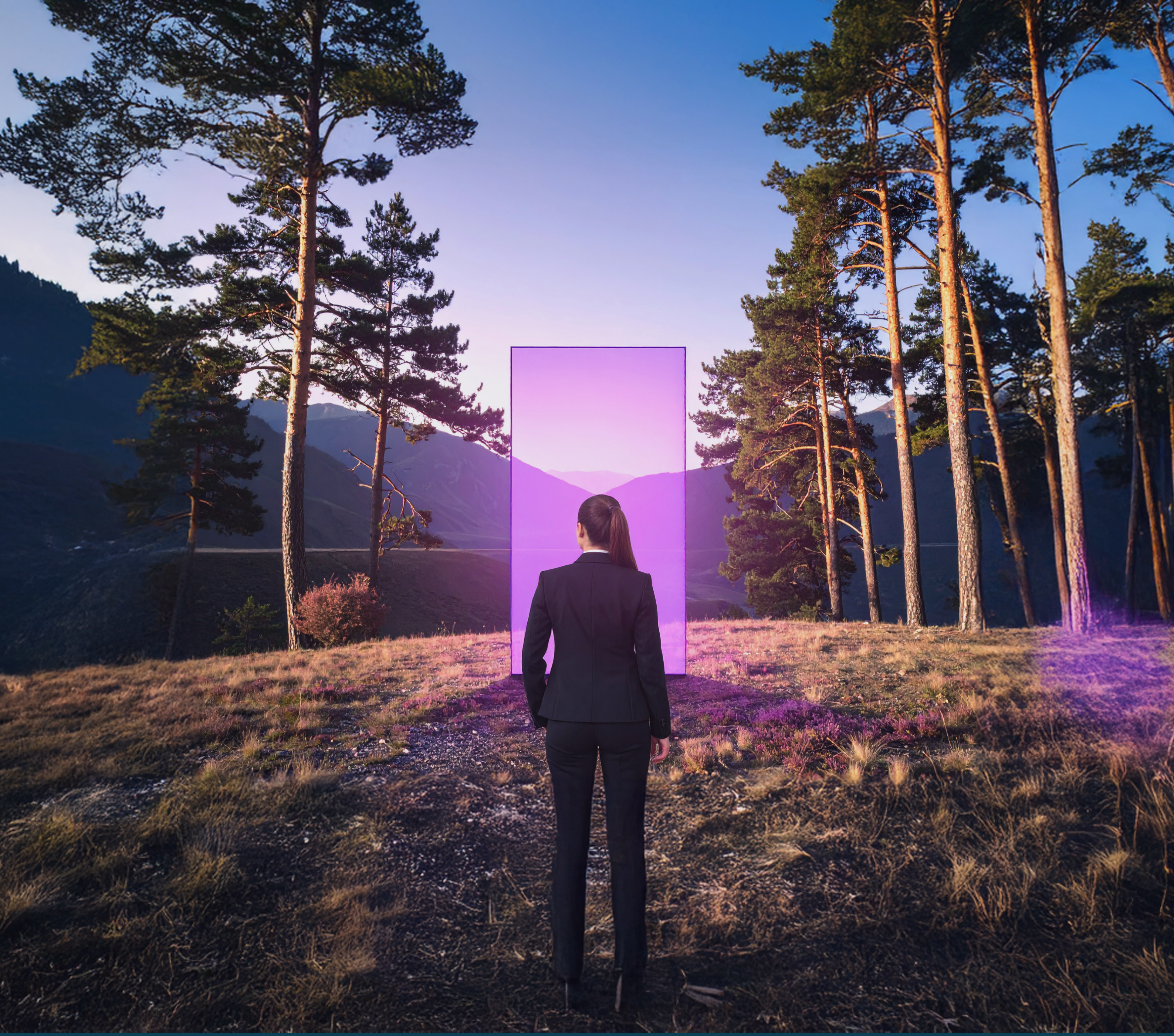


Scaling GenAI with confidence

How organizations can move beyond pilots, measure what matters, and turn innovation into lasting value.

thoughtworks
Design. Engineering. AI.

