O'REILLY®

2023 Generative AI in the Enterprise

Mike Loukides

RADAR REPORT

Generative AI in the Enterprise

Mike Loukides

Beijing • Boston • Farnham • Sebastopol • Tokyo O'REILLY®

Table of Contents

Generative AI in the Enterprise

Generative AI has been the biggest technology story of 2023. Almost everybody's played with ChatGPT, Stable Diffusion, GitHub Copilot, or Midjourney. A few have even tried out Bard or Claude, or run LLaMA[1] on their laptop. And everyone has opinions about how these language models and art generation programs are going to change the nature of work, usher in the singularity, or perhaps even doom the human race. In enterprises, we've seen everything from wholesale adoption to policies that severely restrict or even forbid the use of generative AI.

What's the reality? We wanted to find out what people are actually doing, so in September we surveyed O'Reilly's users. Our survey focused on how companies use generative AI, what bottlenecks they see in adoption, and what skills gaps need to be addressed.

1 Meta has dropped the odd capitalization for Llama 2. In this report, we use LLaMA to refer to the LLaMA models generically: LLaMA, Llama 2, and Llama n, when future versions exist. Although capitalization changes, we use Claude to refer both to the original Claude and to Claude 2, and Bard to Google's Bard model and its successors.

Executive Summary

We've never seen a technology adopted as fast as generative AI—it's hard to believe that ChatGPT is barely a year old. As of November 2023:

- Two-thirds (67%) of our survey respondents report that their companies are using generative AI.
- AI users say that AI programming (66%) and data analysis (59%) are the most needed skills.
- Many AI adopters are still in the early stages. 26% have been working with AI for under a year. But 18% already have applications in production.
- Difficulty finding appropriate use cases is the biggest bar to adoption for both users and nonusers.
- 16% of respondents working with AI are using open source models.
- Unexpected outcomes, security, safety, fairness and bias, and privacy are the biggest risks for which adopters are testing.
- 54% of AI users expect AI's biggest benefit will be greater productivity. Only 4% pointed to lower head counts.

Is generative AI at the top of the hype curve? We see plenty of room for growth, particularly as adopters discover new use cases and reimagine how they do business.

Users and Nonusers

AI adoption is in the process of becoming widespread, but it's still not universal. Two-thirds of our survey's respondents (67%) report that their companies are using generative AI. 41% say their companies have been using AI for a year or more; 26% say their companies have been using AI for less than a year. And only 33% report that their companies aren't using AI at all.

Generative AI users represent a two-to-one majority over nonusers, but what does that mean? If we asked whether their companies were using databases or web servers, no doubt 100% of the respondents would have said "yes." Until AI reaches 100%, it's still in the process of adoption. ChatGPT was opened to the public on November 30,

2022, roughly a year ago; the art generators, such as Stable Diffusion and DALL-E, are somewhat older. A year after the first web servers became available, how many companies had websites or were experimenting with building them? Certainly not two-thirds of them. Looking only at AI users, over a third (38%) report that their companies have been working with AI for less than a year and are almost certainly still in the early stages: they're experimenting and working on proof-of-concept projects. (We'll say more about this later.) Even with cloud-based foundation models like GPT-4, which eliminate the need to develop your own model or provide your own infrastructure, fine-tuning a model for any particular use case is still a major undertaking. We've never seen adoption proceed so quickly.

When 26% of a survey's respondents have been working with a technology for under a year, that's an important sign of momentum. Yes, it's conceivable that AI—and specifically generative AI—could be at the peak of the hype cycle, as Gartner has argued. We don't believe that, even though the failure rate for many of these new projects is undoubtedly high. But while the rush to adopt AI has plenty of momentum, AI will still have to prove its value to those new adopters, and soon. Its adopters expect returns, and if not, well, AI has experienced many "winters" in the past. Are we at the top of the adoption curve, with nowhere to go but down? Or is there still room for growth?

We believe there's a lot of headroom. Training models and developing complex applications on top of those models is becoming easier. Many of the new open source models are much smaller and not as resource intensive but still deliver good results (especially when trained for a specific application). Some can easily be run on a laptop or even in a web browser. A healthy tools ecosystem has grown up around generative AI—and, as was said about the California Gold Rush, if you want to see who's making money, don't look at the miners; look at the people selling shovels. Automating the process of building complex prompts has become common, with patterns like retrieval-augmented generation (RAG) and tools like LangChain. And there are tools for archiving and indexing prompts for reuse, vector databases for retrieving documents that an AI can use to answer a question, and much more. We're already moving into the second (if not the third) generation of tooling. A roller-coaster ride into Gartner's "trough of disillusionment" is unlikely.

What's Holding AI Back?

