ISBSG
Delivering IT Confidence

The global and independent source of data and analysis for the IT industry

Short paper

Benchmarking Artificial Intelligence & Machine Learning Projects

Published by the International Software Benchmarking Standards Group

2026

Introduction

The ISBSG repository contains a vast array of project data and includes data for projects undertaken in an agile way of working. This enables analysis of the differences between traditional projects and agile projects.

ISBSG collects industry data, where output is measured using ISO/IEC standardized and therefore objective, repeatable, auditable methods, such as Nesma, IFPUG and COSMIC function points.

Typical key metrics based on function points are:

- Project Delivery Rate (PDR)[1]: Hours spent per function point
- Cost efficiency: Cost (or Price) per function point
- Quality: Defects per function point (in test and/or 1st month of production)
- Delivery Speed: Function points delivered per calendar month.

Generative artificial intelligence (AI) and machine learning (ML) are no longer confined to coding assistants that help developers write software faster — increasingly, they are the functionality being delivered. Recommendation engines, computer vision, fraud detection, forecasting and conversational AI are becoming ordinary line items in development portfolios.

In this short report we look at what it takes to benchmark these AI/ML application projects themselves, why today's function point standards and the ISBSG repository are not yet equipped to do so, and what the latest industry data on AI outcomes tells us about why that gap matters.

Background

Our February 2026 short report examined how AI coding assistants — tools such as GitHub Copilot, Cursor and Claude Code — affect the productivity and delivery speed of building ordinary software. This report asks a different question. A growing share of new development and enhancement projects are not merely built with AI tools;

[1] The PDR is the inverse of the universal concept of Productivity (output/input) as it is easier to process for human minds, which usually struggles with metrics with many decimals

they exist to deliver AI/ML capability as the product itself - for example, a recommendation engine, a churn-prediction model, a document-classification pipeline, or an agentic customer-service bot. These AI/ML application projects have their own functional requirements, their own effort profile, and, as we show below, their own risk of being measured (or mismeasured) by yardsticks built for conventional software.

That risk is not theoretical. The outcomes of these projects are, by every recent account, highly variable, yet none of the organizations reporting on them can explain that variability in the objective, function-point terms that ISBSG has used to explain software delivery outcomes for almost three decades.

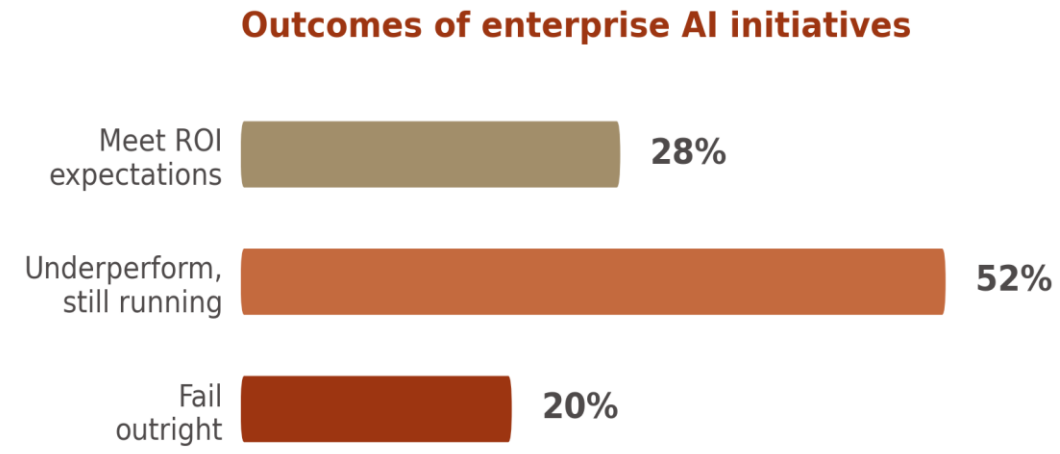
In this short report we bring together the latest external evidence on AI project outcomes with what an ISBSG-style, size-based benchmark of these projects would need to measure. We show why the functional sizing community's own standards are only just starting to catch up.

