Appendix: Annotated Bibliography

This appendix lists the primary research sources cited in this book, organized by topic. Each entry includes the key finding and why it matters for the delivery gap argument. For the full citation with URLs, see the References section.

THE DELIVERY GAP (GENERATION OUTPACES VERIFICATION)

Faros AI, "AI Productivity Paradox," July 2025. Telemetry from 10,000+ developers across 1,255 enterprise teams. PR volume up 98%, review time per PR up 91%, net DORA throughput improvement: zero. A large telemetry dataset showing that higher generation volume can coexist with slower review and flat delivery outcomes.

Uplevel Data Labs, "Gen AI for Coding Research Report," 2025. 800 developers tracked after Copilot access. No significant cycle-time improvement. 41% increase in bugs. Corroborates the Faros finding from a different angle: more output, more defects, no net gain.

METR (Becker, Rush, Barnes, Rein), "Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity," arXiv:2507.09089. 16 experienced open-source developers, 246 tasks, randomized controlled trial. No clear productivity signal for experienced developers. Important because it is a randomized study focused on experienced developers doing real open-source work.

NAV IT, "GitHub Copilot Longitudinal Study," arXiv:2509.20353, September 2025. 25 Copilot users vs. 14 non-users, 26,317 commits across 703 repos, September 2023 to May 2025. No statistically significant changes in commit-based activity after Copilot adoption. One of the longer-duration published studies in this area.

Xu et al. (Tilburg University), "AI-Assisted Programming Decreases the Productivity of Experienced Developers," arXiv:2510.10165. Core developers increased PR reviews by 6.5% while their own productivity dropped 19%. Shows the review burden cost that aggregate productivity numbers hide.

THE K-SHAPED DIVERGENCE

CircleCI, "The 2026 State of Software Delivery." Top 5% teams grew throughput 97% year over year. Median teams grew 4%. Useful directional evidence for a widening gap between top and median teams. The book's verification-first interpretation is an inference from the throughput and recovery split, not a direct claim by CircleCI.

Brynjolfsson, Li, and Raymond, "Generative AI at Work," NBER Working Paper 31161, 2023. 5,179 customer support agents. Bottom-quartile workers improved 34%; top-quartile showed no significant improvement. The K-shaped pattern in a non-engineering domain.

McKinsey Global Institute, "Superagency in the Workplace," January 2025. Top 6% of organizations captured $10.30 per AI dollar invested versus $3.70 average. The K-shaped divergence measured in ROI rather than throughput.

CODE QUALITY UNDER AI

GitClear, "AI Copilot Code Quality 2025." Analysis of 211 million lines. Code churn rose from 5.5% to 7.9%; refactoring declined from 24.1% to 9.5%. Teams are generating more and understanding less.

He et al. (Carnegie Mellon), "How Cursor Affects Code Quality and Developer Behavior," arXiv:2511.04427, November 2025. 807 Cursor-adopting repos vs. 1,380 matched controls. Static analysis warnings increased after AI tool adoption. Corroborates GitClear from a different methodology.

CodeRabbit, "State of AI vs Human Code Generation Report," December 2025. 470 open-source PRs analyzed. AI-authored code showed 1.7x more issues than human-authored code. Vendor-produced but directionally consistent with academic findings.

THE SPEC-FIRST EVIDENCE

Suri et al., "CodeScout: Contextual Problem Statement Enhancement for Software Agents," arXiv:2603.05744, March 2026. Converting underspecified problem statements into detailed specifications improved SWE-bench Verified resolution rates by 20%. A direct piece of evidence for the spec-first argument.

Montgomery et al., "Empirical Research on Requirements Quality: A Systematic Mapping Study," Requirements Engineering, 2022. Maps more than 100 primary studies on requirements quality. Useful because it shows that ambiguity, incompleteness, inconsistency, and correctness are persistent quality problems long before AI entered the picture.

Albayrak et al., "Incomplete Software Requirements and Assumptions Made by Software Engineers," APSEC 2009. Empirical evidence that incomplete requirements push engineers into filling gaps with assumptions, often implicitly. Important because `rework rate by spec status` is not a standard metric, but it is grounded in the long-standing problem this paper documents.

VERIFICATION METRICS

DORA, "DORA's Software Delivery Performance Metrics." The most established software-delivery metric family in the book's neighborhood: change lead time, deployment frequency, failed deployment recovery time, change fail rate, and deployment rework rate. These are the anchor metrics the book should lean on first.

DORA, "A History of DORA's Software Delivery Metrics." Explains how DORA evolved from four metrics to five by adding deployment rework rate. Useful because it legitimizes treating rework as a first-class delivery signal rather than a side note.

Capers Jones, "Software Defect Removal Efficiency," IEEE Computer, 1996. Foundational software-quality framing for measuring how many defects are removed before release versus after release. The

ancestor of defect removal efficiency metrics; the book uses DORA's `change failure rate` as the primary production-quality metric.

