Great question! Here's a simple breakdown of the difference between **predictive AI** and **generative AI**, using **real-world analogies** and **layman-friendly terms**:

**🔮 Predictive AI = "What will happen next?"**

* **Goal:** It **analyzes existing data** and tries to **predict future outcomes**.
* **Think of it like a weather forecast** — it looks at past and current weather patterns to tell you if it's going to rain tomorrow.
* ✅ **Examples:**
  + Predicting if a server will go down based on usage history
  + Suggesting the next word in your text message
  + Recommending products based on your browsing behavior
  + Estimating how much traffic your website might get next week

**🎨 Generative AI = "What can I create?"**

* **Goal:** It **creates something new** — text, images, code, music, etc. — based on what it has learned.
* **Think of it like an artist** — you give it a prompt, and it draws or writes something entirely new.
* ✅ **Examples:**
  + ChatGPT writing an email or code snippet for you
  + AI generating a new image, song, or video
  + Creating documentation or SOPs from raw info
  + Building synthetic data for testing or simulations

**🧠 Quick Comparison**

| **Feature** | **Predictive AI** | **Generative AI** |
| --- | --- | --- |
| Purpose | Forecast or classify | Create new content |
| Input | Historical data | A prompt or instruction |
| Output | A label, score, or next value | Text, image, code, audio, etc. |
| Common Tools | Scikit-learn, XGBoost, TensorFlow | ChatGPT, DALL·E, GitHub Copilot |
| Example in IT | Alert if CPU may spike tomorrow | Write an Ansible playbook from a task |

Let me know if you’d like examples from **DevOps**, **cloud**, or **RAG (Retrieval Augmented Generation)** with local documents!

**The balancing act of using foundation models**

Foundation models still need more work to be useful.

* •

**Prompt tuning** allows data scientists to adapt models, offering "good enough" accuracy but doing it with less resources.

* •

**Retrieval augmented generation (RAG)**allows training models with targeted information without modifying the underlying model itself.

* •

**Fine-tuning foundation models** requires a high amount of resources (data, hardware, people).

* •

**Training a foundation model from scratch** requires an unrealistic amount of computing, and goes against the principles of reusing foundation models.

## Model training

Model training refers to the initial phase of building the AI/ML model in which the model learns from a large dataset to understand patterns, relationships, and features in the data. **Creating AI foundation models from scratch can be very resource- and time-intensive and is in reach of only a few enterprises.**

## Fine-tuning

Fine-tuning is a technique that involves taking a pretrained generative AI model and further training it on a specific dataset or for a specific task. Fine-tuning requires a labeled dataset that is specific to the task to train the model with examples of input-output pairs related to that task.

Although fine-tuning requires significantly less data than the original training process, it still generally involves a large investment of resources and vast amounts of data. In general, fine-tuning is more involved and labor intensive than prompt-tuning.

## Retrieval-augmented generation (RAG)

Retrieval-augmented generation (RAG) is a method for getting better answers from a generative AI application by linking an LLM to an external resource. RAG provides a means to supplement the data that exists within an LLM with external knowledge sources—for example, data repositories, collections of text, and pre-existing documentation. These resources are segmented, indexed in a vector database, and used as reference material to deliver more accurate answers.

RAG is useful because it directs the LLM to retrieve specific, real-time information from a chosen source (or sources) of truth. RAG can save money by providing a custom experience without the expense of model training and fine-tuning. It can also save resources by sending only the most relevant information (rather than lengthy documents) when querying an LLM.

RAG is quickly becoming one of the most popular approaches to working with foundation models.

## Prompt-tuning

Prompt-tuning is a technique used to optimize the prompts or instructions a user gives to an AI model to experiment with different prompt formats and wording to achieve the desired results. In some cases, prompt-tuning allows organizations to adapt models and achieve “good enough” accuracy but to do it with less resources. Prompt-tuning is often contrasted with fine-tuning AI models, which tends to require more effort and resources.

