Quantifying the Expectation–Realisation Gap for Agentic AI Systems

This manuscript (permalink) was automatically generated from slolab/agentic-expectation-realisation-gap@f94562f on February 11, 2026.

Authors

- **Sebastian Lobentanzer** ✉
  https://orcid.org/0000-0003-3399-6695 · slobentanzer
  Institute of Computational Biology, Computational Health Center, Helmholtz Center, Munich, Germany; German Center for Diabetes Research, Munich, Germany

✉ — Correspondence possible via GitHub Issues or email to Sebastian Lobentanzer <sebastian.lobentanzer@helmholtz-munich.de>.

Abstract

Agentic AI systems are deployed with expectations of substantial productivity gains, yet rigorous empirical evidence reveals systematic discrepancies between pre-deployment expectations and post-deployment outcomes. We review controlled trials and independent validations across software engineering, clinical documentation, and clinical decision support to quantify this expectation–realisation gap. In software development, experienced developers expected a 24% speedup from AI tools but were slowed by 19%—a 43 percentage-point calibration error. In clinical documentation, vendor claims of multi-minute time savings contrast with measured reductions of less than one minute per note, and one widely deployed tool showed no statistically significant effect. In clinical decision support, externally validated performance falls substantially below developer-reported metrics. These shortfalls are driven by workflow integration friction, verification burden, measurement construct mismatches, and systematic heterogeneity in treatment effects. The evidence motivates structured planning frameworks that require explicit, quantified benefit expectations with human oversight costs factored in.

Introduction

Agentic AI systems—autonomous software agents that plan, reason, and execute multi-step tasks with limited human oversight—are being adopted across software engineering, clinical medicine, and customer operations with the expectation of transformative productivity gains. Vendor announcements routinely promise multi-minute time savings per encounter, double-digit percentage speedups, or near-expert-level decision accuracy. Procurement and investment decisions follow these expectations, committing substantial resources before deployment-grade evidence is available.

Yet a growing body of controlled trials and independent external validations reveals that realised outcomes frequently fall short of pre-deployment expectations—sometimes dramatically so. This discrepancy, which we term the *expectation–realisation gap*, is not simply a matter of immature technology. It reflects systematic patterns in how agentic systems interact with human workflows, how performance is measured, and how benefits are distributed across user populations.

Understanding and quantifying this gap is a prerequisite for responsible deployment. Without structured, quantified expectations that account for real-world integration costs, organisations risk over-investing in systems that deliver marginal gains at best or impose net costs at worst. This review synthesises the strongest available empirical evidence for the expectation–realisation gap across three domains—software engineering, clinical documentation, and clinical decision support—identifies the mechanistic drivers of shortfalls, and argues that structured planning frameworks with explicit benefit quantification are necessary to close the gap between aspiration and reality.

Evidence from controlled trials

Software engineering copilots

The sharpest illustration of the expectation–realisation gap comes from a randomised controlled trial conducted by METR (Model Evaluation & Threat Research) on 16 experienced open-source developers working in their own mature repositories across 246 real tasks [1]. Before each task, participants forecast that AI assistance would reduce their completion time by 24%. The measured outcome was a 19% *increase* in completion time—a 43 percentage-point calibration error on the time-change scale, and a complete reversal in direction. Tasks took approximately 56% longer than developers expected (realised completion time factor 1.19 versus expected 0.76).

This result contrasts instructively with a pre-release controlled trial of GitHub Copilot on developers recruited via Upwork, where treated participants completed a standardised, self-contained programming task 55.8% faster (95% CI 21–89%) [2]. In that experiment, participants' self-estimated productivity gains averaged approximately 35%, meaning they *underestimated* the realised speedup. The divergence between these two trials is the key finding: on constrained, well-defined tasks, AI coding tools can exceed expectations; in high-context, real-world repositories, the same class of tools can impose net costs that developers fail to anticipate.

Field experiments at Microsoft and Accenture provide intermediate evidence: developers assigned to Copilot completed 12.9–21.8% more pull requests per week at Microsoft and 7.5–8.7% more at Accenture, though the authors emphasise imprecision and threats to inference including low compliance and organisational confounds [3]. Crucially, these throughput metrics do not account for code quality: independent security analyses find that 32.8% of Python and 24.5% of JavaScript snippets generated by Copilot are flagged with security issues [4], and Copilot can replicate known-vulnerable code patterns at rates around 33% [5]. Productivity gains that increase review, remediation, and incident risk are not net gains.

Clinical documentation agents

Ambient AI scribes—systems that listen to clinical encounters and generate draft documentation—represent one of the most actively deployed categories of agentic AI in healthcare. Vendor procurement narratives frequently frame benefits in terms of "minutes saved per encounter"; for instance, Microsoft publicised "5 minutes saved per clinician per encounter on average" for its DAX Copilot product [6].

