

SOFTWARE ENGINEERING

CodeMesh: 95% fewer tokens for AI coding agents

When AI coding assistants moved to token billing, an engineering leader at a global energy-equipment manufacturer asked how to cut the cost. Wexa built CodeMesh: a structural graph of the codebase that coding agents query over MCP instead of reading whole files.

95%

fewer tokens across 98 engineering queries (25.3M → 1.23M)

20×

fewer tokens per query (258,115 → 12,524)

\$23.67 → \$1.65

measured API spend for the benchmark

92–97%

savings held on hard and expert queries

CLIENT

Wexa product, built from a problem raised by a global energy-equipment manufacturer

REGION

Global

WHEN

2026

ENGAGEMENT

Wexa product

WEXA COMPONENTS

CodeMesh on the Wexa memory layer - MCP server - 5 retrieval tools

The challenge

- Coding agents re-read whole files and rebuild context for every query
- Token-based billing makes that cost visible and large
- Agents miss relationships between functions, classes and dependencies

What Wexa built

- ✓ Code knowledge graph of files, functions, classes, imports and git history on the Wexa memory layer
- ✓ Exposed over MCP to Claude Code, Cursor, ChatGPT and custom agents
- ✓ Guard hook redirects file search to the graph; no LLM in the retrieval path
- ✓ Read-only, permission-checked access with audit of every agent query

How it runs on Wexa

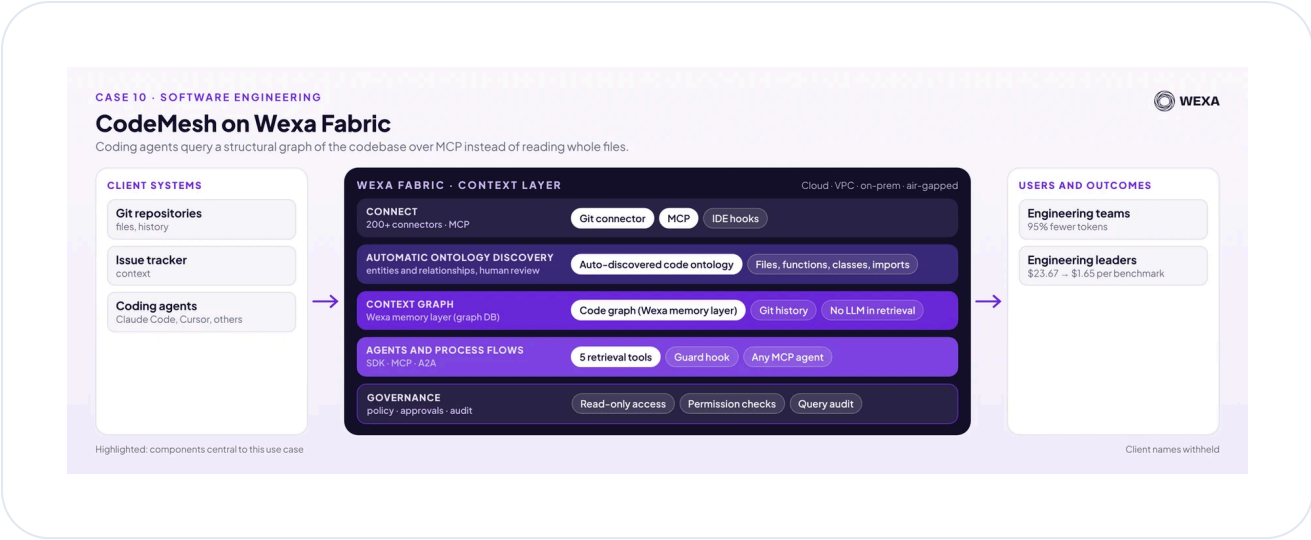


FIGURE How this use case runs on Wexa: client systems connected, ontology discovered, context graph in the Wexa memory layer, agents and process flows, and governance. Highlighted components are central to this use case.