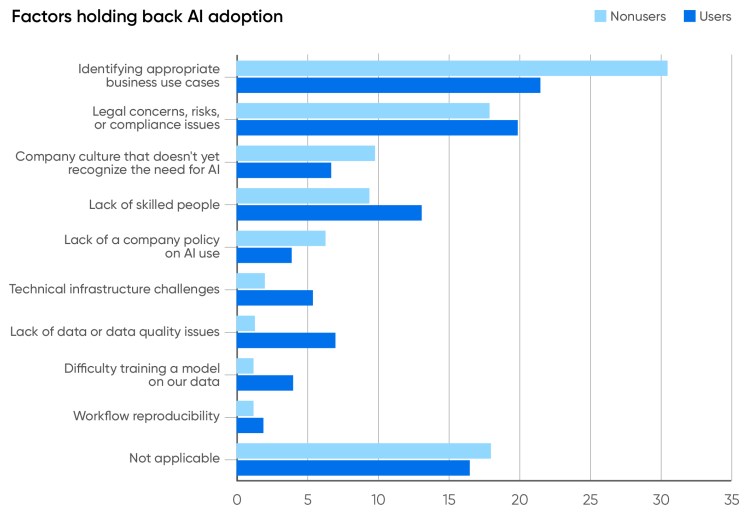
It was important for us to learn why companies aren't using AI, so we asked respondents whose companies aren't using AI a single obvious question: "Why isn't your company using AI?" We asked a similar question to users who said their companies are using AI: "What's the main bottleneck holding back further AI adoption?" Both groups were asked to select from the same group of answers. The most common reason, by a significant margin, was difficulty finding appropriate business use cases (31% for nonusers, 22% for users). We could argue that this reflects a lack of imagination—but that's not only ungracious, it also presumes that applying AI everywhere without careful thought is a good idea. The consequences of "Move fast and break things" are still playing out across the world, and it isn't pretty. Badly thought-out and poorly implemented AI solutions can be damaging, so most companies should think carefully about how to use AI appropriately. We're not encouraging skepticism or fear, but companies should start AI products with a clear understanding of the risks, especially those risks that are specific to AI. What use cases are appropriate, and what aren't? The ability to distinguish between the two is important, and it's an issue for both companies that use AI and companies that don't. We also have to recognize that many of these use cases will challenge traditional ways of thinking about businesses. Recognizing use cases for AI and understanding how AI allows you to reimagine the business itself will go hand in hand.

The second most common reason was concern about legal issues, risk, and compliance (18% for nonusers, 20% for users). This worry certainly belongs to the same story: risk has to be considered when thinking about appropriate use cases. The legal consequences of using generative AI are still unknown. Who owns the copyright for AI-generated output? Can the creation of a model violate copyright, or is it a "transformative" use that's protected under US copyright law? We don't know right now; the answers will be worked out in the courts in the years to come. There are other risks too, including reputational damage when a model generates inappropriate output, new security vulnerabilities, and many more.

Another piece of the same puzzle is the lack of a policy for AI use. Such policies would be designed to mitigate legal problems and require regulatory compliance. This isn't as significant an issue; it

was cited by 6.3% of users and 3.9% of nonusers. Corporate policies on AI use will be appearing and evolving over the next year. (At O'Reilly, we have just put our policy for workplace use into place.) Late in 2023, we suspect that relatively few companies have a policy. And of course, companies that don't use AI don't need an AI use policy. But it's important to think about which is the cart and which is the horse. Does the lack of a policy prevent the adoption of AI? Or are individuals adopting AI on their own, exposing the company to unknown risks and liabilities? Among AI users, the absence of company-wide policies isn't holding back AI use; that's self-evident. But this probably isn't a good thing. Again, AI brings with it risks and liabilities that should be addressed rather than ignored. Willful ignorance can only lead to unfortunate consequences.

Another factor holding back the use of AI is a company culture that doesn't recognize the need (9.8% for nonusers, 6.7% for users). In some respects, not recognizing the need is similar to not finding appropriate business use cases. But there's also an important difference: the word "appropriate." AI entails risks, and finding use cases that are appropriate is a legitimate concern. A culture that doesn't recognize the need is dismissive and could indicate a lack of imagination or forethought: "AI is just a fad, so we'll just continue doing what has always worked for us." Is that the issue? It's hard to imagine a business where AI couldn't be put to use, and it can't be healthy to a company's long-term success to ignore that promise.



We're sympathetic to companies that worry about the lack of skilled people, an issue that was reported by 9.4% of nonusers and 13% of users. People with AI skills have always been hard to find and are often expensive. We don't expect that situation to change much in the near future. While experienced AI developers are starting to leave powerhouses like Google, OpenAI, Meta, and Microsoft, not enough are leaving to meet demand—and most of them will probably gravitate to startups rather than adding to the AI talent within established companies. However, we're also surprised that this issue doesn't figure more prominently. Companies that are adopting AI are clearly finding staff somewhere, whether through hiring or training their existing staff.