The strongest trial evidence tells a different story. A randomised controlled trial at UCLA across 238 physicians in 14 specialties compared two commercial ambient scribe tools (DAX and Nabla) against usual care, with approximately 24,000 encounters per arm [7]. Nabla reduced time-in-note by 9.5% relative to control (95% CI −17.2 to −1.8; P=0.02), while DAX showed no statistically significant effect (−1.7%, 95% CI −9.4 to +5.9; P=0.66). Partial adoption was a key contextual factor: the tools were used in only approximately 30–34% of visits, and roughly 15% of treatment-group physicians never used their assigned scribe at all. Clinicians did not strongly endorse that generated notes were "at least as good as my own," rating this near neutral, and clinically significant inaccuracies were reported as occurring "occasionally."

A peer-matched cohort study of DAX in an integrated delivery system (99 providers, 12 specialties) found documentation EHR time fell from 5.3 to 4.54 minutes per patient—a saving of approximately 46 seconds—while after-hours EHR time *worsened* significantly, suggesting time-shifting rather than uniform savings [8]. A pre/post study of the Abridge ambient listening tool across 332 physicians confirmed sub-minute savings: mean time in notes per note fell from 5.11 to 4.16 minutes (difference 0.95 minutes, 95% CI 0.48–1.42) [9].

Perhaps most revealing is the perception–reality mismatch documented in a study of 252 physicians: 86.5% *perceived* that their documentation time had decreased, yet there was no overall association between perceived reductions and objectively measured time changes (OR 0.975, P=0.144) [10]. The objective effect was modest: each 10 percentage-point increase in AI scribe usage was associated with approximately 30 seconds lower documentation time per scheduled hour.

Clinical decision support

Independent external validation of proprietary clinical AI models provides some of the starkest expectation–realisation gaps. The Epic Sepsis Model, widely implemented across US hospitals, was externally validated in a large academic health system (38,455 hospitalisations) with an area under the receiver operating characteristic curve (AUC) of 0.63 (95% CI 0.62–0.64), while Epic's internal documentation reported AUC values of 0.76–0.83 [11]. At an operational alert threshold, the model achieved only 33% sensitivity, raising questions about clinical utility at scale.

A similar pattern appears in oncology decision support. IBM publicised concordance rates as high as 96% for Watson for Oncology in lung cancer cases relative to a multidisciplinary tumour board [12]. A subsequent peer-reviewed retrospective study in Korea found strict concordance of 48.9% for colon cancer, with "acceptable" concordance of 65.8% and strong heterogeneity by patient age (concordance dropping to approximately 20% among patients aged 70 and older) [13]. This discrepancy reflects both definition dependence—concordance rises substantially when "for consideration" is treated as concordant—and local constraint mismatches in guidelines, reimbursement, and patient demographics that prevent cross-site transferability.

Why expectations overshoot

The empirical evidence points to three recurrent mechanistic drivers that explain why expectations systematically exceed realised outcomes.

Workflow integration friction and partial adoption. Agentic AI systems do not operate in isolation; they must integrate into existing workflows, tools, and team practices. Clinical scribe evaluations repeatedly show partial adoption—tools used in a minority of encounters, with non-trivial drop-off over time [7,8]. Even when per-use effects are real, intention-to-treat estimates are attenuated by low compliance, and the practical benefit to an organisation depends on the adoption rate actually achieved, not the rate assumed during procurement. This is not a temporary onboarding issue; the UCLA RCT's 30–34% utilisation rate was observed over the full study period.

Verification and review burden. Agentic systems generate outputs that require human verification, and this verification cost is rarely accounted for in pre-deployment projections. In the METR software engineering trial, the net slowdown occurred because the time spent reviewing, debugging, and integrating AI-generated code exceeded the time saved in initial generation [1]. In clinical documentation, neutral ratings on note quality and "occasional" clinically significant inaccuracies indicate non-trivial editing and review work that partially or fully offsets time-in-note reductions [7]. The DAX cohort's simultaneous reduction in documentation time and *increase* in after-hours EHR time is a concrete example of time-shifting: the verification and cleanup work does not disappear, it moves [8].

Measurement construct mismatch. Pre-deployment expectations are often framed in metrics that do not correspond to what deployment-grade evaluations actually measure. Vendor claims of "minutes saved per encounter" refer to broader workflow impacts, while trial outcomes measure "time-in-note"—one slice of documentation burden [7]. Developer-reported model performance (AUC 0.76–0.83 for Epic's sepsis model) reflects evaluation choices—time horizons, denominators, and inclusion of post-onset predictions—that systematically inflate apparent discrimination relative to operational alerting performance [11]. The gap between lab-task performance and field performance in software copilots is a measurement construct problem at its core: bounded tasks estimate *tool capability under low-context load*, while field trials estimate *net productivity under realistic verification and integration costs* [1,2].

