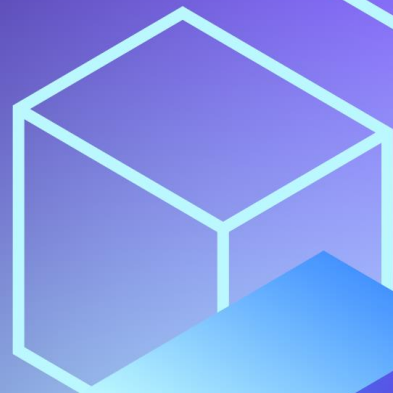
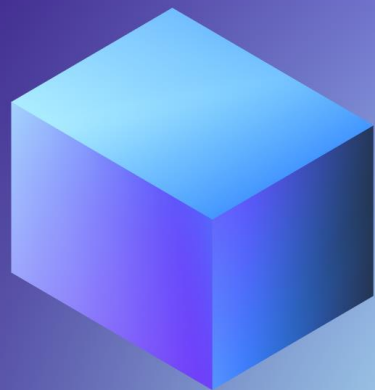




Generative AI with AWS

Lukasz Magiera

Sr. Solutions Architect
AWS



2023

The Year of POCs



What is generative AI?

Is this secure?

Do I need to become a prompt engineer?

How do I choose a model?

Where do I get started?



What does this mean for my business?

What is a Foundation Model?



Which models should we try out?

What is FM?

What is a Large Language Model?



2024

The Year of Production

(FOR SOME)



How do I prioritize my projects?

How can I lower my costs?

How do I make this real?

What customization method should I use?



How I can I scale this?

Which models should I use?

Should I train my own model?

How do I manage risks?



How can we move faster?

