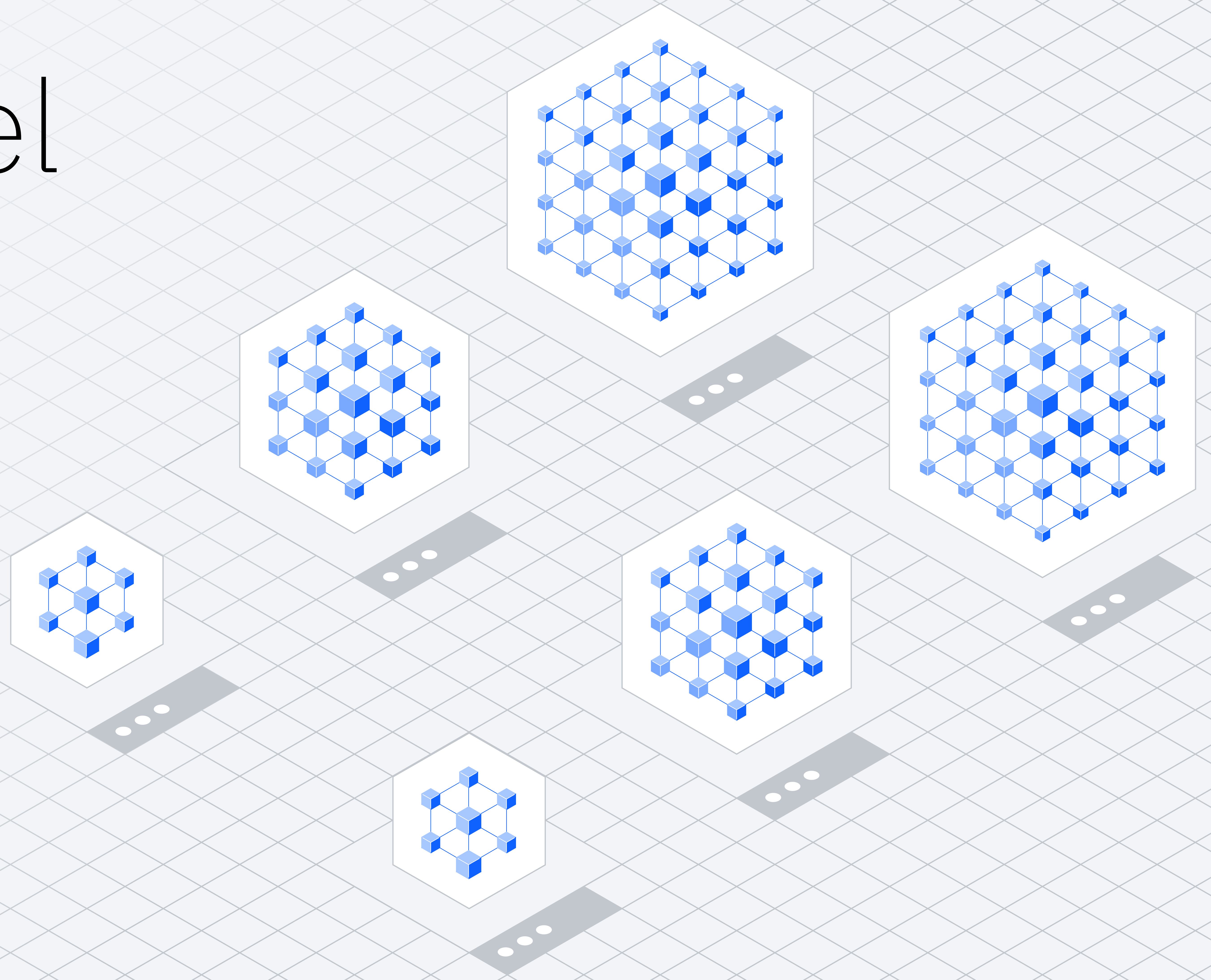
How to choose the right AI foundation model