Heterogeneity as the default

Across every domain reviewed, treatment effects are not uniform. They are systematically moderated by baseline user efficiency, task complexity, and local context. This heterogeneity is not noise—it is the central empirical regularity.

In clinical documentation, objective time savings from AI scribes concentrate among physicians with higher baseline documentation inefficiency; efficient documenters derive minimal benefit [10]. In customer support, a field study of 5,172 agents found an average 15% productivity increase from a generative AI assistant, but gains were heavily concentrated among less experienced and lower-skilled workers, while the most experienced agents saw smaller gains and occasional quality declines [14]. In software engineering, the METR trial specifically selected experienced developers working in familiar repositories—precisely the population most likely to have optimised their workflows already—and this is the population that was slowed [1].

The implication is direct: there is no stable, globally positive treatment effect for agentic AI. Average headline figures (whether from vendors, lab trials, or even well-designed field studies) will systematically misrepresent the benefit realised by any specific user, team, or organisation. Planning that relies on average expected gains without modelling who benefits and who does not will over-invest in low-yield deployments and under-invest in targeted high-yield ones.

Adjacent experimental evidence reinforces this concern on a longer time horizon. A randomised controlled trial in higher education found that students who used ChatGPT as a study aid scored significantly lower on a surprise retention test 45 days later (57.5% vs 68.5%; Cohen's d = 0.68) [15], suggesting that cognitive offloading can trade immediate task completion for degraded durable learning—a dimension of "benefit" that short-term productivity metrics entirely miss.

Implications for structured planning

The evidence reviewed here converges on a clear conclusion: pre-deployment expectations for agentic AI systems are poorly calibrated, and the resulting expectation–realisation gap is large enough to undermine investment decisions, deployment strategies, and trust. This is not an argument against agentic AI—the evidence also shows that real gains exist in specific contexts and for specific user populations. It is an argument for *structured planning that takes the gap seriously*.

Several design principles follow directly from the empirical patterns. First, benefit expectations must be *explicit and quantified*, not framed as vague promises of efficiency. The contrast between "5 minutes saved per encounter" marketing and sub-minute measured reductions illustrates what happens when expectations lack precision. Second, expectations should capture *dual perspectives*—what users expect to gain and what developers assess as technically feasible—because miscalibration occurs on both sides (developers overshoot in internal validation; users overshoot in self-forecasts). Third, *human oversight costs must be deducted* from projected benefits. Every controlled trial reviewed here shows that verification, review, and cleanup absorb a substantial fraction of the gross time savings; ignoring this yields unrealistic net benefit estimates. Fourth, *outcome metrics must link back to initial expectations* in the same units and at the same level of granularity, enabling direct comparison rather than post hoc rationalisation. Fifth, heterogeneity should be *modelled explicitly* by specifying which user populations and task types are expected to benefit, rather than assuming uniform effects.

These principles are implemented in the Agentic Automation Canvas (AAC), a structured framework for designing, governing, and documenting agentic automation projects that captures user expectations as quantified benefit metrics with baseline values, confidence levels from both user and developer perspectives, and explicit human oversight accounting [16]. The canvas formalises the bidirectional contract between stakeholders that the evidence reviewed here shows is necessary: without structured mechanisms for surfacing and testing expectations, the gap between aspiration and reality will persist.

Conclusion

The expectation–realisation gap for agentic AI systems is empirically documented, directionally consistent, and mechanistically explicable. Across software engineering, clinical documentation, and clinical decision support, pre-deployment expectations—whether from user forecasts, vendor claims, or developer-reported metrics—systematically overestimate realised benefits in deployment settings. The drivers are not mysterious: workflow integration friction, verification burden, measurement construct mismatches, and treatment effect heterogeneity are observable, predictable, and in principle addressable.

Closing this gap requires moving from *ad hoc* expectation-setting to structured, quantified planning that accounts for real-world integration costs, models heterogeneity across user populations, and links outcome measurement directly to initial benefit projections. The alternative—continued reliance on benchmark results, marketing claims, and intuitive forecasts—will perpetuate a cycle of over-

promise and under-delivery that erodes trust in systems that, when properly targeted and governed, can deliver genuine value.

