TRENDS REPORT

The Technology Leader's Primer For AI Foundation Models

April 4, 2024

By Charlie Dai, Leslie Joseph, Rowan Curran with Sudha Maheshwari, Srividya Sridharan, Brandon Purcell, Michele Goetz, Karsten Monteverde, Jen Barton

FORRESTER®

Summary

Foundation models are becoming critical for the next generation of AI applications. Trained on massive and diverse datasets, foundation models have capabilities beyond traditional ML models. But to select these models effectively, technology leaders must understand their characteristics, evaluate use case fit, leverage testing environments, consider model deployment and transparency, and assess ecosystem capabilities. Use this report to help you develop evaluation frameworks to effectively select and use foundation models.

Foundation Models Are Crucial For The Next Generation Of AI Applications

Today's advanced AI systems must be creative, accurate, generalizable, and pervasive. They must be designed, built, deployed, and iterated on at increasingly fast rates, and they must engage customers in a natural and intuitive way. But these demands often push companies beyond their current data and model training capabilities. Foundation models are emerging as the answer; they are the essential engines for the next generation of AI applications. Most businesses will use third-party foundation models as the capacity to train them in-house is currently out of reach for most businesses. But to use these models effectively, technology leaders must understand how they work. Forrester defines foundation models as:

Large pre-trained machine learning models that are capable of performing a wide set of tasks within one or more domains.

Foundation models are part of the generative AI (genAI) landscape, though not all foundation models are themselves generative. They are an evolution of deep learning algorithms and transfer learning theories, and can be based on LLMs or other advanced contemporary modelling frameworks and architectures like transformers, generative adversarial networks (GANs), or diffusion models to generate original output from input text and image prompts (see Figure 1). Regardless of the underlying architecture, foundation models can:

- **Enable applications far beyond traditional ML models due to data scale and diversity.** Unlike older models that were trained on smaller, curated datasets, foundation models are trained across massive diverse datasets that may span multiple modalities (e.g. text, image, video) or task-specific/functional domains. The diversity and size of the datasets lets models gain capability far greater than the narrow set of tasks that traditional ML models are trained for. In fact, foundation models are sometimes conceptually characterized as a single model that behaves like a collection of many individual expert models working together. This characteristic is what enables models such as OpenAI's GPT-4 to understand the contents of an uploaded image and answer questions based on it.
- **Be applied to various specific tasks as a result of generalized training.** Compared to prior generations of ML models, foundation models have more advanced performance for generalization. They can effectively respond to a broad range of inputs and require comparatively small amounts of effort to specialize the models in new behaviors. Different model types achieve this in a number of ways:

For example, transformers use a mechanism of “self-attention” that allows each layer in its neural network to understand many inputs at once, rather than developing features linearly. Foundation models efficiently and contextually understand the relationships between pieces of information in their training data. They leverage this understanding to generalize effectively across diverse tasks, both ones the model is explicitly trained for and ones that emerge over time. As a result, foundation models can be trained on data in specific domains without having to know all the tasks that the model will be applied to.

- **Engage in emergent behavior outside of pre-training goals.** Emergence refers to the unpredictable, spontaneous, and complex behaviors that arise because of the model’s training on vast and diverse datasets. As the model learns from the data, it captures intricate relationships, context, and patterns. Emergence is observed when a model, such as an LLM, displays through its training a capacity to generate coherent and contextually relevant text beyond the specific training examples. Essentially, once trained, foundation models can perform tasks in unforeseen ways that sometimes seem to approximate human creative problem solving. But be careful: These unforeseen methods may have unforeseen risks as well.
- **Adapt to tasks with, and without, further training.** Because of their strong core capabilities, foundation models can be adapted to address specific horizontal and vertical business needs by applying additional data to focus the model’s behavior. Different roles can perform this adaptation depending on the technique required. For example, data engineers and business application developers working with LLMs can use retrieval-augmented generation (RAG) and prompt enrichment to focus model behavior and make it more repeatable. Data scientists can use compression techniques, such as quantization, pruning, and distillation, to improve model inferencing performance. Foundation models offer different levels of transparency (e.g. model architecture, weights and parameters from pretraining, or training data) depending upon the provider, which can also extend or limit adaptability.