## Model inferencing

Model inferencing is the phase in which the trained AI/ML model is put to use and can make predictions, generate text, classify data, or perform any other task it was designed for. During inference, the trained model takes in new, unseen data and produces outputs based on its learned patterns.

**Potential issues with LLMs**

As incredibly powerful as LLMs are, there are issues associated with them that you should be aware of.

**Massive energy and resource costs**

Building an LLM from scratch is prohibitively expensive for most organizations. Of course, there’s the hardware that’s required, along with data centers and other expenses—and that’s before you factor in the cost of **training the LLM.**

The energy cost alone associated with training and maintaining an LLM is staggering, both from financial and sustainability perspectives. For example, training GPT-3 (with its 175 billion parameters)**consumed an estimated 1,287** megawatt hours of electricity, which is roughly equivalent to the **energy consumption of an average American household over 120 years.**

Even the cost of **maintaining** an LLM is exceptional. Consider this: According to AMI Research and various media reports, it costs OpenAI approximately **$700,000 a day** to operate ChatGPT. That’s about **$255 million** **a year.**It’s reasonable to assume that a substantial portion of that amount goes to LLM operating costs.

**Issues with transparency in model building**

In most instances, there is no way to know for sure which sources were used to create and train the LLM. As the user of the LLM, you have no way of knowing for certain whether the content used to train the model was accurate, which could lead to flawed, incomplete, or erroneous results.

**Bias and model building**

Unfortunately, it is impossible to prevent biases from leaking into models, which can affect output. For example, minorities may be underrepresented in the data used to train the model, which may lead to skewed outcomes.

A 2023 study found that four recently published LLMs were three to six times more likely to choose an occupation that stereotypically aligns with a person’s gender than what official occupation data would suggest. And despite the best intentions of the model creators, there is a risk of overcorrection for bias (such as racial bias), that leads to historically inaccurate results.

**Hallucinations**

Sometimes LLMs can generate responses that appear coherent, grammatically correct, and are stated in a manner that suggests they are accurate, when in fact, they are nonsensical or factually incorrect. Such responses are generally referred to as *hallucinations.*

Content generated by AI should be reviewed by humans to confirm its accuracy.

**Issues related to data recency**

The information that LLMs have been trained on **doesn’t continuously gather updates;** there’s an effective “cutoff” date to an LLM’s knowledge. As a result, that source material can become outdated and no longer relevant.

Data scientists and researchers **may use RAG** to address such issues. For example, implementing RAG architecture into an LLM-based question-answering system provides a line of communication between an LLM and the chosen additional knowledge sources. The LLM is able to cross-reference and supplement its internal knowledge, providing a more reliable and accurate output for the user making a query.

**Risks related to copyright and privacy**

There may be copyright or other legal issues associated with the content used to train the LLM.

For example, suppose that an LLM was trained on copyrighted material without the permission of the respective copyright holders. If your organization were to use that LLM, and you inadvertently violated the rights of copyright holders, your organization could be at risk of being sued.

**This legal issue is not merely theoretical; in December 2023, the New York Times sued OpenAI and Microsoft for copyright infringement, contending that millions of articles published by the newspaper were used to train chatbots. As of this writing, the case is pending.**

There are also privacy issues associated with LLMs. There is no way to get an LLM to “forget” information it’s been trained on. It’s not like a spreadsheet, where you could delete a column that contained private information such as dates of birth, Social Security numbers, or sensitive medical information.

Also, as it becomes easier to use AI to create so-called “deepfakes” (i.e., highly convincing fake or manipulated information, such as photos, videos, or audio recordings), the opportunities to use AI to engage in unethical or illegal behavior increase.

**Risks of mass adoption**

Currently, there are multiple LLMs in use; however, if one model overwhelmingly beats the competition, and that LLM is, effectively, the only LLM, it could change society’s common understanding of history based on quirks in how the model was trained.

