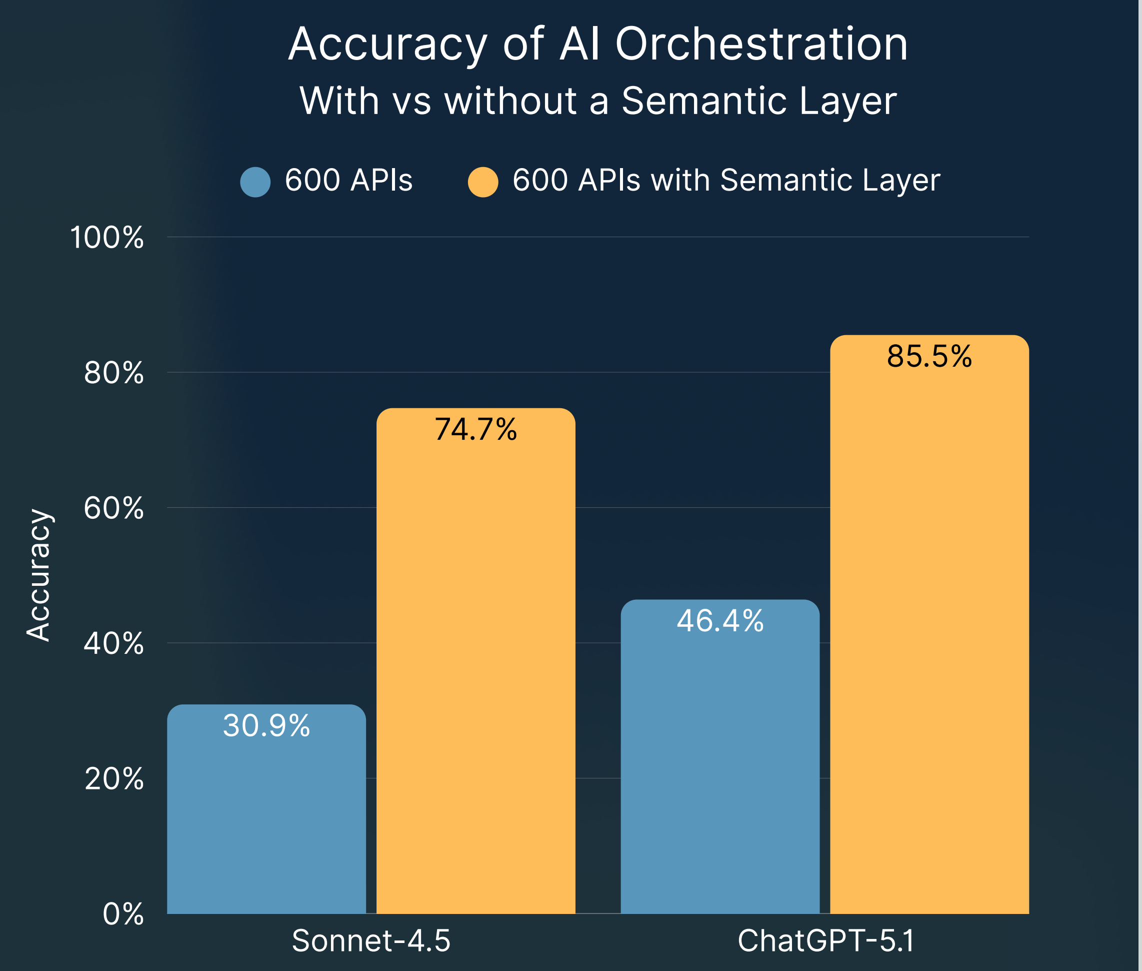


orbital

"Accuracy" was measured as the ability to produce a plan which selected and sequenced the correct endpoints, resolved identifiers, and selected the correct fields for business logic. See the full research paper for detailed breakdown.