Figure 1

Example Types Of Foundation Models

Model type	Description	Example use
Generative adversarial network (GAN)	A type of generative ML model that uses two neural networks (generator and discriminator), to generate new data. The generator generates fake samples of data (images, audio, etc.) and tries to fool the discriminator, which tries to distinguish between the real and fake samples. The generator and the discriminator are both neural networks, and they run in competition with each other in the training phase. Convolutional neural networks (CNNs) are sometimes used within GANs to generate and discern visual and audio content.	Synthetic data generation of images, text, or tabular data
Variational auto-encoder (VAE)	A type of generative model that can generate new data by learning the underlying structure of a data set. These models are used for applications like dimensionality reduction and anomaly detection.	Text, image, or audio generation
Diffusion model (diffusion)	A type of generative model that works by destroying training data through the successive addition of Gaussian noise and then learning to recover the data by reversing this noising process. After training, you can use a diffusion model to generate data by simply passing randomly sampled noise through the learned de-noising process.	Image and video generation
Transformer neural networks (transformer)	A deep learning model distinguished by its adoption of self-attention, differentially weighting the significance of each part of the input (which includes the recursive output) data. It's used in many neural network designs for processing sequential data, such as natural language text, genome sequences, sound signals, or time series data. Large language models are an example of a transformer.	Text generation, conversational bots
Neural radiance field (NeRF)	A type of generative model that takes a set of images of a 3D scene along with their orientations and creates new views of that scene. It uses a combination of CNNs and recurrent neural networks (RNNs) to generate a radiance field, which is a representation of the object's light-field in 3D space. It's used in many neural network designs for processing sequential data, such as natural language text, genome sequences, sound signals, or time series data.	3D object and model generation

Foundation Models Shine With Intelligence, But Sparkle As Domain Experts

Foundation models are often associated with LLMs, but there is an increasingly large array of architectures and datasets underlying today's foundation models. While any foundation model can be applied to an array of different tasks, no one foundation model will support all the needs of an enterprise. When evaluating foundation models as part of your application architectures, consider the different kinds of tasks that foundation models cover:

- **Generalized language tasks.** Most of the prominent use cases for foundation models today are related to language-based tasks. Models such as Llama 2, GPT-4, Falcon, Claude 2, ERNIE 4.0, Tongyi Qianwen, Pangu, Hunyuan, GLM, Baichuan, and Titan allow enterprises to build applications that support tasks like content summarization, transformation, and generation. These models can also support data pipeline tasks like extracting metadata from unstructured content such as topics, entities, and sentiment. Information and analytics provider LexisNexis leverages Anthropic's Claude 2 to perform language tasks such as summarization and provides intelligent legal drafting capabilities to their clients in legal and

professional services.

- **Specific language tasks and code generation.** While generalized language tasks can be executed by LLMs, specialized tasks benefit from more specialized models. For example, models like Llama 2 aren't effective for multilingual translation, but models like XLM-RoBERTa have been pre-trained to excel at multilingual translation. Many current models also lack sufficient training data in some human languages, leading to the development of LLMs specific to these languages, such as Jais for Arabic and ERNIE 4.0 for Chinese. Code generation in a variety of programming languages is also a specialized language task, and there is a plethora of pretrained models for code generation, from Amazon Code Whisperer and Microsoft GitHub Copilot, to specialized company-specific models like Reddit's Ghostwriter. ANZ Bank leverages foundation models embedded in GitHub Copilot to improve developer productivity by over 40%.
- **Specialized knowledge domains.** While generalized language models have strong capabilities for understanding and reproducing language broadly, many domains need more specialized language. We expect significant development in this area in the next several years, but several models are already available, like MedPalm2 for medicine, Galactica for scientific knowledge, BloombergGPT for finance, and Google's TimesFM for time-series forecasting.
- **Computer vision tasks of many kinds.** A foundation model enhances computer vision use cases by allowing users to quickly understand their own data and content without having to train the model to understand their context. Foundation models have been developing in the computer vision space for some years: Models like the Segment Anything Model, You Only Look Once, and CLIP enable companies to understand the semantic content of their images and video, allowing them to extract meaning and make data-based business decisions. The National Aeronautics and Space Administration (NASA) and IBM are collaborating to build and leverage foundation models to analyze NASA's rich libraries of satellite imagery to understand long-term trends related to climate change.