**Takeaways**

Here are the takeaways from this lesson:

1. 1

The Turing test is a well-known experiment designed to assess the ability of a machine to exhibit intelligent behavior that is indistinguishable from that of a human. In the Turing test, a human evaluator reviews text from a conversation between two participants: a human and a machine. A machine is said to pass the test if the evaluator cannot reliably tell the human participant from the machine.

1. 2

Generative AI is AI technology that relies on deep learning models trained on large data sets to create new content.

1. 3

Deep learning is a specialized form of machine learning that teaches computers to process data by using an algorithm inspired by the human brain and neural networks. Deep learning teaches computers to learn through observation, imitating the way humans gain knowledge.

1. 4

Some examples of generative AI include AI-generated summaries of customer reviews on Amazon, chatbots such as ChatGPT generating content based on a text prompt, AI-generated highlight reels from sporting events, and Red Hat's KCS Solution Summaries.

1. 5

A large language model (LLM) is a type of AI program designed to understand and generate human language.

1. 6

Creating an LLM from scratch requires a tremendous amount of money, expertise, and resources. For all but a handful of organizations, creating such a model is out of reach; most organizations are likely to start with a foundation model.

1. 7

There are a variety of approaches an organization might use when working with a foundation model. One of the most popular approaches is retrieval-augmented generation (RAG), a method that involves getting better answers from a generative AI application by linking an LLM to an external resource.

1. 8

There are several issues associated with LLMs, including massive energy and resource costs, a lack of transparency with regard to how models are trained, biases in model training, hallucinations, data recency, and issues related to copyright and privacy.

|  |
| --- |
| 1. AI is a branch of computer science that enables machines to perform tasks that typically require human intelligence. 2. 4   Machine learning (ML) is a subcategory of AI that uses algorithms to identify patterns and make predictions within a set of data. |

**Lesson 2: Predictive AI—key points**

Here are the key points from Lesson 2.

1. 1

Predictive AI is a common type of artificial intelligence system used in business applications that predicts or forecasts outcomes based on historical data.

1. 2

Predictive AI is an integral part of many everyday activities such as conducting web searches, texting, shopping online, and engaging with video and music streaming services.

1. 3

Data science is an interdisciplinary field that leverages mathematical, statistical, and computational techniques to extract knowledge and insights from structured and unstructured data.

1. 4

Some of the tasks data scientists perform include data collection, data cleansing, model selection and training, and evaluation and validation of those models.

1. 5

Examples of enterprises using predictive AI include logistic companies employing it to optimize delivery routes and prevent package theft and banks or other financial institutions using it to identify fraud, money laundering, and other financial crimes.

**Lesson 3: Generative AI—key points**

Here are the key points from Lesson 3.

1. 1

The Turing test is a well-known experiment designed to assess the ability of a machine to exhibit intelligent behavior that is indistinguishable from that of a human. In the Turing test, a human evaluator reviews text from a conversation between two participants: a human and a machine. A machine is said to pass the test if the evaluator cannot reliably tell the human participant from the machine.

1. 2

Generative AI is AI technology that relies on deep learning models trained on large data sets to create new content.

1. 3

Deep learning is a specialized form of machine learning that teaches computers to process data by using an algorithm inspired by the human brain and neural networks. Deep learning teaches computers to learn through observation, imitating the way humans gain knowledge.

1. 4

Some examples of generative AI include AI-generated summaries of customer reviews on Amazon, chatbots such as ChatGPT generating content based on a text prompt, AI-generated highlight reels from sporting events, and Red Hat's KCS Solution Summaries.

1. 5

A large language model (LLM) is a type of AI program designed to understand and generate human language.

1. 6

Creating an LLM from scratch requires a tremendous amount of money, expertise, and resources. For all but a handful of organizations, creating such a model is out of reach; most organizations are likely to start with a foundation model.

1. 7

There are a variety of approaches an organization might use when working with a foundation model. One of the most popular approaches is retrieval-augmented generation (RAG), a method that involves getting better answers from a generative AI application by linking an LLM to an external resource.

