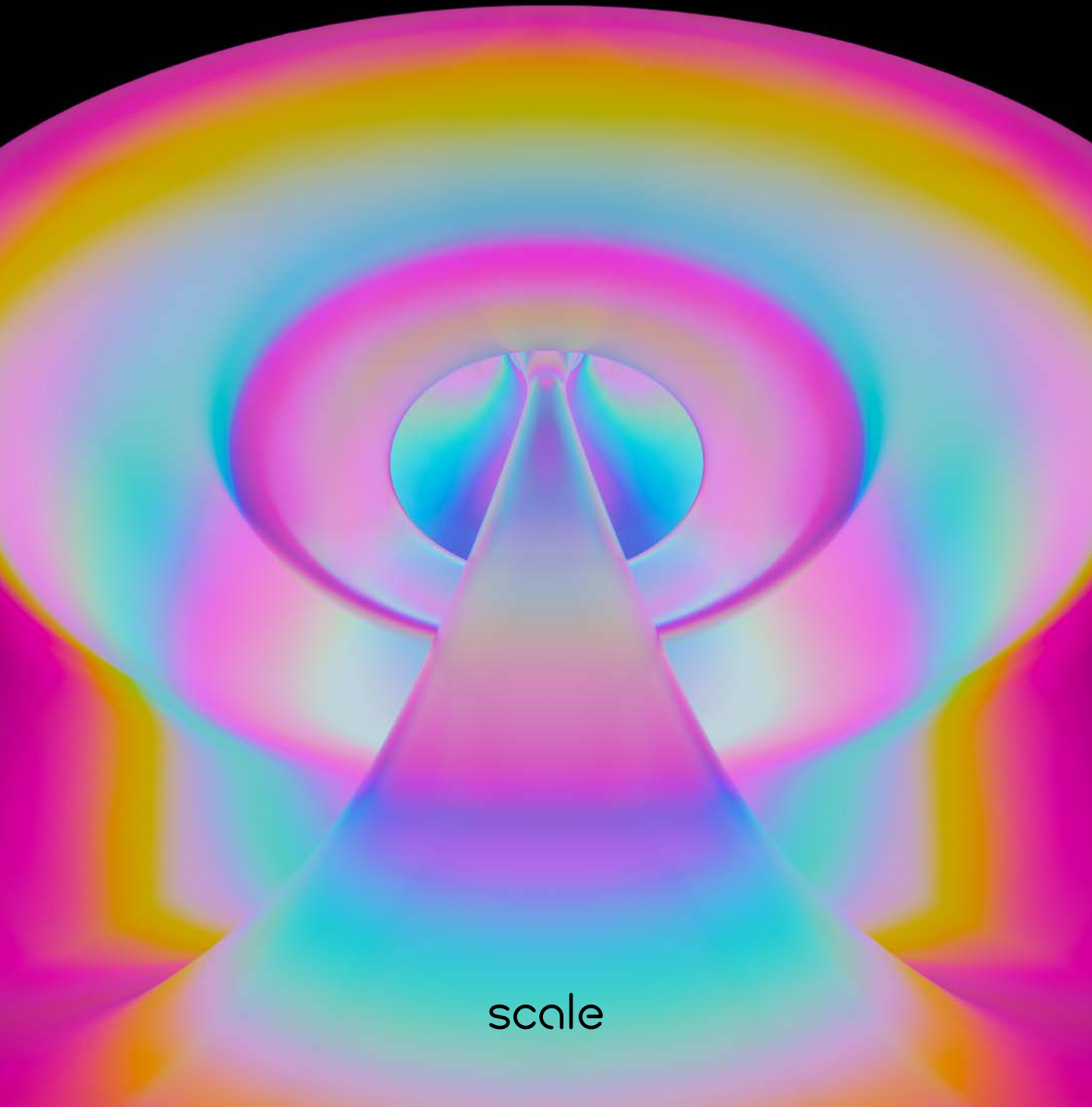


AI READINESS REPORT 2024

zeitgeist



scale

Table of Contents

AI Year in Review	4
Apply AI	13
Build AI	26
Evaluate AI	36
Conclusion	46
Methodology	47

Introduction

The hype for generative AI has reached its peak. Developers continue to push the limits, exploring new frontiers with increasingly sophisticated models. At the same time, without a standardized blueprint, enterprises and governments are grappling with the risks vs. rewards that come with adopting AI.

That’s why in our third edition of Scale Zeitgeist: AI Readiness Report, we focused on what it takes to transition from merely adopting AI to actively optimizing and evaluating it. To understand the state of AI development and adoption today, we surveyed more than 1,800 ML practitioners and leaders directly involved in building or applying AI solutions and interviewed dozens more. In other words, we removed responses from business leaders or executives who are not equipped to know or understand the challenges of AI adoption first-hand.

Our findings show that of the 60% of respondents who have not yet adopted AI, security concerns and lack of expertise were the top two reasons holding them back. This finding seems to validate the “AI safety” narrative that dominates today’s news. Among survey respondents who have adopted AI, many feel they lack the appropriate benchmarks to effectively evaluate models. Specifically, 48% of respondents referenced

lacking security benchmarks, and 50% desired industry-specific benchmarks. Additionally, while 79% of respondents cited improving operational efficiency as the key reason for adopting AI, only half are measuring the business impact of their AI initiatives. And while performance and reliability (each at 69%) were indicated as the top reasons for evaluating models, safety ranked lower (55%), running counter to popular narratives.

This report presents expert insights from Scale and its partners across the ecosystem, including frontier AI companies, enterprises, and governments. Whether you are developing your own models (building AI), leveraging existing foundation models (applying AI), or testing models (evaluating AI), there are actionable insights and best practices for everyone.

“The rapid evolution of AI offers both immense opportunities and challenges. Embracing it responsibly, with robust infrastructure and rigorous evaluation protocols, unlocks the potential of AI while safeguarding against the risks, known and unknown.”

Alexandr Wang,
FOUNDER & CEO, SCALE

Year in Review

Generative AI continues to reshape our world

Advancements in generative AI continued to accelerate in 2023. After the release of OpenAI's ChatGPT in November 2022, the platform reached an estimated 100 million users in just two months. In March 2023, OpenAI released GPT-4, a large language multimodal model that demonstrated human-level performance across industry benchmarks.

Other model builders joined the launch party last year. Google launched Bard, initially running on the LaMDA model and replaced shortly after by PaLM 2 (with improved domain-specific knowledge - including coding and math). Anthropic introduced Claude 2 in the summer with a 100K context window. A week later, Meta unveiled Llama 2 and Code Llama, and included model weights and code for the pretrained model.

Google DeepMind closed out 2023 with the release of Gemini - representing a significant improvement in performance as the first model to outperform human experts on the Massive Multitask Language Understanding (MMLU) test. Newer open source model families like Falcon, Mixtral, and DBRX demonstrated the possibility for local inference while innovating on model architecture to use far less compute. This year, in March 2024, Anthropic launched the family of Claude 3 models, doubling the context window. Just a few days later, Cohere released their Command R generative model - designed for scalability and long context tasks.

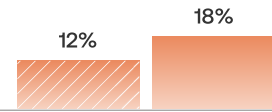
Frontier research underlies many of these model advancements. Some significant advancements include:

1. Open AI achieved improvements in mathematical reasoning through rewarding chain-of-thought reasoning. Scale contributed to the creation of PRM800K, the full process supervision dataset released as part of this paper.
2. Anthropic uncovered an approach for better model interpretability through analysis of feature activation compared to individual neurons.
3. The Microsoft Research team discovered that a model with a smaller number of parameters relative to state-of-the-art models can demonstrate impressive performance on task-specific benchmarks when fine-tuned with high-quality textbook data.

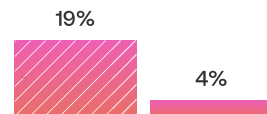
Key findings, 2023 to 2024

To illustrate the evolving landscape, we see the following changes as important trends in AI over the past year.

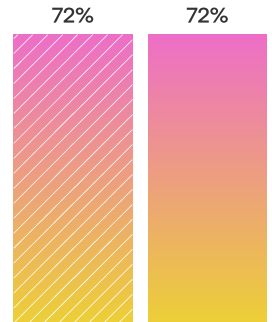
Organizations reporting generative AI forced the creation of an AI strategy:



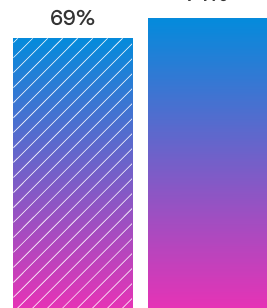
Organizations with no plans to work with generative AI:



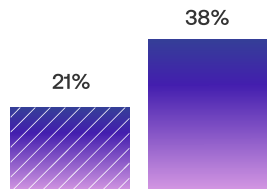
Organizations planning to increase investment in commercial and closed-source models over the next three years:



Organizations that consider AI to be very or highly critical to their business in the next three years:

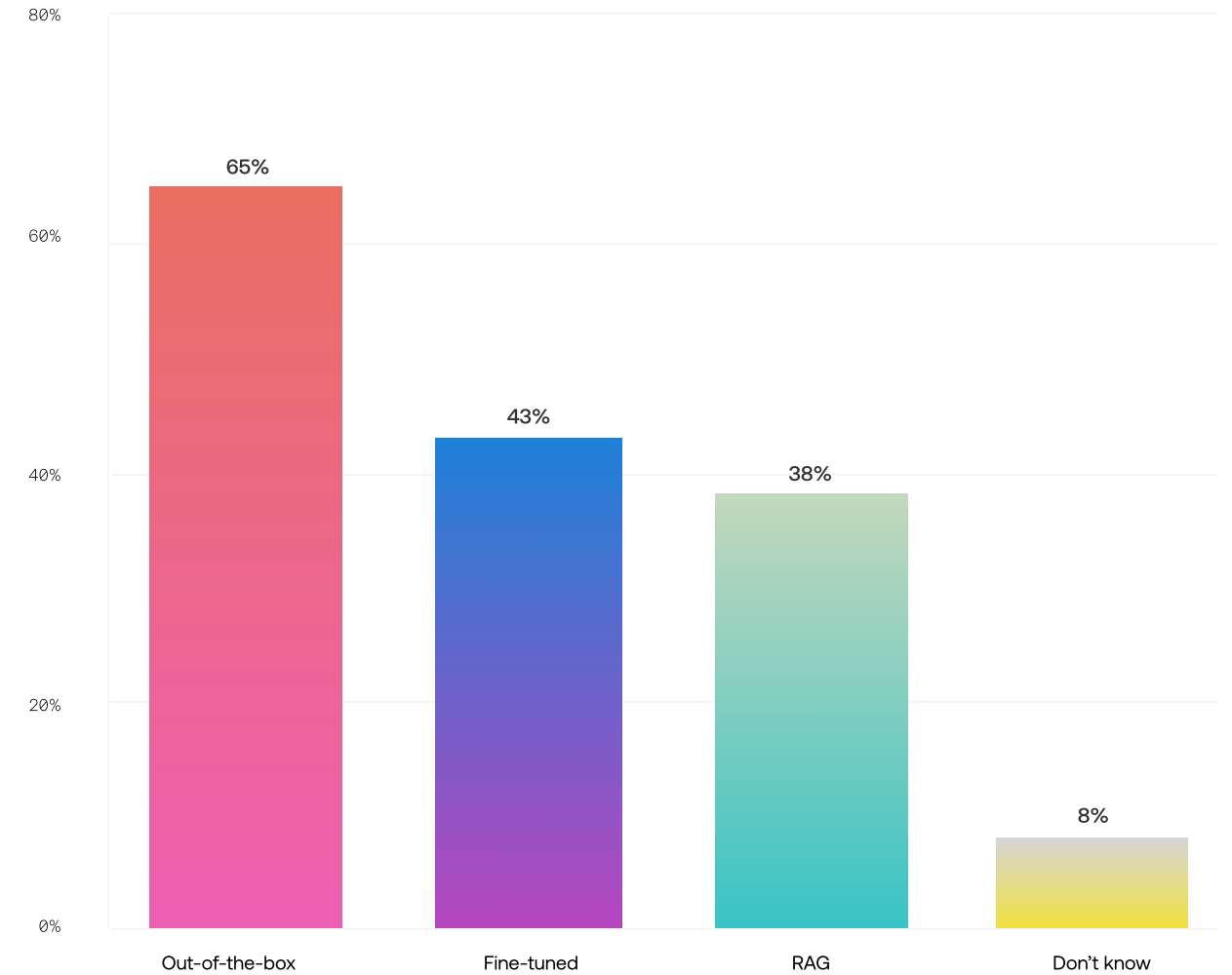


Organizations with generative AI models in production:



