# First Benchmark for Legacy Code Comprehension Shows Specialized AI Approach Outperforms General-PurposeModels

*LegacyCodeBench tests whether AI can understand COBOL well enough to document itaccurately not just generate plausible text*

NEW YORK, NY, UNITED STATES, January 13, 2026 /EINPresswire.com/ -- A new benchmark designed to measure whether AI systems can actuallyunderstand legacy enterprise code shows that specialized approaches significantlyoutperform general-purpose models. LegacyCodeBench, developed by Kalmantic (anapplied AI research lab) in collaboration with Hexaview Technologies, evaluates AIcomprehension of COBOL the language still processing 95% of ATM transactions and $3trillion in daily global transactions.
The benchmark finds that domain-specialized systems like Hexaview's Legacy Insightsachieve 92% accuracy, compared to 86-90% for general-purpose models like GPT-4o andClaude Sonnet 4.

-Why This Matters
Over 220 billion lines of COBOL remain in production worldwide, but the engineers whowrote it are retiring. Modernization projects fail at rates exceeding 60%, and the pattern isusually the same: organizations try to replace systems they never fully understood.

"The risk everyone focuses on is the legacy technology itself, but that's not actually whereprojects fall apart," said Ankit Agarwal, Founder and CTO of Hexaview. "What kills these programs is undocumented business logic. We needed an objective way to measurewhether AI can actually understand these systems well enough to trust the output."

-How It Works
Most AI benchmarks use another LLM to judge output quality, which creates reproducibilityproblems. LegacyCodeBench takes a different approach: it verifies claims against theoriginal program's behavior.The process extracts specific behavioral claims from AI-generated documentation -statements like "PREMIUM is calculated by multiplying BASE-RATE by RISK-FACTOR" - andthen verifies them by executing the original COBOL program with test inputs. If the claimdoesn't match what the code actually does, it fails."We're not testing whether

documentation reads well," said Nikita, co-author of the paper."We wanted to know if you could actually trust it. There's a difference."The benchmark also penalizes gaming. Documentation that avoids making testable claimsscores zero on the behavioral track, which carries 50% of the total weight. And if the AIhallucinates variables that don't exist in the source code, the entire task fails

-Results

| System | LCB Score | Structural | Doc Quality | Behavioral | T1 Basic | T4 Enterprise |
| --- | --- | --- | --- | --- | --- | --- |
| Legacy Insights (Hexaview) | 92% | 94% | 96% | 90% | 96% | 90% |
| Claude Sonnet 4 (Anthropic) | 90% | 96% | 78% | 91% | 92% | 92% |
| AWS Transform Mainframe | 88% | 98% | 68% | 91% | 88% | 87% |
| IBM Granite 13B | 87% | 93% | 72% | 90% | 89% | 84% |
| GPT-4o (OpenAI) | 86% | 92% | 71% | 89% | 91% | 82% |

Specialized systems (Legacy Insights, AWS Transform) outperform general-purposemodels, particularly on documentation quality. All models maintain reasonably strongperformance from basic programs (T1) to enterprise-scale COBOL (T4), though GPT-4oshows the largest drop (9 points).

"General-purpose models have gotten quite good at parsing legacy code, which is realprogress," Agarwal said. "But there's still a gap between understanding the syntax andunderstanding what the code is actually doing in a business context. That's wherespecialization matters."

-Open Source
LegacyCodeBench is fully open source with deterministic evaluation. The publicleaderboard is at legacycodebench.com, and the team welcomes submissions via GitHub

-Resources
• Website: legacycodebench.com
• Paper: Available at legacycodebench.com
• GitHub: github.com/kalmantic/legacycodebench
• Legacy Insights: legacyip.hexaview.ai

-About Hexaview
Hexaview is a strategic implementation partner for regulated enterprises, specializing inlegacy system preservation and modernization. Learn more: hexaviewtech.com

-About Kalmantic Labs Kalmantic is an applied AI research lab studying the challenges that emerge when AI meetsproduction systems. They publish research openly and build tools based on their findings.Learn more: kalmantic.com

LegacyCodeBench is open source under MIT license.

Ankit Agarwal
Hexaview Technologies
+1 845-653-3855
email us here

---