Contents

Introduction

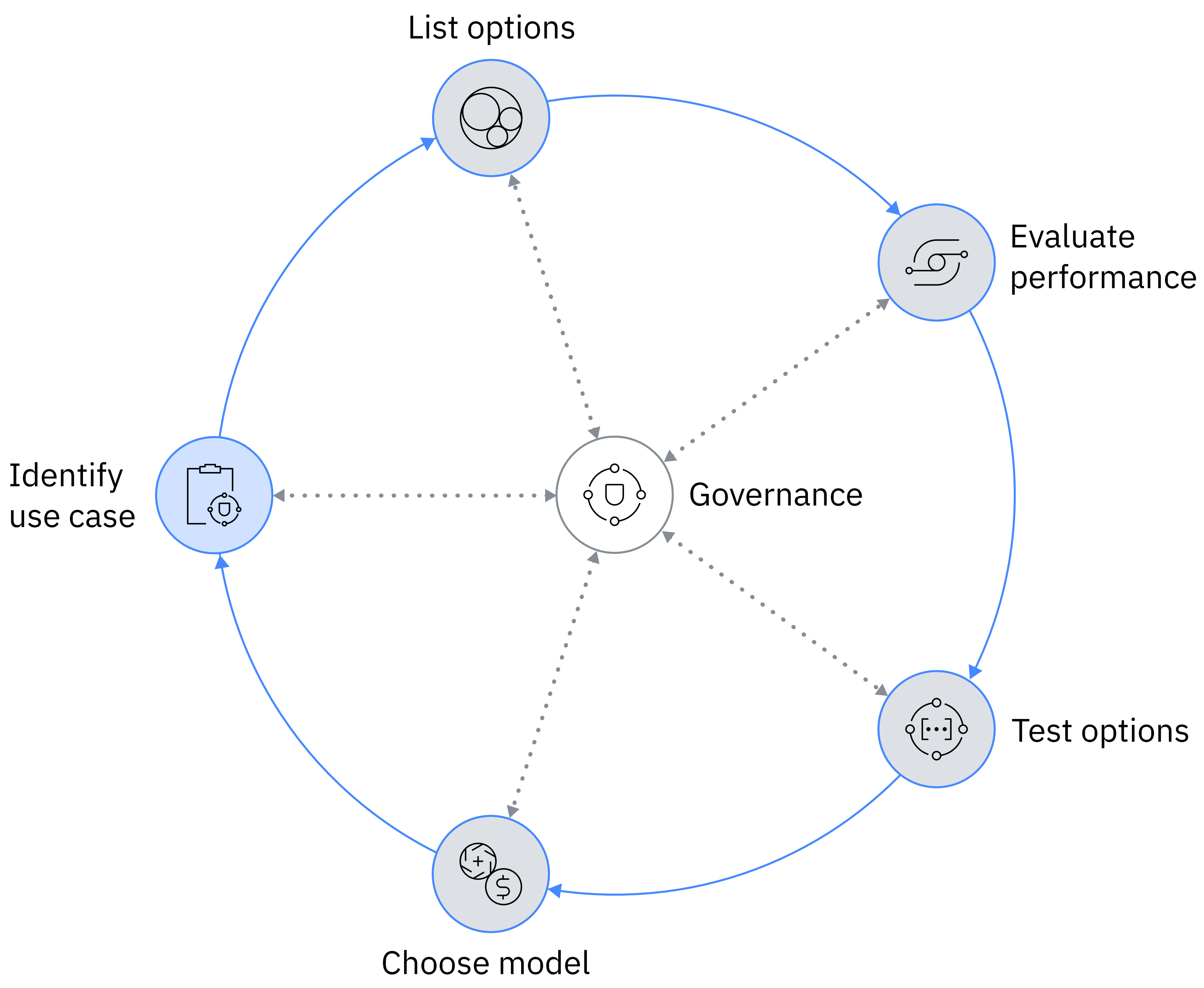
While most organizations are clear about the outcomes they expect from generative AI, what's not so well understood is the way to go about realizing these outcomes. Different outcomes require different approaches—in terms of the data sets you prepare and the AI models you employ.

Choosing the wrong model can severely impact all aspects of business—from finance, strategy and legal to even your workforce. Risks can range from biases originating from the training data or algorithms of a specific model, to a faulty outcome from an upstream model that could snowball into a major issue, potentially leading to lawsuits and reputational damage.

An evaluation framework will help you consider the diverse needs and skill sets of all kinds of AI decision makers and users in your organization. The end users of AI models could range from data scientists and machine learning engineers to business analysts, legal and compliance teams, and decision-makers. It's important to take all of their specific requirements into account so you can identify the right model for each use case and successfully implement AI to drive ROI.