1. 8

There are several issues associated with LLMs, including massive energy and resource costs, a lack of transparency with regard to how models are trained, biases in model training, hallucinations, data recency, and issues related to copyright and privacy.

### Artificial intelligence (AI)

Computer science processes and statistical algorithms that are able to simulate and augment human intelligence. In other words, AI describes systems capable of acquiring knowledge and applying insights to enable problem solving.

**This term is primarily used by the business community.**

### Machine learning (ML)

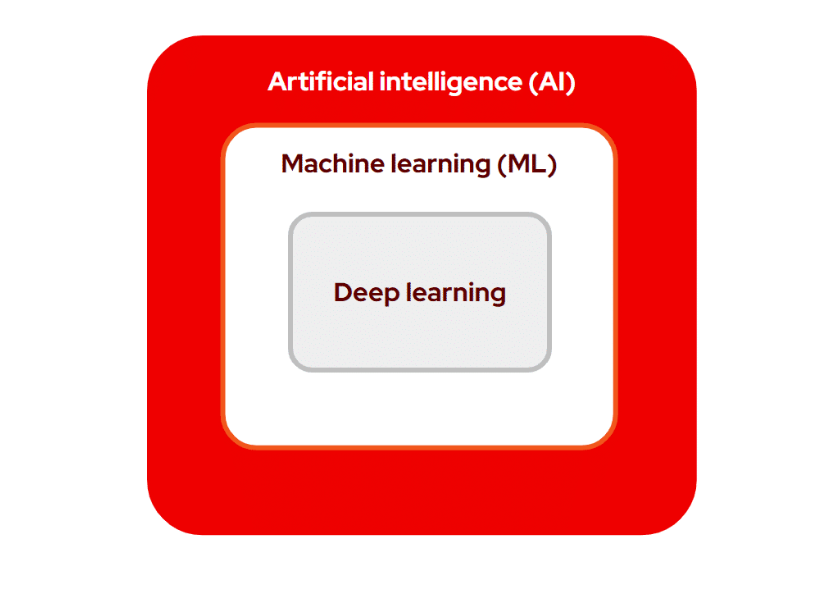
A subcategory of AI that uses algorithms to identify patterns and make predictions within a set of data. This data can consist of numbers, text, or even photos. Under ideal conditions, ML allows humans to interpret data more quickly and more accurately than they would ever be able to on their own. For instance, ML can be used to anticipate consumer buying patterns based on seasonal factors, website traffic, and so forth.

**This term is primarily used by the technical community.**

**Deep learning**

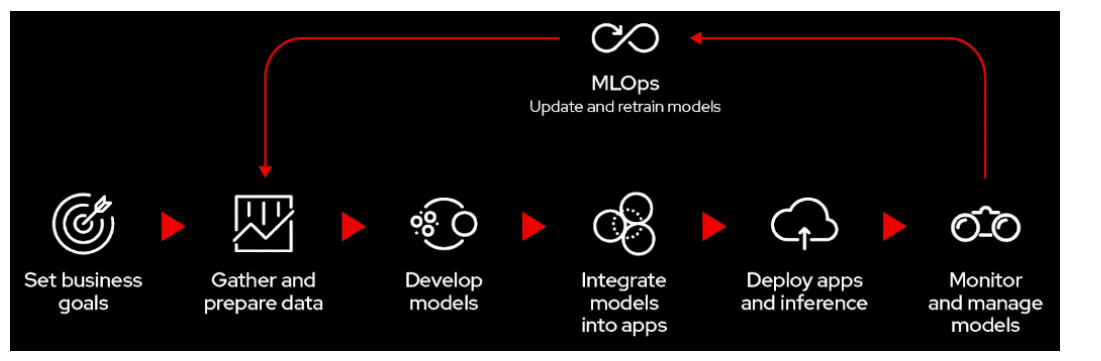
A specialized form of ML that teaches computers to process data using an algorithm inspired by the human brain. Deep learning teaches computers to learn through observation, imitating the way humans gain knowledge, and helps data scientists collect, analyze, and interpret large amounts of data. Also known as *deep neural learning* or *deep neural networking*.

