

Generative AI for Everyone

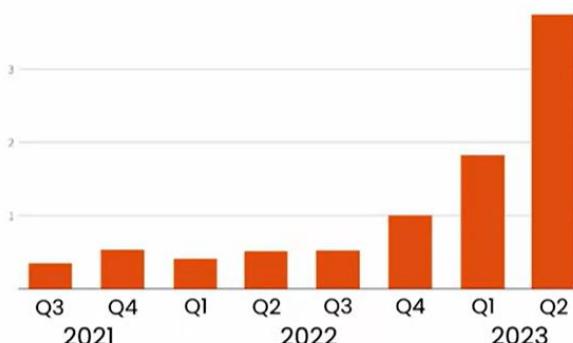
Week 1

Welcome

Rise of Gen AI

The rise of generative AI

Average number of 'AI' mentions per S&P 500 analyst call



Source: Reuters

Generative AI could

- Add \$2.6-\$4.4 trillion annually to the economy¹
- Raise global GDP by 7% in the next 10 years²
- Impact 10% of the tasks carried out daily by 80% of workers³

Sources: (1) McKinsey, (2) Goldman Sachs, (3) OpenAI/UPenn

What is Gen AI?

What is generative AI?

Artificial intelligence systems that can produce high quality content, specifically **text, images, and audio**.

Social post caption

AN Write three captions for a social post about our new line of sunglasses for robots

1. "Shine bright, even if your eyes are digital! Introducing our latest sunglasses line for robots."

2. "Because even bots deserve to look cool in the sun! #RobotChic"

3. "Future's so bright, our robots need shades! Check out our new eyewear collection now."

Regenerate

Send a message

ChatGPT/OpenAI

Bard Experiment

Give me three ideas for t-shirt graphics that would appeal to llamas.

Sure, here are three ideas for t-shirt graphics that would appeal to llamas:

1. A cartoon llama with a funny expression, such as a surprised llama, a grumpy llama, or a silly llama.
2. A llama doing something that is typically associated with llamas, such as eating grass, spitting, or being majestic.

Bard/Google

Microsoft Bing

Find me some recipes for snacks for hungry pirates

Searching for: snacks for hungry pirates

Generating answers for you...

Sure! Here are some fun and tasty pirate-themed snacks that you can prepare for hungry pirates:

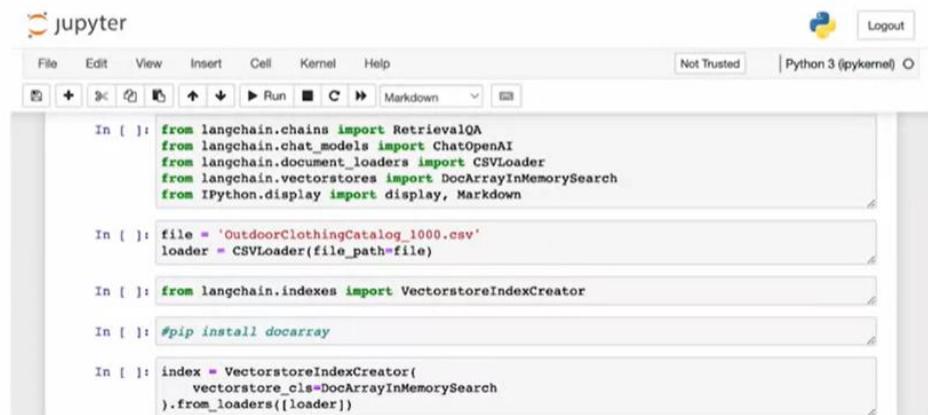
1. **Pirate Mix:** This snack includes pirate pegleg pretzel sticks, Gorilla Munch cannonballs, and "gold" "fish" whole grain Goldfish crackers.

Bing Chat/Microsoft

Developer tool

What is generative AI?