AI model selection framework

This five-step cyclical process for model selection has governance at its core.



Clearly articulate your use case
Text generation, for example. Let's say you want the AI to write personalized emails for a sales or marketing campaign.

List all model options and identify each model’s size, performance and risks
Let's compare two models designed for text generation for this example: A 70B general purpose large model and a 13B specialized general purpose model.

Evaluate model size, performance and risks
One of the downsides of the 70B general-purpose model is its slower speed; however it is highly accurate. The 13B specialized model is faster than the 70B general-purpose model, but accuracy goes down since it's trained on a smaller data set. Risk is also a huge consideration at this phase for either model.

Test options
Select the model that's likely to deliver the desired output. Prompt to see if it works. Assess model performance and the quality of the output using metrics such as perplexity or BLEU score. Try to achieve the same performance with smaller models using techniques such as prompt engineering and model tuning and refine your selection to optimize for cost and your deployment requirements. Add your datasets to the stack and prompt the models to improve accuracy.

Choose the option that provides most value
Whether you need high speed or high accuracy depends on the use case, your cost constraints and deployment methods. In the end you might select The 13B specialized model because the most important aspect is generating text quickly.

In the following chapters, we'll examine in detail the key considerations for selecting an AI model: Use case, size, performance, risk, cost, and deployment needs.

Identify a clear use case

When assessing AI models, one of the key evaluation criteria is how well the model's functionalities intersect with your use case requirements.

The easiest way to identify the right model for your use case is to craft the prompt and ideal answer, and then work backwards from there to find the data needed to provide the desired answer.

Work closely with the product and engineering teams and the business sponsors to understand the actual prompts you'd need to solve the business problems at hand. Factor in the specifics of your business—such as industry terminology and standard definitions—as part of the prompt experience and make the prompts as the specific to the use cases as possible. Can you break down the use case into prompts? Can you build the particulars of your business into the prompts at multiple points?

All of this matters as even subtle nuances in the prompts can make a big difference when it comes to selecting a very specific model. To illustrate this point, consider the prompt "Give me John Doe's ID number." In an HR function, it would mean the employee ID, but in customer service, it could be the customer loyalty number. So, the more specific the prompt is, the more accurate the response. And a more specific prompt could lead you to selecting a model that already knows how to distinguish between different personal IDs, without needing any additional training.

Finding the right fit
When evaluating if a model is right for you, review the model card to see if there's a model that has been trained on data specifically for your purpose. Pre-trained foundation models are fine tuned for specific use cases, such as sentiment analysis or document summarization, enabling your team to use zero-shot prompting to obtain the desired results. This approach also allows for rapid experimentation with targeted, domain-specific models. As it takes less internal training and expertise to adapt the models to your needs, you can achieve accelerated time to value and build competitive advantage.

Generative AI use cases with the highest anticipated promise for organizations, according to IDC.[1]

- Knowledge management (46%)
- Conversational applications (46%)
- Design applications (44%)
- Code generation applications (43%)
- Marketing applications (36%)

Most popular domain applications of AI in the present scenario, according to IDC.[1]

- Software development (29.4%)
- Product development/design (24.7%)
- Customer engagement (23.4%)
- Supply chain (20.5%)
- Finance (18.2%)
- Sales (18%)
- R&D (15.4)
- HR (15.2%)
- Manufacturing (15%)
- Marketing/PR (13.9%)

Ask yourself: Do I need a massive model to execute my task?

↓

The answer, in most cases, is no. The better approach is to right-size the model for the specific use case you have.

Evaluate size, performance and risks

Right-sizing a model to your use case

Large AI models perform so well because they are trained on massive amounts of data. The sheer size of these models' training data enables them to capture the complex patterns and nuanced connections within data and generate high-quality output. However, they need more processing power, cost more to run, and aren't always more accurate. So, the question you have to ask yourself is, do I need a massive model to execute my task? The answer, in most cases, is no.