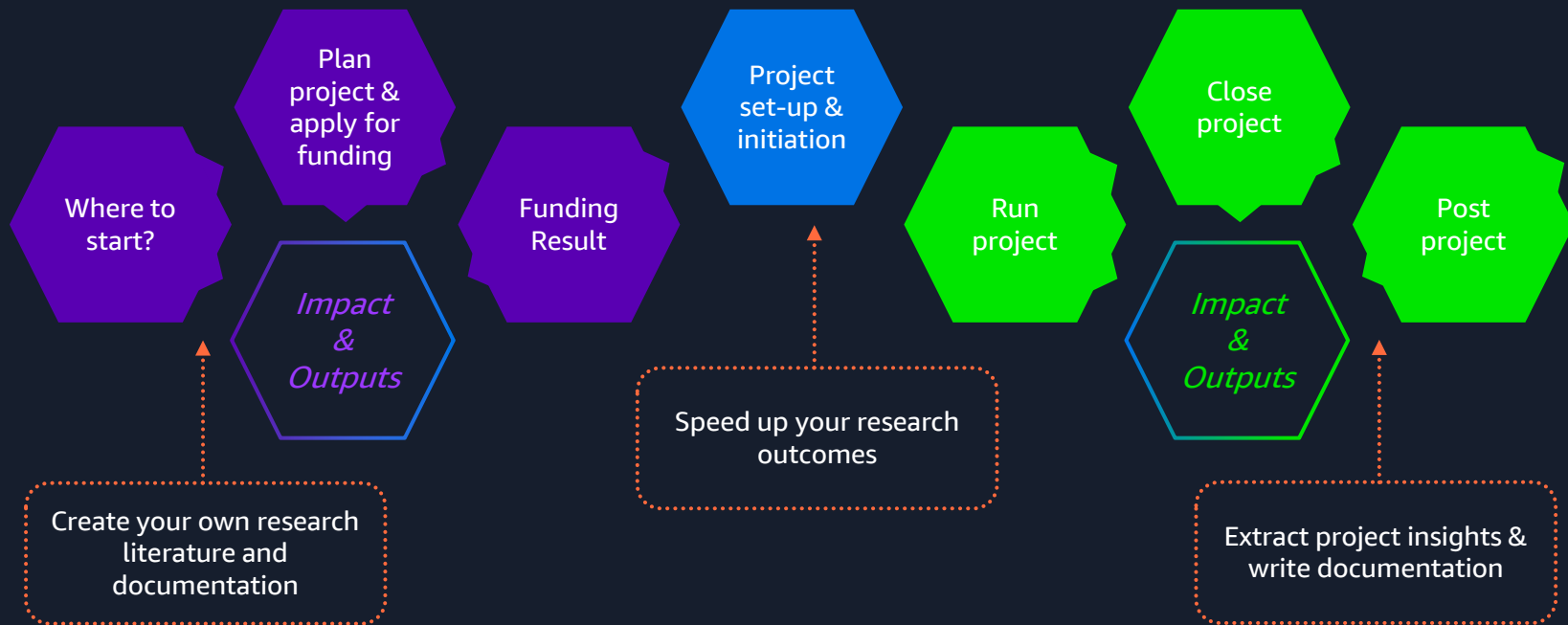


Innovation can
transform industries



GENERATIVE AI

AI and generative AI can help you at different steps in your research lifecycle



Generative AI is used for a wide range of use cases in research



Research and discovery

Analyzing large-scale datasets and identifying patterns, information



Clinical development

Optimizing research protocols



Research funding support

Research application content generation



Researcher support

Chatbots with research specific insights

Literature review

The screenshot displays the Elicit website's search interface. At the top, there is a navigation bar with links for Features, Testimonials, Pricing, FAQ, Careers, Sign In, and a prominent Sign Up button. The main content area is titled 'DISCOVERY' and features a large heading 'Search for research papers'. Below this, a sub-heading asks the user to 'Ask a research question and get back a list of relevant papers from our database of 125 million'. Three key features are listed with icons: 'Get one sentence abstract summaries', 'Select relevant papers and search for more like them', and 'Extract details from papers into an organized table'. The search results are presented in a card format. The top card shows the search query 'What are the benefits of taking L-theanine in combination with caffeine?' and a 'Summary of top 4 papers'. The summary text states that a combination of L-theanine and caffeine has been found to improve cognitive performance and increase subjective alertness, citing several studies. Below the summary are buttons for '+ Add columns', 'Sort', 'Filters', and 'CSV'. A table view is also visible, with columns for 'Paper' and 'Abstract summary'. The first row of the table shows a paper titled 'The combination of L-theanine and caffeine improves cognitive performance and increases subjective alertness' by T. Giesbrecht et al., published in 'Nutritional neuroscience' in 2010, with 74 citations. The abstract summary for this paper is 'L-theanine in combination with caffeine helps to focus attention during a demanding cognitive task.' The second row of the table shows a paper titled 'L-Theanine and caffeine improve task switching but not intersensory attention or subjective alertness' with an abstract summary of 'L-theanine and caffeine in combination can improve attention.'

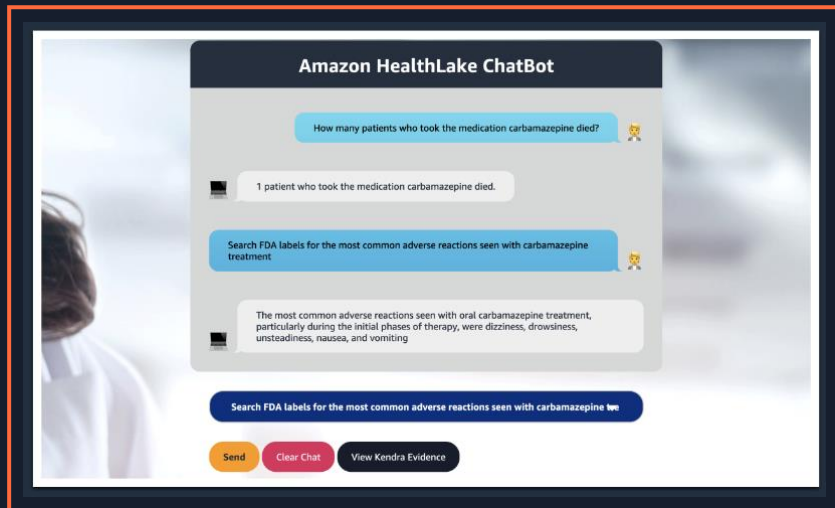
Suggesting New Ideas and Concepts for Innovation