A small percentage (3.7% of nonusers, 5.4% of users) report that "infrastructure issues" are an issue. Yes, building AI infrastructure is difficult and expensive, and it isn't surprising that the AI users feel this problem more keenly. We've all read about the shortage of the high-end GPUs that power models like ChatGPT. This is an area where cloud providers already bear much of the burden, and will continue to bear it in the future. Right now, very few AI adopters maintain their own infrastructure and are shielded from infrastructure issues by their providers. In the long term, these issues may slow AI adoption. We suspect that many API services are being offered as loss leaders—that the major providers have intentionally set prices low to buy market share. That pricing won't be sustainable, particularly as hardware shortages drive up the cost of building infrastructure. How will AI adopters react when the cost of renting infrastructure from AWS, Microsoft, or Google rises? Given the cost of equipping a data center with high-end GPUs, they probably won't attempt to build their own infrastructure. But they may back off on AI development.

Few nonusers (2%) report that lack of data or data quality is an issue, and only 1.3% report that the difficulty of training a model is a problem. In hindsight, this was predictable: these are problems that only appear after you've started down the road to generative AI. AI users are definitely facing these problems: 7% report that data quality has hindered further adoption, and 4% cite the difficulty of training a model on their data. But while data quality and the difficulty of training a model are clearly important issues, they don't appear to be the biggest barriers to building with AI. Developers are learning how to find quality data and build models that work.

How Companies Are Using AI

We asked several specific questions about how respondents are working with AI, and whether they're "using" it or just "experimenting."

We aren't surprised that the most common application of generative AI is in programming, using tools like GitHub Copilot or ChatGPT. However, we *are* surprised at the level of adoption: 77% of respondents report using AI as an aid in programming; 34% are experimenting with it, and 44% are already using it in their work. Data analysis showed a similar pattern: 70% total; 32% using AI, 38% experimenting with it. The higher percentage of users that are experimenting may reflect OpenAI's addition of Advanced Data Analysis (formerly Code Interpreter) to ChatGPT's repertoire of beta features. Advanced Data Analysis does a decent job of exploring and analyzing datasets—though we expect data analysts to be careful about checking AI's output and to distrust software that's labeled as "beta."

Using generative AI tools for tasks related to programming (including data analysis) is nearly universal. It will certainly become universal for organizations that don't explicitly prohibit its use. And we expect that programmers will use AI even in organizations that prohibit its use. Programmers have always developed tools that would help them do their jobs, from test frameworks to source control to integrated development environments. And they've always adopted these tools whether or not they had management's permission. From a programmer's perspective, code generation is just another labor-saving tool that keeps them productive in a job that is constantly becoming more complex. In the early 2000s, some studies of open source adoption found that a large majority of staff said that they were using open source, even though a large majority of CIOs said their companies weren't. Clearly those CIOs either didn't know what their employees were doing or were willing to look the other way. We'll see that pattern repeat itself: programmers will do what's necessary to get the job done, and managers will be blissfully unaware as long as their teams are more productive and goals are being met.

After programming and data analysis, the next most common use for generative AI was applications that interact with customers, including customer support: 65% of all respondents report that their companies are experimenting with (43%) or using AI (22%) for this

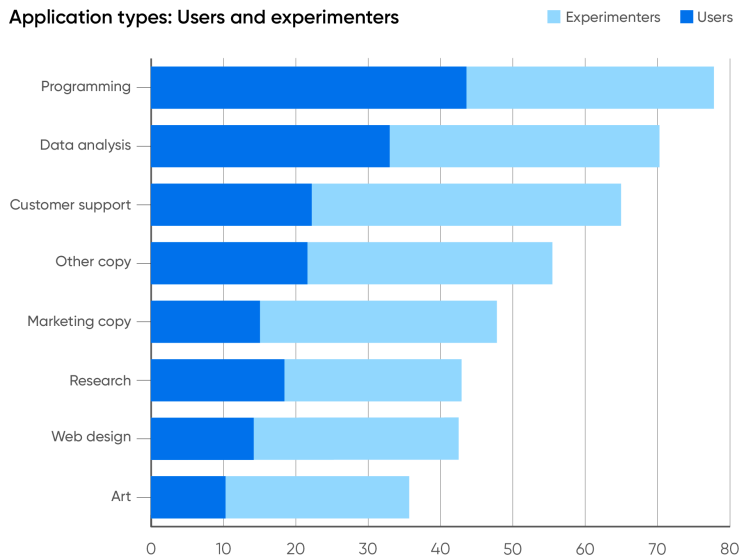
purpose. While companies have long been talking about AI's potential to improve customer support, we didn't expect to see customer service rank so high. Customer-facing interactions are very risky: incorrect answers, bigoted or sexist behavior, and many other well-documented problems with generative AI quickly lead to damage that is hard to undo. Perhaps that's why such a large percentage of respondents are experimenting with this technology rather than using it (more than for any other kind of application). Any attempt at automating customer service needs to be very carefully tested and debugged. We interpret our survey results as "cautious but excited adoption." It's clear that automating customer service could go a long way to cut costs and even, if done well, make customers happier. No one wants to be left behind, but at the same time, no one wants a highly visible PR disaster or a lawsuit on their hands.