Generative AI is also a developer tool



The screenshot shows a Jupyter Notebook interface with the title "jupyter" at the top. The menu bar includes File, Edit, View, Insert, Cell, Kernel, and Help. On the right, there are buttons for "Logout" and "Python 3 (pykernel)". A status bar at the bottom indicates "Not Trusted". The notebook has several code cells:

```
In [ ]: from langchain.chains import RetrievalQA
from langchain.chat_models import ChatOpenAI
from langchain.document_loaders import CSVLoader
from langchain.vectorstores import DocArrayInMemorySearch
from IPython.display import display, Markdown

In [ ]: file = 'OutdoorClothingCatalog_1000.csv'
loader = CSVLoader(file_path=file)

In [ ]: from langchain.indexes import VectorstoreIndexCreator

In [ ]: #pip install docarray

In [ ]: index = VectorstoreIndexCreator(
    vectorstore_cls=DocArrayInMemorySearch
).from_loaders([loader])
```

AI is already pervasive in our lives

AI is already pervasive in our lives

AI technology	Examples
Web search	Google, Bing
Fraud detection	Credit card payments
Recommender system	Amazon, Netflix

Fff

Image, Audio and video generative

Image, Audio and Video generation

A beautiful, pastoral mountain scene.
Landscape painting style (Midjourney)



Two cute kittens playing (DALL-E)



Image, Audio and Video generation

Video generation



What you will learn

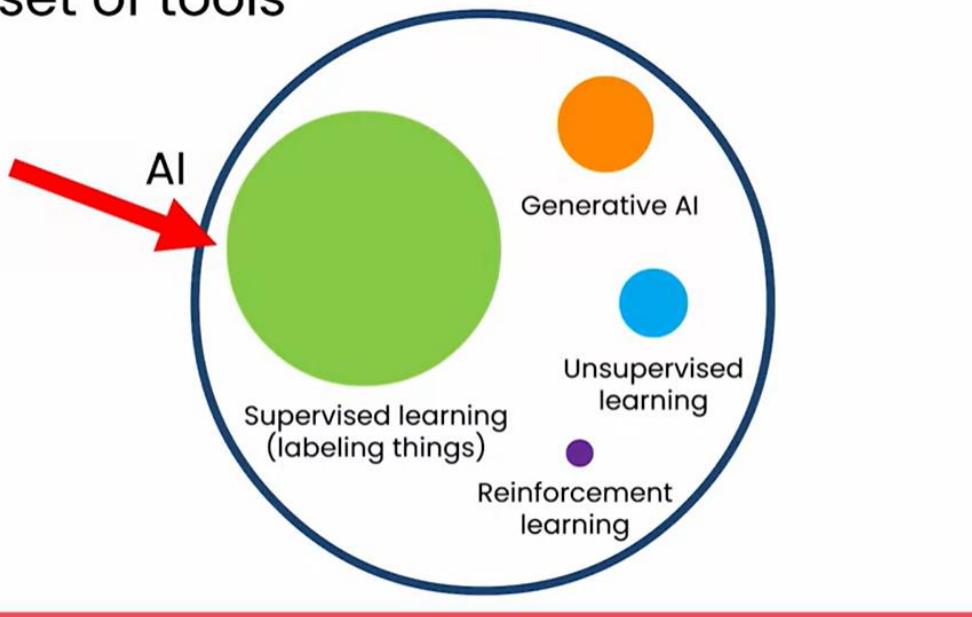
What you'll learn

- How generative AI technology works
 - What it can and can't do
 - Common use cases
- Generative AI Projects
 - Identify and build Generative AI use cases
 - Technology options
- Impact on business and society
 - How teams can take advantage of Generative AI
 - AI risks and responsible AI

