A TECHNOLOGY LEADER'S GUIDE TO

Full-stack generative AI for the enterprise

WRITER

Table of contents

If you're wondering when your company should adopt generative AI solutions, the answer is "yesterday."

You're being hit from all sides: leadership, employees, and the company's board: What's the plan? And you likely know the repercussions of slow adoption. From a business standpoint, your competitors will fly ahead of you if they adopt AI and you don't. And from an investment standpoint, you could have already achieved a return on investment in the timeframe you chose to put off the project.

Companies can't afford to push generative AI adoption into the future. Your strategy can't be a 3-year roadmap. Instead, you want to think about what you can get done in the next six months to a year. You can start with small use cases, in specific groups within the organization, and then expand from there.

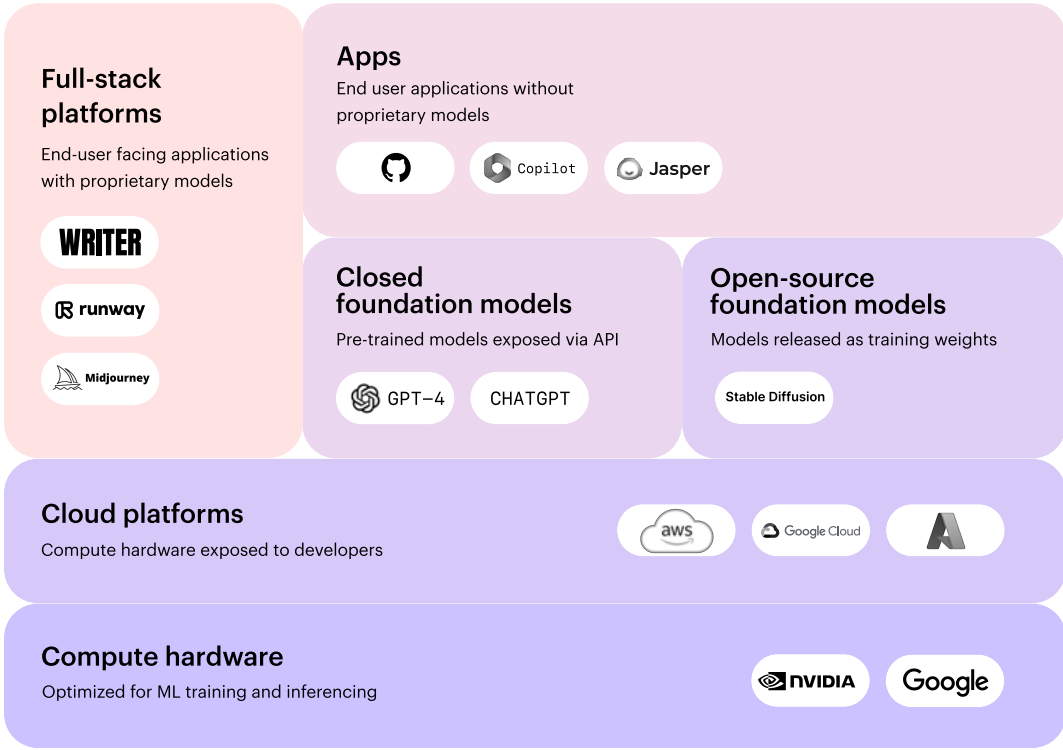
Yet the pace at which generative AI is developing is also overwhelming. You have to find the line between cutting-edge and racing to a solution that ultimately isn't sustainable and won't serve your business needs. It's about balancing scalability and risk: implementing a platform that can meet your data protection and quality needs.

If you're not sure where to start, this guide has got you covered. We'll walk you through the landscape of generative AI technology on the market today, what to consider when evaluating vendors, and how to establish a clear path to ROI.

What are AI foundation models?

You've probably heard — and seen! — that AI is only as good as the data it's trained on. Foundation models are the AI models trained on massive datasets. They're the backbone on which companies build generative AI applications, either for their own internal use or to license to others.

In the world of generative AI, large language models (LLMs) are one type of foundation model that rely on text data. Other foundation models rely on images, audio, and other types of inputs.



01

Open-source foundation models

Open-source foundation models are designed for a wide variety of use cases. Startups and tech companies can take an open-source foundation model, such as Stable Diffusion, and use it to build and train their own AI applications.