A moderate number of respondents report that their companies are using generative AI to generate copy (written text). 47% are using it specifically to generate marketing copy, and 56% are using it for other kinds of copy (internal memos and reports, for example). While rumors abound, we've seen few reports of people who have actually lost their jobs to AI—but those reports have been almost entirely from copywriters. AI isn't yet at the point where it can write as well as an experienced human, but if your company needs catalog descriptions for hundreds of items, speed may be more important than brilliant prose. And there are many other applications for machine-generated text: AI is good at summarizing documents. When coupled with a speech-to-text service, it can do a passable job of creating meeting notes or even podcast transcripts. It's also well suited to writing a quick email.

The applications of generative AI with the fewest users were web design (42% total; 28% experimenting, 14% using) and art (36% total; 25% experimenting, 11% using). This no doubt reflects O'Reilly's developer-centric audience. However, several other factors are in play. First, there are already a lot of low-code and no-code web design tools, many of which feature AI but aren't yet using generative AI. Generative AI will face significant entrenched competition in this crowded market. Second, while OpenAI's GPT-4 announcement last March demoed generating website code from a hand-drawn sketch, that capability wasn't available until after the survey closed. Third, while roughing out the HTML and JavaScript for a simple website makes a great demo, that isn't really the

problem web designers need to solve. They want a drag-and-drop interface that can be edited on-screen, something that generative AI models don't yet have. Those applications will be built soon; tldraw is a very early example of what they might be. Design tools suitable for professional use don't exist yet, but they will appear very soon.

An even smaller percentage of respondents say that their companies are using generative AI to create art. While we've read about startup founders using Stable Diffusion and Midjourney to create company or product logos on the cheap, that's still a specialized application and something you don't do frequently. But that isn't all the art that a company needs: "hero images" for blog posts, designs for reports and whitepapers, edits to publicity photos, and more are all necessary. Is generative AI the answer? Perhaps not yet. Take Midjourney for example: while its capabilities are impressive, the tool can also make silly mistakes, like getting the number of fingers (or arms) on subjects incorrect. While the latest version of Midjourney is much better, it hasn't been out for long, and many artists and designers would prefer not to deal with the errors. They'd also prefer to avoid legal liability. Among generative art vendors, Shutterstock, Adobe, and Getty Images indemnify users of their tools against copyright claims. Microsoft, Google, IBM, and OpenAI have offered more general indemnification.



We also asked whether the respondents' companies are using AI to create some other kind of application, and if so, what. While many of these write-in applications duplicated features already available from big AI providers like Microsoft, OpenAI, and Google, others covered a very impressive range. Many of the applications involved summarization: news, legal documents and contracts, veterinary medicine, and financial information stand out. Several respondents also mentioned working with video: analyzing video data streams, video analytics, and generating or editing videos.

Other applications that respondents listed included fraud detection, teaching, customer relations management, human resources, and compliance, along with more predictable applications like chat, code generation, and writing. We can't tally and tabulate all the responses, but it's clear that there's no shortage of creativity and innovation. It's also clear that there are few industries that won't be touched—AI will become an integral part of almost every profession.

Generative AI will take its place as the ultimate office productivity tool. When this happens, it may no longer be recognized as AI; it will just be a feature of Microsoft Office or Google Docs or Adobe Photoshop, all of which are integrating generative AI models. GitHub Copilot and Google's Codey have both been integrated into Microsoft and Google's respective programming environments. They will simply be part of the environment in which software developers work. The same thing happened to networking 20 or 25 years ago: wiring an office or a house for ethernet used to be a big deal. Now we expect wireless everywhere, and even that's not correct. We don't "expect" it—we assume it, and if it's not there, it's a problem. We expect mobile to be everywhere, including map services, and it's a problem if you get lost in a location where the cell signals don't reach. We expect search to be everywhere. AI will be the same. It won't be expected; it will be assumed, and an important part of the transition to AI everywhere will be understanding how to work when it isn't available.

The Builders and Their Tools

To get a different take on what our customers are doing with AI, we asked what models they're using to build custom applications. 36% indicated that they aren't building a custom application. Instead, they're working with a prepackaged application like ChatGPT,

GitHub Copilot, the AI features integrated into Microsoft Office and Google Docs, or something similar. The remaining 64% have shifted from using AI to developing AI applications. This transition represents a big leap forward: it requires investment in people, in infrastructure, and in education.

Which Model?

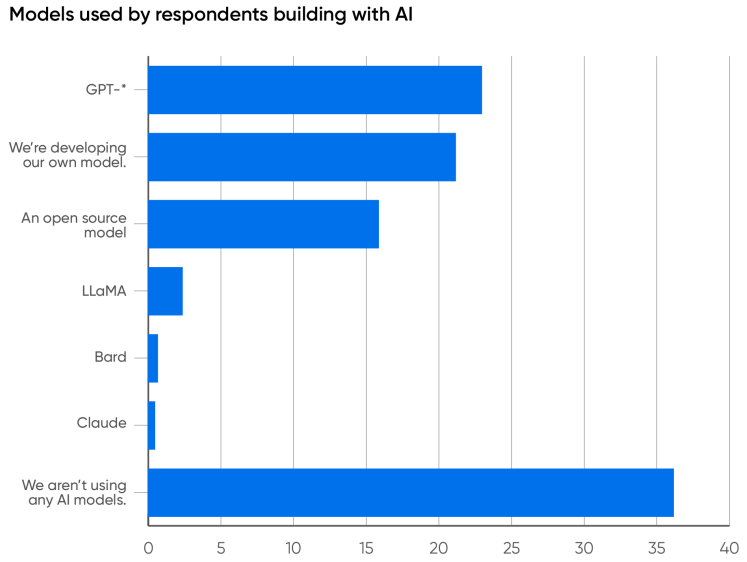
While the GPT models dominate most of the online chatter, the number of models available for building applications is increasing rapidly. We read about a new model almost every day—certainly every week—and a quick look at Hugging Face will show you more models than you can count. (As of November, the number of models in its repository is approaching 400,000.) Developers clearly have choices. But what choices are they making? Which models are they using?