2023 2024

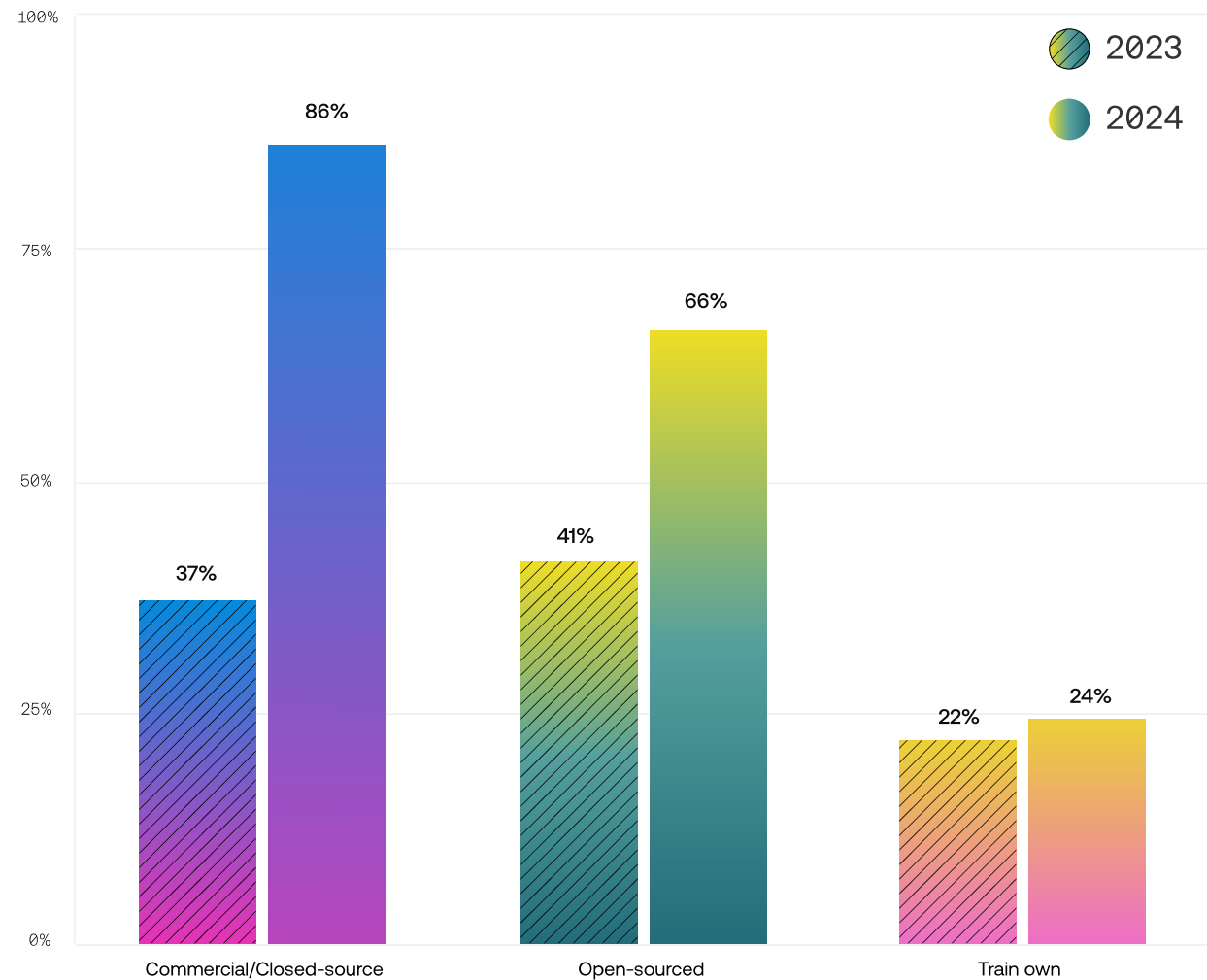
Do you customize generative AI models or use them out of the box?



Organizations applying AI are seeking to extract additional value by optimizing AI through prompt-engineering, fine-tuning models, and retrieval augmented generation (RAG). Despite the desire to optimize foundational models, 65% of organizations use models out-of-the-box, 43% of organizations fine-tune models and 38% use RAG. Fine-tuning can customize models for

specific tasks or datasets, significantly enhancing their performance and accuracy on targeted applications. RAG further enhances this by dynamically incorporating external information during the generation process, enabling the model to produce more relevant and contextually rich outputs.

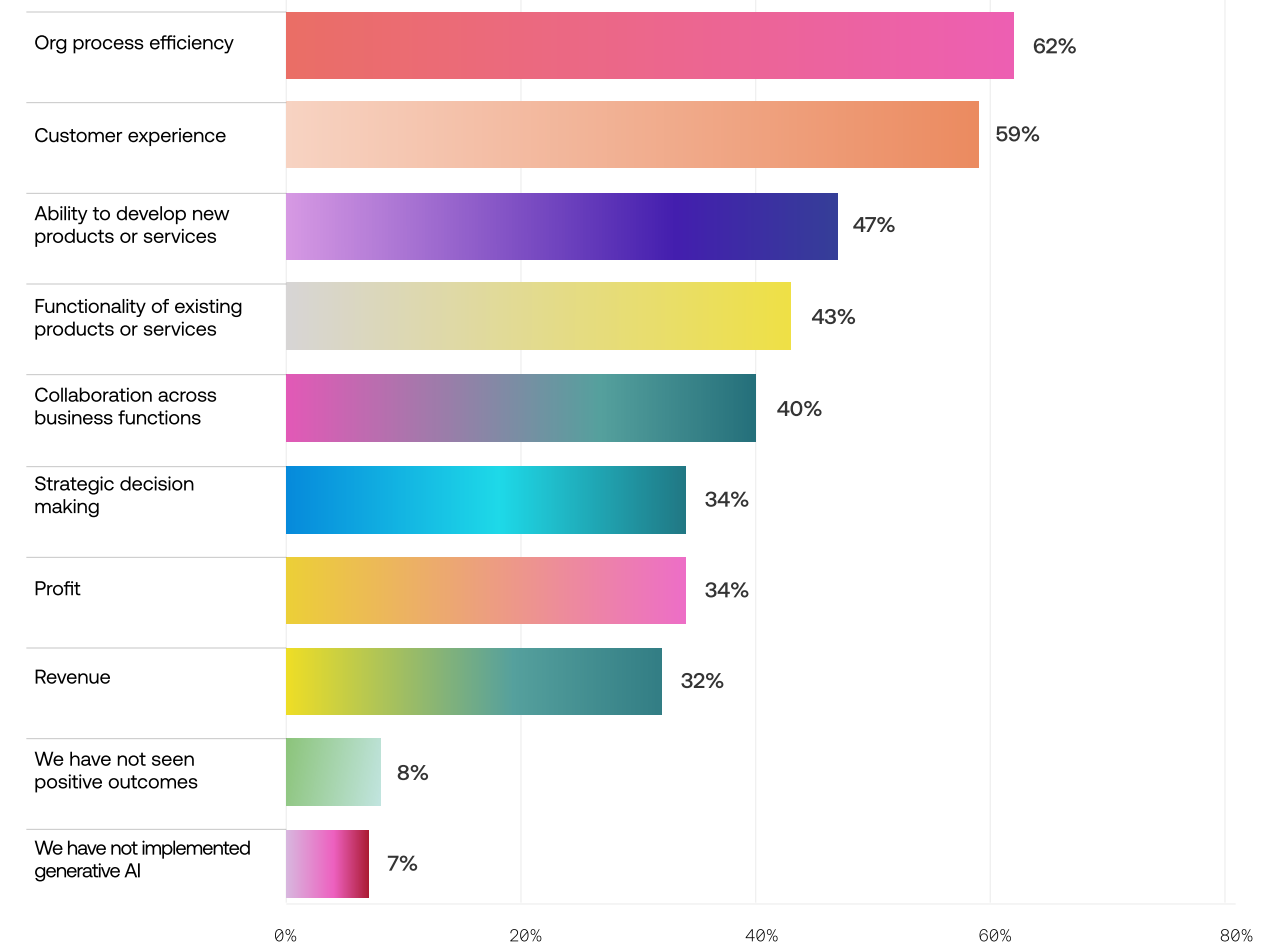
How do you work with generative AI models?



Model preferences continue to evolve and remain a key decision for an organization's AI strategy. The largest increase in usage came from closed-source models with 86% of organizations using these models compared to 37% the year prior. This is likely due to a combination of factors. Many organizations have existing contracts with cloud service providers who in turn have partnerships with closed-source model developers, making

usage of closed-source models easier. Many closed-source models also outperform open source models out-of-the-box. Despite that, open-source model usage still increased from 41% to 66%. This is likely due to the flexibility open-source models provide for fine-tuning and hosting. The smallest change in model preferences were organizations that trained their own models at 24% in 2024.

What positive outcomes have you seen from generative AI adoption?



Similar to last year, 61% of organizations stated improved operational efficiency as the leading driver behind adopting generative AI. Improved customer experience came in second at 55%.

Despite growing adoption, there are still a number of challenges that stall widespread use of generative AI. 61% of respondents cited infrastructure, tooling, or out-of-the-box solutions not meeting their specific needs.

Processes like RAG and fine-tuning introduce the complexity of integrating external data sources in real-time, ensuring the relevance and accuracy of retrieved information, managing additional computational costs, and addressing potential biases or errors. Fine-tuning requires careful selection of data to avoid overfitting and ensuring models remain generalizable to new, unseen information.

Proprietary data is a key ingredient to power performance enhancements for generative AI models. While Scale's machine learning team proved how fine-tuning [can enhance model capabilities](#), 41% of organizations lack the ML expertise to execute the data transformations and measure and evaluate results to justify the initial investment.

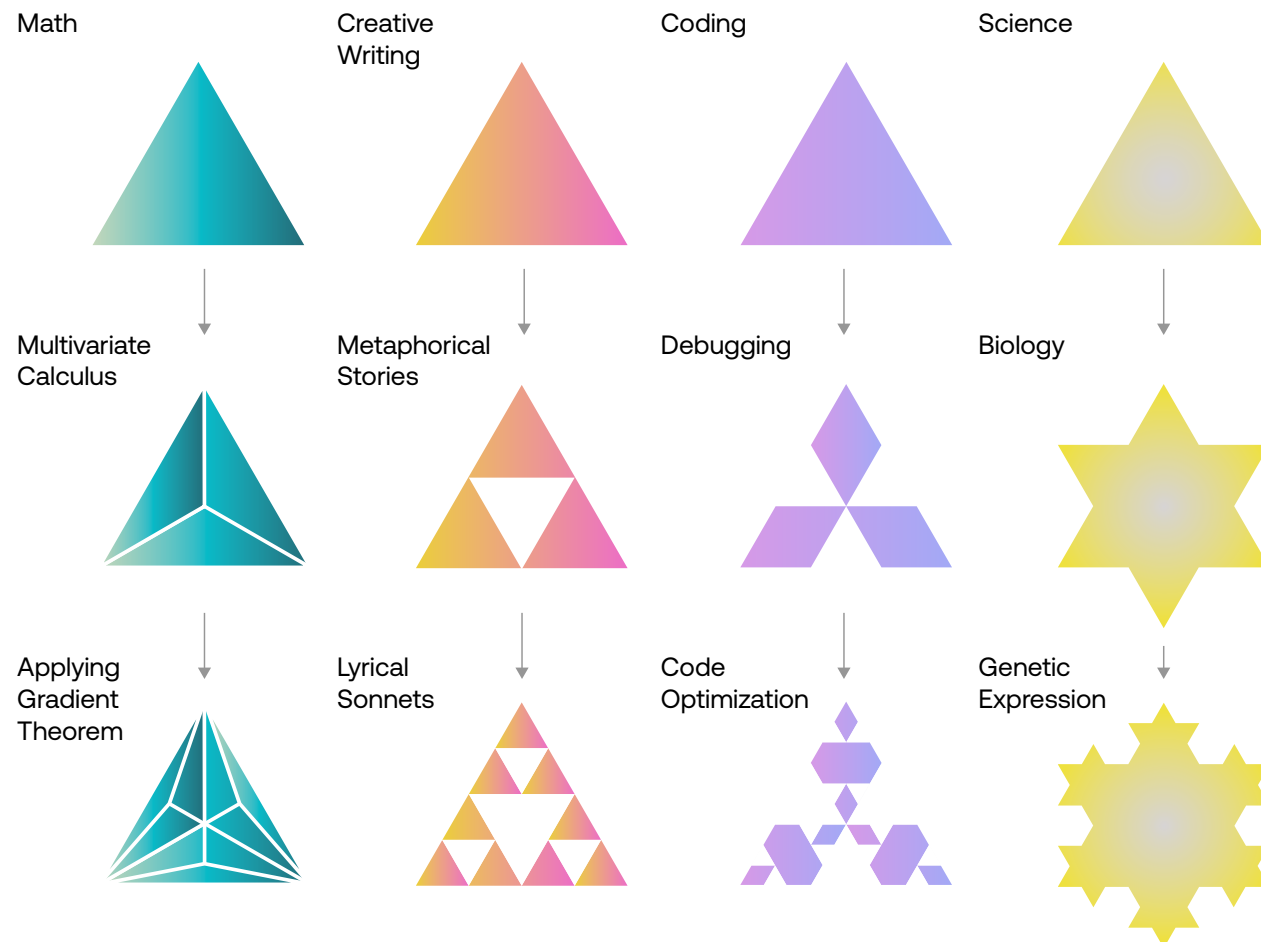
What to Expect in 2024

Increasingly Capable Foundation Models

In the coming year, we expect notable advancements in generative AI foundation models to continue. Models like Claude 3 have demonstrated improved performance on various benchmarks, such as scoring 86.8% on the MMLU dataset and 95.0% on the GSM8K math problem set, indicating enhanced capabilities in reasoning and problem-solving. We also expect to see the emergence of more sophisticated multimodal models that can seamlessly integrate and generate content across various modalities, including text, images, audio, and video as both inputs and outputs.

As researchers continue to refine these models, we can also anticipate improvements in accuracy and reduced latency, making models more reliable and efficient. The size of these foundation models is also likely to grow, allowing them to capture and leverage even more knowledge and nuance from the vast amounts of data they are trained on.

Evolution of generative AI capabilities: domain and functional capabilities are rapidly growing



Expert Insight Will Power Performance Improvements

Human experts will play an increasingly crucial role in model advancements and evaluation. As models start to exhaust the corpus of general information widely available on the internet, models will require additional data to improve their capabilities. While some organizations may look to replace human-generated data with synthetic data for training, [models reliant on synthetic data can be susceptible to model collapse](#). A hybrid human and synthetic data approach can mitigate biases from synthetic data and still reflect nuanced human preferences. The domain-specific knowledge of experts allows them to provide data that captures the nuance, complexity, and diversity to supplement model training. Experts are also critical for testing and evaluation alongside reinforcement learning from human feedback, with the knowledge to identify subtle errors, inconsistencies, or biases in order to provide reliable guidance to preferred model outputs.

While experts are necessary to improve model capabilities, we anticipate organizations defining new roles that are centered around generative AI. Prompt engineers, machine learning researchers, and generative AI experts will collaborate with subject matter experts to ensure AI initiatives are successful. Generative AI will fundamentally change the nature of work.