Analysis and Key Findings

This section brings together the most recent external data on AI/ML project and initiative outcomes. It sets it against what ISBSG-style functional size measurement can and, today, still cannot tell us about why those outcomes occur.

The ROI and Adoption Reality

A Gartner survey of 782 infrastructure & operations leaders (November–December 2025) found that only 28% of AI initiatives meet their organization's return on investment (ROI) expectations. 20% fail outright, and the remaining 52% continue running while underperforming against their business case. See Figure 1.



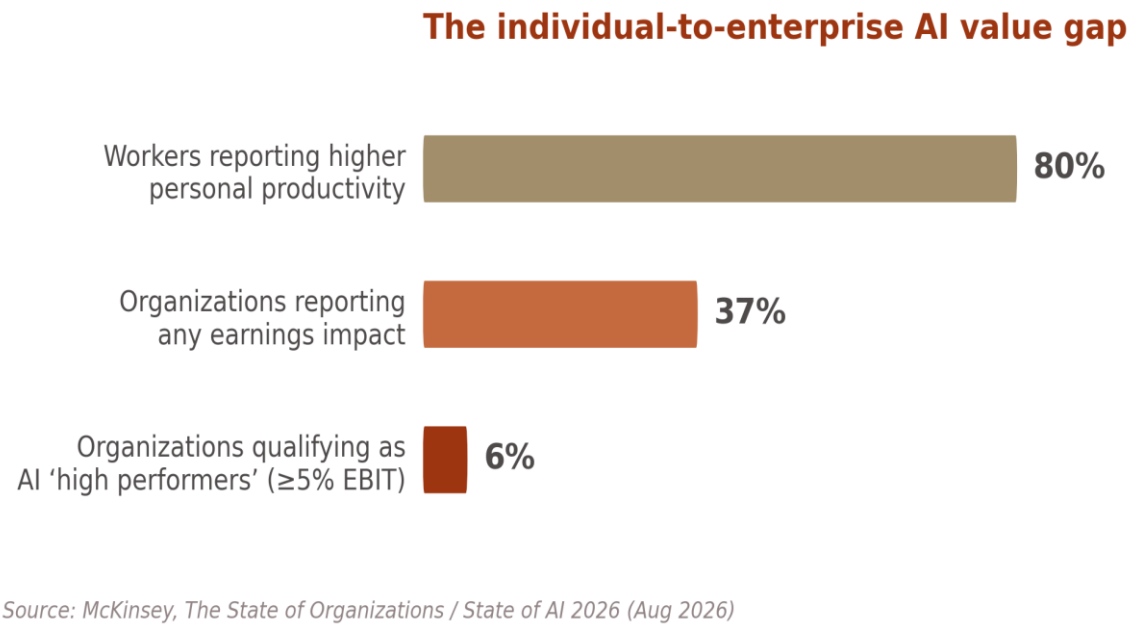
Source: Gartner survey of 782 I&O leaders, Nov–Dec 2025

Figure 1: Outcomes of enterprise AI initiatives

38% of leaders who reported setbacks pointed to a shortage of team expertise, and an equal share pointed to poor data quality or limited data access (not the underlying AI technology itself). Earlier MIT research (August 2025) indicated a worse outcome - roughly 95% of generative AI pilots failed to produce a measurable business outcome.

The Individual-to-Enterprise Value Gap

McKinsey's State of Organizations / State of AI research, published in the final days of August 2026, sharpens the picture further. Eighty percent of workers who use AI report a genuine improvement in their own productivity. However, only 37% of organizations report any enterprise-level earnings impact from AI at all, and just 6% qualify as AI 'high performers' with a measurable (≥5%) EBIT (i.e. earnings before interest and tax) contribution. This is shown in Figure 2. Ninety percent of function-level AI use cases remain stuck in pilot rather than scaled production. The benefit is visible at the desk but disappears somewhere on the way to the balance sheet.



