

# Working with Large Language Models

Mandie Quartly, PhD  
AI / ML Specialist  
Google Cloud

# Agenda

- 01 Introduction to GenAI / LLMs
- 02 Potential use cases & areas for focus
- 03 What does that mean in practise?
- 04 Demo
- 05 Things to consider

# Generative AI

will transform your business

Generative AI

What does **GenAI** mean?

What are **LLMs**?

Large Language Models

**AI** (artificial intelligence) is the broader concept of enabling a machine or system to sense, reason, act, or adapt like a human

**ML** (machine learning) is an application of AI that allows machines to extract knowledge from data and learn from it autonomously

**Generative AI** refers to the use of AI to create new content, like text, images, music, audio, and videos.

# What are large language models?



ML algorithms that can **recognize, predict, and generate** human languages



Pre-trained on petabyte scale text-based datasets resulting in large models with **10s to 100s of billions of parameters**



LLMs are normally **pre-trained on a large corpus of text** followed by fine-tuning on a specific task



LLMs can also be called **Large Models** (includes all types of data modality) and **Generative AI** (a model that produces content)



Go read this huuuuuge pile of books.



So, you've learned about cats and millions of other concepts ... what's a cat?

A cat is a small, domesticated carnivorous mammal.



**Generative language models**

LaMDA, PaLM, GPT-3, etc.

# Why are large language models different?



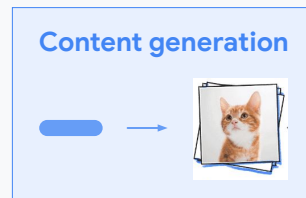
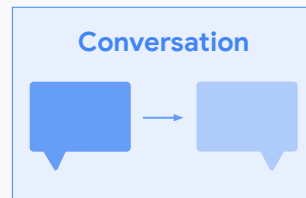
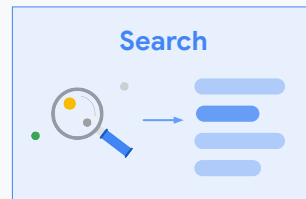
LLMs are characterized by **emergent abilities**, or the ability to perform tasks that were not present in smaller models.



LLMs contextual understanding of human language **changes how we interact** with data and intelligent systems.



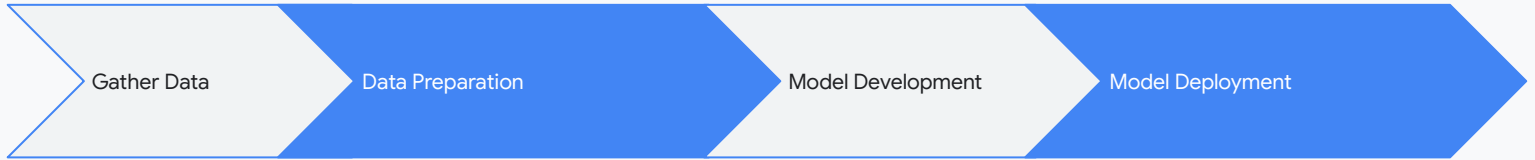
LLMs can find patterns and connections in **massive, disparate data corpora**.



# Potential for much shortened time to deployment

## Traditional AI

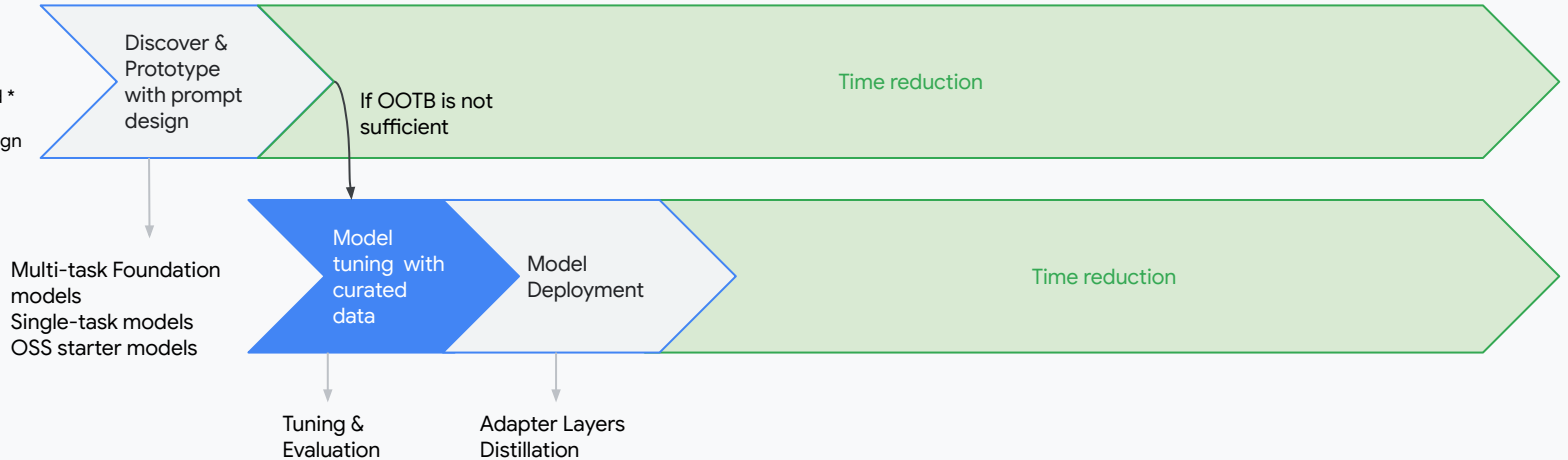
- 1000s of training examples
- ML Expertise
- Frameworks & compute
- Think about minimizing loss function