Systematically Select Foundation Models For Your Insights Architectures

Most enterprises today, and for the foreseeable future, will source their foundation models from third parties, not pre-train their own. These models function as integral components within the broader AI stack, but not all foundation models are created equal (see Figure 2). When creating your evaluation framework for foundation models, make sure you:

- **Determine if a foundation model is right for your use case.** Technology teams and business stakeholders must work together to determine how targeted and esoteric selected use cases are. Use this to determine whether adapting a pretrained foundation model will satisfy the desired outcomes. In some cases, the task may be too specific to efficiently adapt a foundation model to it, or there may be no foundation models in the domain required. For example, for industries with very specific language or content — such as medical, manufacturing, or financial data — specific foundation models are being developed that deeply understand these industries and require less alignment when deployed.
- **Leverage model gardens, libraries, and simulation to test.** Some domains, like generalized language tasks, have many different foundation models for developers to choose from. When including open source options, the choices go from multiple to overwhelming. To manage their options, teams should leverage an emerging tool like Dataiku's Prompt Studio, H2O.ai's GPT testing environment, or evaluation tools from vendors such as Galileo or Arize.ai to test different foundation models against the same task within a single pane of glass. In addition to testing environments, providers of ML platforms and foundation models are also beginning to curate sets of open source foundation models that they have pre-vetted for specific business tasks. Technology leaders should specifically plan what toolsets and approaches they will use to test and evaluate foundation models within their projects.
- **Model deployment and model transparency.** Consider where you want to deploy the model. Most proprietary models can only be deployed in cloud environments, while open source models can also be deployed on your company's own servers. Additionally, there can be a high degree of variability in the information available about how some models were trained and what data they were trained on — even for open source models. Teams should not assume that all models are equally open, or that all open source models have the same licensing.
- **Ecosystem capabilities represent the market potential.** Ecosystem capabilities constitute a critical dimension in the evaluation of foundation models. These encompass several key elements today and are maturing to include more over time. Prompt engineering is integral to enhancing the usability and accuracy of foundation models by efficiently engineering prompts linked to predictions. Technologies like RAG, vector embeddings, and vector databases play a key role. Furthermore, the availability and functionality of plugins and APIs contribute significantly to the versatility and integration potential of foundation models within broader AI and technology ecosystems. Lastly ModelOps ensures the smooth

integration of foundation models into the operational fabric, promoting effective management throughout their lifecycle within the broader AI infrastructure.

Figure 2

Assess Foundation Models Against Important Capability Categories

Capability category	Description
Technical capabilities	
Model modality	The inputs that the model can accept: language, image, speech, multimodal
Model performance	The performance (scores of inferencing, such as accuracy, F1, correlation, etc.) of major datasets used for the training of foundation models
Model alignment	The alignment of foundation models towards security, privacy, explainability, ethics, and other targets
Model adaptation	The breadth, depth, and experience to adapt foundation models towards specific needs of business context/use cases in industries and/or functional domains
Business capabilities	
Open-source support	The open-source coverage across training data, training code, model architecture, model weights and parameters, hyperparameters, and documentation; and the licensing model
Cost effectiveness	Training cost of foundation models
Local availability	The commercial general availability in APAC or specific countries, or coverage for specific languages in APAC
Ecosystem capabilities	
Prompt engineering and RAG support	The support for prompt engineering to direct behavior of the foundation models
Plugins and APIs	The support for plugins and APIs of foundation models to embed into other applications
ModelOps	The full lifecycle support of foundation models: data prep, model building, training, deployment, inferencing, monitoring, CI/CD integration, model governance, etc.
Fine-tuning	The support for adding additional data to the model to change its behavior and make it more precise

FORRESTER®

We help business and technology leaders use customer obsession to accelerate growth.

FORRESTER.COM

Obsessed With Customer Obsession

At Forrester, customer obsession is at the core of everything we do. We're on your side and by your side to help you become more customer obsessed.

Research

Accelerate your impact on the market with a proven path to growth.

- Customer and market dynamics
- Curated tools and frameworks
- Objective advice
- Hands-on guidance

Learn more.

Consulting

Implement modern strategies that align and empower teams.

- In-depth strategic projects
- Webinars, speeches, and workshops
- Custom content

Learn more.

Events

Develop fresh perspectives, draw inspiration from leaders, and network with peers.

- Thought leadership, frameworks, and models
- One-on-ones with peers and analysts
- In-person and virtual experiences

Learn more.

FOLLOW FORRESTER

Contact Us

Contact Forrester at www.forrester.com/contactus. For information on hard-copy or electronic reprints, please contact your Account Team or reprints@forrester.com. We offer quantity discounts and special pricing for academic and nonprofit institutions.

Forrester Research, Inc., 60 Acorn Park Drive, Cambridge, MA 02140 USA
Tel: +1 617-613-6000 | Fax: +1 617-613-5000 | forrester.com