The best approach here is to right-size the model for the specific use case you have. Start out with the biggest or the best-performing model and use a basic prompt to get your ideal performance. Now scale down to a smaller model and use techniques such as prompt tuning to see if you can get the same results. Prompt-tuning a smaller model is far more cost-efficient than fine-tuning that same model—which requires considerable data and computing resources—while still achieving accurate and contextually relevant responses.

If you do choose to go with a larger general-purpose LLM for, say, a resume screening use case, you can apply HR-specific prompting and model tuning to get to the results you need. Effective model tuning techniques include applying prompt tuning with use case-specific data to achieve optimal results. In the earlier example of a general-purpose LLM that doesn't specialize in HR, altering the prompt from "What are our annual onboarding costs over the past 10 years?" to "What are our annual new hire onboarding costs over the past 10 years?" could get you a more accurate and relevant answer. The former prompt could have returned information related to new client onboarding instead of new hire onboarding.

Evaluating model performance

When evaluating a model for performance, the key criteria are accuracy, reliability and speed. The weightage accorded to each of these criteria varies from one use case to another. A trade-off between these factors is often necessary, depending on the specific needs of the use case.

Accuracy
Accuracy denotes how close the generated output is to the desired output and can be measured objectively and repeatedly. It's the primary way of evaluating a model—against industry benchmarks. Choose evaluation metrics that are relevant to your use cases. For instance, content summarization or RAG performance can be assessed by ROUGE (Recall-Oriented Understudy for Gisting Evaluation) while BLEU (Bilingual Evaluation Understudy) indicates the quality of text translation.

Reliability
Reliability is a measure of how best the model generates the same output. It is a function of several factors, such as consistency, explainability and trustworthiness, as well as how well a model avoids toxicity (hate speech, harmful language) and bias. It's a crucial consideration, especially in the case of external-facing applications. Typically, a model that offers transparency into its training methodology and data tends to be more reliable as it addresses key issues including governance, risk and privacy concerns.

Speed
Speed is about how quickly a user gets an answer to a question. It's especially critical in the case of real-time applications that demand a low-latency response. Consider use cases ranging from chatbots to financial trading—where timely answers can make all the difference—compared to financial forecasting where accuracy is more important. The speed-accuracy trade-off is a crucial consideration here and prioritizing one factor over the other is a decision that's dependent entirely on the use case at hand.

