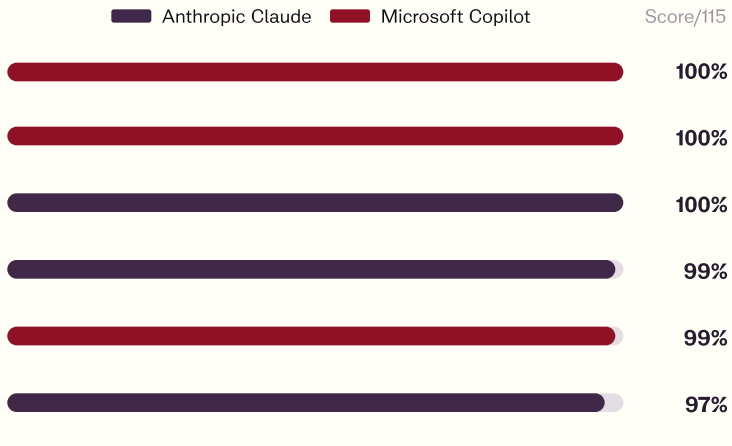
SEI

HEAD-TO-HEAD AI BENCHMARK

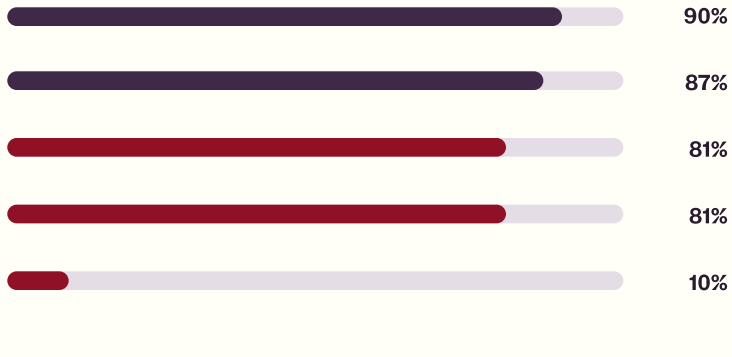
Microsoft Copilot vs. Anthropic's Claude

One everyday spreadsheet task, twelve model-and-settings configurations, 115 scored questions — and a clear lesson about where AI performance actually comes from.

Results By Run



The Performance Cliff



115
QUESTIONS

100K
DATA POINTS

12
CONFIGURATIONS

10-100%
SCORE RANGE

What it means

01. The model matters more than the platform

The strongest Claude models and the "deep-think" ChatGPT models all clustered at 97–100% — whether accessed directly or inside Copilot. The engine matters; the brand on the login screen far less.

02. The defaults are rarely the right setting

Copilot's "Auto" selector scored 81%, and the standard Sonnet default at low effort hit 90% — versus 97% for that same model at high effort.

03. Reasoning effort is the biggest free lever

The same model, moved from low to high effort, swung 7–10 points. It is the single easiest quality gain available — and almost always left switched off.

03. Claude inside Copilot performs on par with Claude direct

At high effort, Claude Opus through Copilot matched Claude used straight from Anthropic — a meaningful result for Microsoft-standardized organizations.

Methodology. One Excel file (100K+ data points), 115 questions, hand-scored against a single rubric across all 12 configurations — designed, executed, and validated by SEI consultant Chris Ventura with no AI assistance. Directional results on moderate, everyday tasks, not a formal published benchmark.