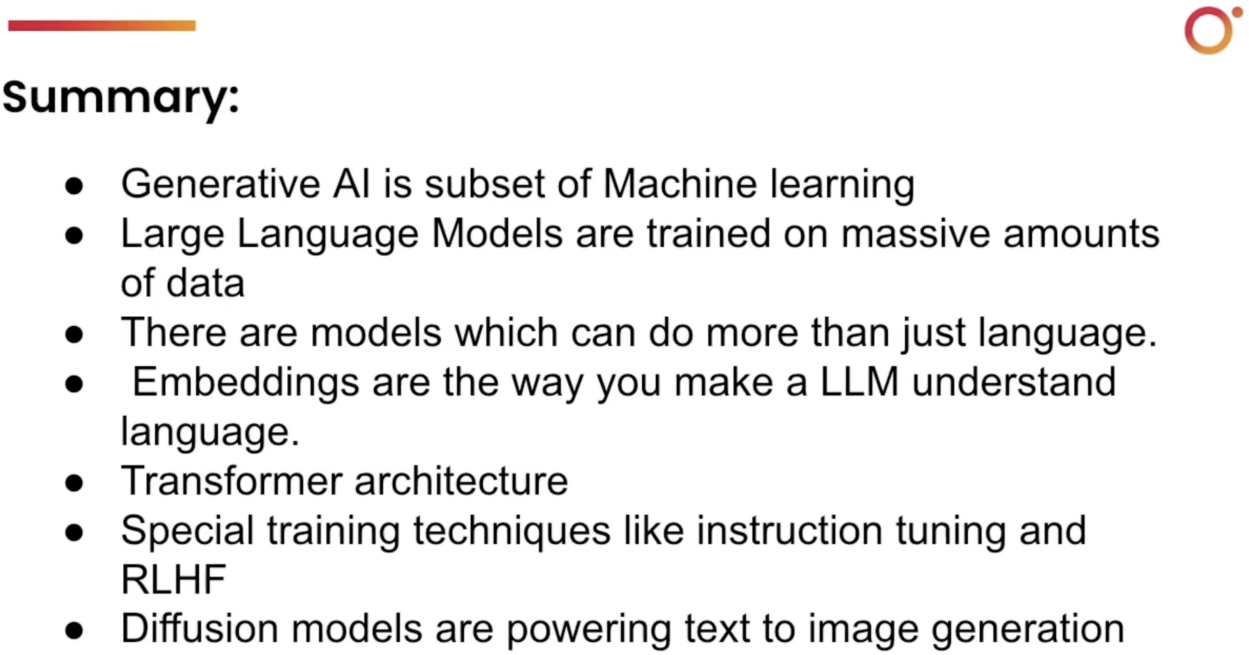
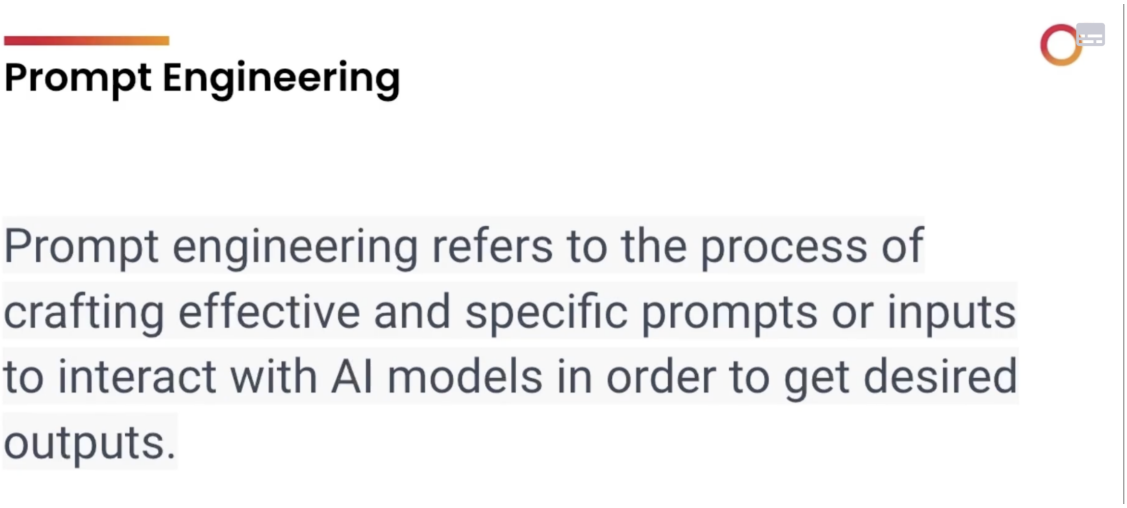
Notes of OReilly course - Generative AI Foundations: Understand Generative AI Applications, Benefits, and Risks