It's no surprise that 23% of respondents report that their companies are using one of the GPT models (2, 3.5, 4, and 4V), more than any other model. It's a bigger surprise that 21% of respondents are developing their own model; that task requires substantial resources in staff and infrastructure. It will be worth watching how this evolves: will companies continue to develop their own models, or will they use AI services that allow a foundation model (like GPT-4) to be customized?

16% of the respondents report that their companies are building on top of open source models. Open source models are a large and diverse group. One important subsection consists of models derived from Meta's LLaMA: llama.cpp, Alpaca, Vicuna, and many others. These models are typically smaller (7 to 14 billion parameters) and easier to fine-tune, and they can run on very limited hardware; many can run on laptops, cell phones, or nanocomputers such as the Raspberry Pi. Training requires much more hardware, but the ability to run in a limited environment means that a finished model can be embedded within a hardware or software product. Another subsection of models has no relationship to LLaMA: RedPajama, Falcon, MPT, Bloom, and many others, most of which are available on Hugging Face. The number of developers using any specific model is relatively small, but the total is impressive and demonstrates a vital and active world beyond GPT. These "other" models have attracted a significant following. Be careful, though: while this group of models is frequently called "open source," many of them

restrict what developers can build from them. Before working with any so-called open source model, look carefully at the license. Some limit the model to research work and prohibit commercial applications; some prohibit competing with the model's developers; and more. We're stuck with the term "open source" for now, but where AI is concerned, open source often isn't what it seems to be.

Only 2.4% of the respondents are building with LLaMA and Llama 2. While the source code and weights for the LLaMA models are available online, the LLaMA models don't yet have a public API backed by Meta—although there appear to be several APIs developed by third parties, and both Google Cloud and Microsoft Azure offer Llama 2 as a service. The LLaMA-family models also fall into the "so-called open source" category that restricts what you can build.

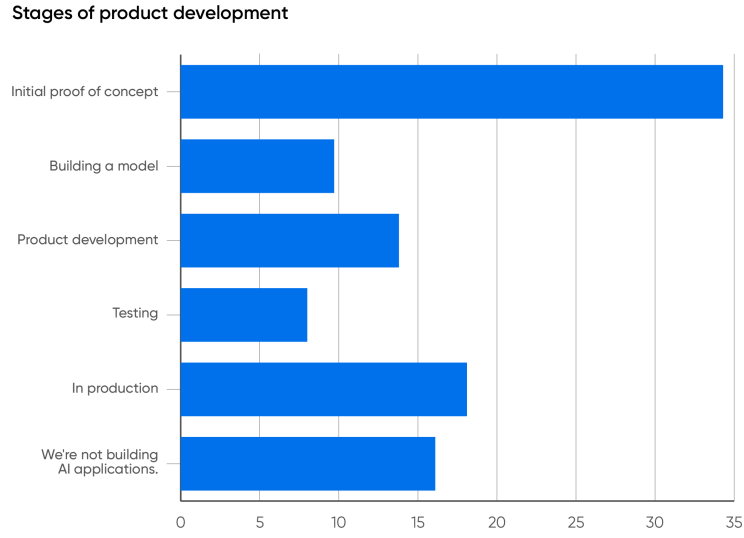


Only 1% are building with Google's Bard, which perhaps has less exposure than the others. A number of writers have claimed that Bard gives worse results than the LLaMA and GPT models; that may be true for chat, but I've found that Bard is often correct when GPT-4 fails. For app developers, the biggest problem with Bard probably isn't accuracy or correctness; it's availability. In March 2023, Google announced a public beta program for the Bard API. However, as of November, questions about API availability are still

answered by links to the beta announcement. Use of the Bard API is undoubtedly hampered by the relatively small number of developers who have access to it. Even fewer are using Claude, a very capable model developed by Anthropic. Claude doesn't get as much news coverage as the models from Meta, OpenAI, and Google, which is unfortunate: Anthropic's Constitutional AI approach to AI safety is a unique and promising attempt to solve the biggest problems troubling the AI industry.

What Stage?