## Generative AI

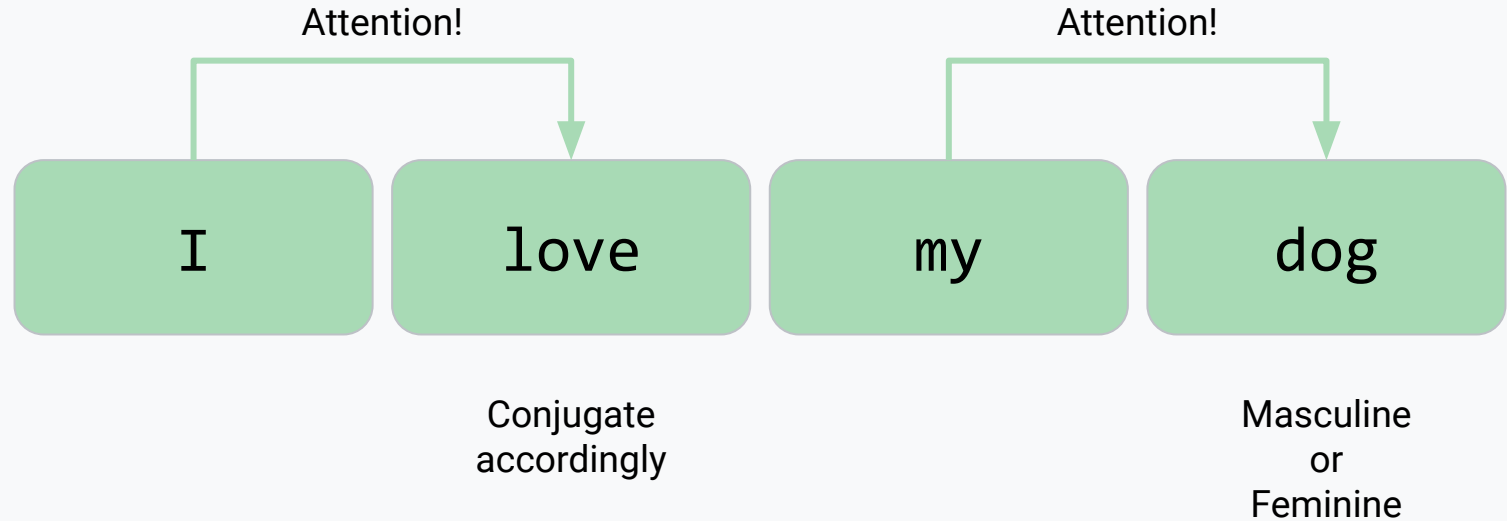
- 0-10 training examples \*
- No ML expertise needed \*
- APIs & natural language
- Think about prompt design

