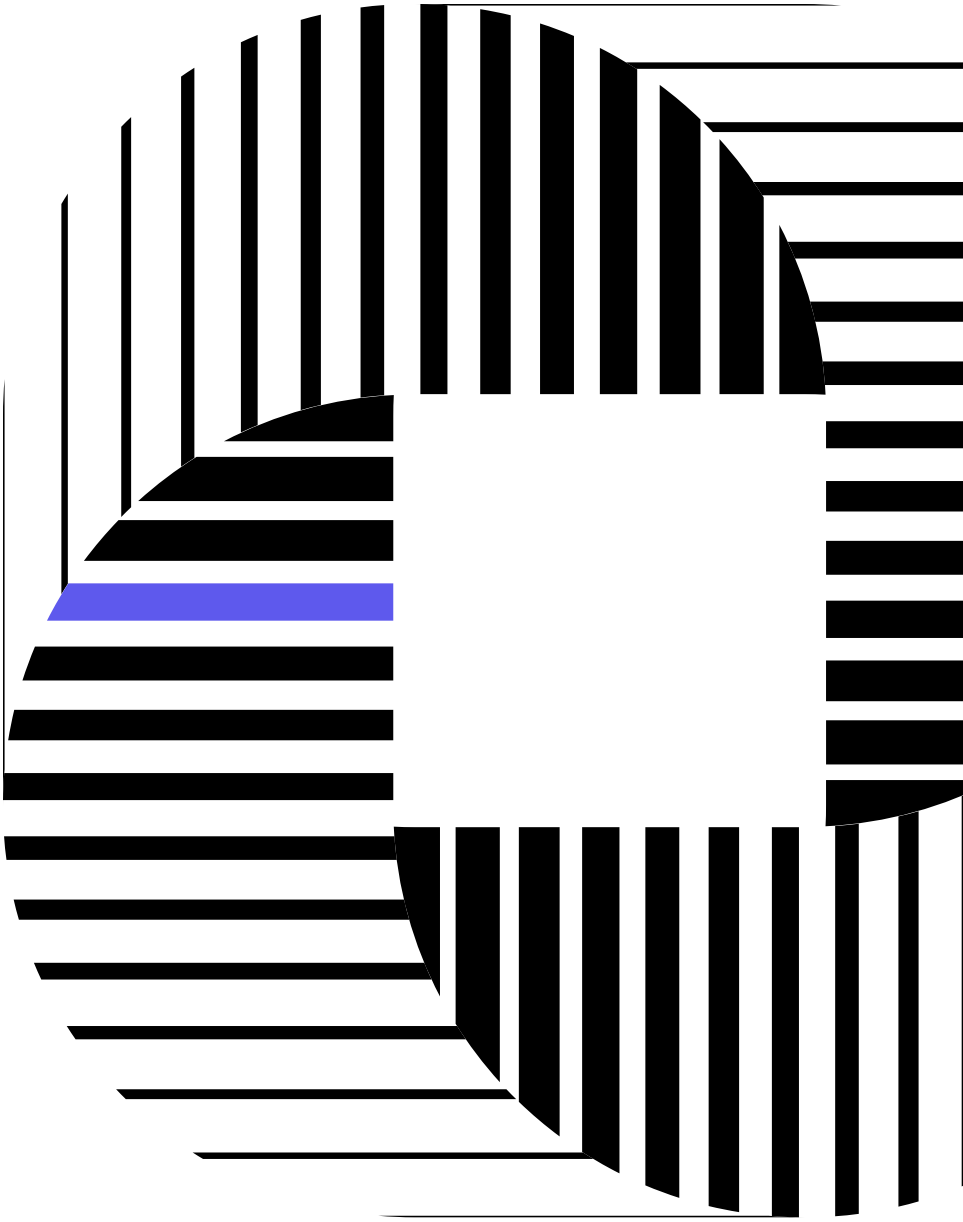

Measuring AI code assistants and agents

Measuring AI code assistants and agents

Abi Noda, Laura Tacho

To thrive in the AI era, organizations must adapt quickly. The DX AI Measurement Framework™ offers research-based metrics for measuring the impact of AI-assisted engineering in your organization.

