NEHIA/HFMA

2025 Compliance & Internal Audit Conference

Wednesday, December 3 - Friday, December 5, 2025

Mystic Marriott Hotel, Groton, CT

Generative AI – Deep Dive

Unlocking the Power of Language: Understanding Generative AI

WHAT is Generative AI | artificial intelligence that creates **original content across various modalities** (e.g., text, images, audio, code, voice, video) that would have previously taken human skill and expertise to create

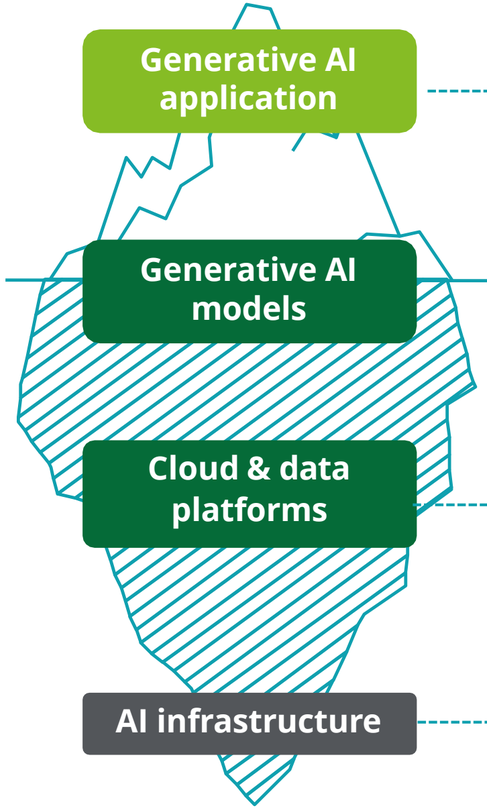
HOW does it work | Generative AI is powered by **large language models (LLMs)** such as OpenAI's GPT-4 , NVIDIA's Megatron, and Google's PaLM, which are trained on **vast amounts of data and computation** to perform a broad range of downstream tasks

WHY now | the **availability of LLMs** and advanced computing power, coupled with the **viral popularity** of publicly released applications have propelled Generative AI into the zeitgeist

WHO is involved | **Big Tech** is building—and enabling access to—foundation models; **start-ups** are developing user applications on these underlying models; **companies** are beginning to adopt and the **average person** have begun to use apps

POTENTIAL BUSINESS IMPACT | the **marginal cost of producing initial versions of knowledge-intensive content**—such as IT code, marketing copy, and creative design—**can decrease significantly**

The Generative AI technology stack is made up of many key players



OUTPUT


Text Generation


Image Generation


Code Generation


Video Generation


Audio Generation

PREDICTION

Fine Tuning
Applies use-case-specific "prompt response" examples to "teach" the model to follow new patterns

MODELS

Foundation Models
Publicly available models – available via API
Open-source models – details available to use, modify, distribute
Private models – purpose-built and tuned versions of open-source models in private servers

DATA

Modern Data Platforms
e.g., Databricks, Snowflake, Google BigQuery, AWS S3, Azure Synapse

COMPUTE POWER

Compute Hardware and Cloud Platforms
e.g., GCP, AWS, Azure, GPUs, TPUs

WHO'S PLAYING IN THE SPACE?

(not exhaustive)

Busy and growing application space

335+ startups[1], in addition to open source and enterprise apps, with new apps being funded every month

OpenAI · NVIDIA · HUGGING FACE

aws · Meta · Google

BioNemo · BloombergGPT

snowflake · Google

databricks · aws

Google Cloud · aws · NVIDIA

[1] Source: CB Insights

Prompt operations

There are 3 fundamental operations which form a key framework for understanding and utilizing prompts

Reductive Operations

These operations aim to produce a smaller output from a larger input. Examples include:

- **Summarization**: Condensing information into a shorter form.
- **Distillation**: Extracting the core underlying principle or fact from a text.
- **Extraction**: Pulling out specific information, such as answering questions or extracting dates and numbers from a text.
- **Characterizing**: Defining the nature of the text or the main topic within the text.

Transformational Operations

In these operations, the input and output are roughly the same size and meaning, but the form or structure is changed. Examples include:

- **Reformatting**: Changing the presentation of content, such as translating data between formats like XML and JSON.
- **Refactoring**: Rewriting code or text in a better or different way while preserving the original meaning.
- **Language Change**: Translating content between natural languages (e.g., English to Portuguese) or programming languages (e.g., C++ to Python).
- **Restructuring**: Altering the order of content, removing or adding sections, and optimizing for logical flow.

Generative Operations

These operations expand a small input into a much larger output. Examples include:

- **Drafting**: Providing instructions for a language model to create a document, such as a code file, legal document, or story.
- **Planning**: Giving a set of parameters to develop plans, such as action plans or project plans.
- **Brainstorming**: Using imagination to list possibilities, generate ideas, explore options, and solve problems.
- **Hypothesizing**: Generating hypotheses based on given information or a prompt.
- **Amplification**: Expanding on a topic to fully explore and explain it.

Common Risks with Generative AI

Bias

Bias in; bias out. If the training data is biased (e.g., over/under-representation of a population cohort, sexism, racism), then outputs generated could exhibit biases as well. Bias reductions in the training data and/or human supervision during model training is needed

Cost

Foundation Models generally offer a pay-as-you-go billing mechanism, and the cost per use of sophisticated models is materially significant. Fine tuning the biggest model and running large documents through several times could easily run up a bill of tens of US $1000s

Ethical Use

Is the AI being used in a manner consistent with the purpose of the overall exercise? Is a human being brought into the loop to decide whether the AI's suggestion needs adjustment before actual use? Submitting an AI-generated essay for a high-school assignment may not be ethical

Hallucination

Models might output facts that are factually false. Sources and citations are unavailable for most models. Users should be conscious that outputs could be inaccurate and should perform due diligence to validate generated content.

IP Protection

SaaS-AI companies may save some or all of prompt payloads for future training. Therefore, confidential data will be used to train future versions of the base model – how will this affect your organization's competitiveness in the market?

Malicious behavior

It is critical to proactively minimize risk from malicious behavior on the network to maintain operations and customer trust. For example, a customer service bot revealing confidential information to a hacker either by prompt or unintentionally

Model Performance

Foundation Models are comprised of billions of parameters (model size) and trained on petabytes of data. In theory, the larger the model, the better the output. Foundation Models take time to produce outputs, which may limit real-time use cases

Privacy

SaaS-AI companies require to submit text as a payload to users' API call. The data could be crossing borders. Is this in accordance with data privacy laws and with your company's policies? Many cloud service providers offer market-leading controls to manage data privacy of Foundation Models

Text Formatting

Models are good at understanding text but struggle when the data are in irregular formats, or when the position of the text on the page (e.g., infographic, PPT presentation slide) is relevant to the context and understanding. Other emphasis generators such as bolded text, font color, etc., don't play a role yet

Token Size Limits

Most models have a 2k token size limit. Some larger ones can process 4k tokens in a single call. 2k tokens are approximately 2-2.5 pages. This limit makes it difficult to process larger documents

Meet your panel

STEPHEN GILLIS
Mass General Brigham

Director, Compliance Coding, Billing & Audit

MIKE CRONIN
Deloitte & Touche LLP

Internal Audit Leader

MICHAEL KOPPELMANN
Deloitte & Touche LLP

Digital Internal Audit Leader

KATIE GONICK
Deloitte & Touche LLP

Our Facilitator

QUESTIONS