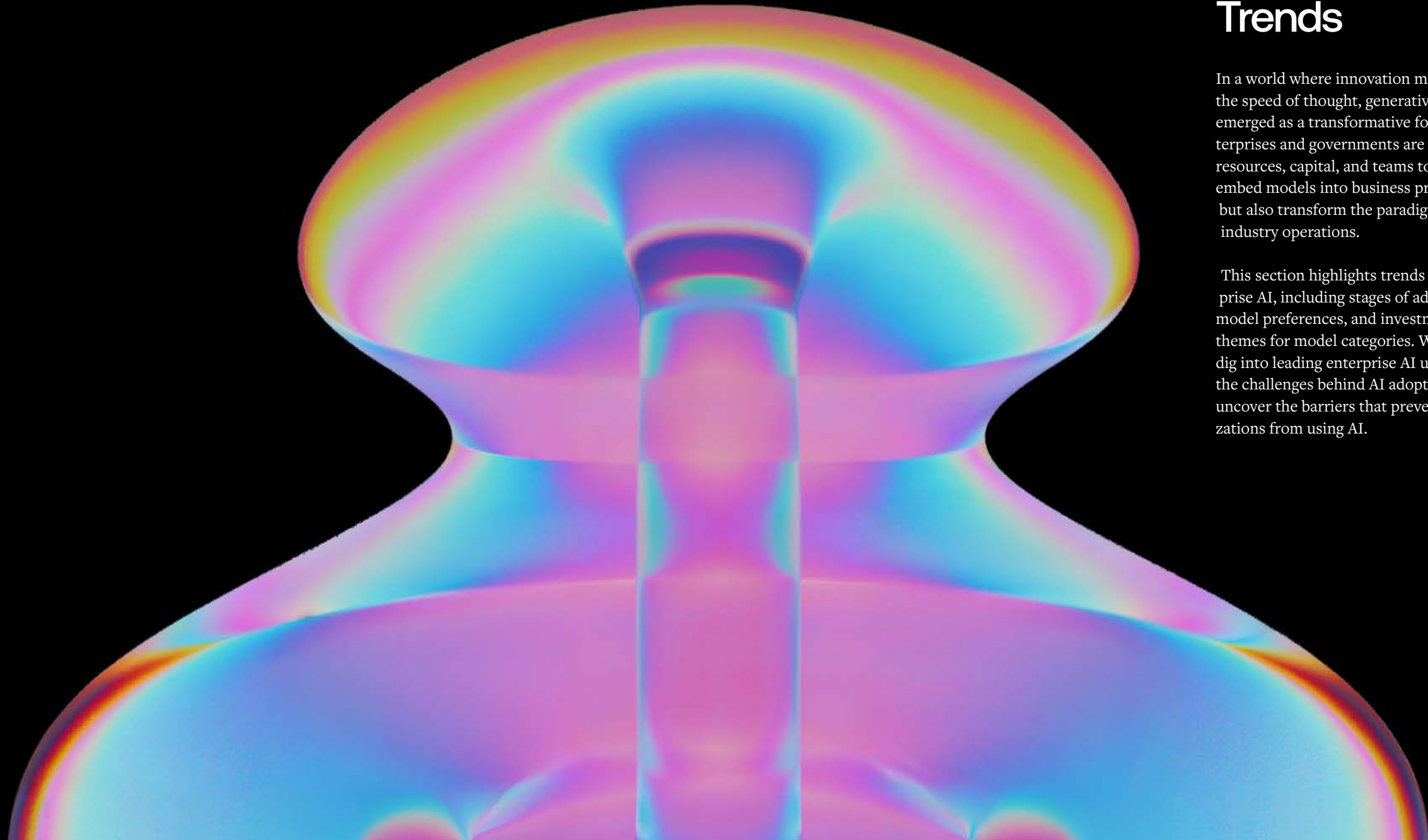
Evolving Proof-of-Concepts to Scaling Production Deployments

Improvements in model performance and capabilities will motivate leaders to quickly iterate from proof-of-concepts to pilots to production deployments. More user friendly RAG and fine-tuning solutions will emerge as on-ramps to improve adoption so that organizations can more easily customize models. As start up costs taper, model effectiveness improves, and more robust evaluation strategies emerge, organizations will be able to more clearly capture and define return on investment.

Increasing Emphasis on Test & Evaluation Practices

Nearly every major model release usurps a different leading model on various benchmarks. Enterprises will want to create their own evaluation methodology consisting of industry benchmarks, automated model metrics, and measures for return on investment to continuously evaluate their preferred model. As model capabilities grow, model builders will place more importance on guardrails, steerability, safety, security, and transparency. Public sector institutions now must consider the [White House's OMB Policy](#) and test and evaluate AI systems to ensure that AI is safe.

Apply AI



Adoption Trends

In a world where innovation moves at the speed of thought, generative AI has emerged as a transformative force. Enterprises and governments are deploying resources, capital, and teams to not just embed models into business processes, but also transform the paradigm of industry operations.

This section highlights trends in enterprise AI, including stages of adoption, model preferences, and investment themes for model categories. We'll also dig into leading enterprise AI use-cases, the challenges behind AI adoption, and uncover the barriers that prevent organizations from using AI.

What is the current stage of your AI/ML project?

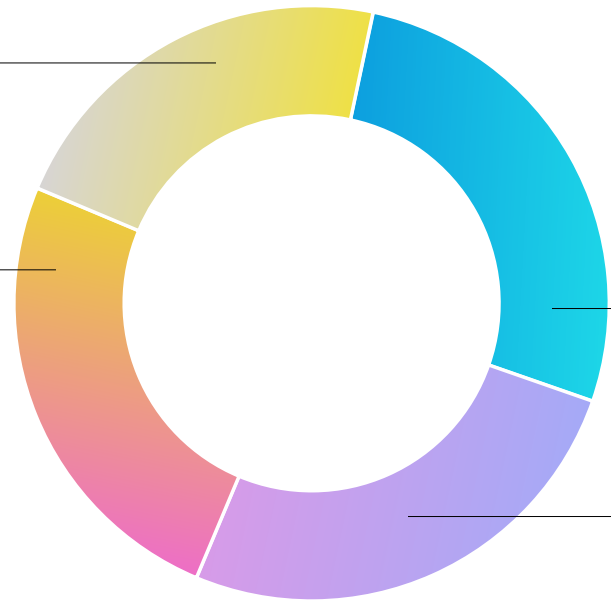
No model deployed to production

25%

Evaluating use cases

26%

Developing the first model/application



One or more models deployed

22%

One model/application deployed to production

27%

Multiple models/applications deployed to production

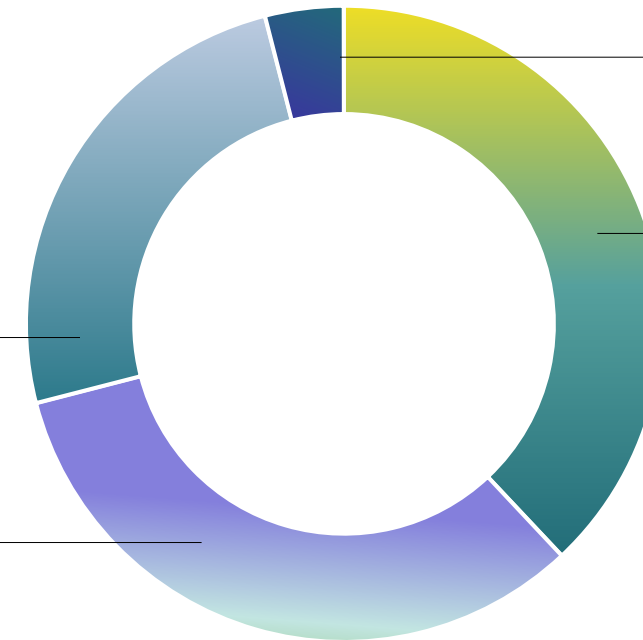
Which of the following describes how your company works with generative AI models?

25%

Plan on working with generative AI models

33%

Experimented with generative AI models



4%

No plans to work with generative AI models

38%

Generative AI models in production

The Evolution of AI Adoption

22% of organizations have one model in production with 27% of total respondents reporting multiple models in production. Deploying multiple generative AI models in production allows organizations to leverage specialized capabilities, avoid vendor lock-in, and scale multiple use-cases. By comparing performance across models and maintaining flexibility,

businesses can adapt to evolving requirements while mitigating risks associated with relying on a single model. The growing number of models in production reflects the progression of proof-of-concepts to production deployments.

49% of organizations are still either evaluating use cases or developing the first model or application.

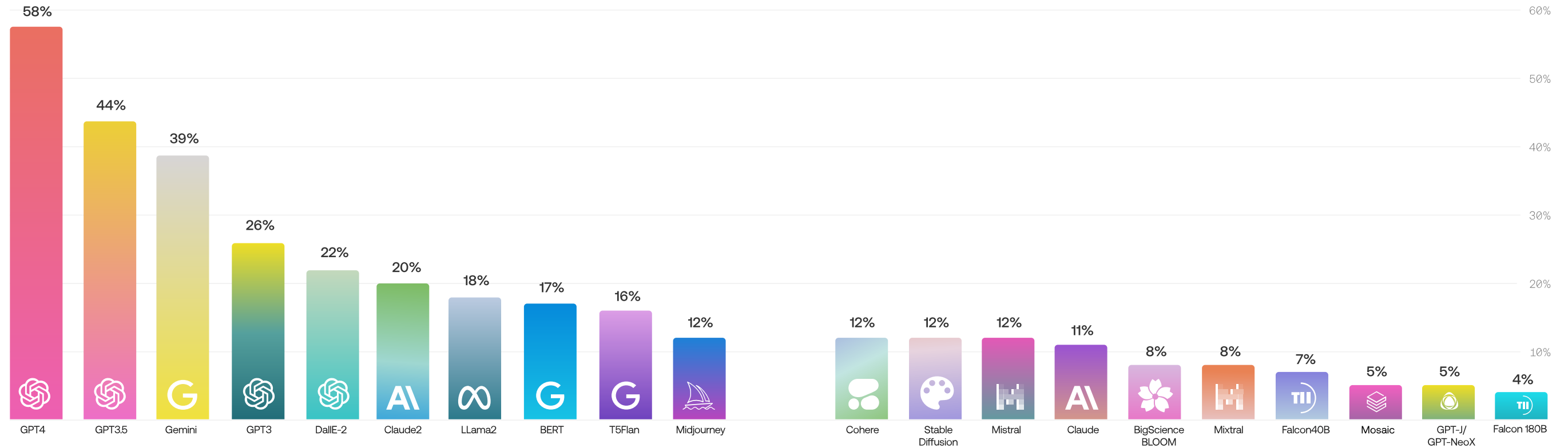
Many organizations are increasingly dedicating time to evaluating use cases to ensure alignment with business objectives. Thorough use case evaluation allows companies to identify applications with high ROI potential, assess feasibility and risk, and prioritize implementation efforts.

Application and model development follows use case selection. Deploying generative AI in an enterprise setting involves a multi-step process, including data preparation and pre-processing, model selection and architecture design, hyperparameter tuning and training, API development for integration, monitoring feedback, and test and evaluation.

Technical organizations are ahead of the curve with generative AI adoption. Software and internet companies are leading the pack with 48% of organizations reporting generative AI models in production. Conversely, only 24% of government and defense entities have generative AI models in production.

Which generative AI models do you work with?

Note - at the time of the survey, Claude 3, Grok, and DallE3 were not released and thus not included in the survey.



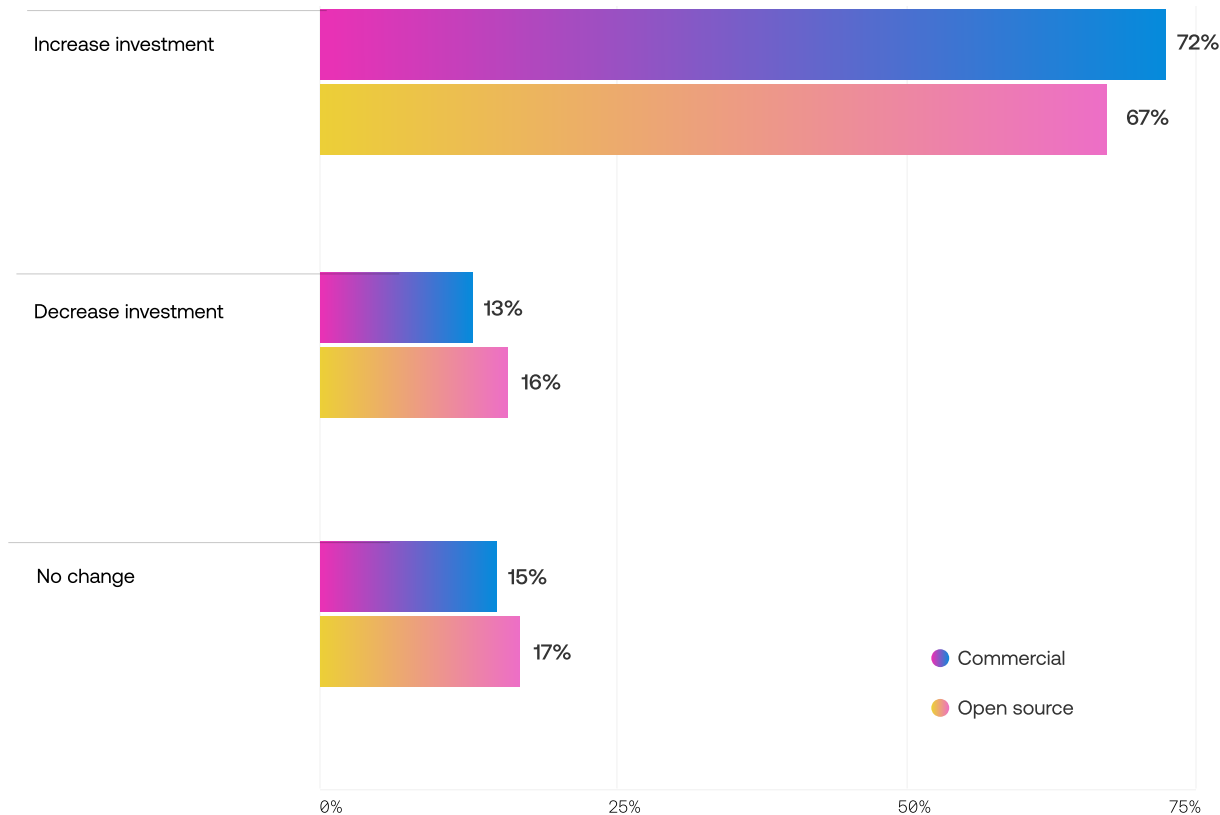
Model Preferences

Model selection is critical for generative AI development, as it determines the system's performance, scalability, and alignment with specific task requirements, data characteristics, computational resources, and trade-offs between model complexity and inference speed. Organizations also evaluate model selection through cost trade-offs - comparing investments tied to

infrastructure, managed services, and per token inputs and outputs. OpenAI is overwhelmingly the preferred model vendor. Virality and the ongoing rollout of advanced features positioned OpenAI as the preferred model vendor even as other models demonstrate comparable performance.