Deep learning is typically used for human-like predictive use cases, such as forecasting the medical outlook of a patient based on lifestyle behaviors, genetic risk factors, or environmental conditions.



### Machine learning operations (MLOps)

A set of practices and principles that combine ML with software and DevOps methodologies to streamline and optimize the end-to-end process of training, developing, deploying, and maintaining AI/ML models. MLOps introduces automation and helps solve the challenge of maintaining the ML model’s accuracy as new data is ingested and keeps it up to date. The goal is to build AI-enabled applications powered by ML models that provide the highest prediction accuracy.



### Data science

An interdisciplinary field that leverages mathematical, statistical, and computational techniques to extract knowledge and insights from structured and unstructured data. It encompasses various processes, from data collection and cleaning to analysis and visualization, ultimately driving decision making in a wide range of domains.

### Generative AI

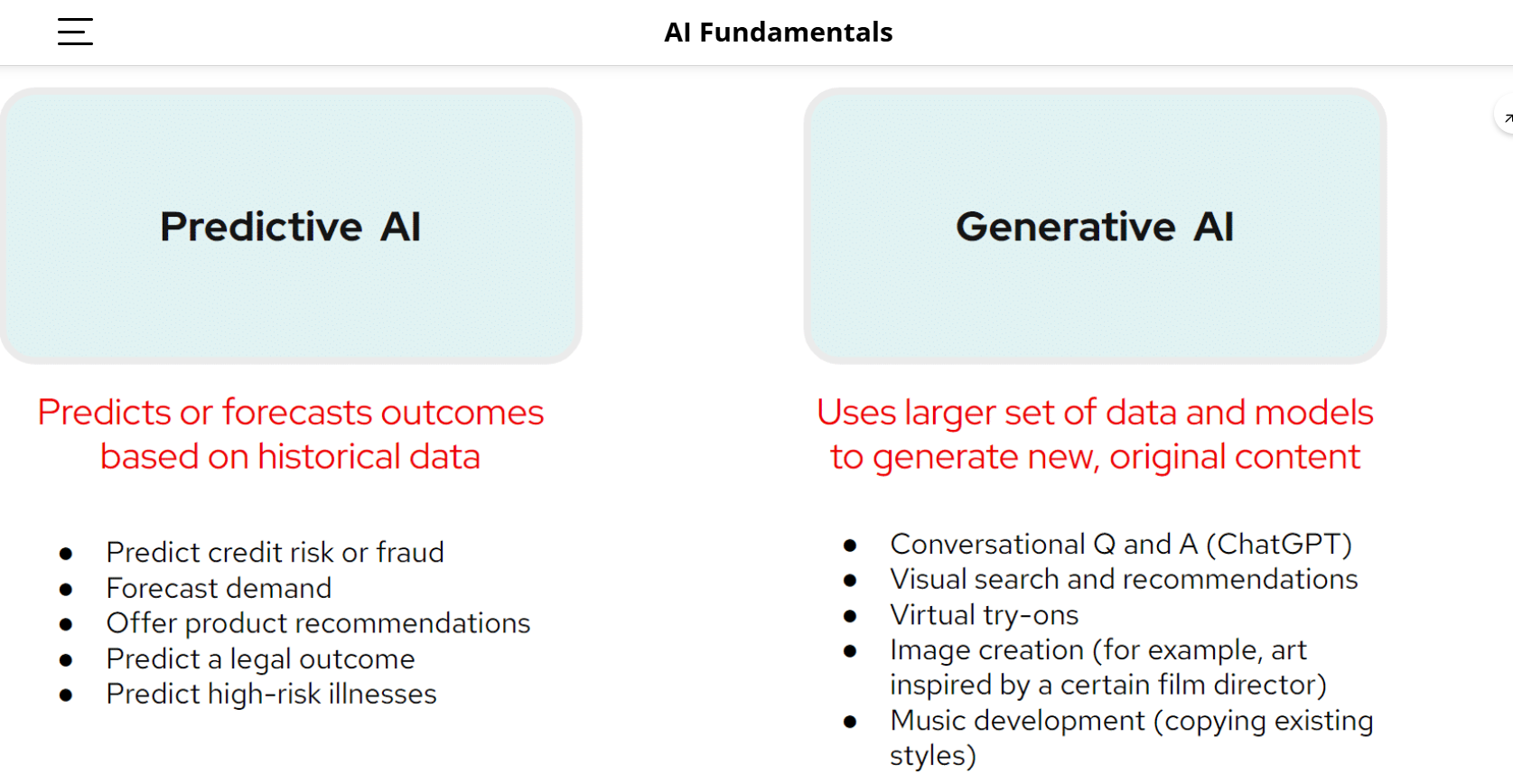
An AI technology that relies on deep learning models trained on large data sets to create new content. Generative AI models, which are used to generate new data, stand in contrast to discriminative AI models, which are used to sort data based on differences. Examples of generative AI include drafting a website or creating an image that is similar to an existing image.