\*to get started



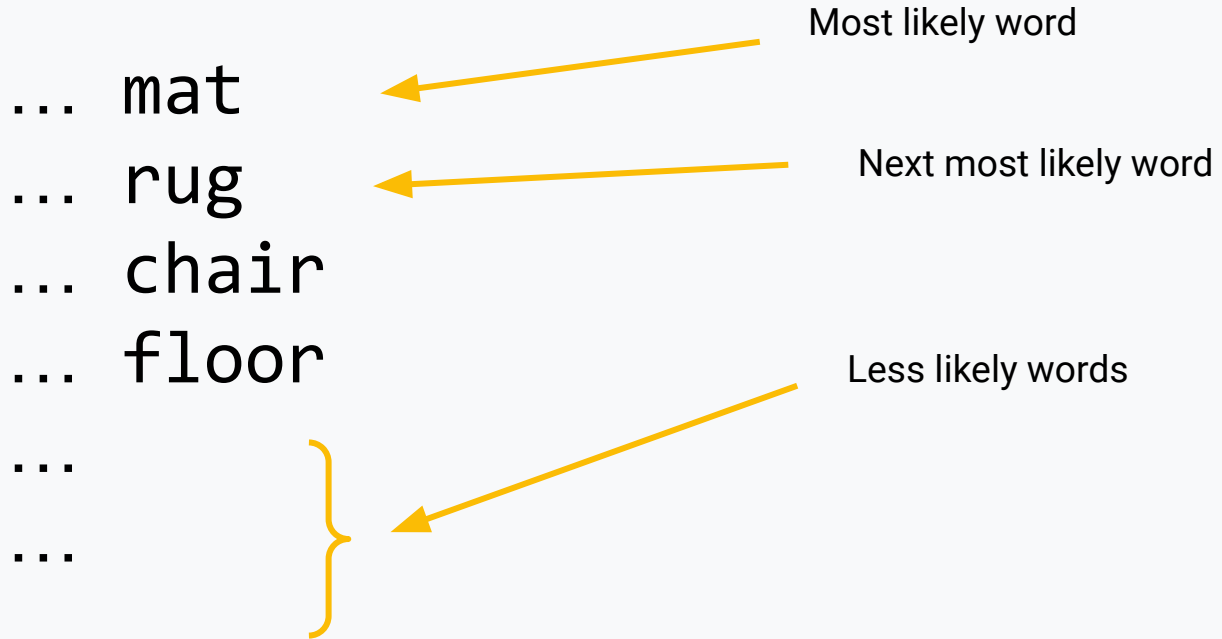


# The transformer architecture behind LLMs



“Attention is all you need” <https://arxiv.org/pdf/1706.03762.pdf>

# The cat sat on the ...



# Consumers & enterprises have different needs....

## Consumers

**Plan** a 3-day trip to Patagonia

**Create** a valentine poem

**Generate** a picture of a panda playing yahtzee

## Enterprises

How do we **control** our data?

How will we manage **costs**?

How to ensure data **accuracy** and **security**?