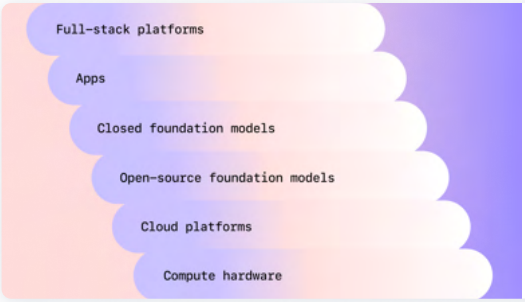
Writer makes its family of foundation models, Palmyra LLMs, inspectable. Additionally, smaller Palmyra LLMs are available open-source through Hugging Face.

02

Closed-source foundation models

Closed-source foundation models have developed their own technology. They make their models available via API for other companies to use. The most well-known closed-source foundation model is OpenAI's GPT-4.

The difference between open-source and closed-source is transparency. With open-source, you can dig into (and customize) the source code that powers the AI foundation model. With closed-source, you can make use of the data, but can't see the "nuts and bolts" of how the company trained the model and what data was used.



Learn more: Making sense of AI foundation models

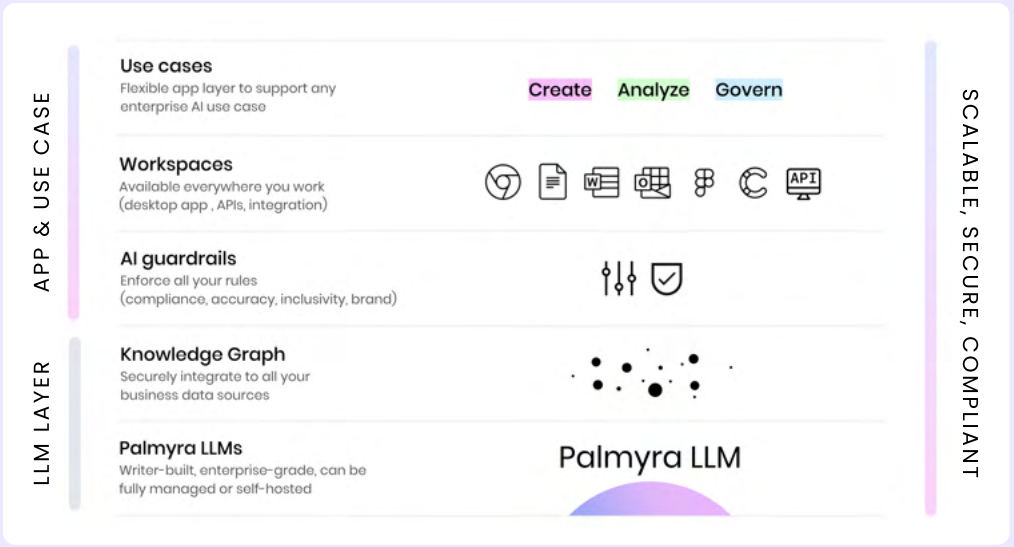
What's full-stack generative AI for enterprise?

A full-stack generative AI platform combines a foundation model and knowledge retrieval solution with an integrated, secure application layer. It's the front-end and the back-end (versus using a closed-source or open-source foundation model and creating an application on top of that). With full-stack platforms, the developer has full control of both the model's output and the user's interaction with the app.

Many generative AI tools, such as Jasper AI, GitHub Copilot, and Microsoft Copilot, are in a category many experts call “AI wrappers” — a front-end application built on top of an existing, closed LLM back-end (OpenAI’s GPT4). They’re not full-stack platforms, because developers can only control the user interfaces and not the LLM itself. AI wrappers are difficult to customize and scale for enterprise use cases due to this lack of control and transparency.

Full-stack generative AI has to cover the entire depth of business needs. It goes beyond simply a foundation model plus an application. A full-stack platform can handle the use cases of an organization, including its data, security, and other operational needs.

Full-stack generative AI



01 A large language model

Generative AI for enterprise companies shouldn't be the same as LLMs for the general population.

You need to start with an LLM that's trained on just formal and business writing rather than the mediocrity of the internet. It should have filtered out datasets that shouldn't be included, such as copyrighted material. Business-grade LLMs should contain industry-specific language and meet compliance requirements. They're also built for advanced skills, rather than a breadth of knowledge.

Further, an LLM designed for enterprise companies will be highly accurate. None of the "AI hallucinations" that have become fodder for internet jokes at best and draw intense skepticism at worst. These LLMs will also mitigate bias and toxic output.