When asked what stage companies are at in their work, most respondents shared that they're still in the early stages. Given that generative AI is relatively new, that isn't news. If anything, we should be surprised that generative AI has penetrated so deeply and so quickly. 34% of respondents are working on an initial proof of concept. 14% are in product development, presumably after developing a PoC; 10% are building a model, also an early stage activity; and 8% are testing, which presumes that they've already built a proof of concept and are moving toward deployment—they have a model that at least appears to work.



What stands out is that 18% of the respondents work for companies that have AI applications in production. Given that the technology is new and that many AI projects fail,[2] it's surprising that 18% report that their companies already have generative AI applications in production. We're not being skeptics; this is evidence that while most respondents report companies that are working on proofs of concept or in other early stages, generative AI is being adopted and is doing real work. We've already seen some significant integrations of AI into existing products, including our own. We expect others to follow.

Risks and Tests

We asked the respondents whose companies are working with AI what risks they're testing for. The top five responses clustered between 45 and 50%: unexpected outcomes (49%), security vulnerabilities (48%), safety and reliability (46%), fairness, bias, and ethics (46%), and privacy (46%).

It's important that almost half of respondents selected "unexpected outcomes," more than any other answer: anyone working with generative AI needs to know that incorrect results (often called hallucinations) are common. If there's a surprise here, it's that this answer wasn't selected by 100% of the participants. Unexpected, incorrect, or inappropriate results are almost certainly the biggest single risk associated with generative AI.

We'd like to see more companies test for fairness. There are many applications (for example, medical applications) where bias is among the most important problems to test for and where getting rid of historical biases in the training data is very difficult and of utmost importance. It's important to realize that unfair or biased output can be very subtle, particularly if application developers don't belong to groups that experience bias—and what's "subtle" to a developer is often very unsubtle to a user. A chat application that doesn't understand a user's accent is an obvious problem (search for "Amazon Alexa doesn't understand Scottish accent"). It's

2 Many articles quote Gartner as saying that the failure rate for AI projects is 85%. We haven't found the source, though in 2018, Gartner wrote that 85% of AI projects "deliver erroneous outcomes." That's not the same as failure, and 2018 significantly predates generative AI. Generative AI is certainly prone to "erroneous outcomes," and we suspect the failure rate is high. 85% might be a reasonable estimate.

also important to look for applications where bias isn't an issue. ChatGPT has driven a focus on personal use cases, but there are many applications where problems of bias and fairness aren't major issues: for example, examining images to tell whether crops are diseased or optimizing a building's heating and air conditioning for maximum efficiency while maintaining comfort.

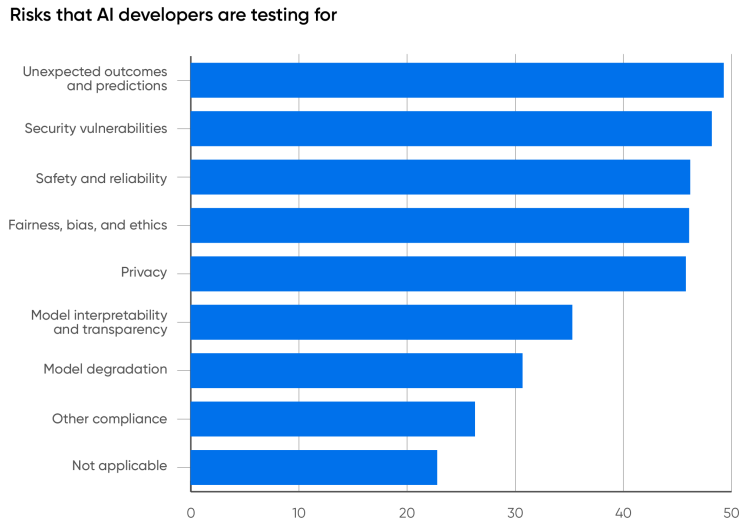
It's good to see issues like safety and security near the top of the list. Companies are gradually waking up to the idea that security is a serious issue, not just a cost center. In many applications (for example, customer service), generative AI is in a position to do significant reputational damage, in addition to creating legal liability. Furthermore, generative AI has its own vulnerabilities, such as prompt injection, for which there is still no known solution. Model leeching, in which an attacker uses specially designed prompts to reconstruct the data on which the model was trained, is another attack that's unique to AI. While 48% isn't bad, we would like to see even greater awareness of the need to test AI applications for security.

Model interpretability (35%) and model degradation (31%) aren't as big concerns. Unfortunately, interpretability remains a research problem for generative AI. At least with the current language models, it's very difficult to explain why a generative model gave a specific answer to any question. Interpretability might not be a requirement for most current applications. If ChatGPT writes a Python script for you, you may not care why it wrote that particular script rather than something else. (It's also worth remembering that if you ask ChatGPT why it produced any response, its answer will not be the reason for the previous response, but, as always, the most likely response to your question.) But interpretability is critical for diagnosing problems of bias and will be extremely important when cases involving generative AI end up in court.

Model degradation is a different concern. The performance of any AI model degrades over time, and as far as we know, large language models are no exception. One hotly debated study argues that the quality of GPT-4's responses has dropped over time. Language changes in subtle ways; the questions users ask shift and may not be answerable with older training data. Even the existence of an AI answering questions might cause a change in what questions are asked. Another fascinating issue is what happens when generative models are trained on data generated by other generative models.