Vertex AI



Duet AI

AI ecosystem

# PaLM 2



# Foundation Models

## Imagen

Generate and customize studio-grade images at scale

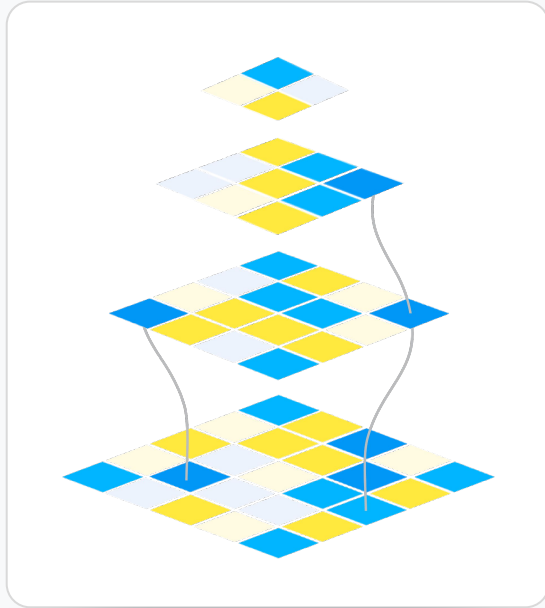
## Chirp

Speech-to-Text language captioning and voice assistance

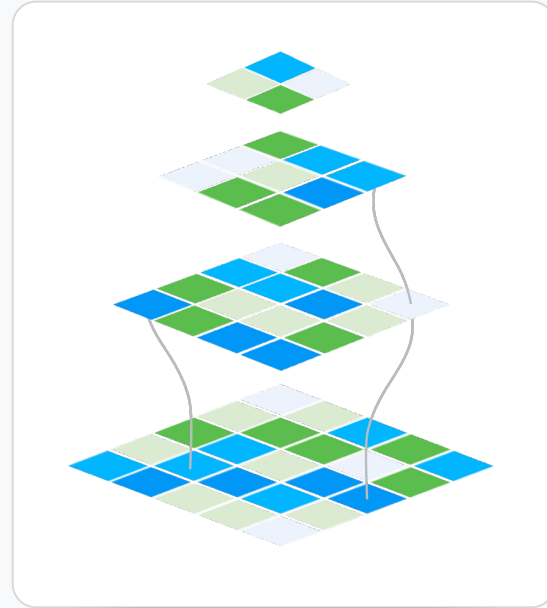
## Codey

Customized code generation and completion

# Domain Specific Models



**Med-PaLM**  
Life Science and Healthcare



**Sec-PaLM**  
Cybersecurity

# Tools to enhance the capabilities of foundation models



Grounding



Extensions



Tuning



# How to **tune** foundation models

Simple, cost efficient

Complex, more expensive



Prompt  
design



Adapter  
tuning



Reinforcement  
learning with  
human feedback



Full fine  
tuning

Where should I start?

# Gen AI will reshape **processes**



## Online interactions made conversational

---

- Conversational journeys
- Customer service automation
- Knowledge access



## Complex data, intuitively accessible

---

- Product / content catalog discovery
- Product / content recommendation
- Business process automation
- Content search and synthesis








## Content generation at the click of a button

---

- Creative assistance
- Documentation generation
- Developer efficiency

# You can benefit from Generative AI in multiple ways today

	Out of the box				DIY
	Assistants	Solutions	Applications	APIs	Models
What?	“Help me write”	“Summarize content of 100 docs in 100 words”	“Search for answers in my data”	“Build my own chatbot using a pre-trained model”	“Customize a model”
For who?	 Every employee	 Business users	 Developers	 Developers & AI Practitioners	 AI Practitioners
How?	Zero-code	Low-code	Low-code	Pro-code	ML expertise
Google products	<b>Duet AI for Workspace</b> Gmail, Docs, Slides, Sheets	<b>AI Agents</b> Contact Center AI Document AI <b>Duet AI for Cloud</b> Chat & code assistance	<b>Gen AI App Builder</b> Enterprise Search Conversational Chatbots Digital Assistants	<b>Vertex AI</b> Gen AI Studio Model Garden	<b>Vertex AI</b> ML Platform ML Operations

# Which generative AI use cases do I start with?

Here are three common uses:

## Research and information discovery

---

- Summarize content from several documents
- Create a conversational FAQ

## Content generation and prototyping

---

- Create images for presentations in minutes
- Turn messaging docs into engaging social media posts

## Improve developer efficiency

---

- Accelerate software prototyping
- Coach junior coders

Other common uses alongside reference architectures:

[cloud.google.com/use-cases/generative-ai](https://cloud.google.com/use-cases/generative-ai)

# Gen AI will transform every industry



## Retail and CPG

Creative Assistance

Conversational Commerce

Customer Service Automation

New Product Development

Improving Employee Productivity

Supply Chain Advisor



## Financial Services

Financial Document Search & Synthesis

Enhanced Virtual Assistant

Capital Markets Research

Regulatory Code Change Consultant

Personalized Financial Recommendations



## Healthcare & Life Sciences

Digital Patient Concierge

Public & Private Contextual Search

Expedite Prior Authorization Letter

Clinical Trial Report Generation

Customer Service Agent



## Media & Entertainment

Media Content Discovery

Creative Assistance

Internal Document/Media Search

Branded Consumer Interactions

Content Summarization and Metadata



## Manufacturing

Machine Generated Events Monitoring

Customer Service Automation

Document Search & Synthesis

Product/Content Catalog Discovery

Supply Chain Advisor



## Communications Service Providers

Customer/Employee Service Automation

Network Planning & Operations\*

Advertising & Content/Creative Assistance

Employee Knowledge Search

Test/Code Script Generation

Contract Analysis & Negotiation

\*Network Planning & operations use cases include Network Capacity Planning, Network Root Cause Analysis, and Post Mortem Creation

# Fast, dynamic access to insights is increasingly critical in **finance**, and **AI** is the key to unlocking them.

- 01. Geopolitical Uncertainty
- 02. Climate Risk
- 03. Changing Market Conditions

Three of the top five risks that bank executives expect to most influence their industry in the coming decade.