Our respondents indicate that their preferred model is the latest version of OpenAI GPT- 4 with 58% of enterprises using the latest version and 44% of enterprises using GPT-3.5. Trailing closely behind, 39% of enterprises use Google Gemini. There's a notable drop-off in model selection following these three models with OpenAI GPT-3 at 26%.

How does your company plan on investing in generative AI over the next 3 years?



Model Investment

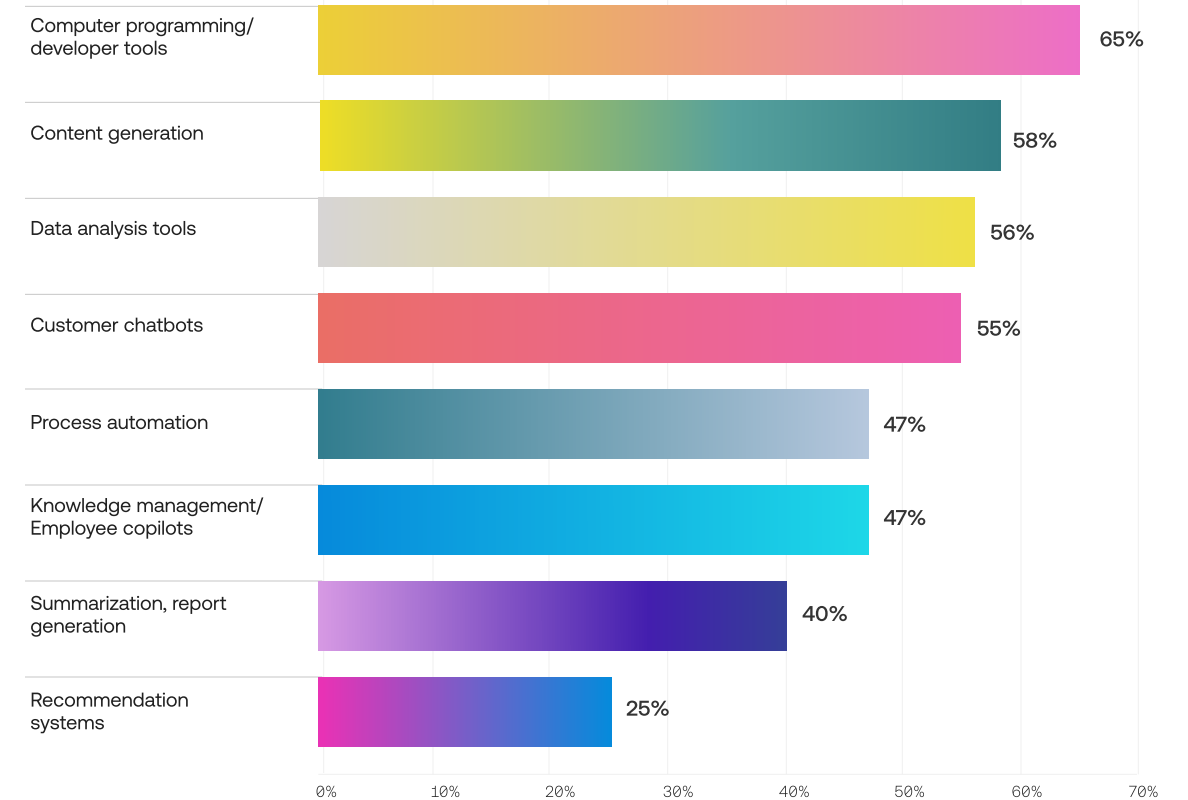
Just as the leading preferred models are closed-source commercial models, planned investments in these categories of models reflect usage trends. 72% of organizations plan to increase investments in commercial closed-source models. A lower percentage of organizations plan to invest in open-source models at 67%. While open-source models provide organizations with greater control, many

leading commercial closed-source models are closely tied to leading cloud-service providers. Enterprises can draw down from cloud spend commitments through use of partner models (e.g., Amazon and Anthropic, Microsoft and Open AI).

Last year, organizations referenced the ability to develop new products or services as the leading reason to adopt generative AI. This year,

improved operational efficiency is the key driver behind adopting generative AI. Generative AI use cases reflect this shift in priorities. The leading use-cases for generative AI adoption are computer programming and content generation..

In which ways has your company implemented AI?



Deploying and Customizing AI Use Cases

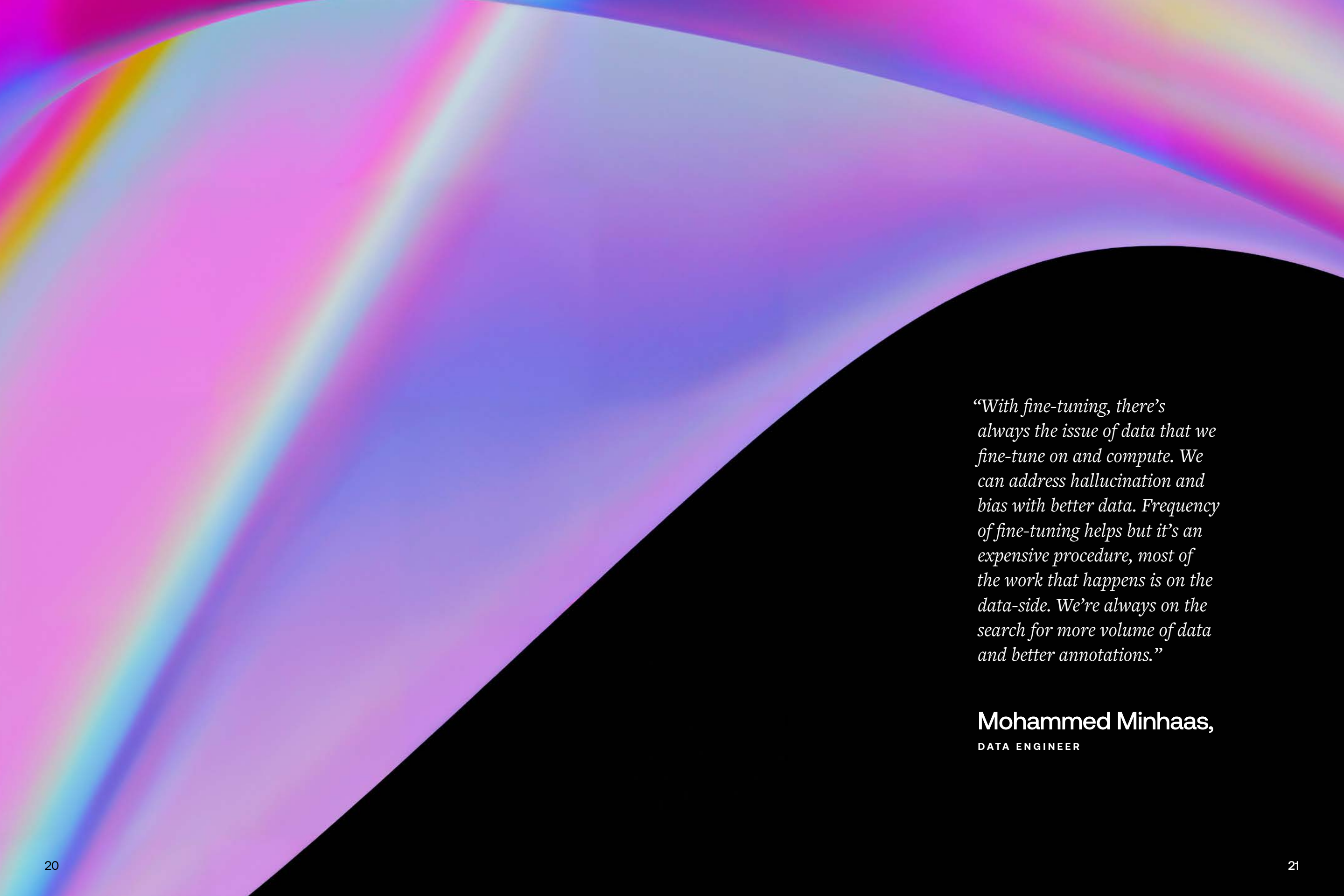
Coding copilots are becoming mainstream with technical users being early adopters of solutions like GitHub Copilot, CodeLlama, and Devin. Model vendors have responded to demand for content generation with prompt templates that guide users to effective content creation questions for functions including Marketing, Product Management, and Public Relations.

Organizations can optimize generative AI models for specific use cases through the following techniques:

- **Prompt-engineering** - guiding the model's output through carefully crafted input prompts
- **Fine-tuning** - training the model on domain-specific data
- **Retrieval-Augmented Generation**

(RAG) - enhancing the model's knowledge by integrating information from external sources during the generation process.

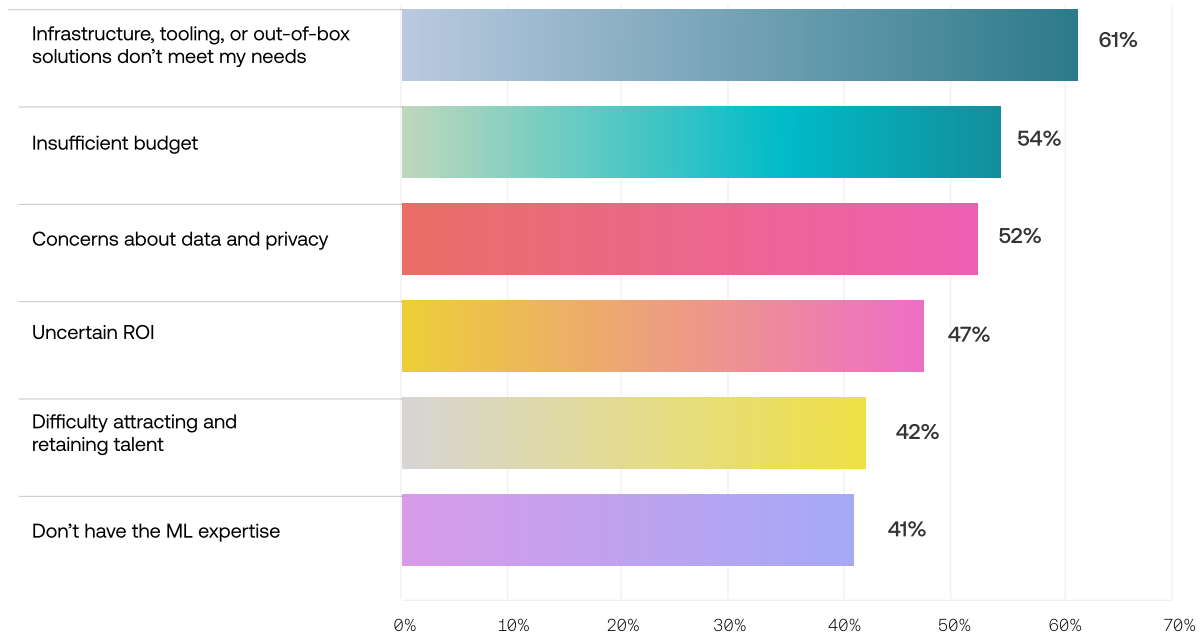
Teams are likely to maximize their AI investments by adopting these techniques. For organizations that already fine-tune their own models, 39% saw improved performance on domain-specific tasks compared to out-of-the-box models.



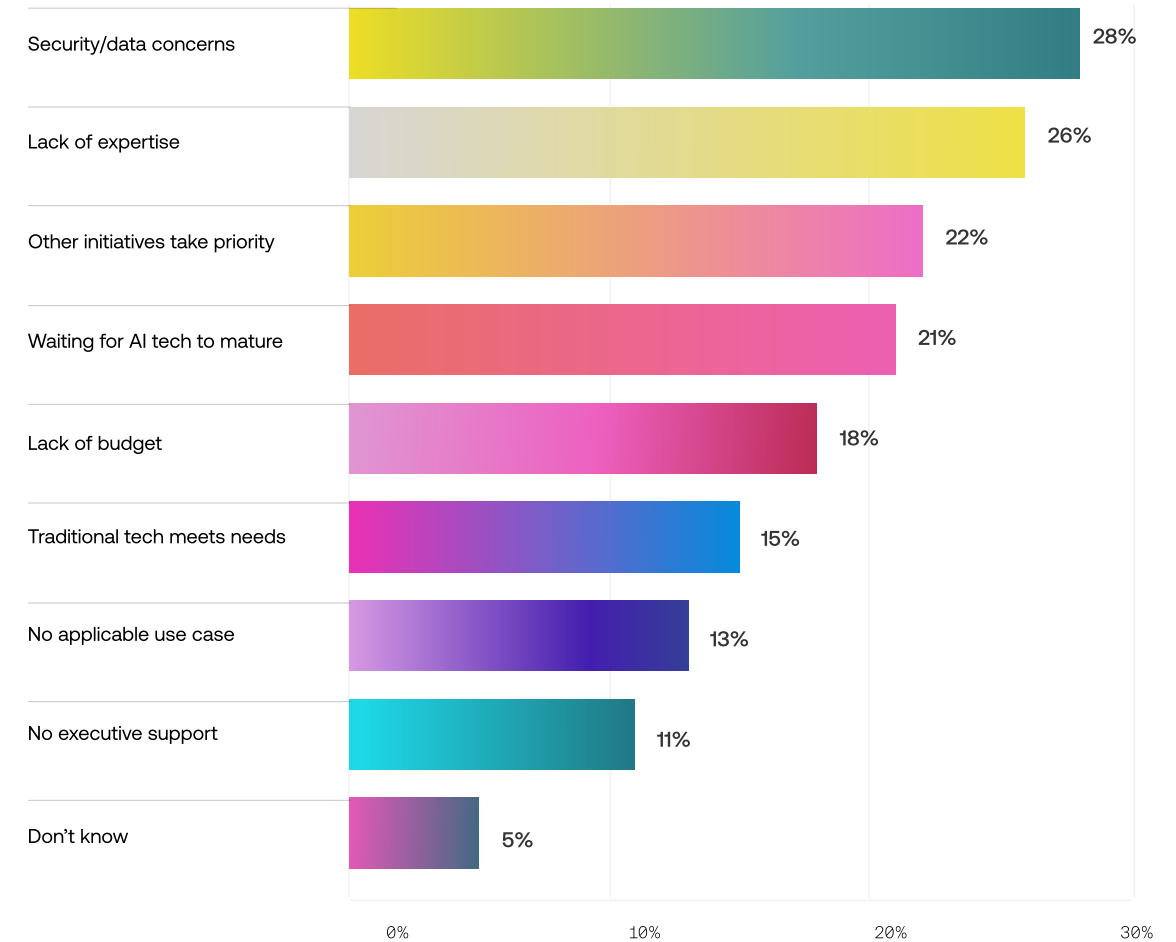
“With fine-tuning, there’s always the issue of data that we fine-tune on and compute. We can address hallucination and bias with better data. Frequency of fine-tuning helps but it’s an expensive procedure, most of the work that happens is on the data-side. We’re always on the search for more volume of data and better annotations.”