3 Adopting a semantic layer improves LLM planning accuracy by 73-142%[1]

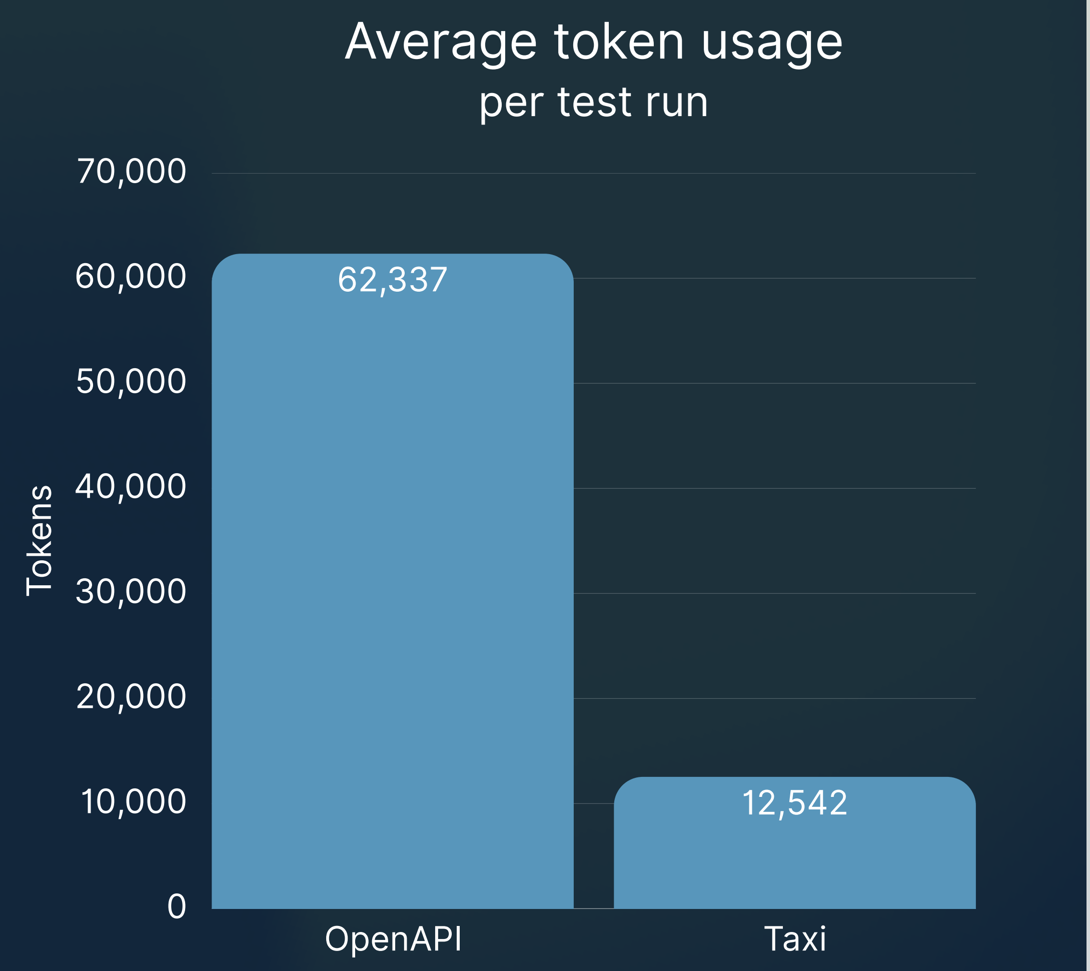
AI agents using declarative query languages (like TaxiQL) to express intent dramatically outperformed LLMs generating orchestration code directly



4 Using Taxi to describe APIs reduces token usage by up to 80%, driving cost savings

Taxi is a compact schema format that represents the same information as OpenAPI.

Converting API specs to Taxi before sending to LLMs is an automated change teams can adopt, which materially reduces token cost while improving accuracy.



[1] Each test was run 30 times per model; scores are averaged

Key Learnings

1 Adopting Taxi drives token savings

- **Low Risk**
- **Immediate ROI**

Convert OpenAPI specs to Taxi before sending to LLMs. A low risk change which delivers immediate cost reduction. Can be easily automated, and no major architecture changes required

2 Semantic annotations improve accuracy

- **Incremental adoption**
- **Compounding benefits**

Annotating OpenAPI specs with semantic types reduces ambiguity and improves LLM planning accuracy. Can be adopted incrementally using existing API specs

3 For production-grade agentic orchestration, use a declarative query layer

- **Highest payoff**
- **Critical at scale**

For teams building agents that need reliable enterprise data access, a semantic layer (like Taxi) and a declarative query layer (like TaxiQL or GraphQL) delivers the largest accuracy gain - turning 30% accuracy into >75%.

Summary

At enterprise scale, AI agent accuracy requires solid architectural foundations, with semantic layers emerging as a critical component.

Contact us to discuss how to apply these findings to your API estate, or run a similar benchmark against your own API endpoints

About this research

This study was conducted by Orbital using a scenario designed to mirror real trading workflows at tier-one financial institutions.

We presented LLMs with a realistic financial services task: build a pre-trade impact checker requiring orchestration of 5 APIs selected from a larger population.

Models needed to identify correct endpoints, resolve different identifier schemes across systems, and select appropriate fields from complex response objects.

Full methodology and detailed results are available in the complete research paper, available on request.

Contact

marty.pitt@orbitalhq.com orbitalhq.com

orbital

hello@orbitalhq.com

orbitalhq.com