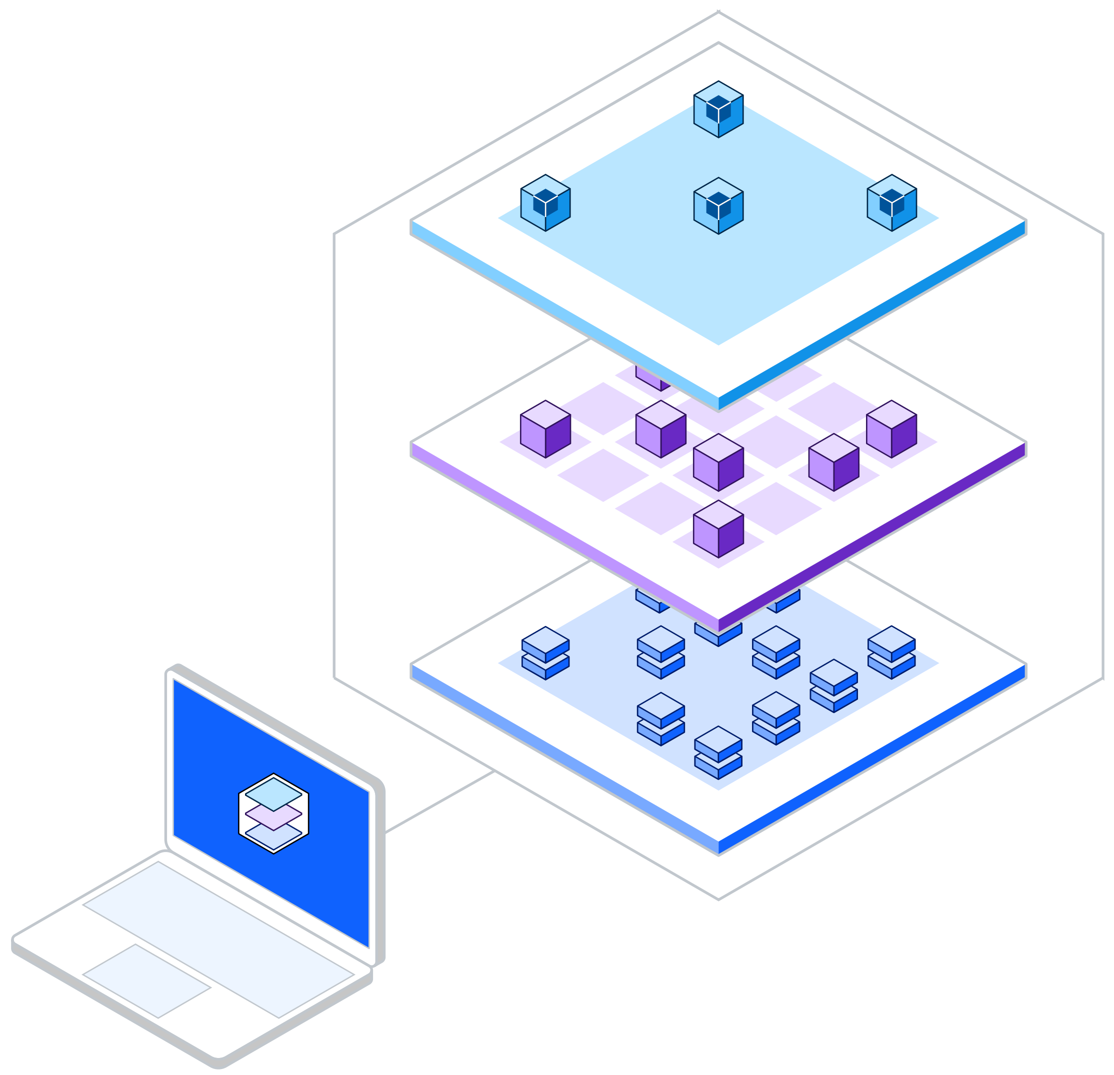
Evaluating risks and governance

For any organization, data privacy and security are key considerations, irrespective of the use case. Similarly, transparency and traceability of the training data, and accuracy and reliability of the output—which should be free of bias, toxicity and hallucination—are all crucial aspects in model selection.

As a result, AI governance applies throughout the whole selection process—from performance evaluation and optimization through prompt engineering to validation and cost control. It's that one essential ingredient you need to select the right model for your use case with attributes such as risk, transparency, reliability and trustworthiness, and it should remain an ongoing effort as the model goes through continuous monitoring and evaluation.

In almost all use cases, full transparency into the training methodology is critical to enable responsible deployment of the models and to address key issues including governance, risk assessment, privacy concerns and bias mitigation. A model that's highly transparent will enable users to achieve greater reliability and trustworthiness as they can easily monitor and optimize for performance and operational risks. And then there are models that go one step further and include the advantage of contractual protection for use or indemnification.