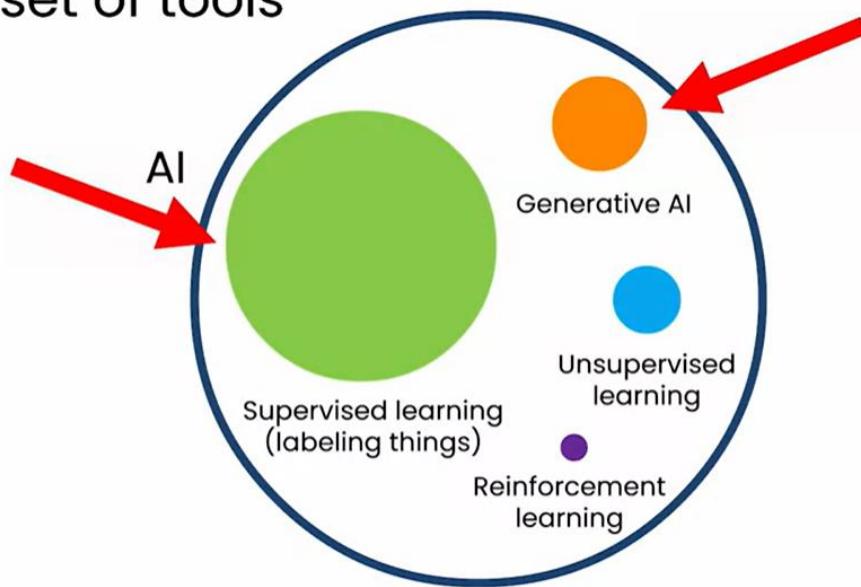
How Generative AI works

AI is a set of tools

AI is a set of tools



AI is a set of tools



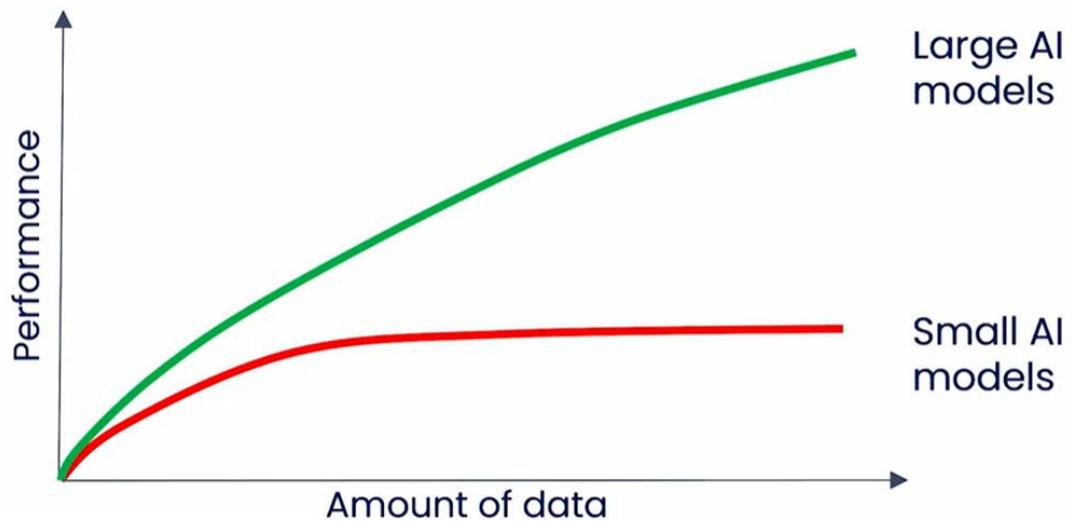
Supervised learning (labelling things)

Supervised learning (labeling things)

Input (A)	Output (B)	Application
Email	Spam? (0/1)	Spam filtering
Ad, user info	Click? (0/1)	Online advertising
Image, radar info	Position of other cars	Self-driving car
X-ray image	Diagnosis	Healthcare
Image of phone	Defect? (0/1)	Visual inspection
Audio recording	Text transcript	Speech recognition
Restaurant reviews	Sentiment (pos/neg)	Reputation monitoring

2010-2020 – Large scale supervised learning

2010–2020: Large scale supervised learning



Generative text using Large Language Models (LLMs)

Generating text using Large Language Models (LLMs)

Text generation process



How LLM work

- Based on Supervised training

How Large Language Models (LLMs) work

LLMs are built by using supervised learning (A→B) to repeatedly predict the next word.