The rise of AI is reshaping how engineering organizations must chart their path to success. Companies are no longer as limited by the number of engineers they can hire, but rather, the degree to which they can augment them with AI to gain leverage.

In order to navigate this shift, leaders need metrics on the efficacy of AI code tools and agents. But today, many leaders struggle to answer pressing questions: Which tools are working? How are they being used? What's actually driving value?

The DX AI Measurement Framework™ includes AI-specific metrics to enable organizations to track AI adoption, measure impact, and make smarter investments—all while continuing to roll out and experiment with AI tools at a rapid pace. When combined with the DX Core 4, which measures overall engineering productivity, leaders gain deep insight into how AI is providing value to their developers, and what impact AI is having on organizational performance.

Our approach has been developed in partnership with leading companies, researchers, and AI vendors. Booking.com used this framework to deploy AI tools to over 3,500 engineers and, within several months, achieved a 16% increase in throughput. Block, with over 4,000 engineers, leverages this data-driven approach to guide its AI engineering strategy—including the development of their AI agent, codename goose.

The framework

Effectively measuring AI code assistants and agents requires focusing on three key dimensions: utilization, impact, and cost. These dimensions align closely with the natural lifecycle of AI adoption—where teams first prioritize adoption and usage, then shift to measuring impact, and eventually focus on governance, standardization, and cost efficiency.

Table 1: DX AI Measurement Framework

* Metrics for autonomous AI agents

Utilization How much are developers adopting and utilizing AI tools?	**Impact** How is AI impacting engineering productivity?	**Cost** Is our AI spend and return on investment optimal?

• AI tool usage (DAUs/WAUs)
• Percentage of PRs that are AI-assisted
• Percentage of committed code that is AI-generated
• Tasks assigned to agents *
	• AI-driven time savings (dev hours/ week)
	• Developer satisfaction
	• DX Core 4 metrics, including:
	  • PR throughput
	  • Perceived rate of delivery
	  • Developer Experience Index (DXI)
	  • Code maintainability
	  • Change confidence
	  • Change fail percentage
	• Human-equivalent hours (HEH) of work completed by agents *
		• AI spend (both total and per developer)
		• Net time gain per developer (time savings - AI spend)
		• Agent hourly rate (HEH / AI spend) *

Driving successful adoption and utilization of AI tools is a top priority for organizations today. Never before has such tangible impact been so closely tied to the adoption of a specific tool. For example, by nearly doubling adoption of AI code assistants, Intercom—one of the world's leading AI customer service companies—achieved a 41% increase in AI-driven developer time savings. Metrics and tracking were key to Intercom's ability to drive adoption of new AI workflows.

Adoption is just the beginning—real impact comes from using data to inform strategic enablement, skill development, and high-leverage use cases. Yet measuring productivity, and AI's contribution to it, remains difficult for most organizations. Based on our experience, the most reliable approach combines direct and indirect metrics rather than relying on any one single measure.

We recommend starting by measuring impact with direct metrics like AI-driven time savings (e.g., time saved per developer per week). These direct metrics offer immediate signals to evaluate the effectiveness of specific tools. Indirect measurements, through regression and longitudinal analysis of DX Core 4 metrics (including PR throughput, Perceived Rate of Delivery, and the Developer Experience Index), help surface longer-term benefits and hidden risks.

Once past tool selection and rollout, tracking cost becomes essential—not just to monitor usage, but to identify high-ROI use cases worth replicating. This is also the stage where standardization and governance matter most: setting model configurations, usage guidelines, and security protocols to ensure scalable, compliant AI adoption. Without these frameworks, organizations risk inconsistent outcomes, security gaps, and missed opportunities to scale impact.

Balance velocity with quality and maintainability

While AI tools can deliver impressive speed gains in the near term, organizations must balance these efficiency measures with quality metrics to avoid undermining long-term velocity. For example, code generated by AI may be less intuitive for human developers to understand, potentially creating bottlenecks when issues arise or modifications are needed.

However, AI tools can also empower developers to work confidently with unfamiliar or complex code they might otherwise avoid touching. By tracking both immediate AI-driven improvements and longer-term, larger scope metrics, organizations can identify the right balance where AI enhances both speed and sustainable code quality.