Link: https://learning.oreilly.com/course/generative-ai-foundations/0642572079628/

What is Gen AI



Prompt Engineering



The C.R.E.A.T.E framework

Practice	What is it	Importance	What it does
Character	Crafting prompts with a distinct tone, persona, or specific attributes that align with your brand's identity.	Establishes a consistent brand voice, making AI-generated content feel cohesive and on-brand.	Infuses AI-generated content with a unique personality, reinforcing brand recognition and loyalty.
Request	Constructing clear and specific prompts that explicitly convey the desired outcome or information from the AI model.	Reduces ambiguity, ensuring the AI generates content that meets your intended purpose and requirements.	Guides the AI towards producing accurate and relevant outputs, saving time and improving efficiency.
Examples	Providing relevant examples within the prompts to give the AI context and a reference for generating appropriate content.	Helps the AI understand the desired output style or format, resulting in content that matches expectations.	Enhances the AI's ability to generate content that aligns with the provided examples and context.
Additions	Including specific instructions or keywords that guide the AI to focus on certain aspects or elements in the output.	Allows you to control the AI's creative direction and highlight specific points in the generated content.	Directs the AI's attention to particular details, ensuring the generated content is comprehensive.
Type of Output	Specifying the type of content or format you're seeking from the AI, whether it's a summary, list, explanation, etc.	Tailors the AI's output to match your intended presentation or communication style for better relevance.	Shapes the AI's response to match the desired content structure and format, improving communication.
Extras	Including additional context, relevant background information, or constraints that help the AI better understand the task.	Provides the AI with a complete picture, reducing potential confusion and leading to accurate outputs	Equips the AI with necessary information to generate content that aligns with your specific needs.

A good prompt

You are a highly experienced writer who writes concise and readable text without stop words, filler words or jargon. I want you to summarise the following text, highlighting the most important concepts. Deliver this as a short paragraph of 100 words. Then list the most important points as a bullet-point list. Finally, follow it with a one sentence summary. The text I want you to summarise is "[TEXT]"

Additional learning

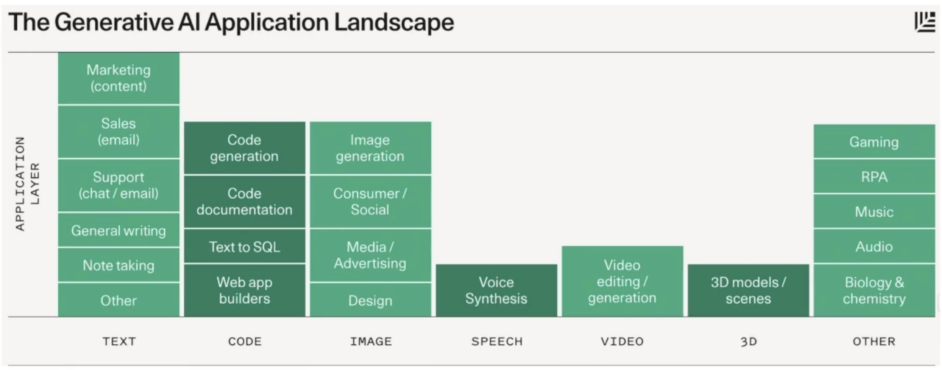
https://www.promptingguide.ai/

Applications of Gen AI

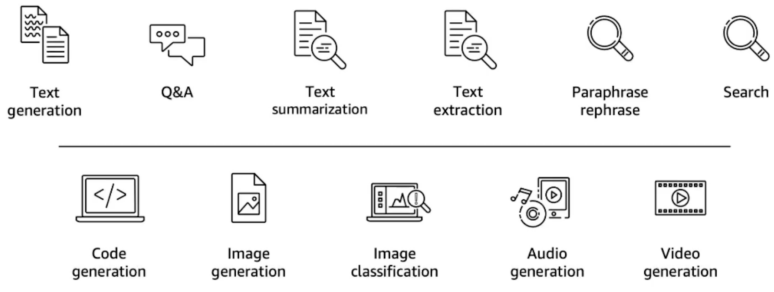
What Generative AI does today?

Read Understand	Generate	Research & Synthesize	Interact

Use Cases and Applications



Use Cases and Applications



Applications and usage

Built into product we use:

Office 365
Dynamics 365
Teams
Power Platform
Zoom

Enhanced Info Synthesis

Question answering on private documents.
Summarizing, ranking, commenting.