Mohammed Minhaas,
DATA ENGINEER

What are the top challenges in implementing AI technologies at your company?



If you have not yet adopted AI, why have you not adopted it?



Barriers to AI Adoption and Implementation

Despite rapid advancements in the field, organizations still face challenges with AI implementation. 61% of organizations specified that infrastructure, tooling, or out-of-the-box solutions don't meet their needs. Insufficient tooling for tasks such as data preparation, model training, and deployment, combined with the lack of standardized frameworks for integrating generative AI into existing

systems, can hinder the scalability and efficiency of AI implementations, leading to increased complexity and higher costs.

54% of organizations struggle with insufficient budget. Finding a home on the balance sheet for new generative AI projects limits the pace of adoption. 52% also have concerns about data privacy. Fine-tuning can

use vast amounts of potentially sensitive training data. The risk of data breaches, unauthorized access, or misuse of personal information during the data collection, storage, and processing stages can expose organizations to legal liabilities and reputational damage, particularly in industries with stringent data protection regulations. For example, certain health and human service providers

“RAG aims to address a key challenge with LLMs - while they are very creative, they lack factual understanding of the world and struggle to explain their reasoning. RAG tackles this by connecting LLMs to known data sources, like a bank’s general ledger, using vector search on a database. This augments the LLM prompts with relevant facts.

However, implementing RAG presents its own challenges. It requires creating and maintaining the external data connection, setting up a fast vector database, and designing vector representations of the data for efficient search. Companies need to consider if they require a purpose-built database optimized for vector search.

Keeping this vectorized representation of truth up-to-date is tricky. As the underlying data sources change over time and users ask new questions, the vector database needs to evolve as well. Deciding if and how to incorporate user assumptions into the vector representations is a philosophical question that also has practical implications for implementation. The industry is still grappling with how to design RAG systems that can continually improve over time.”

Jon Barker,

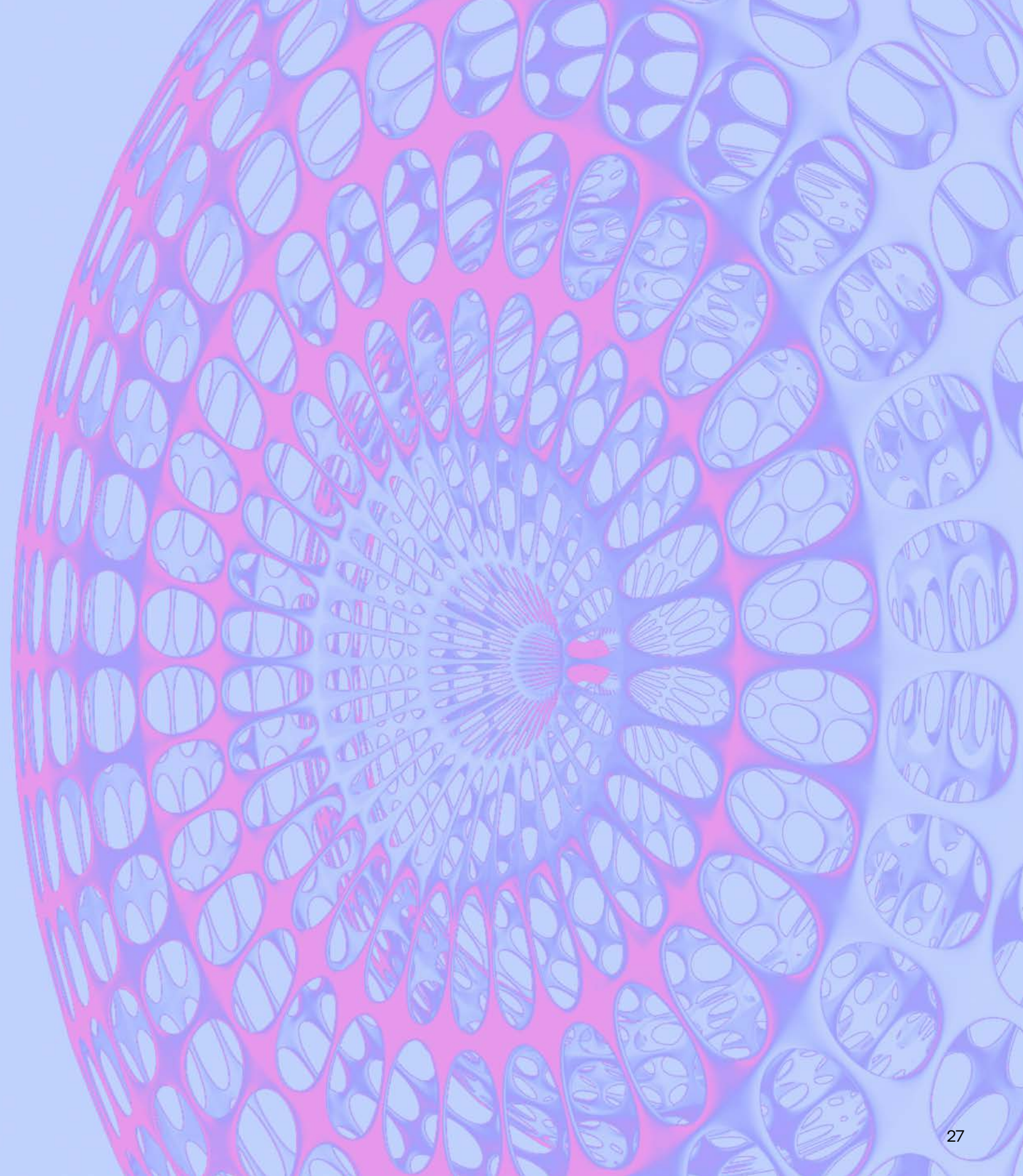
**CUSTOMER ENGINEER,
GOOGLE**

Build AI

Pushing the Boundaries: AI's Rapid Advancement Across Domains

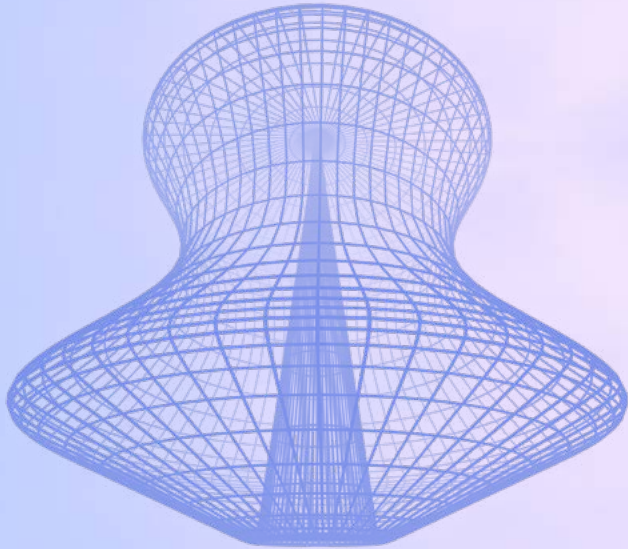
As highlighted in the Year In Review section of this report, we've seen a significant leap in model capabilities in the past year. The latest models have revolutionized programming, writing clean, efficient code from natural language prompts with an almost human-like understanding of intent.

But the advancements don't stop there. We're not far away from a world where AI agents effortlessly communicate across language barriers, solve complex mathematical equations, explain scientific concepts, and even make new discoveries. Moreover, AI is rapidly advancing in its ability to perceive and generate content across multiple modalities, including text, images, audio, and video.

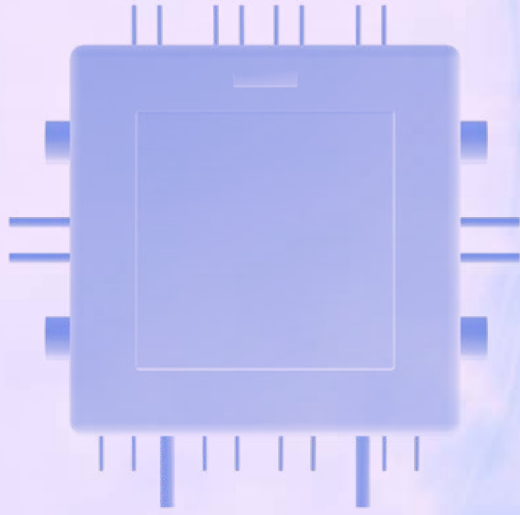


The key pillars of effective AI models

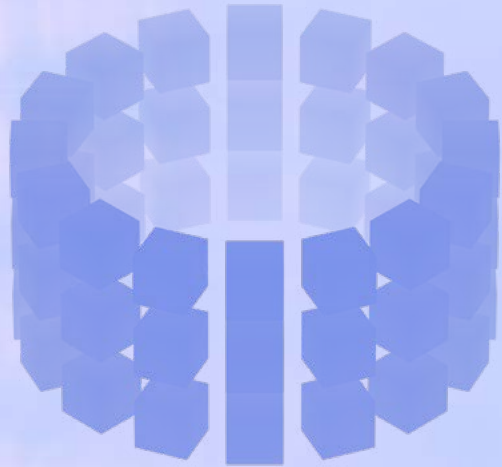
Developing industry-leading AI requires a combination of:



**THOUGHTFUL
MODEL ARCHITECTURES**



**VAST COMPUTATIONAL
RESOURCES**



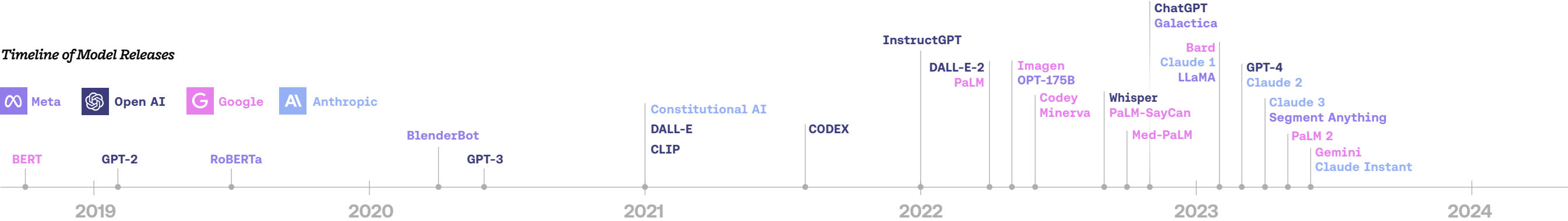
**CAREFULLY
CURATED DATASETS**

The race between leaders like OpenAI, Anthropic, Google, Meta, and others is driving the rapid advancement of foundation models. Each lab is pushing the boundaries of what’s possible, releasing new models that leapfrog the capabilities of predecessors.

However, the pace of releases is not constant. The survey data reveals that it typically takes companies three to six months to develop a model and deploy it to production. For the top labs, major releases are often spaced six to nine months apart, waiting until achieving a significant step-change in performance before unveiling a new model. We expect this six to nine month release cadence to continue over the coming year. However, the pace could decelerate as organizations encounter data limitations and struggle to achieve meaningful improvements over current models’ performance.