Generative AI for Stimulating Creativity and Innovation

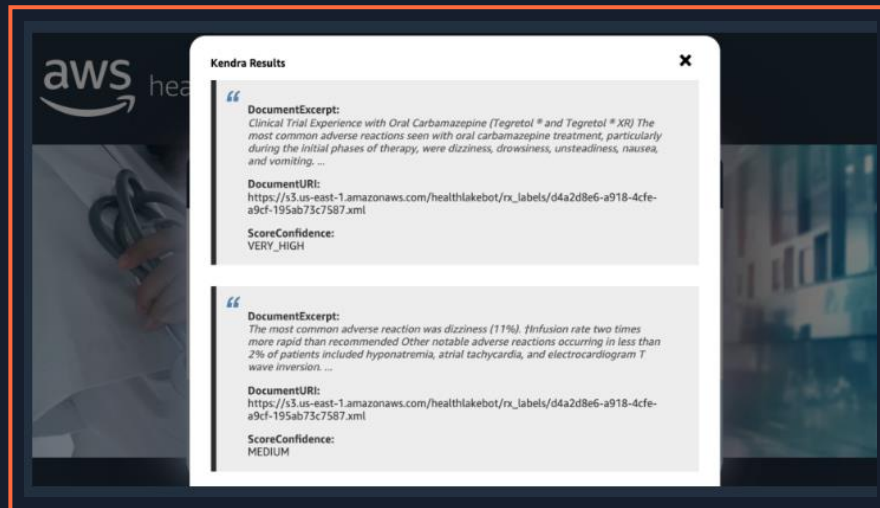
- Combines and recombines existing knowledge in novel ways
- Generates new concepts, hypotheses, and potential solutions
- Serves as a starting point for researchers to explore unconventional approaches
- Provides context and justification for each suggested idea
- Exposes researchers to a wide array of AI-generated concepts
- Helps break mental barriers and encourages innovative thinking



Search and summarize: Scientific and clinical trials



Simpler accessibility to research
and trial data



Traceability into decision making
and data generation

Predict molecular properties with specialized FMs

Extensible Architecture Supports Nine Modules

MSA-Based Structure Prediction

- AlphaFold 2 from DeepMind
- OpenFold from Columbia University

pLM-Based Structure Prediction

- OmegaFold from Helixon US
- ESMFold from Facebook AI Research

Orchestration

- Nextflow from Seqera Labs



Protein Design

- RFDesign and RFDiffusion from the University of Washington
- ProteinMPNN from the University of Washington

Virtual Screening

- DiffDock from MIT

Easy integration into your current tools OR Build new pipelines and UX

Architecture Diagram

Download the architecture diagram PDF [PDF](#)



Step 1
[AWS CloudFormation](#)
deploys the infrastructure in your AWS account.

Implementation Resources

The sample code is a starting point. It is industry validated, prescriptive but not definitive, and a peek under the hood to help you begin.

[Open sample code on GitHub](#)

The screenshot shows the AWS Drug Discovery Workbench interface. The main heading is 'AWS Drug Discovery Workbench' with the sub-heading 'Run bio algorithms at cloud scale'. Below this, there is a section titled 'How it works' with a large graphic that says 'Upload & Manage Protein Sequences'. The interface includes a sidebar with navigation options like 'Manage Proteins', 'Protein Sequences', 'MSA based structure prediction', 'pLM Based structure prediction', 'Protein Structure Design', 'Docking', and 'Workflow'. The main content area displays a 'Getting started' section with a 'View Tutorial' button and a 'More resources' section with a 'View Docs' button.