Current trends:

Role playing:
Question answering
Utility (reading, filtering etc)
Developer Productivity (50%)

Enterprise systems

Enhanced Intelligence

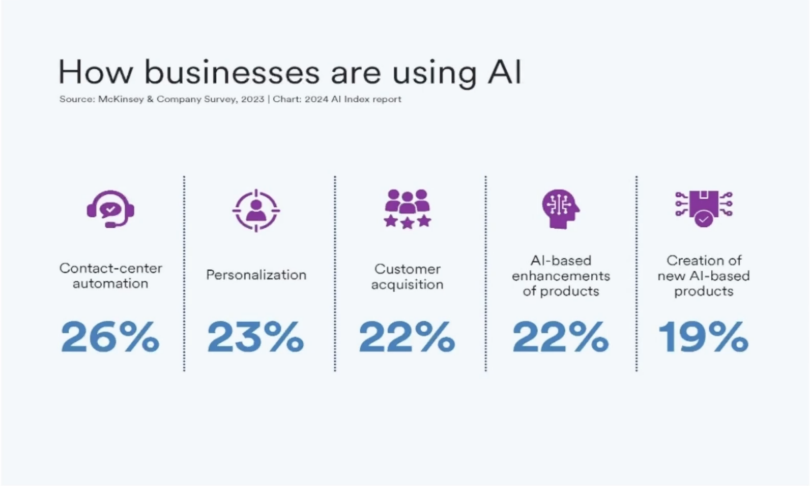
Content Creativity

Marketing, Insurance, Legal, Test Cases, Code snippets, Ideation,
Creative writing, illustrations,,Media

Experimental

Multi-step agents
Hyper automation / personalisation

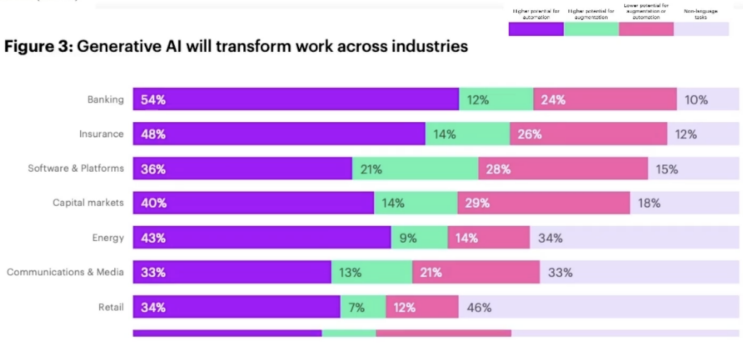
Use Cases of GenAI



Impact on Business

40% of working hours across industries can be impacted by Large Language Models (LLMs)

Figure 3: Generative AI will transform work across industries



Top enterprise use cases for Gen AI

Written and augmented content	Producing a “draft” output of text in a desired style and length
Question answering and discovery	Enabling Knowledge base using private data / In-context learning
Summarization	Text manipulation, to soften language or professionalize text
Tone	Shortened versions of conversations, articles, emails and webpages
Simplification	Breaking down titles, creating outlines and extracting key content
Classification of content	Sorting by sentiment, topic, etc.
Chatbot performance improvement:	Sentiment classification and generation of journey flows from general descriptions
Software coding:	Code generation, translation, explanation and verification

LLM use cases

Economic Impact of Gen AI

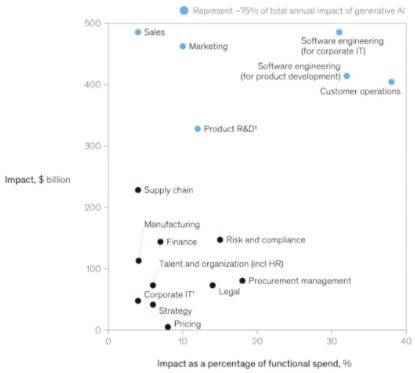
$2.6 - $4.4 trillion global economy

Four key areas:

1. Customer operations,
2. Marketing and sales,
3. Software engineering,
4. Research and Development (R&D).

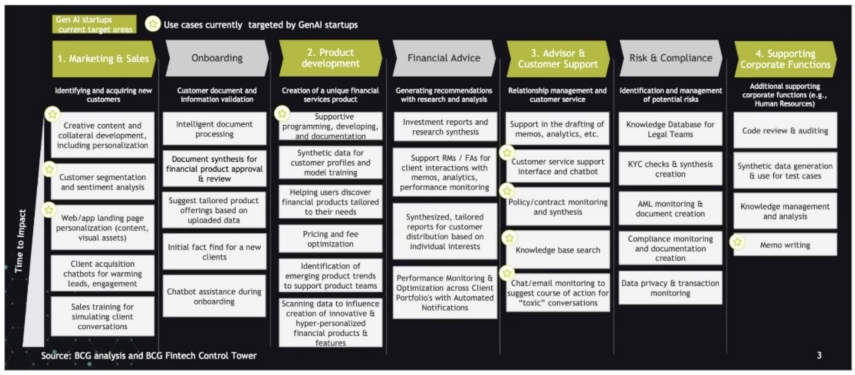
In banking we could potentially see an extra $200 billion to $340 billion in value each year if we fully implement generative AI solutions.

Generative AI is changing the way we work by automating tasks that currently take up 60-70% of employees' time.



Financial services

Generative AI use cases



Financial services : Stories

01	JP Morgan	IndexGPT to select Securities
02	Goldman sachs	POCs, classification docs, Dev team productivity
03	DBS	Personalise product recommendations
04	Alibaba	Tongyi Qianwen, LLM
05	Morgan Stanley	Own chatbot for 16000 advisors

Limitations

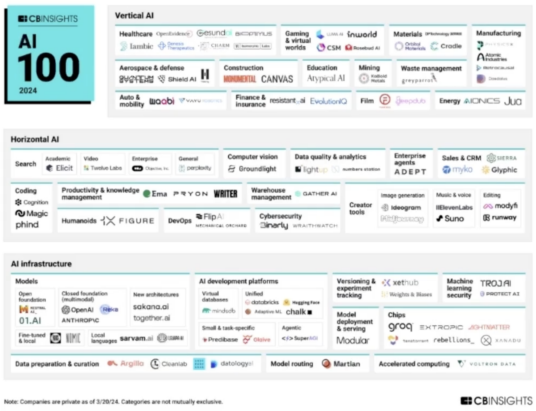
Limitations and Risks

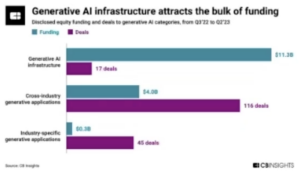
- Lack of Accuracy
- Bias and Fairness
- Misinformation / hallucinations
- Lack of Interpretability
- Ethical concerns
- Data Privacy and Security

Responsible AI

Companies and Tools

Companies



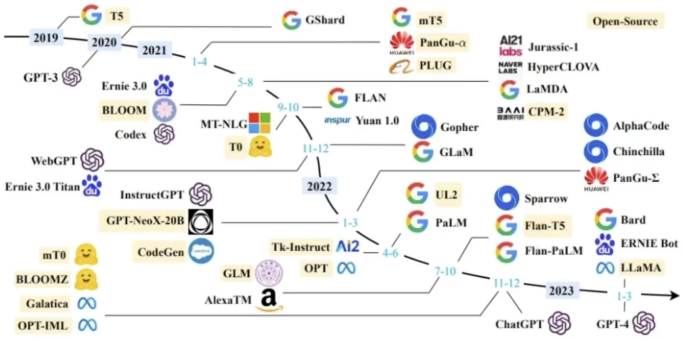


Tools

Tool	Company	Description
Chatgpt, GPT-4	OpenAI	A large language model that can generate text, translate languages, write different kinds of creative content, and answer your questions in an informative way.
Midjourney	Midjourney	A text-to-image AI tool that can create realistic images from text descriptions.
Stable Diffusion	Stable Diffusion	A generative model that can create high-quality images from text descriptions.
GitHub Copilot	Microsoft	A tool that helps programmers write code. It is powered by OpenAI's GPT-3 language model and can suggest code completions, documentation, and other helpful information.
DALL-E 3	OpenAI	An image generation tool that can create realistic images from text descriptions.
Adobe Firefly	Adobe	A tool that can generate realistic images, videos, and 3D models from text descriptions.
D-ID	D-ID	A tool that can create realistic video avatars connected with audio and language models.
Runway.ml	Runway	A platform that allows users to create and deploy generative AI model generating videos, enhancing images from just text.