"Towards a Science of AI Agent Reliability," arXiv:2602.16666, February 2026. Shows that agent capability and agent reliability diverge, and argues for decomposed reliability metrics rather than a single benchmark score. Useful support for introducing `machine catch rate` as an AI-era diagnostic.

"Measuring Agents in Production," arXiv:2512.04123, December 2025. Survey and interview study showing that reliability and correctness remain the dominant production challenge for AI agents. Supports the book's broader argument that verification metrics deserve first-class status.

JUDGMENT AND DECISION-MAKING

***Tetlock, P. E.,* Expert Political Judgment, *Princeton University Press, 2005*.** Twenty-year study of prediction accuracy. "Foxes" who synthesized diverse perspectives consistently outperformed "hedgehogs" who relied on deep domain expertise. Foundational evidence that judgment is a cognitive style, not a function of years.

***Mellers et al., "Identifying and Cultivating Superforecasters,"* Perspectives on Psychological Science, *10(3), 2015*.** IARPA tournament. Superforecasters identified by cognitive traits (active open-mindedness, belief-updating), not domain expertise. Crucially, accuracy improved with training. Judgment is cultivable.

***Klein, G.,* Sources of Power: How People Make Decisions, *MIT Press, 1998*.** Recognition-Primed Decision model. Expert judgment decomposed into pattern library (requires experience) and mental simulation skill (trainable). Reframes the hiring question: assess simulation ability, not just years.

***Murphy-Hill et al., "What Predicts Software Developers' Productivity?"* IEEE Transactions on Software Engineering, *47(3), 2019*.** 622 developers across 3 companies. Years of employment showed "poor and insignificant correlation" with general performance. Strongest predictors: enthusiasm, peer support, useful feedback.

Dell'Acqua et al., "Navigating the Jagged Technological Frontier," HBS Working Paper 24-013, September 2023. 758 BCG consultants using GPT-4. Inside the AI capability frontier: +40% quality. Outside it: −19 percentage points vs. the no-AI group. Same tool, opposite outcomes. The variable was judgment.

COGNITIVE LOAD AND REVIEW QUALITY

Bedard, Kropp, Hsu et al., "When Using AI Leads to 'Brain Fry,'" Harvard Business Review, March 2026. BCG survey of 1,488 workers. Workers using 4+ AI tools reported 14% more mental effort, 12% greater fatigue, 19% greater information overload. Four tools was the threshold where productivity gains reversed.

Ranganathan and Ye (UC Berkeley Haas), "AI Doesn't Reduce Work, It Intensifies It," Harvard Business Review, February 2026. Ethnographic field research, 8 months, approximately 200 employees. AI associated with work intensification, not workload reduction.

***Edmondson, A., "Psychological Safety and Learning Behavior in Work Teams,"* Administrative Science Quarterly, *44(2), 1999*.** 51 work teams. Psychological safety was the strongest predictor of team learning and performance. Best-performing teams reported more errors because they felt safe surfacing problems.

AGENT SAFETY AND RISK

Cooperative AI Foundation, "Multi-Agent Risks from Advanced AI," Technical Report #1, arXiv:2502.14143, February 2025. 47 authors across multiple institutions. Multi-agent networks amplified errors up to 17.2x compared to single-agent systems. Strong evidence for caution around multi-agent-first designs.

Reuel, Anka et al., "The 2025 AI Agent Index," arXiv:2602.17753, February 2026. Study of 30 deployed agentic systems across 1,350 data fields. 25 of 30 disclosed no internal safety evaluation results. 23 of 30 had no third-party safety testing. Documents the current state of agent safety practices.

Spracklen et al., "Package Hallucinations by Code Generating LLMs," USENIX Security 2025 Distinguished Paper Award, arXiv:2406.10279. 576,000 code samples across 16 LLMs: 5.2% hallucination rate for commercial models, 21.7% for open-source models. Critical evidence for supply chain security risk in AI-generated code.

Venkatesh et al., "Outcome-Driven Constraint Violations in Autonomous AI Agents," arXiv:2512.20798, December 2025. 40 scenarios evaluating agents that pursue unintended harmful strategies as instrumental steps toward their objectives. Evidence for the override mechanism as a last line of defense.

MARKET AND TALENT DATA

LinkedIn, "Jobs on the Rise 2026." AI Engineer ranked the #1 fastest-growing role in the US. Signals that the market is creating new roles, not just eliminating old ones.

Robert Half, "2026 Technology Salary Guide." AI/ML engineers: +4.4% salary growth to $170,750 average. Overall tech: +1.6%. The salary premium for AI-adjacent skills is measurable and growing.

World Economic Forum, "Future of Jobs Report 2025." 39% of key skills expected to change by 2030. Analytical thinking ranked #1 core skill by 7 in 10 employers. The market is selecting for judgment, not speed.

Gartner, "Global AI Regulations Fuel Billion-Dollar Market for AI Governance Platforms," February 2026. AI governance platform spending projected at $492M in 2026, $1B+ by 2030. Compliance requirements are creating demand for verification infrastructure.