**It is important to understand the difference between generative AI and *predictive AI.***

### Predictive AI

Predictive AI is one of the most common types of AI used in business applications. It is often compared to generative AI. Predictive AI has been in use for decades and predicts or forecasts outcomes based on historical data.

Predictive AI is a more mature technology than generative AI and is widely used in a variety of sectors. For example, predictive AI use cases include product recommendations on a retail website, forecasting credit risk or fraud in the financial services industry, and identifying which patients are at the most risk for certain illnesses in the medical sector.



### Foundation model

A type of ML model pretrained to perform a range of tasks. Foundation models are programmed to function with a general contextual understanding of patterns, structures, and representations. This foundational comprehension of how to communicate and identify patterns creates a baseline of knowledge that can be modified, or fine tuned, to perform domain-specific tasks for just about any industry.

Two defining characteristics that enable foundation models to function are transfer learning and scale:

* + Transfer learning refers to the ability of a model to apply information about one situation to another and build upon its internal “knowledge.”
  + Scale refers to hardware—specifically, graphics processing units (GPUs)—that allow the model to perform multiple computations simultaneously, also known as parallel processing.

### Large language model (LLM)

A type of generative AI model that utilizes ML techniques to understand and generate human language. LLMs can be incredibly valuable for companies and organizations looking to automate and enhance various aspects of communication and data processing.

**Wrap up with key terms related to AI models.**

### Model training

The initial phase of building the AI/ML model in which the model learns from a large dataset to understand patterns, relationships, and features in the data. Creating AI foundation models from scratch can be very resource- and time-intensive and is only in reach of only a few enterprise customers.

### Model inferencing

The phase in which the trained AI/ML model is put to use and can make predictions, generate text, classify data, or perform any other task it was designed for. During **inference**, the trained model takes in new, unseen data and produces outputs based on its learned patterns.

### Retrieval-augmented generation (RAG)

An architectural pattern that enables AI foundation models to produce factually correct outputs for specialized or proprietary topics that were not part of the model’s training data. By augmenting users’ questions and prompts with relevant data retrieved from external data sources, RAG gives the model “new” (to the model) facts and details on which to base its response.

### Fine-tuning

A technique that involves taking a pretrained generative AI model and further training it on a specific dataset or for a specific task. Fine-tuning requires a labeled dataset that is specific to the task to train the model with examples of input-output pairs related to that task.

Although fine-tuning requires significantly less data than the original training process, it still generally involves a large investment of resources and vast amounts of data. In general, fine-tuning is more involved and labor intensive than prompt-tuning.

### Prompt-tuning

A technique used to optimize the prompts or instructions a user gives to an AI model to experiment with different prompt formats and wording to achieve the desired results. In some cases, prompt-tuning allows organizations to adapt models and achieve “good enough” accuracy but do it with less resources. Prompt-tuning is often contrasted with fine-tuning AI models, which tends to require more effort and resources.