Solution architecture

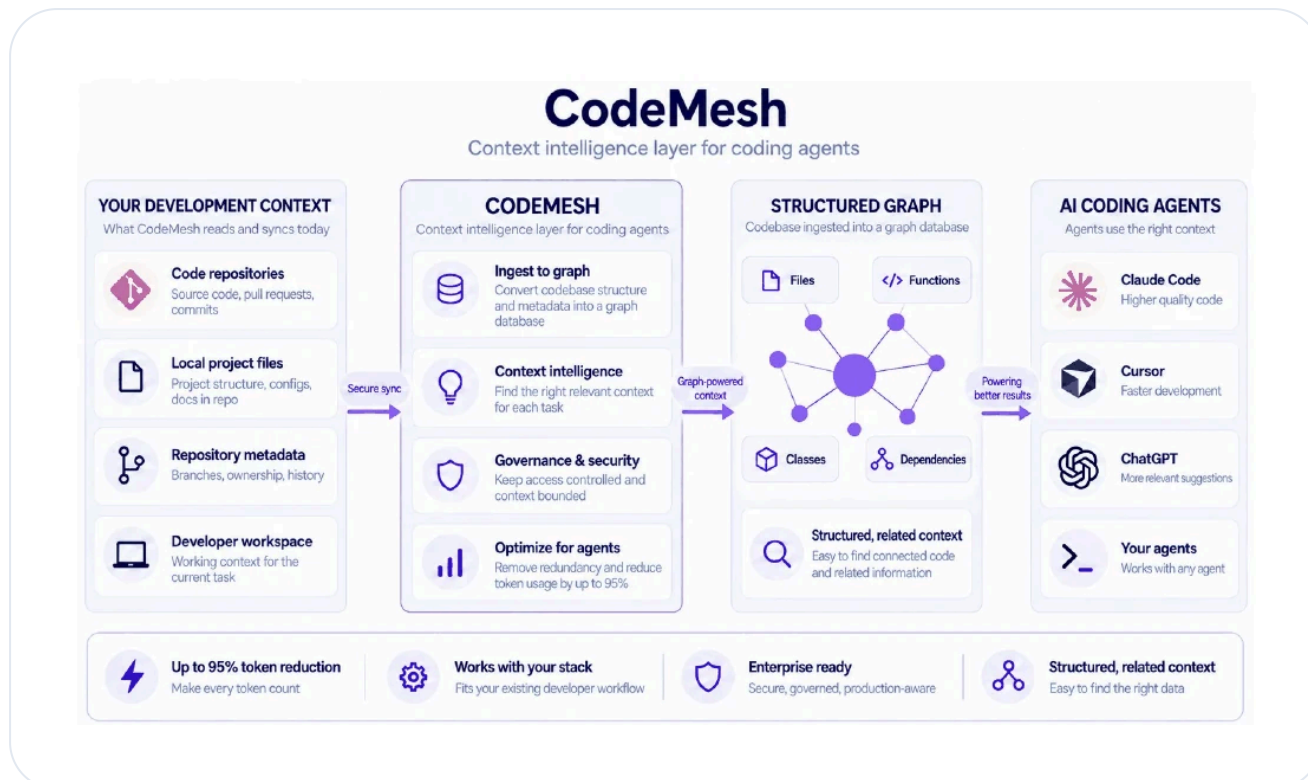


FIGURE CodeMesh architecture: development context → CodeMesh → structural graph → AI coding agents.