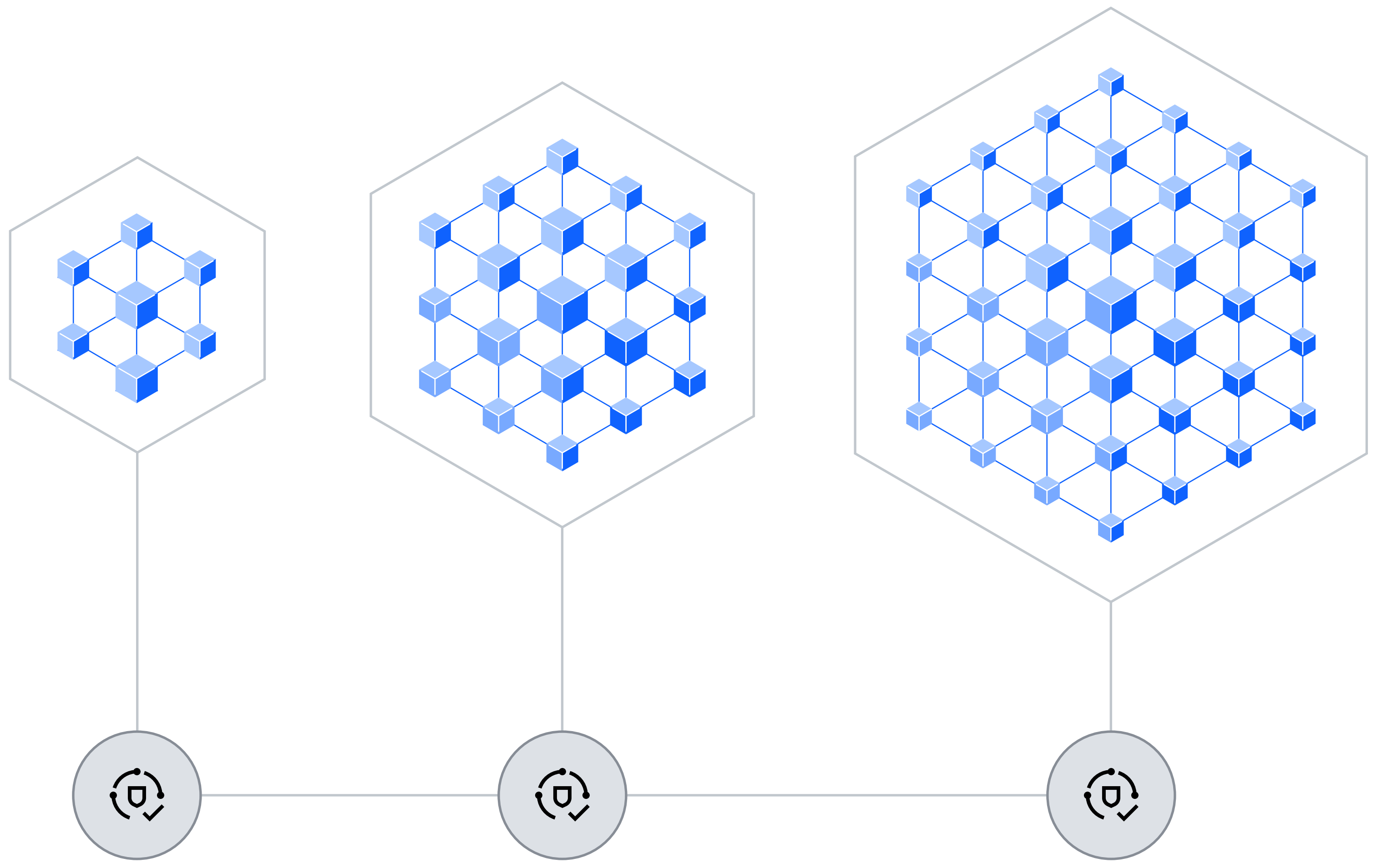
Lastly if there are data gaps in your ecosystem or data privacy risks based on your use case, you'll want to look for a model that can generate synthetic tabular data to help address the gaps and enable users to train the model with minimal risk of privacy or bias.



Boston Scientific applied off-the-shelf open-source AI models to its stent inspection process, resulting in USD 5 million in direct savings based on a modest budget of roughly USD 50,000, as well as accuracy that exceeded the existing inspection process.[2]

Refine selection based on cost and deployment needs

In model selection, while the key criteria are the use case, the size and performance of the model and the mode of deployment, there's a common throughline of costs and ROI.



Typically, the objective is not just about accomplishing task accuracy, it's also about figuring out the practicalities of ROI and cost effectiveness. The cost factor is important as you discern between models. A more expensive, larger model might be slightly more accurate than a significantly smaller, cheaper one. However, the question is, will the expensive model provide you the ROI to justify its use for that particular use case? The answer isn't always yes.

It really comes down to finding the sweet spot between performance, speed and cost. A smaller, less expensive model may not offer performance or accuracy metrics on par with an expensive one, but it would still be preferable over the latter if you consider any additional benefits the model might deliver like lower latency and greater transparency into the model inputs and outputs. For example, the smaller model could be scaled across your organization for multiple use cases, increasing its overall value. On the flip side, if your use case requires a high degree of accuracy and precision in your outputs, you might opt for the larger model—but there's a point where you might hit diminishing returns with ROI, so it's more a game of tradeoffs in the end.

Another way to approach the size-performance tradeoff is to use prompt tuning techniques on a smaller model to achieve the same or better results than prompt engineering a larger, expensive model. Prompt tuning is an efficient, low-cost method of adapting a model to a specific use case as it does not require retraining the model and can be performed using your existing resources.