November 2025

Introduction

Generative AI is moving rapidly from experimentation to mainstream ambition, yet many organizations still struggle to turn pilots into enterprise-wide impact. Almost two-thirds lack a clear strategy for assessing reliability and moving GenAI into production, while others find it difficult to measure business value or adapt models to their needs. Without solid foundations, initiatives often remain confined to isolated use-cases, limiting their potential.

At the same time, the rewards of scaling responsibly are becoming clearer - from lower operational costs and improved resilience to faster deployment cycles and higher customer engagement. Case studies, such as work Thoughtworks delivered with a major airport operator and a leading digital property platform, show how clear pipelines, feedback loops and a focus on operational excellence can translate AI innovation into measurable business gains.

This executive summary, the latest in our ongoing series, explores how organizations can move from AI pilots to production at scale, highlighting the gaps that hold them back and the practices that turn innovation into lasting value.

Insight one:
The strategic deficit

Generative AI is moving rapidly from early trials into the fabric of business operations, with sixty percent saying that measuring the reliability of GenAI applications and successfully moving them into production is vital to their organization's overall success - yet many still struggle to turn ambition into impact.

Nearly two thirds (64%) do not have a fully developed and optimized strategy for measuring the reliability of GenAI applications and models, nor for successfully transitioning them into production environments. Without that foundation, promising ideas risk stalling before they deliver meaningful results.

A major obstacle lies in understanding and measuring impact. Around a third (34%) of organizations say they struggle to measure the impact of their GenAI applications, while a similar number report challenges tailoring models to domain-specific needs (38%) or meeting evolving regulatory requirements (33%). These gaps leave outputs at risk of being irrelevant or non-compliant, eroding confidence in scaled adoption.

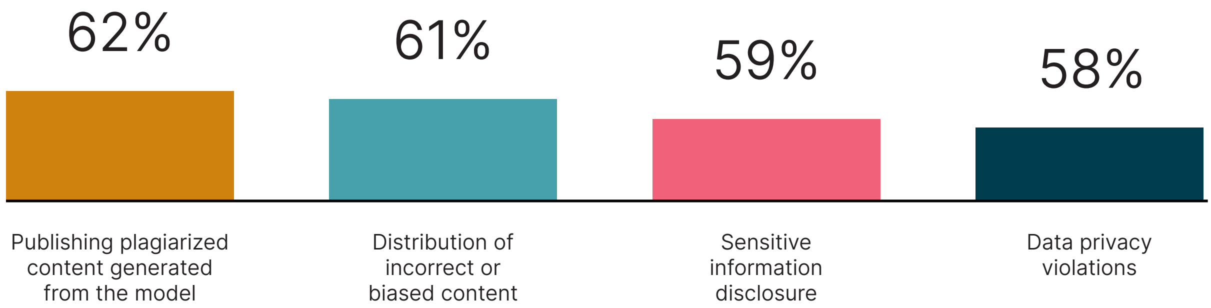
Much of today's AI use focuses on isolated, individual tasks. While helpful, these incremental gains rarely transform how teams, processes, and value streams operate – a potential reason why many efforts stall before they achieve meaningful business impacts. Unlocking larger rewards means shifting to coordinated deployment, underpinned by clear evaluation and governance that connect AI to enterprise priorities.

Closing this deficit requires an organization-wide approach: embedding robust measurement, aligning initiatives with business goals, and building trust in responsible deployment. With those elements in place, enterprises can move beyond pilots and turn GenAI into a sustained source of competitive advantage.

Insight two:
Risk, regulation and the road to trust

As organizations push GenAI into mission-critical workflows, concerns about security, bias, privacy and intellectual property become clear.