How the result was measured

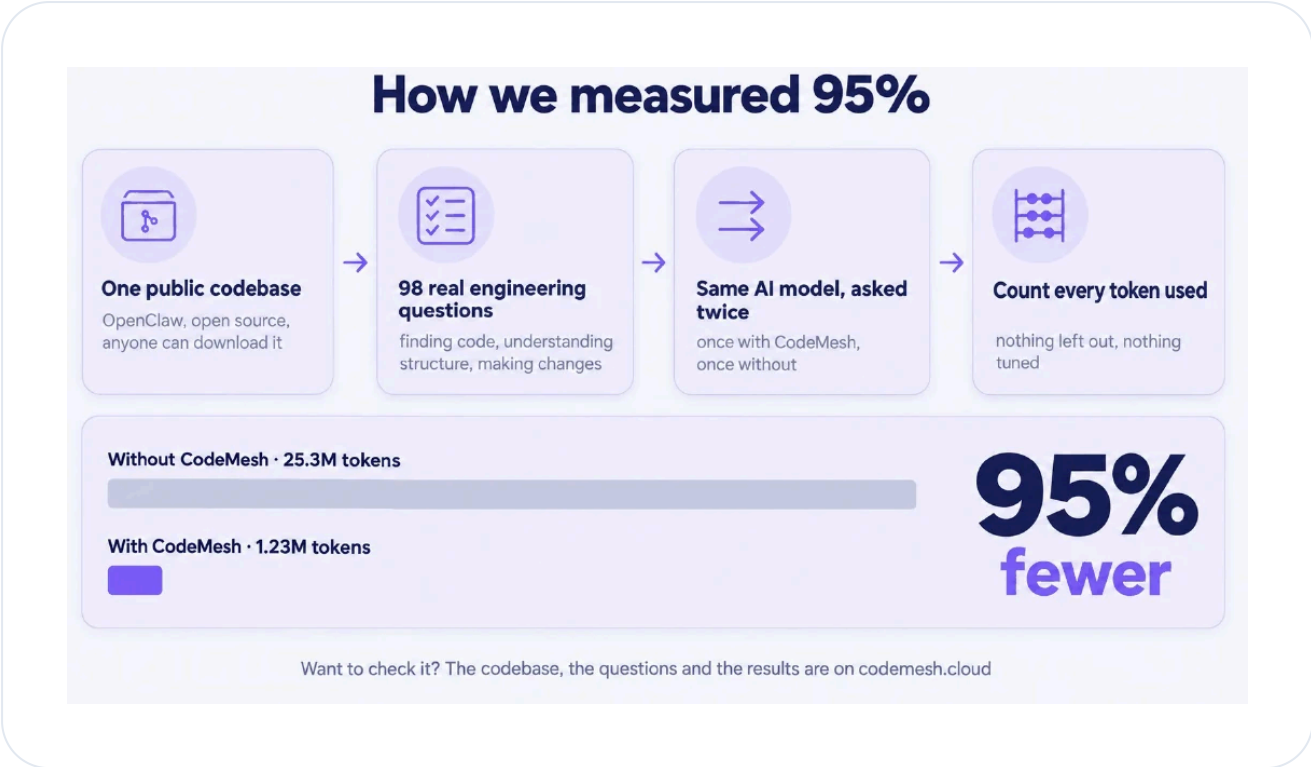


FIGURE Benchmark method: one public codebase, 98 real engineering questions, the same model asked twice, every token counted.

Results

MEASURE	BEFORE / WITHOUT WEXA	WITH WEXA
Coding-agent token spend (team of 50 developers)	\$300K a year	\$21K a year
Annual cost of the process	\$300K	\$21K
Time to implement the AI solution	4 months	1 day
AI build cost (one-time)	\$192K	\$4K

MEASURE	BEFORE / WITHOUT WEXA	WITH WEXA
AI operating cost per year (models, infrastructure, upkeep, platform)	\$141K	\$66K
Total AI cost in year 1	\$333K	\$70K

WITH WEXA

\$213K

net annual value to the client after Wexa's
own operating cost

16×

faster to implement

54%

lower AI operating cost

\$263K

less AI spend in year 1

- ✓ Public benchmark on an open-source repository; method and questions published for anyone to reproduce
- ✓ On a private production repository, a 10-question test measured **82–90%** input-token reduction
- ✓ Prospects include a global IT-services firm and a US hospital group

Benchmark source: CodeMesh benchmark (codemesh.cloud/benchmarks). Same model and questions in both runs.