Open vs Closed Source Models

Open Source Models



Open Source Models

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Choosing Closed vs Open Source

Closed Source (Private)	Open Source (Self hosted)
Effective for Low usage	Effective for high usage
Min Infra setup	High deployment and setup costs
High quality responses	Average quality responses
Sensitive for Privacy	Insensitive to Privacy
External dependency as a black box	Full Control over code and the model.
High Speed of implementation	Requires time and resources to implement

Models

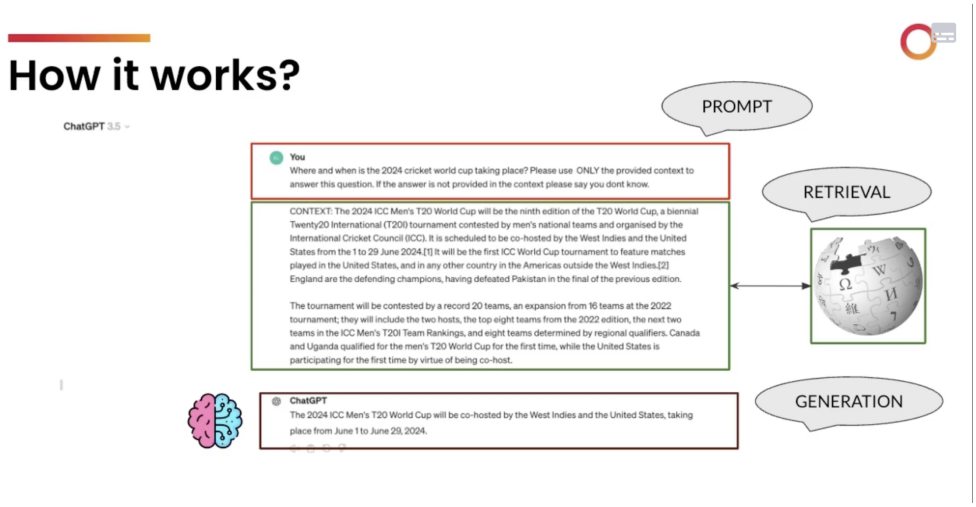
Open Source Models	Closed Source Models
Llama 3.1	GPT4o, GPT-4, GPT 3.5, GPT-3
Mistral	Claude
BLOOM	Cohere
BERT	Gemini
Stable Diffusion	Dall-E

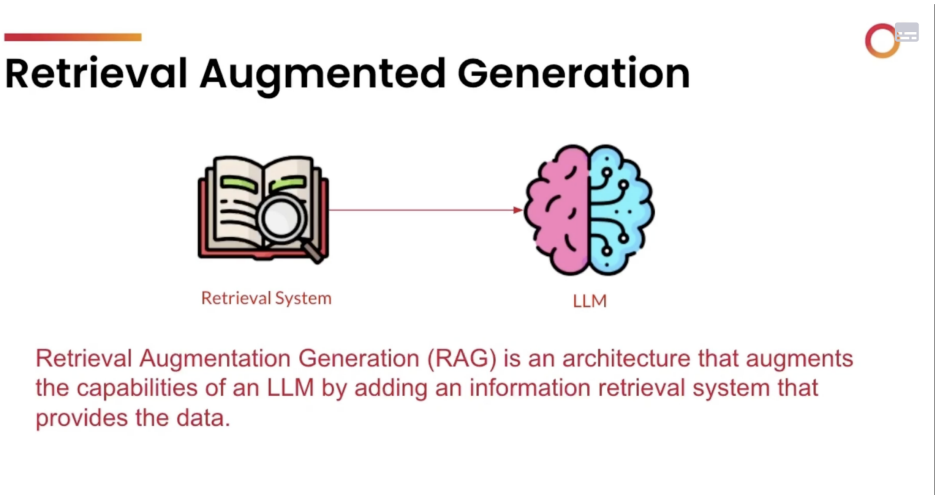
Home of open source models : https://huggingface.co/