Lastly, one other critical factor that influences overall costs and energy consumption is the model deployment method and the corresponding infrastructure and GPU requirements.

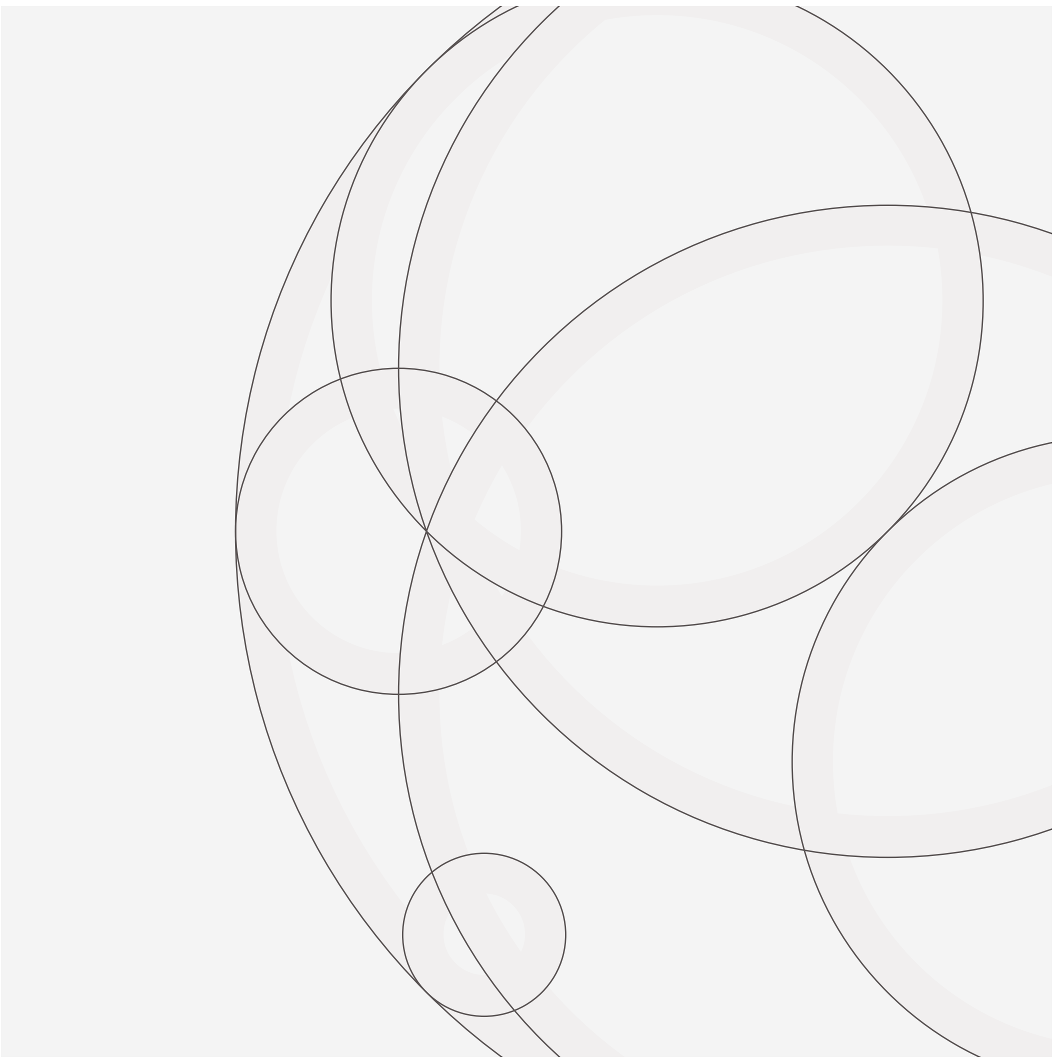
LLMs, no doubt, are resource intensive. So, how do you incorporate these models into your business without substantially impacting your ESG goals? It's important to consider how the models can be deployed sustainably to keep the total cost of ownership lower.

The deployment decision
Another important consideration in evaluating model selection costs is where and how you want the model and data to be deployed. The AI models you select should be compatible with your applications, existing vendor and partner footprint, and your overall AI and data architecture—whether it's on premises, in the cloud or in a hybrid environment.