Guidance for
protein folding on AWS

Simplified UI with AWS
Drug Discovery Workbench

Supports multiple algorithms within a shared user interface



MIT News

ON CAMPUS AND AROUND THE WORLD

Speeding up drug discovery with generative models

MIT researchers built DiffDock, a model that makes drugs faster than traditional methods and reduces side effects.

Alex Ouyang | Abdul Latif Jameel Clinic for Machine Learning
March 31, 2023



Electronics IoT & Wireless Sensors Embedded

EMBEDDED

Generative AI speeds tedious aspects of drug discovery process

By Dan O'Shea · Apr 19, 2023 02:55pm

...and this is **just** the beginning.

Science News

from research organizations

Using AI to create better, more potent medicines

Novel framework could offer chemists greater drug options

Date: May 30, 2023

Source: Ohio State University

Summary: While it can take years for the pharmaceutical industry to create medicines capable of treating or curing human disease, a new study suggests that using generative artificial intelligence could vastly accelerate the drug-development process.

Share: [f](#) [t](#) [p](#) [in](#) [✉](#)

Intelligence is Revolutionizing

Generative AI use cases across industries

ENHANCE CUSTOMER EXPERIENCES

CHATBOTS
VIRTUAL ASSISTANTS
CONVERSATION
ANALYTICS
PERSONALIZATION

BOOST EMPLOYEE PRODUCTIVITY & CREATIVITY

CONVERSATIONAL
SEARCH
SUMMARIZATION
CONTENT CREATION
CODE GENERATION
DATA TO INSIGHTS

OPTIMIZE BUSINESS PROCESSES

DOCUMENT PROCESSING
DATA AUGMENTATION
FRAUD DETECTION
PROCESS OPTIMIZATION

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs



TOOLS TO BUILD WITH LLMs AND OTHER FMs



Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs

TOOLS TO BUILD WITH LLMs AND OTHER FMs

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron





CG1

NVIDIA Tesla
M2050 "Fermi"
GPUs

G2

NVIDIA GRID
GK104
"Kepler" GPUs

P2

NVIDIA
K80
GPUs

G3

NVIDIA
Tesla M60
GPUs

P3

NVIDIA V100
Tensor Core
GPUs

G4

NVIDIA T4
Tensor Core
GPUs

P4

NVIDIA A100
Tensor Core
GPUs

G5

NVIDIA A10G
Tensor Core
GPUs

G5g

NVIDIA T4G
Tensor Core
GPUs

P5

NVIDIA H100
Tensor Core
GPUs

Innovating at the silicon level

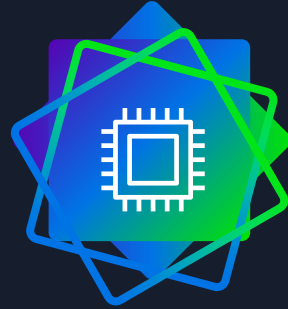
AWS Trainium 2

4x

Faster than
AWS Trainium

65

Exaflops of on-
demand
supercomputing
performance



AWS Inferentia 2

4x

Higher throughput

10x

Lower latency



Amazon SageMaker

Build, train, and deploy ML models
at scale

Automatic model fine-tuning & distributed
training

Flexible model deployment options

Tools for ML operations

Built-in features for responsible AI

Select from 250+ publicly available FMs

AI21 labs

MODELS

Jurassic-2 Ultra, Mid
Contextual answers
Summarize
Paraphrase
Grammatical error
correction

Meta AI

MODELS

*Llama 3 8B, 70B
*Llama 2 7B, 13B, 70B
Code Llama 7B, 13B,
34B, 70B
Llama Guard

cohere

MODELS

Command R, R+
*Generate
*Command-Light
Embed v3
Embed Light v3
Rerank

stability.ai

MODELS

Stable Diffusion XL 1.0
*2.1 base
Upscaling
Inpainting
StableLM (Japanese)