Measure agents as extensions of teams and individuals

One of the most thought-provoking challenges in the AI era is how to measure the impact of autonomous agents. Should they be treated as independent contributors—or as extensions of the teams that deploy and manage them?

In our experience, the most effective approach is to treat agents as extensions of the developers and teams that oversee their work. For example, when assessing a team's PR throughput, it's important to include both human-authored pull requests and those authored by agents operating under that team's direction.

This reflects a broader shift we anticipate: every developer will increasingly operate as a "lead" for a team of AI agents, and the skills of the human operator will be an important factor. Developers will increasingly be measured similar to how managers are measured today—based on the performance of their teams.

We note, however, that agentic tooling is still in its early stages. As these tools mature, measurement strategies and working models must evolve in parallel. Our guidance on measuring agentic AI models will continue to be refined according to new research and field experience.

Set goals based on real industry data

One of the key challenges for leaders today is reconciling the astronomical performance claims seen online with the results they see in their own organizations. Among peers, researchers, and experienced leaders, there's a shared understanding that these numbers often don't reflect reality.

For instance, DX research shows that even leading organizations are only reaching around 60% active usage of AI tools. Yet the landscape is evolving rapidly—our data also shows a significant rise in developer sentiment and AI-driven time savings over the past twelve months.

article, the opportunity for industry-wide benchmarking becomes clear. At DX, we've gathered over four million benchmark samples across hundreds of organizations, helping leaders contextualize their performance, set realistic expectations, and ensure they're staying competitive as AI accelerates.

Expand the definition of "developer"

AI is not just accelerating the work of tenured engineers—it's reshaping who gets to participate in software creation. Product managers, designers, and business analysts are increasingly using AI tools to generate working software, blurring the lines between technical and non-technical roles.

As this shift unfolds, the definition of "developer" must evolve accordingly—and so should the ways we measure impact. However, it remains critical to distinguish between production-grade contributions and disposable, AI-generated prototypes from tools like Lovable. Accurate measurement depends on understanding both who is contributing and what kind of contribution they're making.

How to roll out metrics

As with any measurement effort, leaders must be intentional about how metrics are introduced and communicated. Measuring developer activity—especially in the context of AI—can be a sensitive topic. The hype surrounding AI, combined with the growing telemetry surfaced by AI tools, has only intensified the pressure teams feel.

In this environment, we strongly caution against top-down mandates or using metrics for individual performance evaluation. Metrics like code generation volume are particularly susceptible to gaming. Encouraging behavior that optimizes for the metric, rather than the outcome, risks malicious compliance—undermining team trust and rendering the data meaningless.

Proactive communication is essential. Without it, speculation and fear can fill the void. When rolling out metrics related to AI usage, we recommend clearly emphasizing:

- These metrics will not be used in individual performance evaluations, and reinforcing this by pointing to the existing performance review process.
- The purpose of measurement is to understand how AI-assisted work affects developer experience and software quality, not to micromanage output.
- Data is necessary to guide organizational investment, helping teams determine which tools and workflows deliver real value and which do not.

Don’t lose sight of the bigger picture

While AI is becoming a central force in software development, it’s not the only lever that matters. Many organizations are seeing significant gains in development speed from AI—but they’re also recognizing that their biggest bottlenecks often lie elsewhere: in the outer loop, or in human factors like collaboration, alignment, and the ability to do deep, focused work.

That’s why it’s critical to pair the AI-specific recommendations in this article with continued measurement of overall developer productivity. The DX Core 4 provides a comprehensive view of productivity across four key dimensions, while this new framework focuses specifically on measuring the adoption and impact of AI tools.

Early data from companies show that AI can provide lift across all four Core 4 dimensions. By combining broader productivity metrics with targeted AI measures, organizations can effectively track progress—and adapt their strategies—as the role of AI in software development continues to evolve.

About the authors

Abi Noda is co-founder and CEO of DX where he leads the company’s strategic direction and R&D efforts. Laura Tacho is CTO at DX and leads the company’s executive advisory practice.

Engineering Intelligence

Learn more at getdx.com