Source: Economist Impact: [Banking in 2035 \(2022\)](#)

## Potential impact

for banks from AI applications is estimated at **\$447 billion by 2023**

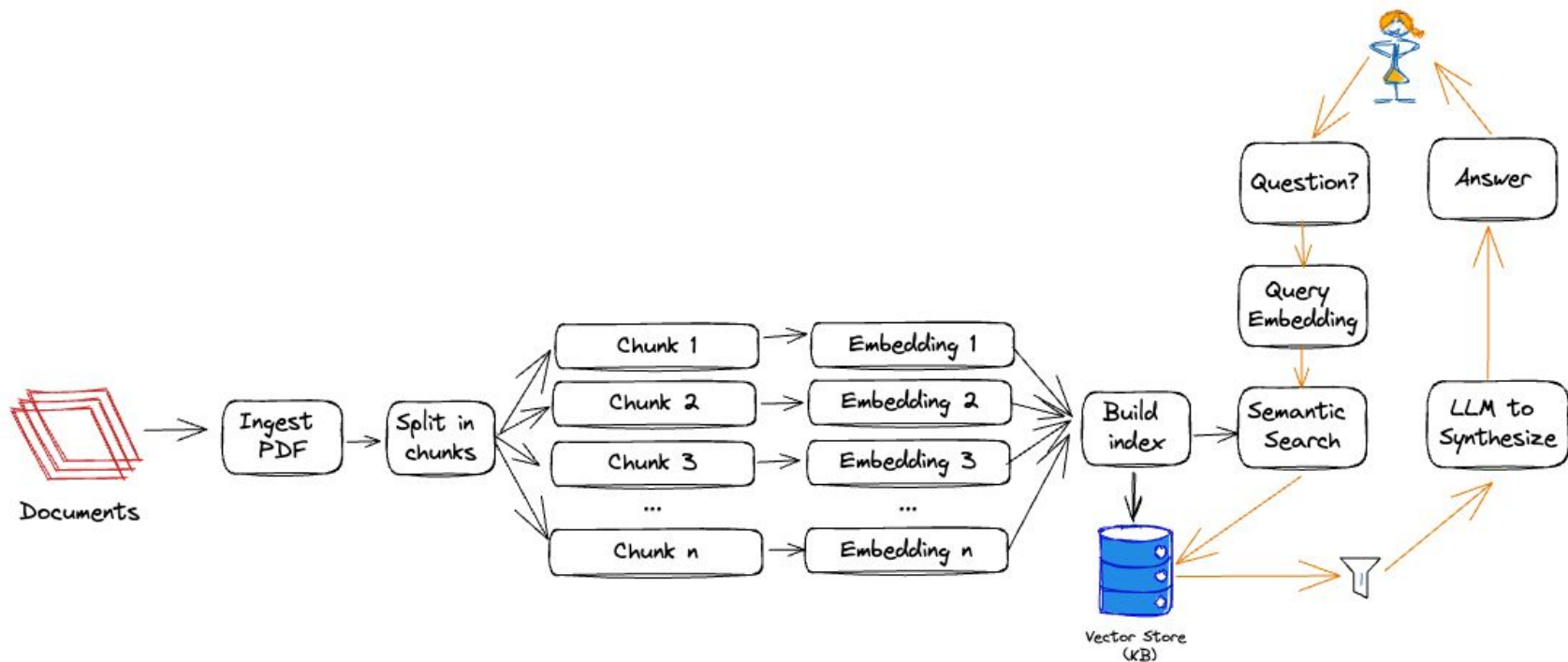
Source: Insider Intelligence: [AI in Finance \(2022\)](#)

Demonstration | hands on | the fun stuff

**Demo** time



# RAG (Retrieval Augmented Generation)



**Bold + Responsible**

# Questions to consider

What's the problem you are trying to solve?  
And what's the impact of doing so?

---

Do you actually need AI / GenAI to solve  
this problem? And how involved in the  
building blocks do you want to be?

---

What about data governance | privacy |  
security | guardrails | MLOps?

**Thank you**

# Useful links

Generative AI on Vertex AI Google Cloud:

<https://cloud.google.com/vertex-ai/docs/generative-ai/learn/overview>

Sample code and notebooks for getting started:

<https://github.com/GoogleCloudPlatform/generative-ai>

Google's AI Principles: <https://ai.google/responsibility/principles/>

Shared fate: Protecting customers with generative AI indemnification:

<https://cloud.google.com/blog/products/ai-machine-learning/protecting-customers-with-generative-ai-indemnification>

Five generative AI use cases for the financial services industry:

<https://cloud.google.com/blog/topics/financial-services/five-generative-ai-use-cases-financial-services-industry>