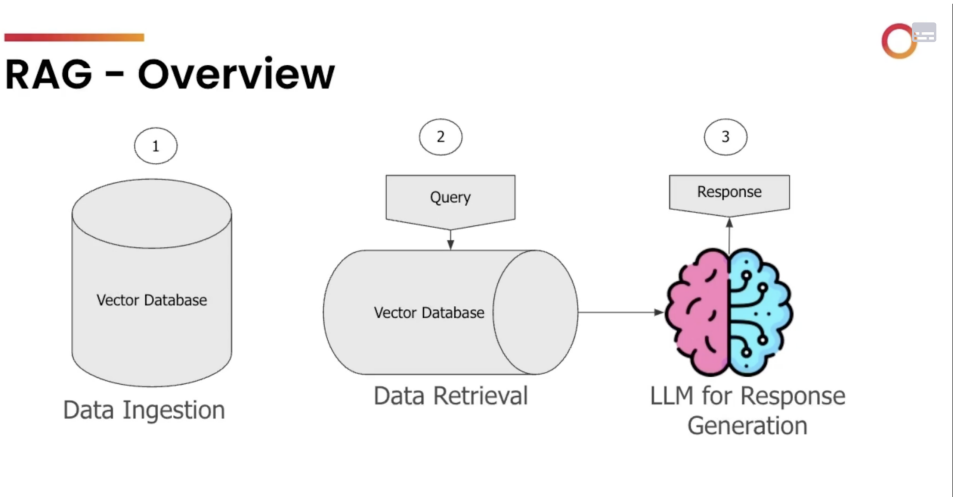
Self hosting vs Vendor LLM

You need	You Choose
More Privacy	Self hosted LLM
Quick Development	OpenAI
Low costs	Open AI
Minimal Infra	Open AI
Full control	Self hosted LLM
No external dependencies	Self hosted LLM
Top quality	Open AI

RAG (Retrieval Augmented Generation)

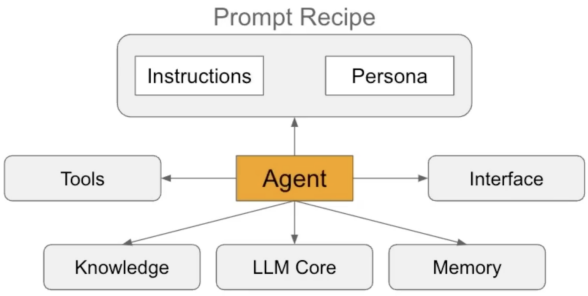

How it works?
ChatGPT 3.5
PROMPT
You
Where and when is the 2024 cricket world cup taking place? Please use ONLY the provided context to answer this question. If the answer is not provided in the context please say you dont know.
RETRIEVAL
CONTEXT: The 2024 ICC Men's T20 World Cup will be the ninth edition of the T20 World Cup, a biennial Twenty20 international (T20I) tournament contested by men's national teams and organised by the International Cricket Council (ICC). It is scheduled to be co-hosted by the West Indies and the United States from the 1 to 29 June 2024.[1] It will be the first ICC World Cup tournament to feature matches played in the United States, and in any other country in the Americas outside the West Indies.[2] England are the defending champions, having defeated Pakistan in the final of the previous edition.
The tournament will be contested by a record 20 teams, an expansion from 16 teams at the 2022 tournament; they will include the two hosts, the top eight teams from the 2022 edition, the next two teams in the ICC Men's T20I Team Rankings, and eight teams determined by regional qualifiers. Canada and Uganda qualified for the men's T20 World Cup for the first time, while the United States is participating for the first time by virtue of being co-host.
ChatGPT
The 2024 ICC Men's T20 World Cup will be co-hosted by the West Indies and the United States, taking place from June 1 to June 29, 2024.
GENERATION


Retrieval Augmented Generation
Retrieval System
LLM
Retrieval Augmentation Generation (RAG) is an architecture that augments the capabilities of an LLM by adding an information retrieval system that provides the data.


RAG - Overview
1
Vector Database
Data Ingestion
2
Query
Vector Database
Data Retrieval
3
Response
LLM for Response Generation

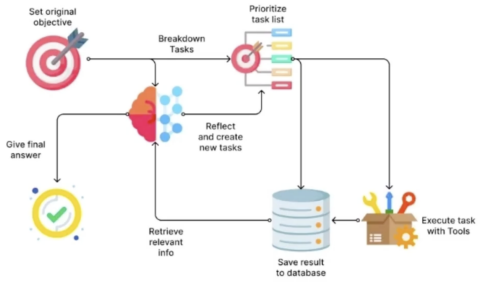
Agents

Agents - Core concept



Agents - BabyAGI

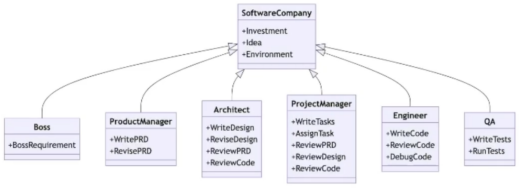
Apply iterative work to accomplish high level goals