Source: McKinsey, The State of Organizations / State of AI 2026 (Aug 2026)

Figure 2: The individual-to-enterprise AI value gap

Why AI/ML Projects Don't Fit Today's Size-Based Benchmarks (Yet)

This individual-to-enterprise gap is, at its core, a measurement problem — a familiar one to anyone who has worked in software metrics. Thirty years ago, organizations could not reliably compare the productivity of two development teams because they had no common, objective unit of output. Function point analysis solved exactly that

problem for traditional software. AI/ML application projects are in a strikingly similar position today.

COSMIC — one of the ISO/IEC-standardized functional sizing methods the ISBSG repository itself relies on — has set up a dedicated taskforce to define measurement procedures for sizing AI and algorithmic functionality. It ran workshops in 2024 with a further session planned for 2026. Its own assessment is candid: "a considerable amount of new, highly specialized knowledge must be developed" before AI software can be sized with the same rigor as conventional applications.

Neither IFPUG nor Nesma has yet published equivalent counting guidance, and the ISBSG Development & Enhancement questionnaire itself has no explicit 'AI/ML' category in its Application Type taxonomy. Projects submitted today cannot be reliably isolated for this kind of analysis.

Given that gap, this report does not claim to deliver a repository-verified productivity benchmark for AI/ML application projects. The data to do so does not yet exist in a taggable form. What ISBSG data does show, consistently, is that application types carrying high requirements uncertainty and iterative build cycles skew their effort distribution toward Specify and Design at the expense of Build. AI/ML projects share those characteristics by nature — problem framing, data sourcing and preparation, and defining acceptable model behavior all happen before a single prediction is ever shipped. This is a reasonable basis for expecting the same shift, even though it cannot yet be confirmed against a tagged AI/ML sample.

Key Findings

1. Outcome variability is high and well documented, but not explained: Gartner shows a wide spread between ROI-meeting, underperforming and failed AI initiatives. However, neither Gartner nor McKinsey ties these outcomes to objective delivery metrics such as size, effort or duration — the same blind spot ISBSG data eliminated from traditional software decades ago.

2. The productivity gain is real at the individual level and evaporates at the enterprise level: an 80-point gap between personal productivity gains and high-performer status at the organizational level points to a systemic delivery and scaling issue, not to the underlying technology.

3. Functional size measurement standards are actively being extended to AI but are not yet usable. COSMIC's own taskforce describes its guidance as still in the initial design phase. IFPUG and Nesma have not yet published AI-specific counting rules. ISBSG's Application Type taxonomy has no AI/ML category.

4. Early proxies exist, but only partially. ISBSG Application Type Grouping values such as Business Intelligence / Decision Support likely already contain AI/ML-

adjacent projects (recommendation logic, forecasting, scoring engines) submitted under a generic label. Until an explicit tag exists, any ISBSG-based benchmark of AI/ML projects will undercount and be sample-biased. However, it is the closest a repository-based, function-point-driven analysis can currently get.

Discussion and Implications

For estimators and benchmarkers, the practical takeaway is not to wait for a finished standard before treating AI/ML application projects as a distinct class. Even without an ISBSG-blessed 'AI/ML' tag, project teams can and should record functional size, effort by phase, and delivery duration for these projects exactly as they would for any other development effort. They should count what the AI capability does for the user (a prediction returned, a document classified, a recommendation displayed) rather than treating the model itself as unsizeable, and expect the effort split to lean further toward Specify and Design than a conventional application of the same size.