My favorite food is a bagel with cream cheese

Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

When we train a very large AI system on a lot of data (hundreds of billions of words), we get a Large Language Model like ChatGPT.

LLM as a thought partner

A new way to find information

A new way to find information

What is the capital of South Africa?

South Africa has three capital cities:
1. Pretoria (administrative capital)
2. Cape Town (legislative capital)
3. Bloemfontein (judicial capital)

What does LLM stand for?

LL.M. stands for "Legum Magister" in Latin, which translates to "Master of Laws" in English.

What about in the context of AI?

In the context of AI, LLM refers to "Large Language Model".

Writing Partner

Writing partner

Rewrite this for clarity:

Students all around the world are realizing learning has to happen not just in school but throughout life including after graduation.

Students worldwide understand that learning extends beyond school and continues throughout life.

Write a 300 word story involving trucks for a 3 year old to encourage them to brush their teeth.

Once upon a time in a town called Toothville, there lived a happy tow truck named Timmy.

One day Timmy noticed that his shiny bumper had a dull spot. Worried, he went to the doctor [...]

Web search or using an LLM?

Web search or using an LLM?

What should I do to treat a sprained ankle?

The screenshot shows the Mayo Clinic website. At the top, there's a navigation bar with links for "Diseases & Conditions", "Request an Appointment", "Login", and a menu icon. Below the header, a large blue banner displays the title "Sprained ankle". Underneath the banner are three buttons: "Symptoms & causes", "Diagnosis & treatment", and "Doctors & departments". The main content area starts with a section titled "Diagnosis", which includes a detailed paragraph about how doctors examine the ankle. There are also sections for "Treatment" and "Doctors & departments".

Diagnosis

During a physical, your doctor will examine your ankle, foot and lower leg. The doctor will touch the skin around the injury to check for points of tenderness and move your foot to check the range of motion and to understand what positions cause discomfort or pain. If the injury is severe, your doctor may recommend one or more of the following imaging scans to rule out a broken bone or to evaluate in more detail the extent of ligament damage:

Mayo Clinic

The screenshot shows the Harvard Health Publishing website. At the top, there's a header with the Harvard logo and the text "Harvard Health Publishing HARVARD MEDICAL SCHOOL". Below the header, a section titled "RECENT BLOG ARTICLES" features a thumbnail image of a bowl of produce and the text "Produce prescriptions may promote better heart health". A navigation bar below the article includes arrows for "PAIN" and "RECENT BLOG ARTICLES".

Recovering from an ankle sprain

June 21, 2021

All it takes is a simple misstep, and suddenly you have a sprained ankle. An ankle sprain is one of the most common musculoskeletal injuries in people of all ages, athletes and couch potatoes alike. The injury occurs when one or more

Harvard Medical School

What should I do to treat a sprained ankle?

I'm not a doctor, but if you believe you have a sprained ankle, the following general steps are often recommended:

- **Rest:** Avoid putting weight on the injured ankle
- **Ice:** Apply ice wrapped in cloth to injured area

[...]

Web search or using an LLM?

pineapple pie recipe

Google pineapple pie recipe

Videos No bake Images Old fashioned With condensed milk Easy Custard

About 58,100,000 results (0.38 seconds)

Recipes :

Pineapple Pie	Pineapple Pie	Tropical Pineapple Pie Recipe
Allrecipes 4.7 ★★★★★ (136) 55 min Crushed pineapple, double crust pie, milk, lemon juice,	Mother Thyme 4.6 ★★★★★ (20) 5 min Cool whip, sour cream, bought graham cracker, crushed	House of Nash Eats 5.0 ★★★★★ (1) 1 hr Pie crust, crushed pineapple, lemon juice, egg, lemon zest

Give me a pineapple pie recipe

Certainly! Here's a simple pineapple pie recipe:

Ingredients:

1 prepared pie crust
2 cups of crushed pineapple (drained)
1 cup of granulated sugar
2 tablespoons cornstarch [...]