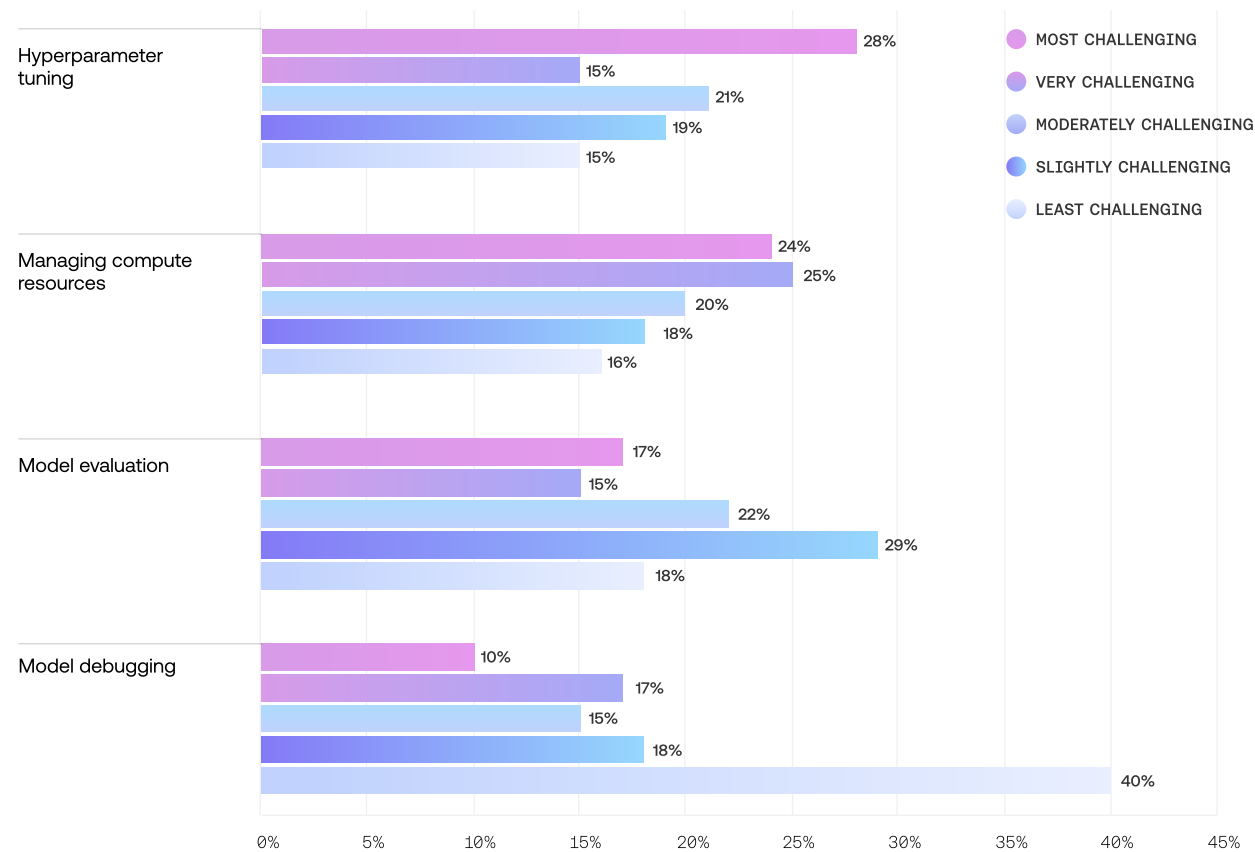
The following sections will explore the key pillars needed to build effective models, including model architecture innovations, computational resource trends, and the high-quality data imperative. We’ll also discuss future investments and priorities in the AI landscape providing insights into the advancements shaping the future of AI.

Timeline of Model Releases



Model Architecture

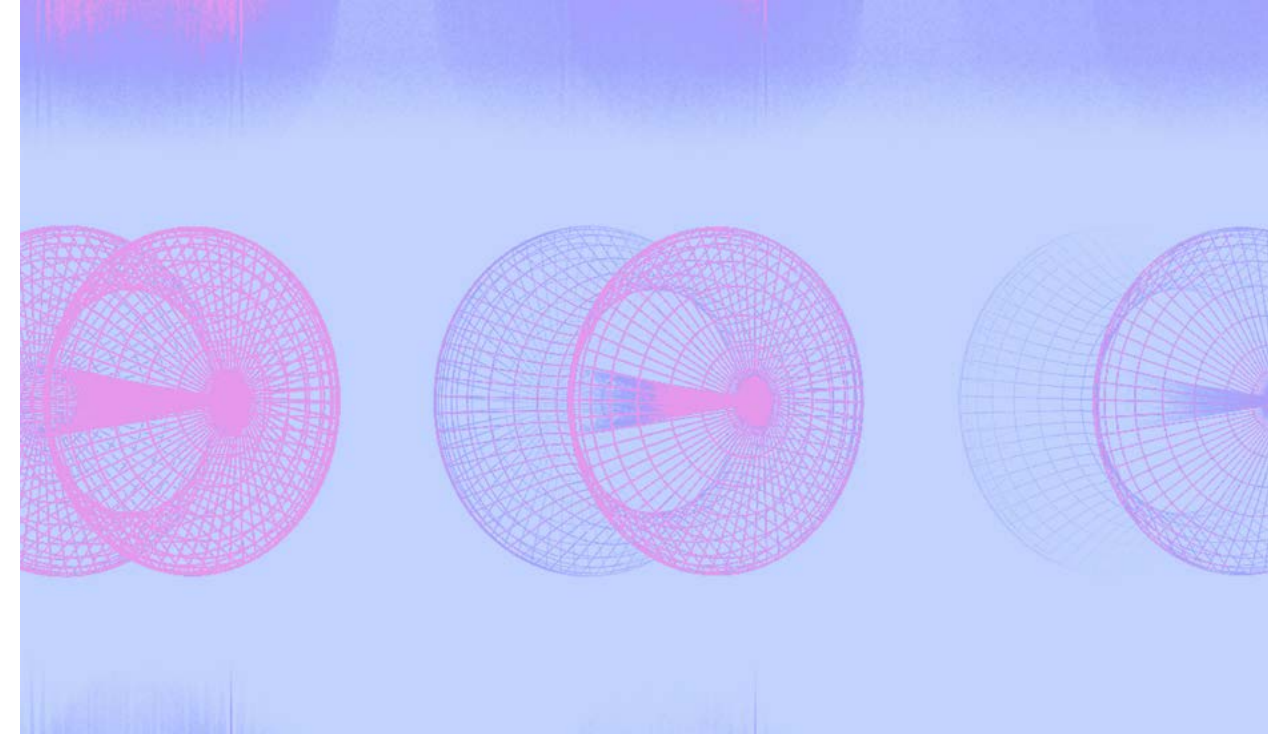
Key challenges in training and developing advanced AI models.



New neural network designs and techniques are enabling the development of larger, more capable models that can tackle increasingly complex tasks.

One new promising approach is the use of sparse expert models, which allows for efficient training of massive networks by activating only relevant subsets of neurons for each input. This enables models to specialize in different domains while still maintaining the ability to generalize across tasks. Recent open-source

models like [Falcon](#), [Mixtral](#), and [DBRX](#) demonstrate the potential of these architectures, scoring high on performance benchmarks with significantly fewer parameters and computational resources when compared to traditional models. Similarly, AI21 Labs' Grok model showcases the power of sparse expert models in natural language processing, excelling across a wide range of language tasks while maintaining high efficiency.



Computational Resources Trends

Demand for compute continues to grow, with model training requiring huge clusters of specialized accelerators like GPUs and TPUs. However, the industry is undergoing a significant shift away from traditional CPUs towards these accelerator architectures optimized for AI workloads. This transition brings significant challenges in terms of infrastructure, tooling, and resource management.

The survey highlights the magnitude of this shift, with over 48% of respondents rating compute resource management as “most challenging” or “very challenging”.

“CPUs consume about 80% of IT workloads today. GPUs consume about 20%. That’s going to flip in the short term, meaning 3 to 5 years. Many industry leaders that I’ve talked to at Google and elsewhere believe that in 3 to 5 years, 80% of IT workloads will be running on some type of architecture that is not CPU, but rather some type of chip architecture like a GPU.”

- Jon Barker, Customer Engineer, Google

This rapid transition towards more costly GPU and TPU-centric workloads presents a number of challenges. While these accelerators offer unparalleled

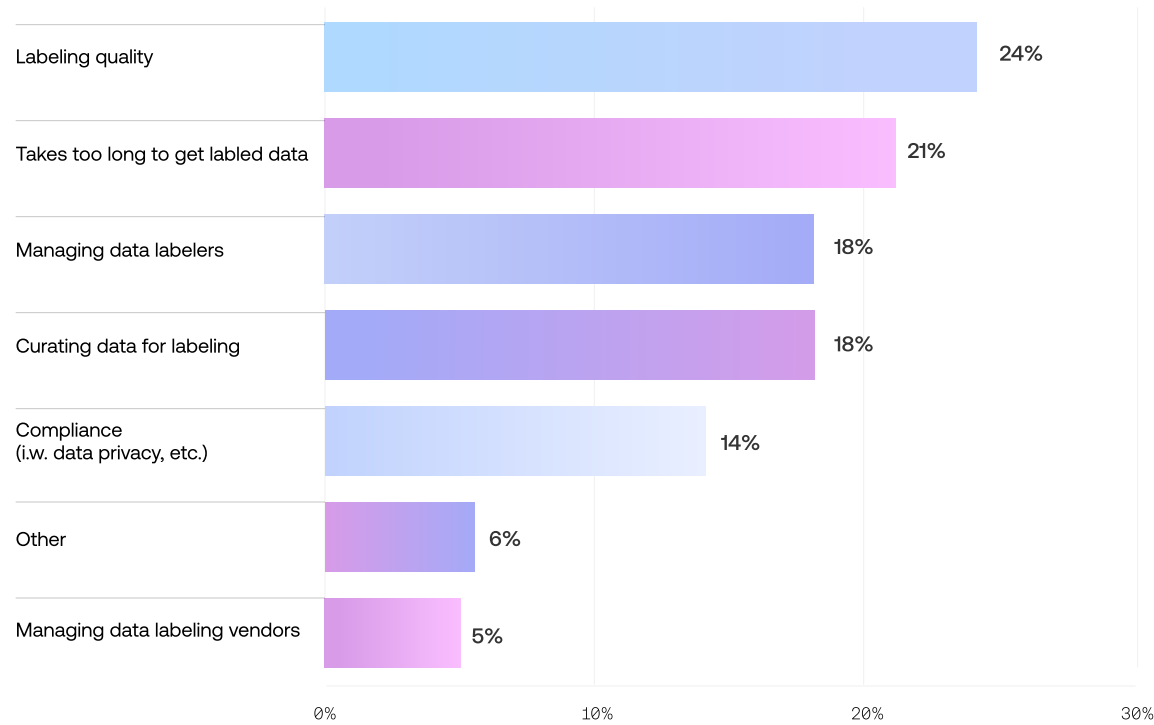
performance for AI tasks, they also require a different programming model, tooling ecosystem, and set of optimization techniques compared to traditional CPU-based workloads. Further, large models are usually trained across many accelerators and distributed across many machines in parallel, requiring complex orchestration frameworks.

To address these challenges, PyTorch introduced the [Fully Shared Data Parallel](#) (FSDP). FSDP is a data parallelism paradigm that shards model parameters, gradients, and optimizer states across data-parallel workers, enabling more efficient memory usage and training of larger models.

In addition to the challenge of compute resource management, model builders also face obstacles due to a lack of suitable tools and frameworks. 38% of respondents indicated that the absence of AI-specific libraries, frameworks, and platforms is a major challenge holding back their AI projects. These tools are crucial for abstracting away the complexities of distributed computing and accelerator programming, allowing researchers to focus on model development and experimentation.

Unlocking AI Potential: Domain-Specific, Human-Generated Datasets

Top challenges in preparing high-quality training data for AI models.



Data is the fuel that powers AI models, and the quality, quantity, and diversity of that data is critical to building effective, unbiased systems. The survey results highlight the importance of high-quality datasets, with labeling quality as the top challenge in preparing data for training models. Obtaining extremely high-quality labels while minimizing the time required to get that labeled data is a significant hurdle for model builders. This highlights the need for efficient data labeling processes and tools that can maintain high standards while expediting the labeling process.

Large, web-scraped datasets have been instrumental in pre-training foundation models. The next leap in capabilities will require more targeted, domain-specific data that captures the nuances and edge cases that only human experts can provide. The advent of generative AI and large language models (LLMs) has fundamentally changed what it means to create high-quality training and evaluation data. For open-ended use cases, such as

question answering, coding, and agentic use cases, advancements in AI capabilities will be bottlenecked by the supervision we can feed into these models.

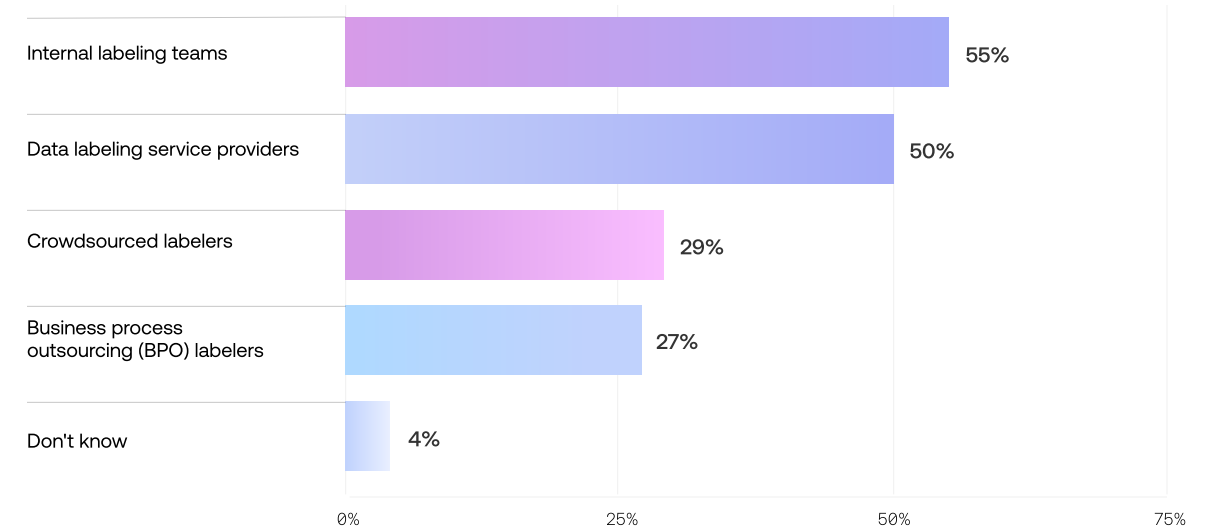
Even if you train long enough with enough GPUs, you'll get similar results with any modern model. It's not about the model, it's about the data that it was trained with. The difference between performance is the volume and quality of data, especially human feedback data. You absolutely need it. That will determine your success.

- Ashiqur Rahman, Machine Learning Researcher, Kimberly-Clark

Human-labeled data plays a critical role in [aligning models with user preferences](#) and real-world requirements. Techniques like reinforcement learning from human feedback (RLHF) can help guide models towards desired behaviors and outputs, but they require a steady stream of high-quality, human-generated labels and rankings.