Is "model collapse" real, and what impact will it have as models are retrained?

If you're simply building an application on top of an existing model, you may not be able to do anything about model degradation. Model degradation is a much bigger issue for developers who are building their own model or doing additional training to fine-tune an existing model. Training a model is expensive, and it's likely to be an ongoing process.

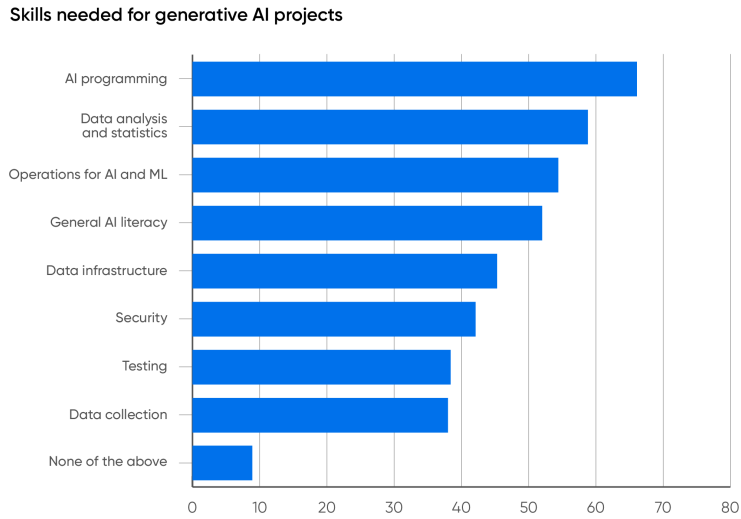


Missing Skills

One of the biggest challenges facing companies developing with AI is expertise. Do they have staff with the necessary skills to build, deploy, and manage these applications? To find out where the skills deficits are, we asked our respondents what skills their organizations need to acquire for AI projects. We weren't surprised that AI programming (66%) and data analysis (59%) are the two most needed. AI is the next generation of what we called "data science" a few years back, and data science represented a merger between statistical modeling and software development. The field may have evolved from traditional statistical analysis to artificial intelligence, but its overall shape hasn't changed much.

The next most needed skill is operations for AI and ML (54%). We're glad to see people recognize this; we've long thought that operations was the "elephant in the room" for AI and ML. Deploying and managing AI products isn't simple. These products differ in many ways from more traditional applications, and while practices like continuous integration and deployment have been very effective for traditional software applications, AI requires a rethinking of these code-centric methodologies. The model, not the source code, is the most important part of any AI application, and models are large binary files that aren't amenable to source control tools like Git. And unlike source code, models grow stale over time and require constant monitoring and testing. The statistical behavior of most models means that simple, deterministic testing won't work; you can't guarantee that, given the same input, a model will generate the same output. The result is that AI operations is a specialty of its own, one that requires a deep understanding of AI and its requirements in addition to more traditional operations. What kinds of deployment pipelines, repositories, and test frameworks do we need to put AI applications into production? We don't know; we're still developing the tools and practices needed to deploy and manage AI successfully.

Infrastructure engineering, a choice selected by 45% of respondents, doesn't rank as high. This is a bit of a puzzle: running AI applications in production can require huge resources, as companies as large as Microsoft are finding out. However, most organizations aren't yet running AI on their own infrastructure. They're either using APIs from an AI provider like OpenAI, Microsoft, Amazon, or Google or they're using a cloud provider to run a homegrown application. But in both cases, some other provider builds and manages the infrastructure. OpenAI in particular offers enterprise services, which includes APIs for training custom models along with stronger guarantees about keeping corporate data private. However, with cloud providers operating near full capacity, it makes sense for companies investing in AI to start thinking about their own infrastructure and acquiring the capacity to build it.



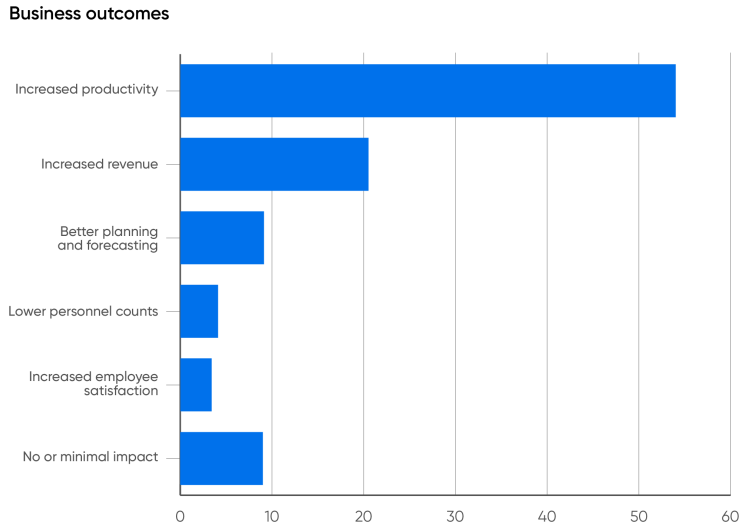
Over half of the respondents (52%) included general AI literacy as a needed skill. While the number could be higher, we're glad that our users recognize that familiarity with AI and the way AI systems behave (or misbehave) is essential. Generative AI has a great wow factor: with a simple prompt, you can get ChatGPT to tell you about Maxwell's equations or the Peloponnesian War. But simple prompts don't get you very far in business. AI users soon learn that good prompts are often very complex, describing in detail the result they want and how to get it. Prompts can be very long, and they can include all the resources needed to answer the user's question. Researchers debate whether this level of prompt engineering will be necessary in the future, but it will clearly be with us for the next few years. AI users also need to expect incorrect answers and to be equipped to check virtually all the output that an AI produces. This is often called critical thinking, but it's much more like the process of discovery in law: an exhaustive search of all possible evidence. Users also need to know how to create a prompt for an AI system that will generate a useful answer.