Web search or using an LLM?

coffee-infused pineapple pie recipe

Google coffee infused pineapple pie

Shopping Images Recipe With condensed milk Old fashioned Videos News

About 5,770,000 results (0.32 seconds)

Recipes :

Pineapple Pie	Pineapple Coffee Cake	Coffee Pie
Mexico In My Kitchen 4.9 ★★★★★ (85) 1 hr 30 min Pie crust, fresh pineapple, eggs, flour, white sugar	Bunny's Warm Oven 5.0 ★★★★★ (1) Sour cream, pineapple tidbits, egg, baking soda, baking	Southern Plate No reviews 3 hr 45 min Sweetened condensed milk, self rising flour, graham cracker

Give me a coffee-infused pineapple pie recipe

Sure! Here's a coffee-infused pineapple pie recipe:

Ingredients:

1 prepared pie crust
2 cups crushed pineapple (drained)
2 tablespoons strong brewed coffee
1 cup granulated sugar [...]

AI as general purpose technology

Electricity

AI is a general purpose technology

Similar to electricity, AI is useful for many tasks

Input (A)	Output (B)	Application
Email	Spam (0/1)	Spam filtering
Ad, user info	Click? (0/1)	Online advertising
Image, radar info	Position of other cars	Self-driving car
X-ray image	Diagnosis	Healthcare
Image of phone	Defect? (0/1)	Visual inspection
Audio recording	Text transcript	Speech recognition
Restaurant reviews	Sentiment (pos/neg)	Reputation monitoring

Example if tasks LLMs can carry out

- Writing - Brain storming – Web search
- Reading – Summarize – Large automating software - app based
- Chatting – company policy – app based

Examples of tasks LLMs can carry out

Writing	Reading	Chatting
Suggest three names for our new analytics platform	<i>I love my new llama T-shirt! The fabric is so soft.</i>	Welcome to BettaBurgers!
Sure, here are three ideas: 1. NexasInsight 2. Quantivue 3. AnalytiSphere	Complaint: NO Department: Apparel	I'd like a cheeseburger for delivery!
Is there parking for staff?	<i>I wore my llama T-shirt to a friend's wedding, and now they're mad at me for stealing the show.</i>	Awesome. Anything else?
Yes, employees can park on levels 1 and 2 of the office lot. You can get a sticker at [...].	Complaint: YES Department: Apparel	That's it thanks!
		Ok, it'll be there in 20 minutes. Enjoy!

Web – based vs software application using LLMs

Web-based vs. software application use of LLMs



Web-based interface
applications e.g. ChatGPT,
Bard, or Bing Chat



Software-based
applications e.g. email
routing, document search

QUIZ What is Generative AI

The screenshot shows a Coursera quiz interface. At the top, there's a navigation bar with the Coursera logo, a search bar, and user profile information for 'Aroh Shukla'. Below the navigation, the page title 'What is Generative AI?' is displayed. The main content area contains a question about generative AI. On the left, there's a section for submitting an assignment with a due date of 'Nov 12, 11:59 PM +08'. On the right, there's a 'Start assignment' button. In the center, there's a box for receiving a grade, with the text 'To Pass: 80% or higher' and a red rectangular highlight around it. At the bottom, there are social sharing icons for 'Like', 'Dislike', and 'Report an issue'.

1. Which of these is the best definition of "Generative AI"? 1/1 point

A form of web search.
 Artificial intelligence systems that can map from an input A to an output B.
 Any web-based application that generates text.
 AI that can produce high quality content, such as text, images, and audio.

Correct
Generative AI refers to a collection of tools that can generate high quality text, images, and audio, including large language models (LLMs) and diffusion models for image generation.

2. Which of these is the most accurate description of an LLM? 0/1 point

It generates text by repeatedly predicting words in random order.
 It generates text by ~~repeatedly predicting the next word~~.
 It generates text by using supervised learning to carry out web search.
 It generates text by finding a writing partner to work with you.