For example, you might want to work with an open-source model like Flan-UL2. If you've trained it with your own enterprise data, though, you might need to deploy it on premises. Deploying on-premises gives you greater control and more security benefits compared to a public cloud environment. But it's an expensive proposition–especially when factoring model size and compute power, including the number of GPUs it takes to run a single large language model. So when deciding on the right model and deployment method, it all comes down to balancing the model hosting costs, performance, security and governance aspects.

How the right AI and data products help

To harness the true potential of AI, enterprises should stop looking at AI as an add-on to existing systems and instead embed it into the core of their business.



However, this is easier said than done. There are several challenges to implementing AI, ranging from data quality and availability to data security and privacy, interoperability with existing systems, scalability for enterprise-wide adoption and of course, cost. While selecting a model, how do you monitor its accuracy and trustworthiness—especially if it's a closed, proprietary model? What can you do to minimize the cost and energy consumption of large-scale models? How do you make your AI responsible, transparent and explainable?

To address these concerns and to drive AI adoption enterprise wide, you need the right AI and data products. IBM® watsonx™ is a portfolio of AI products that enables you to scale and accelerate the impact of AI with trusted data.

With watsonx, you can train, tune and deploy AI across your business using critical data, wherever it resides. These watsonx products can help:

- IBM® watsonx.ai™, a studio for foundation models, generative AI and machine learning. The watsonx.ai studio offers a variety of foundation models—including proprietary, open source and third party.
- IBM watsonx.data™, a fit-for-purpose data store built on an open data lakehouse architecture.
- IBM watsonx.governance™, a solution that provides organizations with the toolkit they need to manage risk, embrace transparency, and anticipate compliance with future AI-focused regulation.

The watsonx.ai studio is designed for a multi-model approach. To make the selection and scaling of models easier, the studio offers a hybrid, full-stack approach that includes:

- Library with IBM-developed foundation models and select third-party, and open-source models from Hugging Face.
- Prompt lab to experiment with foundation models and build prompts for various use cases and tasks.
- Tuning studio to help tune your foundation models with labeled data for better performance and accuracy.
- Data science and MLOps toolkit to build machine learning (ML) models automatically with model training, development, visual modeling and synthetic data generation.

The watsonx.ai studio offers a variety of model sizes, trained to serve multiple use cases. These models can be deployed on premises, in the cloud or in a hybrid environment, providing great flexibility. You can explore the benefits of working with the studio, the foundation models currently available and their specific use cases.

Explore foundation models within the watsonx →

Conclusion

Not all AI models are the same, and neither are your use cases.

Each specific use case demands an AI model that's a right match, which explains why a multi-model approach is pivotal to achieving success with generative AI. Ultimately, you need trusted, performant and cost-effective foundation models that enable you to optimize for various parameters, such as cost, performance and risk, based on your use cases.

With the flexibility to curate the right model mix, you can:

– **Reduce the total cost of ownership** across model training, inferencing, tuning, hosting, compute and production.
– **Optimize compute and costs** as well as scalability of models across use cases and domains for optimal ROI.
– **Make use of intuitive interfaces** that facilitate human-in-the-loop learning to improve the relevancy and accuracy scores and the performance of models based on your needs.
– **Choose models that provide transparency** into the training methodology and offer contractual IP protection to enable responsible deployment and use.
– **Curate models with built-in guardrails** and establish best practices to address key issues including governance, risk assessment, privacy concerns and bias mitigation—and deliver outputs that can be trusted for optimal performance, accuracy, safety and reliability.

Adoption of generative AI can shape business strategy, product roadmaps, talent management, customer experiences and several other areas of business. However, model selection and optimization are by no means a one and done process. It's important that you continually revisit each AI use case in terms of relevancy, model size and performance, and deployment methods to achieve optimal ROI and business outcomes.

Next steps

The AI and automation experts at IBM can work with you to curate an AI model mix—and operationalize models—to address your specific use cases and business requirements. You can start by learning more about the models offered on the wastonx.ai studio and the advantages of choosing IBM watsonx, a portfolio of AI products built for business.

Book a demo →

Learn more about watsonx.ai →

1. Future Enterprise Resiliency & Spending Survey Wave 2, IDC, March 2023
2. How to create business value with AI: 12 stories from the field, IBM IBV 2022-08-16