MetaGPT: Collaboration with Agents

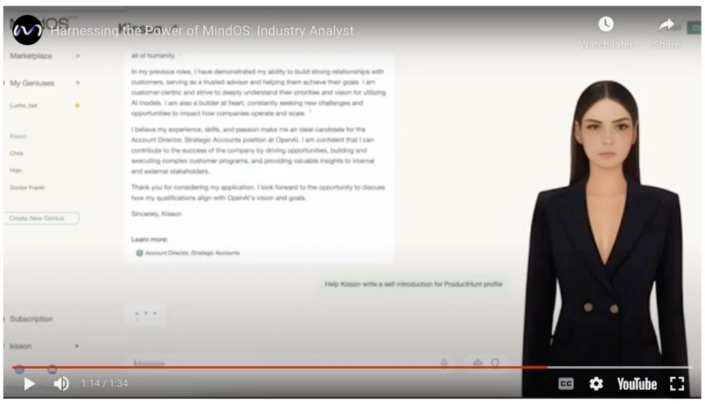
Assign different roles to GPTs to form a collaborative software entity for complex tasks.

1. MetaGPT takes a **one line requirement** as input and outputs **user stories / competitive analysis / requirements / data structures / APIs / documents, etc.**
2. Internally, MetaGPT includes **product managers / architects / project managers / engineers.** It provides the entire process of a **software company along with carefully orchestrated SOPs.**
   i. `Code = SOP(Team)` is the core philosophy. We materialize SOP and apply it to teams composed of LLMs.

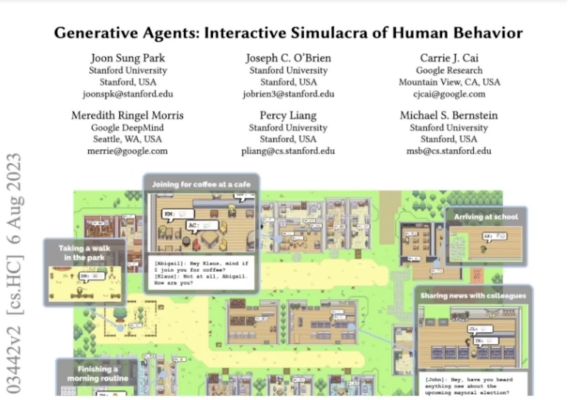


Software Company Multi-Role Schematic (Gradually Implementing)

MindOS: Agent marketplace in action



Collaboration with Agents



Generative Agents: Interactive Simulacra of Human Behavior

Joon Sung Park
Stanford University
Stanford, USA
joonspk@stanford.edu

Joseph C. O'Brien
Stanford University
Stanford, USA
jobrien3@stanford.edu

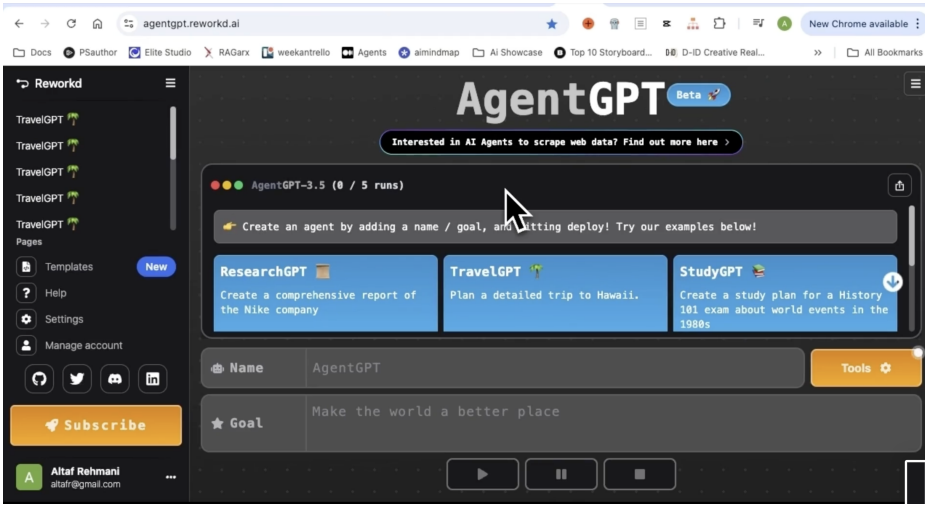
Carrie J. Cai
Google Research
Mountain View, CA, USA
cjcai@google.com