Finally, the Business

So what's the bottom line? How do businesses benefit from AI? Over half (54%) of the respondents expect their businesses to benefit from increased productivity. 21% expect increased revenue, which might

indeed be the result of increased productivity. Together, that's three-quarters of the respondents. Another 9% say that their companies would benefit from better planning and forecasting.

Only 4% believe that the primary benefit will be lower personnel counts. We've long thought that the fear of losing your job to AI was exaggerated. While there will be some short-term dislocation as a few jobs become obsolete, AI will also create new jobs—as has almost every significant new technology, including computing itself. Most jobs rely on a multitude of individual skills, and generative AI can only substitute for a few of them. Most employees are also willing to use tools that will make their jobs easier, boosting productivity in the process. We don't believe that AI will replace people, and neither do our respondents. On the other hand, employees will need training to use AI-driven tools effectively, and it's the responsibility of the employer to provide that training.



We're optimistic about generative AI's future. It's hard to realize that ChatGPT has only been around for a year; the technology world has changed so much in that short period. We've never seen a new technology command so much attention so quickly: not personal computers, not the internet, not the web. It's certainly possible that we'll slide into another AI winter if the investments being made in generative AI don't pan out. There are definitely problems that need to be solved—correctness, fairness, bias, and security are among

the biggest—and some early adopters will ignore these hazards and suffer the consequences. On the other hand, we believe that worrying about a general AI deciding that humans are unnecessary is either an affliction of those who read too much science fiction or a strategy to encourage regulation that gives the current incumbents an advantage over startups.

It's time to start learning about generative AI, thinking about how it can improve your company's business, and planning a strategy. We can't tell you what to do; developers are pushing AI into almost every aspect of business. But companies will need to invest in training, both for software developers and for AI users; they'll need to invest in the resources required to develop and run applications, whether in the cloud or in their own data centers; and they'll need to think creatively about how they can put AI to work, realizing that the answers may not be what they expect.

AI won't replace humans, but companies that take advantage of AI will replace companies that don't.

Appendix

Methodology and Demographics

This survey ran from September 14, 2023, to September 27, 2023. It was publicized through O'Reilly's learning platform to all our users, both corporate and individuals. We received 4,782 responses, of which 2,857 answered all the questions. As we usually do, we eliminated incomplete responses (users who dropped out part way through the questions). Respondents who indicated they weren't using generative AI were asked a final question about why they weren't using it, and considered complete.

Any survey only gives a partial picture, and it's very important to think about biases. The biggest bias by far is the nature of O'Reilly's audience, which is predominantly North American and European. 42% of the respondents were from North America, 32% were from Europe, and 21% percent were from the Asia-Pacific region. Relatively few respondents were from South America or Africa, although we are aware of very interesting applications of AI on these continents.

The responses are also skewed by the industries that use our platform most heavily. 34% of all respondents who completed the survey were from the software industry, and another 11% worked on computer hardware, together making up almost half of the respondents. 14% were in financial services, which is another area where our platform has many users. 5% of the respondents were from telecommunications, 5% from the public sector and the government, 4.4% from the healthcare industry, and 3.7% from education. These are still healthy numbers: there were over 100 respondents in each group. The remaining 22% represented other industries, ranging from mining (0.1%) and construction (0.2%) to manufacturing (2.6%).

These percentages change very little if you look only at respondents whose employers use AI rather than all respondents who completed the survey. This suggests that AI usage doesn't depend a lot on the specific industry; the differences between industries reflects the population of O'Reilly's user base.

About the Author

Mike Loukides is vice president of content strategy for O'Reilly Media, Inc. He's edited many highly regarded books on technical subjects that don't involve Windows programming. He's particularly interested in programming languages, Unix and what passes for Unix these days, and system and network administration. Mike is the author of *System Performance Tuning* and a coauthor of *Unix Power Tools*. Most recently, he's been fooling around with data and data analysis, exploring languages like R, Mathematica, and Octave, and thinking about how to make books social. Mike can be reached on Twitter as @mikeloukides and on LinkedIn.