References

1. **Measuring the Impact of Early-2025 AI on Experienced Open-Source Developer Productivity**
   Joel Becker, Nate Rush, Elizabeth Barnes, David Rein
   arXiv (2025-07-28) https://arxiv.org/abs/2507.09089

2. **The Impact of AI on Developer Productivity: Evidence from GitHub Copilot**
   Sida Peng, Eirini Kalliamvakou, Peter Cihon, Mert Demirer
   arXiv (2023-02-14) https://arxiv.org/abs/2302.06590

3. **The impact of generative AI on software developer productivity: evidence from two field experiments**
   Kevin Cui, Deepak Paramanand, Robert Sloyan
   MIT GenAI Impact (2025) https://mit-genai.pubpub.org/pub/v5iixksv

4. **Security Weaknesses of Copilot-Generated Code in GitHub Projects: An Empirical Study**
   Yujia Fu, Peng Liang, Amjed Tahir, Zengyang Li, Mojtaba Shahin, Jiaxin Yu, Jinfu Chen
   arXiv (2025-02-07) https://arxiv.org/abs/2310.02059

5. **Is GitHub's Copilot as Bad as Humans at Introducing Vulnerabilities in Code?**
   Owura Asare, Meiyappan Nagappan, N Asokan
   arXiv (2024-01-09) https://arxiv.org/abs/2204.04741

6. **DAX Copilot: new customization options and AI capabilities for even greater productivity**
   Microsoft
   Microsoft Industry Blogs (2024-08-08) https://www.microsoft.com/en-us/industry/blog/healthcare/2024/08/08/dax-copilot-new-customization-options-and-ai-capabilities-for-even-greater-productivity/

7. **Ambient AI Scribes in Clinical Practice: A Randomized Trial**
   Paul J Lukac, William Turner, Sitaram Vangala, Aaron T Chin, Joshua Khalili, Ya-Chen Tina Shih, Catherine Sarkisian, Eric M Cheng, John N Mafi
   NEJM AI (2025-12) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12768499/
   DOI: 10.1056/aioa2501000 · PMID: 41497288 · PMCID: PMC12768499

8. **The impact of nuance DAX ambient listening AI documentation: a cohort study**
   Tyler Haberle, Courtney Cleveland, Greg L Snow, Chris Barber, Nikki Stookey, Cari Thornock, Laurie Younger, Buzzy Mullahkhel, Diego Ize-Ludlow
   Journal of the American Medical Informatics Association : JAMIA (2024-04-03)
   https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10990544/
   DOI: 10.1093/jamia/ocae022 · PMID: 38345343 · PMCID: PMC10990544

9. **Ambient listening implementation in primary care and changes in electronic health record documentation metrics: Pre-post study of an ambient listening tool**
   Frederick North, Marc R Matthews, Asif Iqbal, Jason A Post, Jon O Ebbert
   Digital health (2025-11-26) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC12657781/
   DOI: 10.1177/20552076251403211 · PMID: 41323090 · PMCID: PMC12657781

10. **Subjective and objective impacts of ambulatory AI scribes.**
    Julia Adler-Milstein, Orianna DeMasi, Hossein Soleimani, Sarah Beck, Maria E Byron, Aris Oates, Robert Thombley, Jinoos Yazdany, Sara G Murray
    The American journal of managed care (2026-01)
    https://www.ncbi.nlm.nih.gov/pubmed/41592210

DOI: 10.37765/ajmc.2026.89869 · PMID: 41592210

11. **The Epic Sepsis Model Falls Short—The Importance of External Validation**
Anand R Habib, Anthony L Lin, Richard W Grant
JAMA Internal Medicine (2021-08-01) https://doi.org/g99mhd
DOI: 10.1001/jamainternmed.2021.3333 · PMID: 34152360

12. **At ASCO 2017, clinicians present new evidence about Watson cognitive technology and cancer care**
IBM
IBM Newsroom (2017-06-01) https://uk.newsroom.ibm.com/2017-06-01-At-ASCO-2017-Clinicians-Present-New-Evidence-about-Watson-Cognitive-Technology-and-Cancer-Care

13. **Assessing Concordance With Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea.**
Won-Suk Lee, Sung Min Ahn, Jun-Won Chung, Kyoung Oh Kim, Kwang An Kwon, Yoonjae Kim, Sunjin Sym, Dongbok Shin, Inkeun Park, Uhn Lee, Jeong-Heum Baek
JCO clinical cancer informatics (2018-12) https://www.ncbi.nlm.nih.gov/pubmed/30652564
DOI: 10.1200/cci.17.00109 · PMID: 30652564

14. **Generative AI at Work**
Erik Brynjolfsson, Danielle Li, Lindsey Raymond
arXiv (2024-11-07) https://arxiv.org/abs/2304.11771

15. **The hidden cost of AI: ChatGPT use impairs long-term knowledge retention in university students**
Muhammad Farrukh Shahzad
International Journal of Educational Research Open (2025)
https://www.sciencedirect.com/science/article/pii/S2590291125010186

16. **The Agentic Automation Canvas: a structured framework for agentic AI project design**
Sebastian Lobentanzer
(2025)