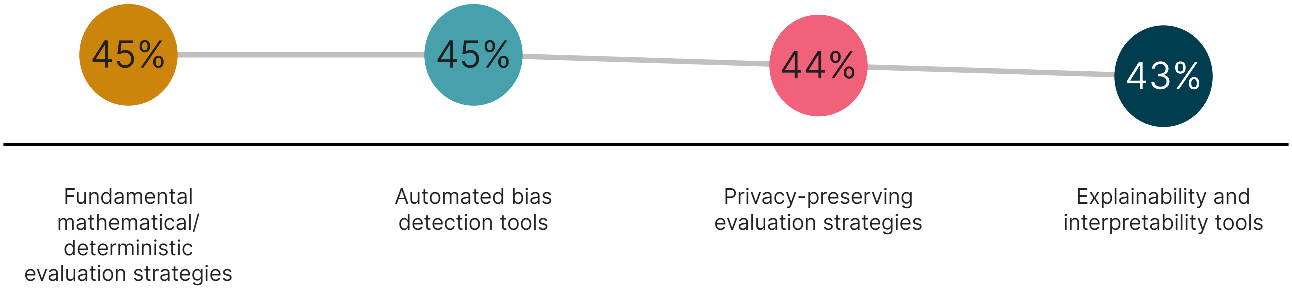
Concerns deploying large language models:



Safeguards such as audits, security reviews, monitoring and independent evaluations can help manage these risks, but may not keep pace with the speed and complexity of GenAI. Clearer, more transparent ways to understand how models behave, why they produce certain results and where vulnerabilities lie will be essential as use expands.

Attention is therefore shifting toward more advanced evaluation methods. Approaches that detect bias, preserve privacy and explain model decisions can strengthen accountability and align systems with emerging regulatory expectations. Innovation in these areas - from deterministic testing to bias detection, privacy-preserving techniques and interpretability tools - will be key to building confidence and scaling GenAI responsibly.

Areas where GenAI/LLM evaluations could benefit from further innovation:



Ultimately, trust is the decisive factor in moving from pilots to enterprise-wide deployment. Embedding evaluation and governance into operating models ensures security, fairness and data protection are treated as core measures of success, enabling organizations to unlock GenAI's potential with confidence.

Insight three:
Execution at a crossroads

Many organizations are still struggling to turn GenAI ambition into sustainable value. Businesses are currently more likely to report a negative return (43%) than a positive one (40%) on efforts to evaluate and move GenAI into production.

Encouragingly, there is clear recognition that current approaches need to evolve. Nearly three-quarters of decision makers (72%) believe their organization's efforts to measure and deploy GenAI requires a significant overhaul or major improvements. This self-awareness could signal readiness for a more disciplined path - one built on clear objectives, robust evaluation and governance that links AI initiatives to tangible outcomes.

Thoughtworks worked with a leading online real estate platform and showed what structured execution looks like in practice. By embedding robust pipelines, automated testing and rapid feedback, the organization deployed machine-learning models up to five times faster and with greater reliability.