02 Safe access to business data

In addition to the LLM's standard dataset, you should be able to supplement and train the LLM with your business data.

You should be able to connect the AI platform to your company's data, including knowledge bases, databases, cloud storage platforms, and more. This increases the accuracy of the output, since it adds your business context. Hundreds (if not thousands) of files can be stitched together, resulting in the ability to analyze, summarize, and repurpose your business content.

A popular methodology to populate LLMs with data from external sources is retrieval-augmented generation, or RAG. The LLMs can generate better output because RAG provides relevant information in the prompt to LLMs. For enterprise generative AI, RAG can be used to access your internal knowledge and business data. RAG feeds the data to the LLM and the LLM is more accurate. It solves a common business challenge of accessing business knowledge, particularly in large organizations with a lot of documentation.

Since the AI platform will be accessing potentially sensitive information, it's also imperative that your data isn't intermingled with another client's training model and isn't used for model training. Your data should remain separate and only accessible to your users.

03 Safe access to business data

If the platform's application isn't easy to use, your users will be less likely to adopt it. There are two areas of consideration when discussing ease of use in generative AI for the enterprise:

- **A flexible application layer that meets any bespoke use case.** This could mean a chat interface or other UI options that are tailored to your use cases. For example, a retailer may have very structured requirements for generating product descriptions, which will call for bespoke input/output criteria in the form of a custom template.
- **An ecosystem of extensions and APIs.** Extensions for products like Chrome, Google Docs, Microsoft Word, and Outlook will bring generative AI to the places where people already work.

Flexibility will meet the needs of your team and their use cases, whether it's an extension, a chat interface, a custom template, or connecting other applications through an API. With a wide range of generative AI use cases, there's not a one-size-fits-all solution. A chat interface, for example, can't be your only option: it doesn't serve all use cases and doesn't scale.

04 Guardrails for AI output

If AI output doesn't meet your business's legal, regulatory, and brand guidelines, it would require a lot of additional editing from your users. In addition to offsetting any time saved, it could create legal issues or damage your brand's reputation.

Full-stack generative AI should cross-check the output against internal data. Anything that seems misaligned for your industry, or specific legal or regulatory requirements, should be flagged.

Further, you should be able to configure the solution with your brand's style guide and specific terminology. This ensures that everyone's work is accurate and on-brand. It also removes manual editing and lengthy review cycles.

05 A way to ensure security, privacy, and governance

When you're integrating sensitive business data into an LLM, data privacy and security have to be at the forefront. A full-stack application should be able to validate its claims of data governance with certifications, such as SOC 2 Type II, HIPAA, PCI, GDPR, and Privacy Shield.

You should be able to govern secure access to a full-stack platform with controls like SSO, role-based permissions, and multi-factor authentication. Activity audits and reports will verify user activity and their access within the product.

You want to govern acceptable use cases within the organization and make sure that users don't use the product for something outside of those use cases.

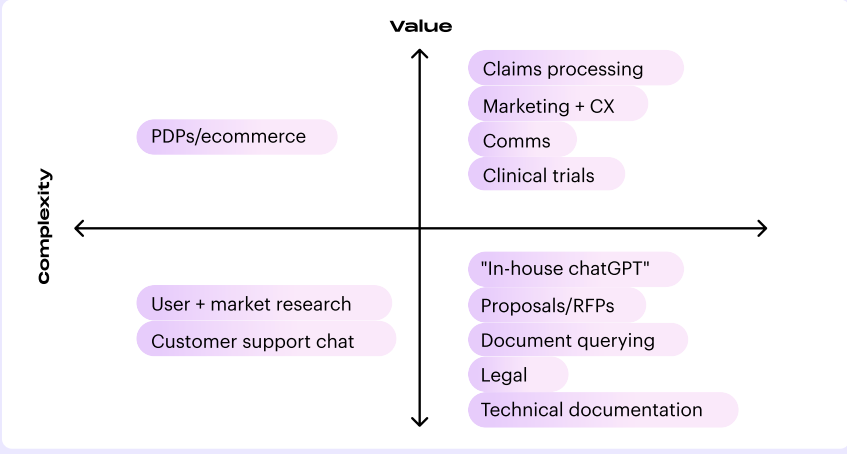
Learn more: What is full-stack generative AI?

How should enterprise teams evaluate use cases for AI?

If you think about every potential use case for generative AI within your organization, you'll be overwhelmed. Full AI adoption across every team is daunting: a future state, rather than a starting point.