For the standards community, ISBSG, Nesma, IFPUG and COSMIC alike, the Gartner and McKinsey numbers make the case for urgency, not just a research agenda. We would recommend three concrete steps:

- That ISBSG add an explicit 'AI/ML embedded' flag and a small set of AI-specific effort categories (data preparation, model training and tuning, evaluation) to the Development & Enhancement Questionnaire well ahead of COSMIC's own guidance being finalized, so submitted projects can at least be identified even before they can be perfectly sized.

- That organizations already running AI/ML delivery contribute early case data to the COSMIC AI sizing taskforce so its procedures are grounded in real delivered projects rather than in theory alone.

- That once a critical mass of tagged projects accumulates, ISBSG publish the first cross-industry, function-point-based benchmark of AI/ML application delivery. The project-level counterpart to the enterprise-level numbers Gartner and McKinsey are already publishing.

For organizations running their own AI/ML initiatives today, the lesson from the 80%-to-6% gap is to instrument delivery now, in whatever standardized terms are available, rather than assessing AI value purely through anecdote or vendor-reported productivity claims. An organization that tracks its own AI/ML projects by functional size, effort and delivery speed, however roughly, will be able to say, with evidence, whether it is one of the 6% capturing real enterprise value or one of the 63% that is not, well before an industry-wide benchmark exists to tell it so.

Conclusions

This short report has made the case for benchmarking AI/ML application projects. This is software whose functionality is itself powered by AI or machine learning and it is separate from the AI-assisted development tooling covered in our February 2026 report.

The outcome is unambiguous: Gartner finds that only 28% of enterprise AI initiatives meet ROI expectations and 20% fail outright, while McKinsey's newest research shows an 80% individual productivity gain shrinking to just 6% of organizations achieving measurable enterprise-level impact.

What neither study can yet do is explain these outcomes in the objective, function-point-based terms ISBSG has for conventional software. This is because the functional sizing community itself is still developing the standards to do so. COSMIC's AI sizing taskforce is active but unfinished, and no ISBSG Application Type category yet exists for AI/ML projects. That gap is closing, not permanent.

As COSMIC's guidance matures, as IFPUG and Nesma follow, and as ISBSG extends its own data collection to tag AI/ML application projects explicitly, the ISBSG repository will be positioned to do for enterprise AI what it has done for every other software delivery paradigm before it. It will replace anecdote and vendor claims with an objective, auditable, industry-wide benchmark. Organizations that start measuring their own AI/ML delivery in these terms today will be ready to compare themselves against that benchmark the moment it exists.

If you wish to do your own analysis, or if you are interested to use the ISBSG data for cost estimation, benchmarking, performance measurement, procurement, etc., please subscribe to the data here: https://www.isbsg.org/project-data/

The International Software Benchmarking Standards Group (ISBSG)

The ISBSG is a not-for-profit organization founded in 1997 by a group of National Software Metrics Associations. Their aim was to promote the use of IT industry data to improve software processes and products.

ISBSG is an independent international organization that collects and provides industry data for software development projects and maintenance & support activities. It aims to help all organizations (commercial and government, suppliers and customers) in the software industry to understand and to improve their performance and decision making.

ISBSG sets the standards of software data collection, software data analysis and software project benchmarking processes and is considered to be the international thought leader in these practices.

The ISBSG mission is to support commercial and public organizations to improve the estimation, planning, control and management of IT software projects and/or maintenance and support contracts.

To achieve this, ISBSG maintains and grows 2 repositories of IT software development/maintenance & support data. This data originates from trusted, international IT organizations and can be obtained for a modest fee from the website www.isbsg.org/project-data/

Help us to collect data

ISBSG is always looking for new data. In return for your data submission, we issue a free benchmark report that shows the performance in your project or contract against relevant industry peers.

Please submit your data through one of the forms listed on https://isbsg.org/submit-data/

A specific Agile/Scrum data collections questionnaire can be downloaded here: https://cutt.ly/4vnuXVT

Partners

This page will help you to find an ISBSG partner in your country: https://www.isbsg.org/board/