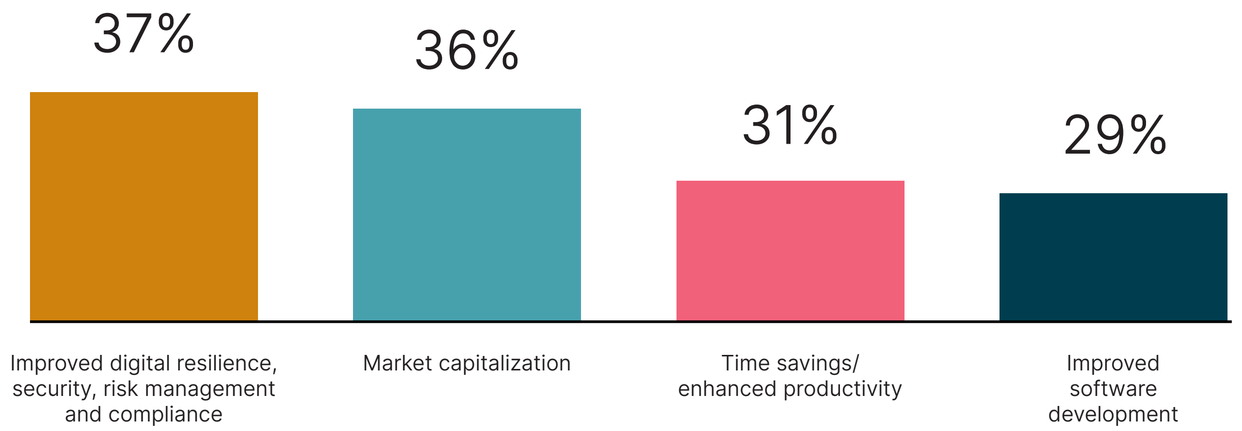
Insight four:
Measuring what matters

Although many organizations are still working through the challenges of scaling GenAI, others are beginning to unlock its potential - not by moving faster, but by focusing on how they measure success.

Consistent, well-structured measurement gives teams the clarity to move solutions from experimentation into production, ensuring innovation delivers sustained value. It also helps to pinpoint where GenAI can extend impact beyond individual users - strengthening team performance, streamlining processes and unlocking value across entire workflows.

A clear example comes from work Thoughtworks delivered with a major airport operator, where applying AI alongside robust evaluation significantly reduced operational costs and improved day-to-day efficiency.

Organizations have seen the most value as a result of measuring the reliability of GenAI applications and models through:



Measurement shouldn't be seen as a box-ticking exercise or technical chore. When organizations build clear evaluation into the way GenAI is delivered, they start to see tangible benefits: stronger resilience and risk management, meaningful productivity gains, smoother software development and even growth in market value. Framing these outcomes as part of a wider strategy turns measurement into a genuine source of competitive advantage, helping businesses scale GenAI with confidence and translate innovation into lasting value.

Conclusion:

Scaling AI is essential for business success, with sixty percent of organizations recognising that measuring the reliability of GenAI applications and successfully moving them into production is vital to their overall success. Yet many still struggle to turn ambition into impact. Without a clear framework for measurement, deployment and risk management, even the most promising pilots struggle to deliver sustainable value.

Experience from organizations already putting GenAI into production shows what is possible. When clear objectives, disciplined evaluation and robust delivery are in place, GenAI can achieve tangible results: lower operational costs, greater resilience, faster deployment and stronger customer engagement. Success depends on moving beyond isolated experiments to enterprise-wide capability, supported by reliable tools and well-defined operating models.

Establishing these foundations will shape the next phase of AI adoption. Organizations that align strategy, governance and measurement – leveraging GenAI within core business processes - will be best placed to scale with confidence and turn innovation into lasting advantage.

About the research

We spoke to 1,000 senior decision-makers from organizations across five key markets: US (300), Australia (175), Singapore (175), UK (175) and Germany (175). 25% were C-suite executives and 75% were senior executives at Director, VP, and ‘Head of’ levels.

Organizations had an annual global revenue between $500 million to over $50 billion and an average annual IT budget of $390 million. They were from a range of industries, including: Public Sector (excluding Healthcare); Automotive and Manufacturing; Technology and Business Services; Healthcare and Life Sciences; Retail and Consumer Goods; Banking, Financial Services and Insurance; Energy and Utilities; and Travel and Transportation.

Please note, the stats we refer to in this document are about measuring the reliability of GenAI applications and models, and successfully moving them into production. We use shortened wording or ‘Scaling AI’ as an umbrella phrase where appropriate.

About Thoughtworks

We are a global technology consultancy that delivers extraordinary impact by blending design, engineering and AI expertise.

For over 30 years, our culture of innovation and technology excellence has helped clients strengthen their enterprise systems, scale with agility and create seamless digital experiences.

We're dedicated to solving our clients' most critical challenges, combining AI and human ingenuity to turn their ambitious ideas into reality.

www.thoughtworks.com

About Vanson Bourne

Vanson Bourne is an independent specialist in market research for the technology sector. Their reputation for robust and credible research-based analysis is founded upon rigorous research principles and their ability to seek the opinions of senior decision makers across technical and business functions, in all business sectors and all major markets.

www.vansonbourne.com

Click here for the full report.

Or visit https://www.thoughtworks.com/insights/reports/readiness-index

/thoughtworks
Design. Engineering. AI.