Instead, you should start with a single use case that can prove the effectiveness of generative AI.

Mapping use cases to tech *in the enterprise*



01

Identify a strong AI use case

Not every AI use case is a good test case. As you make a list of potential use cases, you'll want to cross out those that might only impact a few employees or be really complicated to implement. You can tackle those further down the road, once your teams are more comfortable with AI.

A strong AI use case has high value and low complexity. You want a specific outcome (such as time saved) without overly complex technical requirements.

02

Determine the right solution for that use case

Based on your use case, you'll identify the type of AI technology needed (in this case, generative AI).

The right solution will integrate with your current tools via extensions or APIs. This will make it easy for users to adopt AI into their workflows.

A solution should also have output that matches your business needs. Do you work in Microsoft Word? Need some Google Slides or reports? You shouldn't need to change your current output based on the limitations of the solution.

03

Start with a pilot project

Even if your initial AI use case will impact a large team, you should limit the number of people involved (at least in the beginning). Select a group of users who are eager to try something new.

An initial pilot project should not be a lengthy process. You want to gather feedback from your test group quickly so you can iterate on the results. You should also clearly identify the outcome, such as faster content generation or a shorter review cycle.

Once you've completed the pilot project, you can roll out the initial use case to the rest of the team and start to implement additional use cases.

Learn more: How enterprise companies can map generative AI use cases for fast, safe implementation

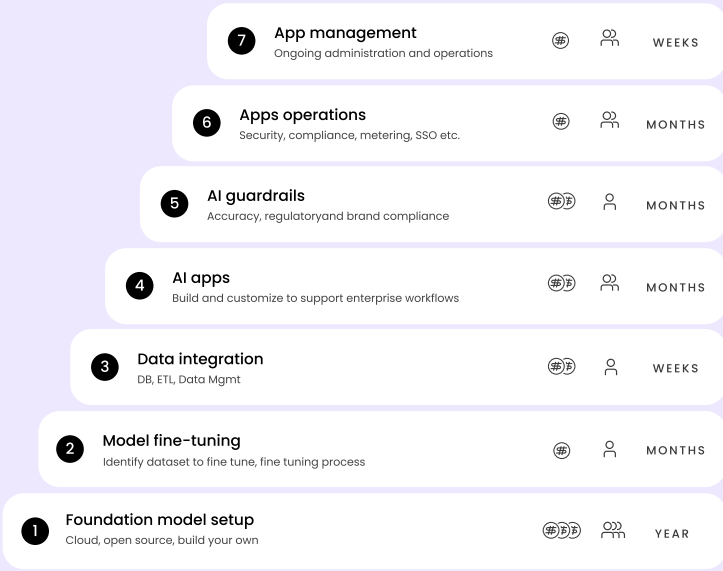
BUILD VS BUY

Should enterprise companies build their own AI solutions?

If your enterprise company has a lot of highly specific AI use cases, you might be considering building a custom solution, versus licensing a full-stack solution.

While this is certainly possible with some of the open-source and closed-source foundation models available, the actual development of a custom solution has a lot of complexities. You'd need to securely integrate your business data with the LLM, and then train the model on that data, in addition to its own dataset.

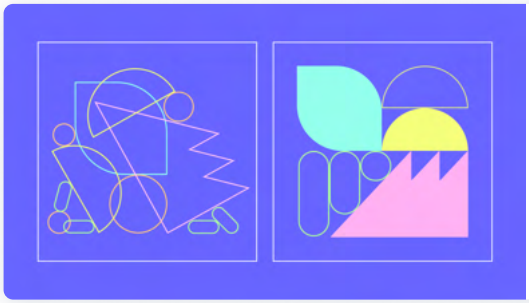
Total cost of ownership analysis: DIY vs Writer full-stack AI solution



Building an application would involve several additional components, such as hosting, logging, API plugins, and more — on top of the foundational LLM you choose. You'd need to add the guardrails for legal and compliance requirements, alongside your brand guidelines. Micro-features or microservices would need to be integrated into the application to meet these specific use cases.

All of this would require a lot of technical expertise, specifically with AI applications. Since AI has flooded the market, many tech companies and startups are eager to hire engineers with AI expertise. Finding the right talent for the human capital and labor involved could be tricky.

If a full-stack platform can customize your business use cases and workflows, you can implement generative AI in short order. The hard work has already been done for you: you'll save the headache and cost of trying to build your own.