Future Investments & Priorities

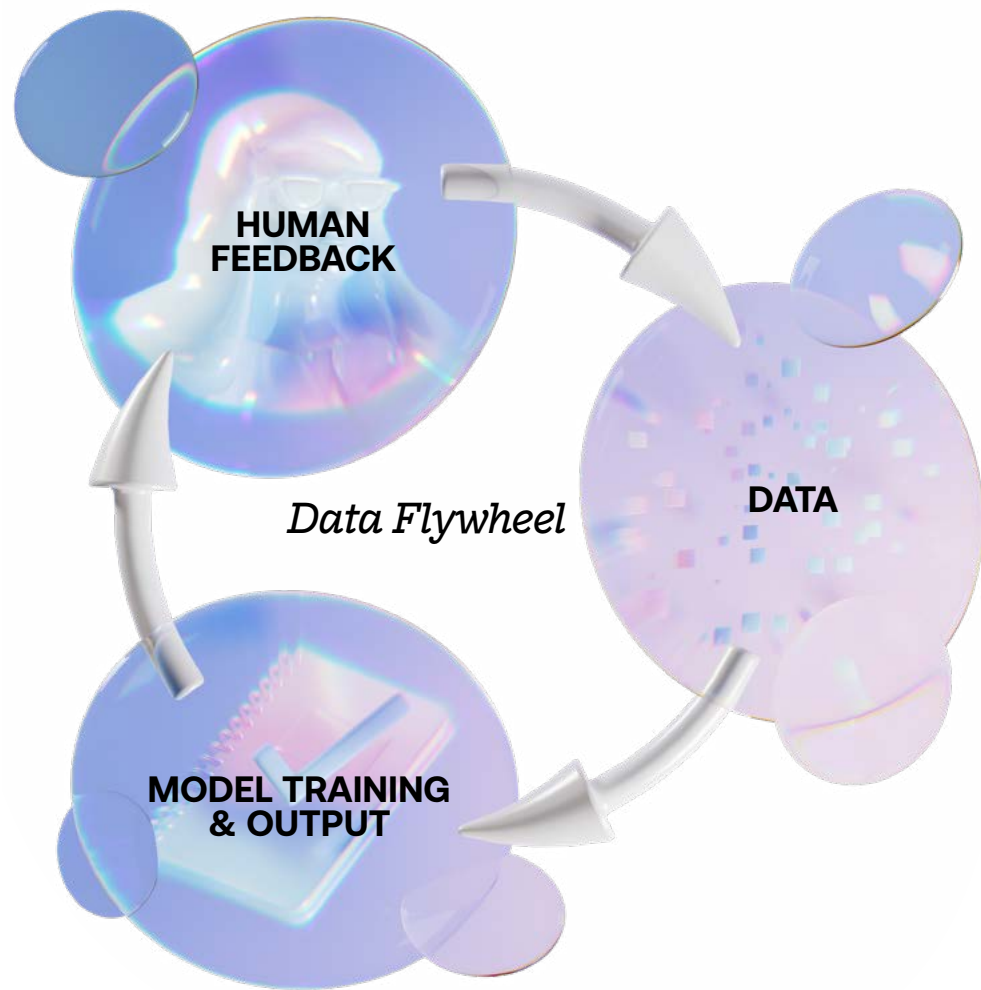
Common approaches for data annotation.



69% of respondents rely on unstructured data like text, images, audio, and video to train their models. However, data quality emerges as the top challenge in acquiring training data, ranked as the largest obstacle by 35% of respondents.

To address this, 55% of organizations are leveraging internal labeling teams, while 50% engage specialized data labeling services and 29% leverage crowdsourcing. Organizations are scaling their annotation efforts with managed labeling services, with 40% of users receiving high-quality labeled data within one week to one month.

Managed labeling services allow companies to scale up labeling operations, reduce overhead, and access expert annotators on-demand. Managed labeling services also handle project management, quality assurance, annotator recruiting, and increasingly offer specialized expertise in areas like coding, mathematics, and languages.



The demand for specific types of [Scale's Data Streams](#) provides insights into the priorities and use cases driving AI development. Among the most sought-after Data Streams are:

1. **Coding, Reasoning, and Precise Instruction Following**
2. **Languages**
3. **Multimodal Data**

Going forward, we expect to see increased adoption of human-in-the-loop pipelines that leverage subject matter experts to refine model outputs and provide targeted feedback. This creates a virtuous “data flywheel” effect, where model usage results in new high-quality training data for continuous improvement.

Multimodal data collection spanning text, speech, images, and video will also be a key priority as organizations seek to build AI systems that can perceive, reason and interact more naturally.

One new notable trend is the [acquisition of proprietary data from platforms like Reddit](#), as exemplified by the recent multi-year data partnership between Reddit and Google. This deal, reportedly valued at \$60 million per year, emphasizes the value placed on unique, human-generated content for training the next generation of models.

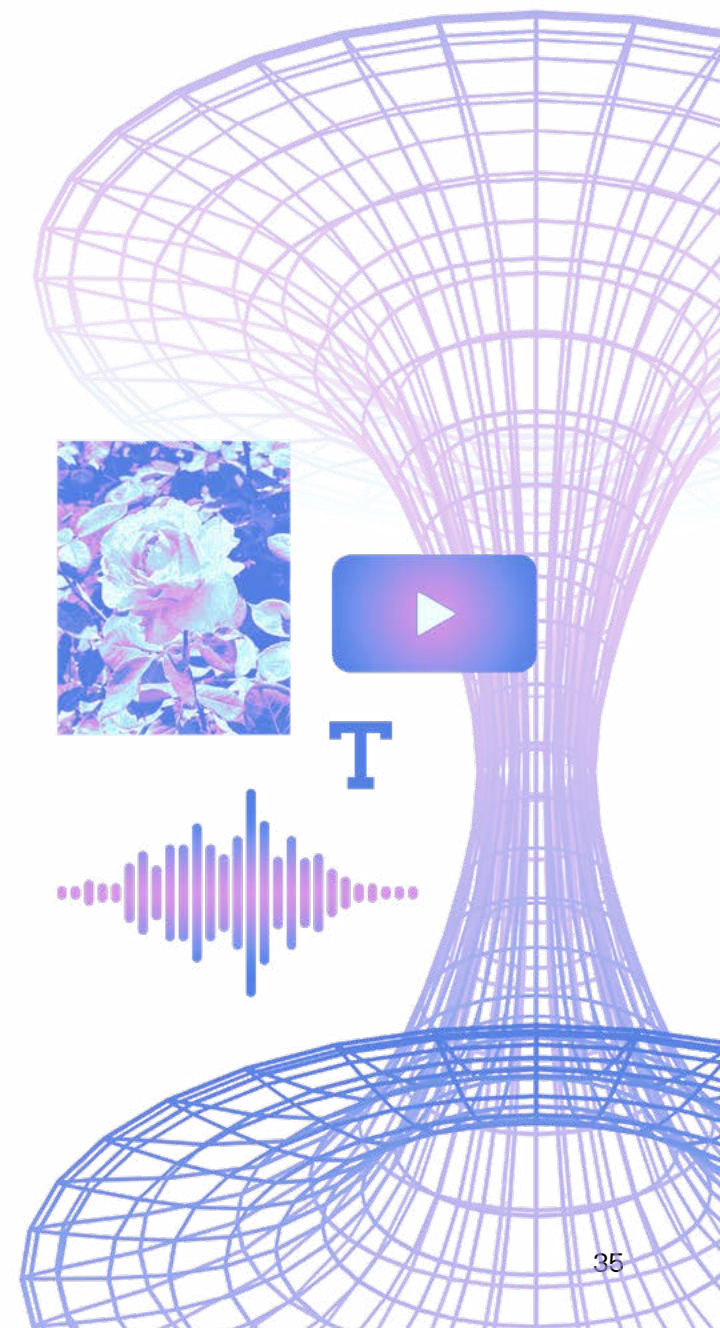
However, simply acquiring vast amounts of data is not enough. To truly stay ahead of the curve, organizations must also invest in robust human-in-the-loop (HITL) pipelines that can process and label data across an ever-expanding range of modalities. As AI systems become more sophisticated, they will require not just text, but also speech, images, video, and even more complex data types like 3D scenes and sensor data.

Moreover, the rise of reinforcement learning from human feedback (RLHF) has fundamentally changed how models are evaluated. RLHF requires “on-policy” human supervision, where human raters provide feedback on the actual outputs generated by the model during the training process.

Additionally, traditional evaluation methods that rely on fixed sets of labels are no longer sufficient. Instead, organizations must conduct side-by-side comparisons of their old and new model responses across a large number of prompts before each release. This approach captures the nuances and edge cases that emerge as models become more sophisticated and ensures that improvements are aligned with user expectations.

Building scalable labeling programs that address multimodal capabilities is a critical challenge for model builders. It will require a combination of advanced tooling, specialized annotator training, and close collaboration between domain experts and machine learning teams. Managed labeling services with expertise across a wide range of modalities will be increasingly sought after to help organizations navigate this complex landscape.

By fusing diverse input modalities and investing in human-in-the-loop pipelines, models can develop richer, more contextual representations that mirror how humans process information and engage with their environments. Organizations that can effectively harness multimodal data and scale their labeling capabilities will be well-positioned to unlock new frontiers in AI.



Evaluate AI

Evaluating Model Performance

Evaluation criteria for models in use

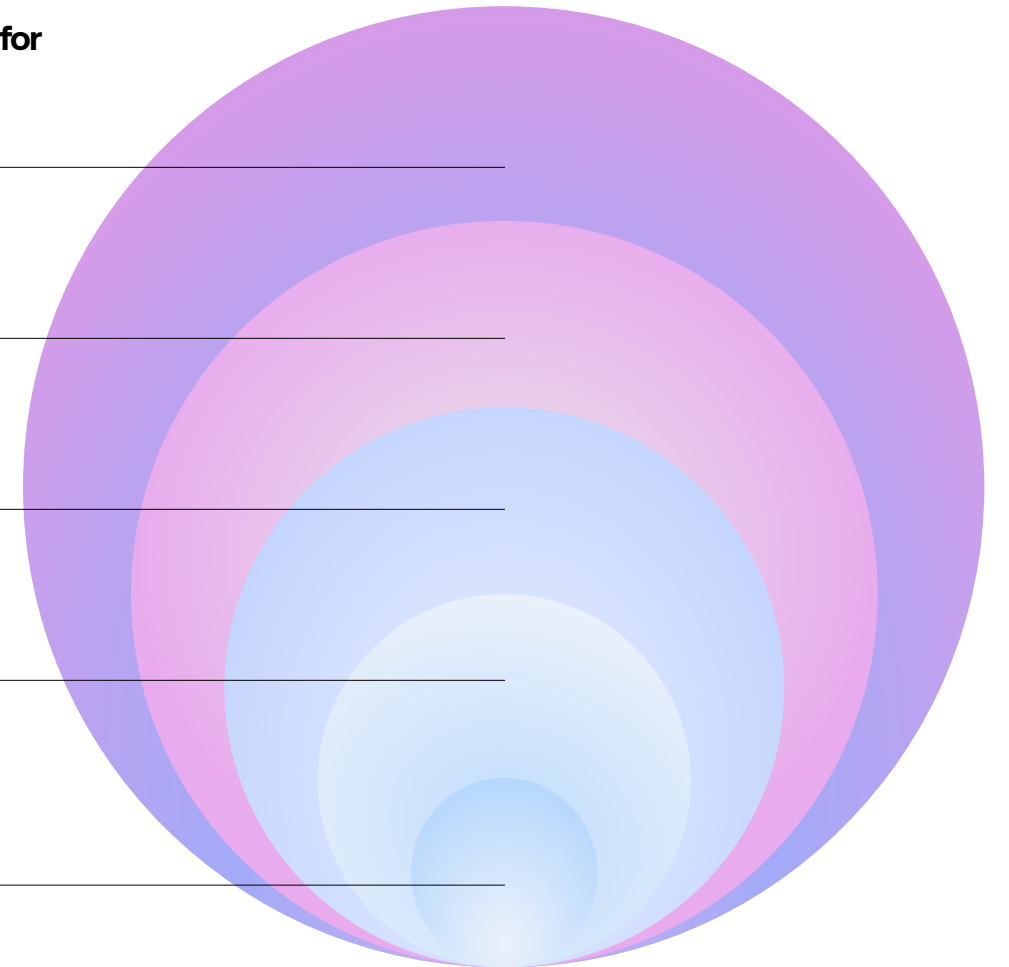
68%
Reliability

67%
Performance

62%
Security

54%
Safety

6%
N/A

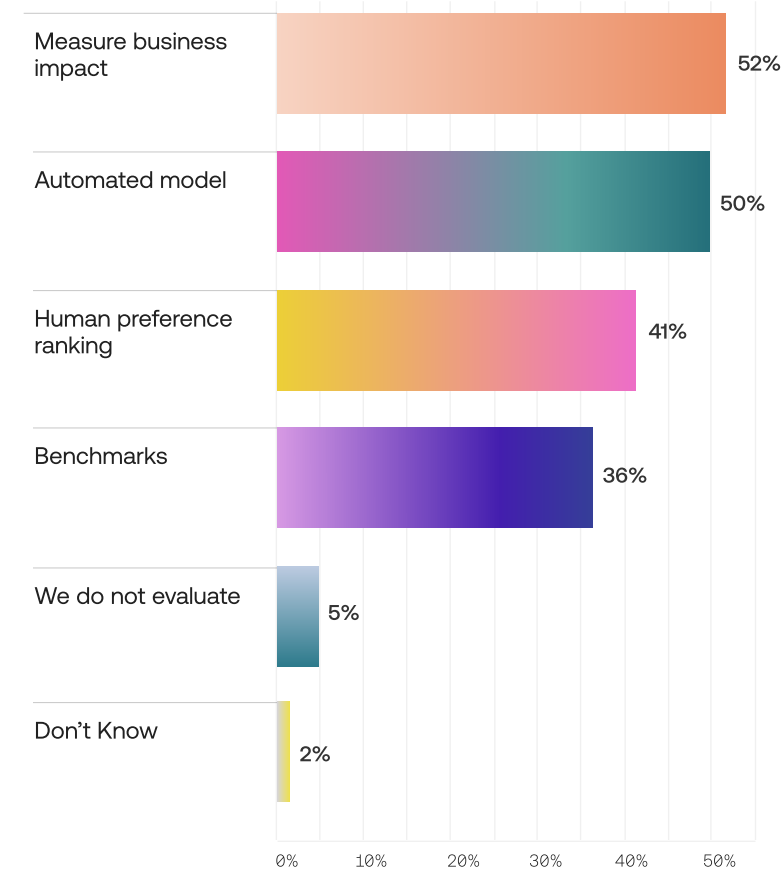


As foundation models grow in capability and impact, comprehensive model evaluation has become paramount whether you are building or applying models. In contrast to common headlines, assessing foundation models is not just about safety. In fact, performance, reliability, and security were indicated as the top three reasons survey respondents evaluate models - with safety ranking as a lower priority.

