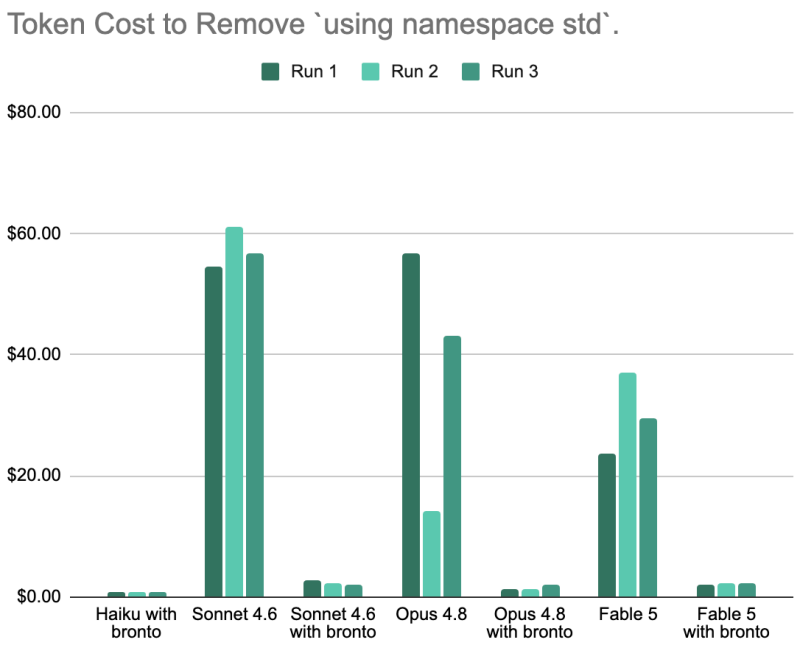
BrontoSource: Reduce Refactoring Token Costs by 14-26x

The Challenge: Agentic refactoring with LLMs are prone to "cost bloat". Unaided agents scale quadratically, spending the majority of their budget on iterative trial-and-error cycles, recompilation, and managing fragmented subagents.

The Solution: A hybrid approach that combines the high-level orchestration of Claude with the deterministic precision of our **`bronto-refactor`** refactoring tool.



- **Tool:** `bronto-refactor` provides the robustness lacking in pure LLM approaches. By building on deterministic compiler based tooling, we eliminate hallucinations and guarantee correctness.
- **Skills:** The `/bronto:refactor` Claude plugin acts as a high-level router, mapping user intent to deterministic static analysis rather than relying on token burn to cover model weakness.

The Impact: Capable models ***reduce token costs by 14-26x*** weaker models ***gain new capabilities***.

Model	Avg. Cost (Unaided)	Avg. Cost (w/ Plugin)	ROI (Cost Reduction)
Haiku 4.5	Failure	$0.81	Impossible → Possible
Sonnet 4.6	$57.45	$2.32	25.2x Reduction
Opus 4.8	$38.05	$1.53	26.4x Reduction
Fable 5	$30.05	$2.16	13.9x Reduction

Data Source: Controlled experiment on a small (~50k line) C++23 codebase.