Hugging Face

MODELS

*Falcon-7B, 40B, 180B
*Mistral 7B, 7B Instruct
*Mixtral 8x7B,
*8x7B Instruct
Gemma
Open Llama
*RedPajama
MPT-7B
*BloomZ
*Flan T5
DistilGPT2
*GPT NeoXT
StarCoder
Whisper
BGE Large
E5 Large

Lightn

MODELS

Lyra-Fr
10B, Mini

databricks

MODELS

DBRX
Dolly

alexa

MODELS

AlexaTM 20B

*Fine-tunable



Overcoming the barriers to ML adoption

Disparate tools for data science & business analyst



Integrated ML tools in a single interface

Build, train, and deploy models using IDEs

Significant effort to build your own FMs from scratch



Choice of FMs

Access 250+ publicly available models FMs that can be fine tuned easily

Improve accuracy and relevancy of models



Comprehensive set of HIL capabilities

Use human feedback across the ML lifecycle to create and evaluate high quality models

Managing underlying infrastructure



Fully managed ML infrastructure

Purpose-built ML accelerators for FM training and inference

Tedious, manual ML operations



Built-in MLOps

Automate and standardize MLOps practices

Challenging to govern ML projects efficiently



Out-of-box ML governance tools

Simplify access control and enhance transparency across ML lifecycle

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs






TOOLS TO BUILD WITH LLMs AND OTHER FMs

Amazon Bedrock

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

 GPUs  Trainium  Inferentia  SageMaker

 UltraClusters  EFA  EC2 Capacity Blocks  Nitro  Neuron



Amazon Bedrock

The simplest way to build and scale generative AI applications with LLMs and other FMs

Choice of industry-leading FMs from AI21 Labs, Amazon Web Services, Anthropic, Cohere, Meta, and Stability AI

Customize FMs using your organization's data

Enterprise-grade security and privacy

Amazon Bedrock

Broad choice of models

AI21labs

amazon

ANTHROPIC

cohere

Mistral AI

Meta

stability.ai

JURASSIC-2

AMAZON TITAN

CLAUDE

COMMAND + EMBED

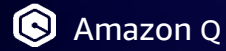
Mistral 7B
Mixtral 8x7B

LLAMA 2

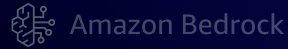
STABLE DIFFUSION XL

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs



TOOLS TO BUILD WITH LLMs AND OTHER FMs



Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE



GPUs



Trainium



Inferentia



SageMaker



UltraClusters



EFA



EC2 Capacity Blocks



Nitro



Neuron

Reinvent work with **Amazon Q**

BUSINESS

 Amazon Q
Business

KNOWLEDGE SEARCH

SUMMARIZATION

CONTENT CREATION

EXTRACT INSIGHTS

RESEARCH & ANALYSIS

 Amazon Q in
QuickSight

UNDERSTAND DATA

BUILD & REFINE VISUALS

BUILD CALCULATIONS

EXECUTIVE SUMMARIES

CREATE DATA STORIES

DEVELOPERS

 Amazon Q
Developer

PLAN APPLICATION

CODE GENERATION

UNIT TESTING

SECURITY SCANNING

CODE REMEDIATION

CODE MIGRATION

TROUBLESHOOTING

DEVELOPER KNOWLEDGE

SPECIALIZED USERS

 Amazon Q in
Connect

AGENT ASSIST

 Amazon Q in
AWS Supply Chain

SUPPLY CHAIN

Generative AI Stack

APPLICATIONS THAT LEVERAGE LLMs AND FMs

 Amazon Q  AWS App Studio






TOOLS TO BUILD WITH LLMs AND OTHER FMs

 **Amazon Bedrock**

Guardrails | Agents | Studio | Customization | Custom Model Import | Amazon Models

INFRASTRUCTURE FOR FM TRAINING AND INFERENCE

 GPUs  Trainium  Inferentia  SageMaker

 UltraClusters  EFA  EC2 Capacity Blocks  Nitro  Neuron