Despite this focus on evaluation, developing robust evaluation frameworks is an evolving challenge. Models must be assessed holistically, accounting for performance on real-world use cases as well as potential risks. Traditional academic benchmarks are generally not representative of production scenarios, and models have been overfitted to these existing benchmarks due to their presence in the public domain. Leading or-

ganizations are moving towards comprehensive private test suites that probe model behavior across diverse domains and capabilities. Universally agreed upon 3rd party benchmarks are crucial for objectively evaluating and comparing the performance of large language models. Researchers, developers, and users can select models based on standardized transparent metrics.

Evaluation practices for model performance.



To understand current evaluation practices, the survey asked respondents how they measure model performance. The top approaches are illustrated in the figure, left.

The data shows that automated model metrics and human preference ranking are the fastest ways to identify issues, with over 70% of respondents discovering problems within one week. This highlights the value of quantitative and qualitative evaluation approaches to rapidly surface model performance problems.

The prevalence of human evaluations is notable (41%), reflecting the importance of subjective judgments in assessing generative outputs. Techniques like preference ranking, where human raters compare model samples, can capture nuanced quality distinctions.

The survey results suggest that a multi-faceted evaluation strategy is necessary, as no single approach dominates. While automated metrics and business impact assessments are widely used, the data indicates the need to incorporate a variety of quantitative and qualitative techniques to comprehensively evaluate models.

When asked why they conduct model evaluations, 69% of respondents selected performance, another 69% selected reliability and 63% selected security as main objectives. Stress testing models is an important defense against failure modes such as hallucination and bias.

87%

Model builders who apply AI indicated that they evaluate models or applications.

72%

Enterprises who apply AI indicated that they evaluate models or applications.



Techniques like red teaming, where expert testers try to elicit unsafe behaviors, can surface vulnerabilities. Careful prompt engineering can also help assess models' resilience against malicious prompts or out-of-distribution inputs.

The results highlight the importance of continuous monitoring, as models can degrade or exhibit new issues over time. Over 40% of respondents evaluate their models following any changes or prior to major releases, highlighting the shift towards a continuous evaluation that goes beyond one-time assessments.

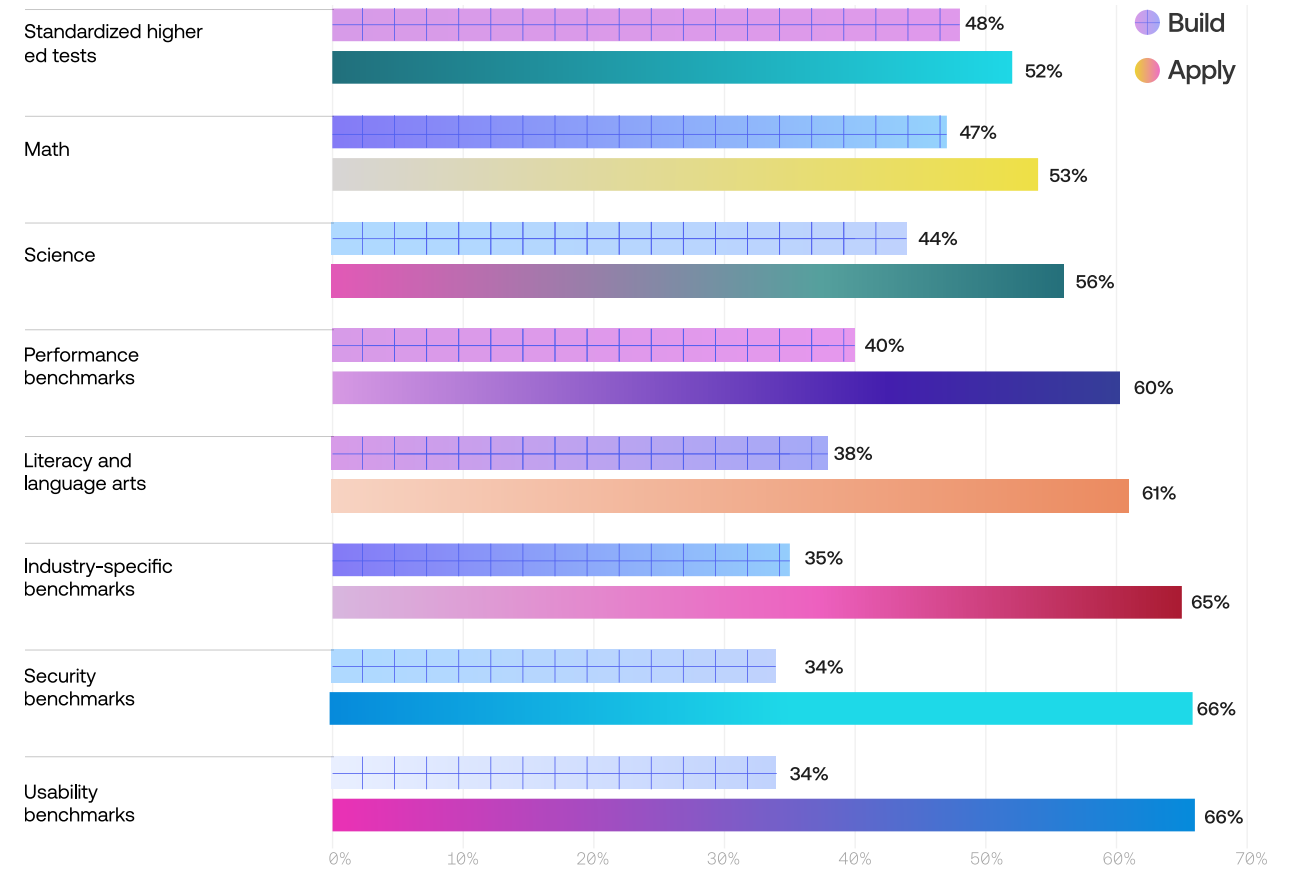
While model evaluation plays a crucial role in measuring AI performance, leaders responsible for applying AI in their organizations must also demonstrate tangible business outcomes. Almost half of respondents evaluate models based on their direct impact on KPIs like operational efficiency or customer satisfaction. Grounding evaluations in downstream outcomes ensures that models are not just technically proficient but actually valuable in practice.

“Evaluating generative AI performance is complex due to evolving benchmarks, data drift, model versioning, and the need to coordinate across diverse teams. The key question is how the model performs on specific data and use cases... Centralized oversight of the data flow is essential for effective model evaluation and risk management in order to achieve high acceptance rates from developers and other stakeholders.”

Babar Bhatti,

IBM, AI CUSTOMER SUCCESS LEAD

Model evaluation challenges: gaps in benchmarking for model builders and enterprises applying AI



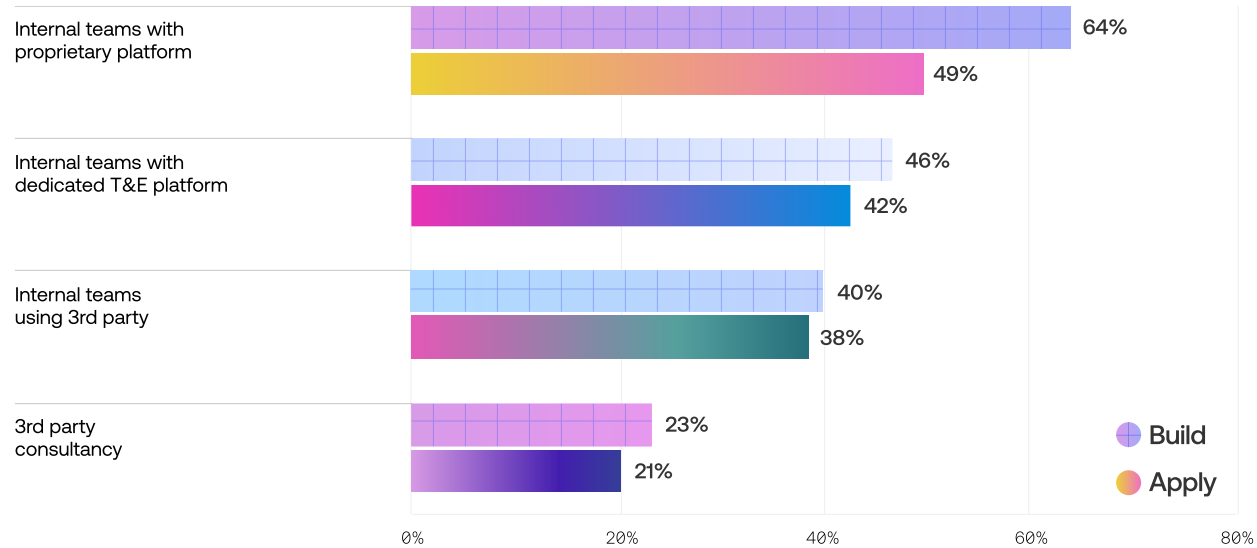
Challenges with model evaluation today

Despite progress, many gaps remain in current model evaluation practices.

Performance and usability benchmarks are critical to ensure models meet rising user expectations while vertical-specific standards will be key as AI permeates different sectors. Industry groups like the National Institute of Standards and Technology (NIST) are working to define comprehensive evaluation standards. Scale’s Safety, Evaluations, and Analysis Lab (SEAL) is also working to develop robust evaluation frameworks.

The data reveals room for improvement in measuring the business impact of AI models. For key outcomes like revenue, profitability, and strategic decision-making, only half of the organizations are assessing business impact. This represents an opportunity for enterprises to more clearly link model performance to tangible business results, ensuring that AI investments are delivering real value.

Practices for evaluating AI systems in production



Evaluating AI Systems in Production

Robust evaluation practices are essential not just during model development, but also when deploying and monitoring AI systems in real-world production environments.

The survey highlights how both model builders and enterprises are investing in evaluation capabilities. On the “Build” side, organizations recognize the importance of comprehensive evaluations and employ a combination of internal dashboards and external platforms to gain a holistic understanding of model performance. 46% of organizations have internal teams with dedicated test and evaluation platforms, while 64% leverage internal proprietary platforms. Adoption of third-party evaluation consultancies (23%) and platforms (40%) is also prevalent, demonstrating the value of external expertise and tools in the evaluation process.

For enterprises focused on “Applying” AI, the investment patterns are similar but with a blend of internal and external solutions. 42% have internal teams using

external evaluation platforms, 49% use proprietary internal platforms, 38% adopt third-party platforms and 21% engage external consultants.

These results underscore the complexity of validating AI system performance, safety, and alignment with real-world operating conditions and business objectives. Effective evaluation requires a blend of skilled in-house teams, robust tools and frameworks, and external specialist support.

Looking ahead, evaluation methodology must evolve in lockstep with AI capabilities. Multidisciplinary research at the intersection of machine learning, software engineering, and social science is needed to define rigorous standards. Scalable infrastructure for human-in-the-loop evaluation pipelines will also be critical. With sustained effort and investment, the industry can build generative models that are not only powerful but truly reliable and beneficial.

“As AI systems become more advanced and influential, it’s crucial that we prioritize AI safety. The rapid progress in large language models and generative AI is both awe-inspiring and sobering - while these technologies could help solve some of humanity’s greatest challenges, they also pose catastrophic risks if developed without sufficient safeguards. At the Center for AI Safety, our research focuses on the important problem of AI safety: mitigating the various risks posed by AI systems.”

We also need proactive governance strategies to navigate the high-stakes landscape of powerful AI, including establishing international cooperation, safety standards, and regulatory oversight. While the era of advanced AI presents tremendous potential, we must not underestimate the risks and challenges ahead. It’s crucial that the AI community comes together to prioritize safety, so we can chart a course towards a future where AI is a profound positive force for the world.”

Dan Hendrycks,
CENTER FOR AI SAFETY (CAIS)

Conclusion

Whether you are building or applying AI, model optimization and evaluation is key to unlock performance and ROI.

The pace of innovation for generative AI continues to accelerate. While the 2023 AI Readiness Report focused on how enterprises could adopt AI, this year's report examined challenges and best practices to apply, build, and evaluate AI. The two most significant trends to emerge in our analysis are:

1. The growing need for model evaluation frameworks and private benchmarks.
2. The continued challenges of optimizing models for specific use cases without sufficient tooling for data preparation, model training, and deployment.

At Scale, our mission is to accelerate the development of AI applications. The Scale Zeitgeist: AI Readiness Report supports that mission. We will continue to shed light on the latest trends, challenges, and what it really takes to build, apply, and evaluate AI.

About Scale

Scale is fueling the generative AI revolution. Built on a foundation of high-quality data and expert insight, Scale powers the world's most advanced models. Our years of deep partnership with every major model builder enables our platform to empower any organization to apply and evaluate AI.

scale.com

Methodology

This survey was conducted online within the United States by Scale AI from February 20, 2024, to March 29, 2024. We received 2,302 responses from ML practitioners (e.g., ML engineers, data scientists, development operations, etc.) and leaders involved with AI in their companies. Participants who reported no involvement in AI or ML projects were excluded from the dataset, resulting in a final sample size of 1800 respondents.

A quarter of the respondents identified themselves as belonging to the Software and Internet/Telecommunications industry (28%), with the Financial Services/Insurance Industry following closely behind at 15%. Business Services accounted for 7%, while the Government and Defense Industry represented 4% of the respondents. Among these industries, a majority of respondents specified their employment within the Information Technology department (33%).

In terms of seniority within their organizations, nearly a quarter of respondents (24%) identified themselves as Team Leads, 22% as department heads, and 5% as owners. Sixty-six percent (66%) of respondents report involvement in AI model application and customization (applying AI), while 34% are directly engaged in developing foundational generative AI models (building AI). Consequently, a significant portion of respondents (46%) represent organizations at an advanced stage of AI/ML adoption, with one to multiple models deployed to production and undergoing regular retraining.

Approximately 26% are in the process of developing their inaugural model, while 23% are in the phase of evaluating potential use cases, underscoring the significance and enthusiasm for AI/ML project development.