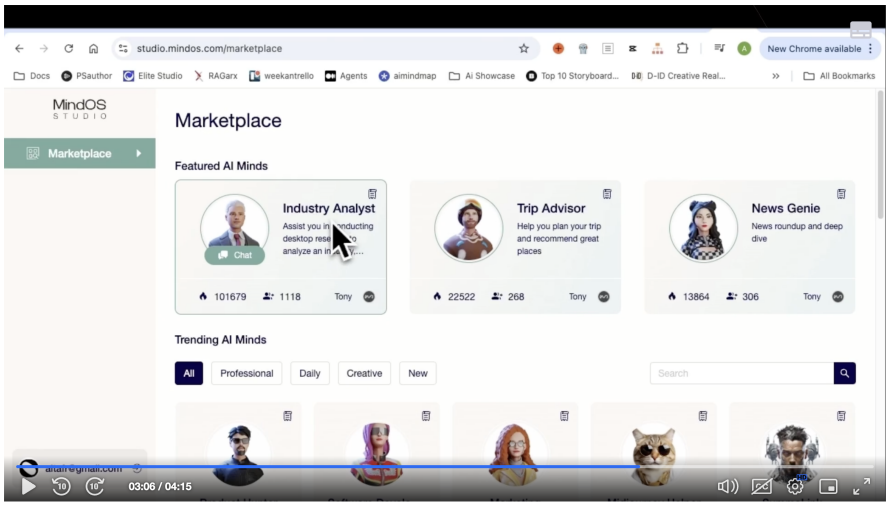
Meredith Ringel Morris
Google DeepMind
Seattle, WA, USA
merrie@google.com

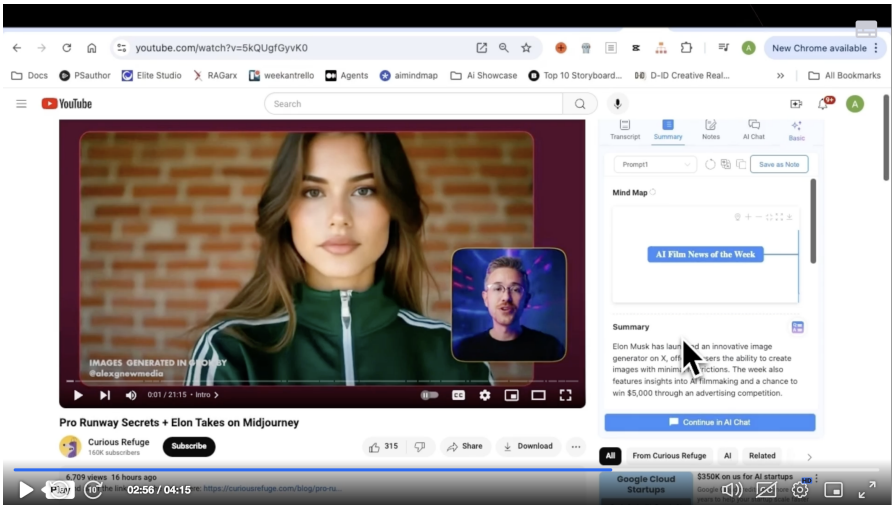
Percy Liang
Stanford University
Stanford, USA
pliang@cs.stanford.edu

Michael S. Bernstein
Stanford University
Stanford, USA
msb@cs.stanford.edu

Some Agents







How to get started / Key Take Aways

How to get Started

Adoption Essentials

1	2	3	4	5	6
Dive in, with a business-driven mindset	Take a people-first approach	Get your proprietary data ready	Invest in a sustainable tech foundation	Accelerate ecosystem innovation	Level-up your responsible AI

Key Takeways

- LLMs are next word predictors based on networks which have “attention” and are autoregressive
- Move towards Factual Information, built into everyday experiences
- Wide range of Applications in every industry
- MultiModal Models - Images, Audio,Video, 3D
- Rise of Open Source Models & RAG
- Responsible AI to mitigate Risks and Limitations
- Agents: The Future of Autonomous Tasks

Future

Are you ready?

Before long, GenAI will greatly impact product development, customer experience, employee productivity and innovation. We predict that:

By 2025, 70% of enterprises will identify the sustainable and ethical use of AI among their top concerns.

By 2025, 35% of large organizations will have a chief AI officer who reports to the CEO or COO.

By 2025, the use of synthetic data will reduce the volume of real data needed for machine learning by 70%.

By 2025, 30% of outbound marketing messages from large organizations will be synthetically generated. That’s up from less than 2% in 2022.

Through 2026, despite all the advancements in AI, the impact on global jobs will be neutral — there will not be a net decrease or increase.

By 2030, AI could reduce global CO2 emissions by 5 to 15% and consume up to 3.5% of the world’s electricity.

By 2030, decisions made by AI agents without human oversight will cause $100 billion in losses from asset damage.

By 2033, AI solutions will result in more than half a billion net-new human jobs.