Learn more: Build vs buy: what's the best solution for your enterprise generative AI program?

How should generative AI vendors be evaluated?

Once you've identified a strong initial AI use case, you can start evaluating solutions.

As you create an RFP and begin discussions with potential vendors, here are some things to keep in mind.

Remember: your initial AI use cases should guide your vendor selection, in terms of business needs. But you should also consider scalability across the organization. Your chosen vendor needs to be able to support additional use cases. If the vendor is too narrow in its application or only serves a single function, you could end up needing additional vendors (or switching platforms — a giant headache for everyone involved) in the future.

Learn more: How to evaluate LLM and generative AI vendors for enterprise solutions

- What is the technical architecture of the solution?
- How is it deployed?
- How does the vendor approach data lifecycle management?
- How does the vendor use data to train the LLM?
- Does the product have enterprise-grade security?
- Can the product meet legal and regulatory compliance requirements?
- How does it support your internal data?
- How does it support your range of use cases?
- What customization and integration features are available?
- What monitoring and reporting options are available?
- What is the cost of the solution?
- What support will the vendor provide to your organization and users?

How can you measure ROI?

As you evaluate your AI use cases, the cost of a full-stack generative AI solution, and the business impact, you'll be wondering: how can we measure ROI?

Your ROI will be based, in large part, on your effectiveness in implementing your AI use cases. It's not simply a matter of what you do, but how you do it.

Develop detailed project planning

You've already outlined your AI use cases. Next, you'll identify the steps and timeline to implement AI – the "nuts and bolts" within that specific pilot project.

Think about the who, the what and the when: who will the AI solution in this pilot project, what will they be replacing within their current process, and when in their workflow will this occur.

Enable all employees to use AI

While your first use case may be part of a single unit, you want to find ways to add AI throughout the organization.

You need to give employees the guidelines and resources they need to use an AI solution. Otherwise, they might use AI tools on their own that don't comply with your company's security and compliance requirements.

You'll also want to schedule trainings and workshops. Not all employees will feel comfortable diving in and testing out potential use cases. Give employees the opportunity to share how they've used AI so they can learn from each other.

Provide change management guidance

Make it clear to employees that your AI use cases on Day One and Day One Hundred will be different.

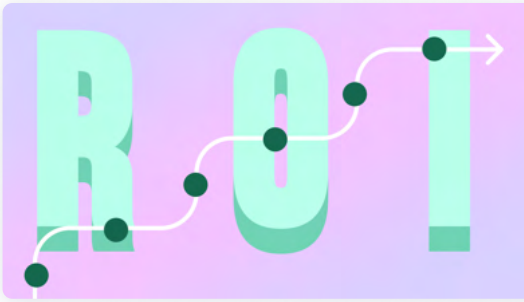
As AI is rolled out across the organization, you'll continue to guide employees through implementation. Provide internal resources that clarify how AI will be used across different business units. These resources should be kept up-to-date as additional use cases are adopted.

Prepare ROI impact analysis

Once you've implemented an AI use case, you can measure ROI. Some key factors include:

- Time saved
- Reduced reliance on other teams, such as legal or compliance review
- Increased output without sacrificing quality
- Accelerated time to market

You can also evaluate the number of employees using AI and the volume of output — both of which should increase over time.



Learn more: The six-step path to ROI for generative AI

Adopt an agile strategy and mindset

A traditional method of software development is called “waterfall.” You move through a project in a linear progression, finishing one before starting another.

Now it’s more common for teams to be agile: multiple, related projects in shorter cycles, all working toward the same goal.

Borrowing these concepts, AI adoption should be an agile approach. Waterfall won’t work: you can’t afford the wasted time on a slow, linear pace.

By following the considerations in this guide, you can take an agile approach to adopting full-stack generative AI across multiple teams within your organization.

WRITER

About Writer

Writer is the full-stack generative AI platform for enterprises. We empower your entire organization to accelerate growth, increase productivity, and ensure compliance.

Writer transforms work by delivering high-quality outputs that are accurate, compliant, and on-brand. Our platform consists of Writer-built LLMs, a Knowledge Graph that connects our models to your internal data sources, AI guardrails to enforce your rules, a flexible application layer, and an ecosystem of robust APIs and integrations. Our enterprise-grade platform doesn't use your data in model training and complies with SOC 2 Type II, HIPAA, PCI, GDPR, and Privacy Shield.