Incorrect
LLMs do not predict words in random order. Instead, they repeatedly predict the next word in a sequence of words.

3. True or False. Because an LLM has learned from web pages on the internet, its answers are always more trustworthy than what you will find on the internet. 1/1 point

True
 False

Correct
Because LLMs can hallucinate (make up facts), it is best to fact-check the response from an LLM before using it in situations where factual accuracy is important.

4. Why do we call AI a general purpose technology? 1/1 point

Because it is useful for many different tasks.
 Because it can chat.
 Because it includes both supervised learning and generative AI.
 Because it can be accessed via the general web.

Correct
General purpose technologies are, by definition, designed to be versatile and useful for a wide range of tasks. This broad utility across various applications is what characterizes AI as a general purpose technology.

Writing

Brainstorming products names

Brainstorming product names

Brainstorm 5 creative names for peanut butter cookies

Sure thing! Here are some ideas:

1. Nutty Nirvana Nibbles
2. ButterBite Bliss
3. Peanut Paddle Pucks
4. ButterNutter Crunchies
5. Nutty Nom-Nom Nuggets

Develop sales strategy

Developing sales strategy

Brainstorm five ideas for increasing cookie sales in Q4

1. Limited-time only seasonal flavors.
2. Beautifully packaged gift sets.
3. Buy more, save more deals.
4. Interactive holiday displays
5. Recipe cards

Write a press release

Writing a press release

Write a press release announcing the hire of a new COO

[Company Name] Welcomes [New COO's Full Name] as Chief Operations Officer

[City, State, Date] – [Company Name], a leading [brief description of the company's industry or niche], is thrilled to announce [New COO's Full Name] as its new Chief Operations Officer (COO). [New COO's first name] brings a wealth of experience [...]

Write a press release – Improved prompt

Writing a press release – improved prompt

Write a press release announcing the hire of our new COO using the following information:

COO bio: Nadiya Grenner received her MBA from Cornell...

Company info: General Robotics, headquartered in Boston, MA, is a leader in robotics research and manufacturing...

General Robotics Welcomes Nadiya Grenner as Chief Operations Officer

Boston, MA – General Robotics, a leading robotics research and manufacturing company, is thrilled to announce Nadiya Grenner as its new COO. A graduate of Cornell's [...]

Translation

Translation

Translate the following into formal Hindi:

Welcome to the Oceanview Hotel – we hope you enjoy your stay. If you need anything, please contact the front desk.

ओशनव्यू होटल में आपका स्वागत है - हम आशा करते हैं कि आपका ठहराव यहाँ आनंदमय हो। अगर आपको कुछ चाहिए होता है, तो कृपया फ्रंट डेस्क से संपर्क करें।

Translation

Translate the following into formal spoken Hindi:

Welcome to the Oceanview Hotel – we hope you enjoy your stay. If you need anything, please contact the **front desk**.

ओशनव्यू होटल में आपका स्वागत है। हम उम्मीद करते हैं कि आप यहाँ अच्छा समय बिताएंगे। अगर आपको कुछ भी चाहिए हो, तो कृपया **रिसेप्शन** पर संपर्क करें।

Reading

Proofreading

Proofreading

Proofread the following text, intended for a website selling children's stuffed toys, for spelling and grammatical errors, and rewrite it with corrections:

Meet this adorable stuffed panda! Made from super **snuggle** materials and easy **to wash in the machine**, this toy is the perfect friend for your little one.

Meet this adorable stuffed panda! Made from super **snuggly** materials that are **machine-washable**, this toy is the perfect friend for your little one.

Summarize an article – web

Summarizing an article



E. Brynjolfsson, 2022



Summarize the following paper in 300 words or fewer:

In 1950, Alan Turing proposed an “imitation game” as the ultimate test of whether a machine was Intelligent [...]

Human-like AI offers benefits but risks worker power. Augmenting humans creates more value than imitating them [...]

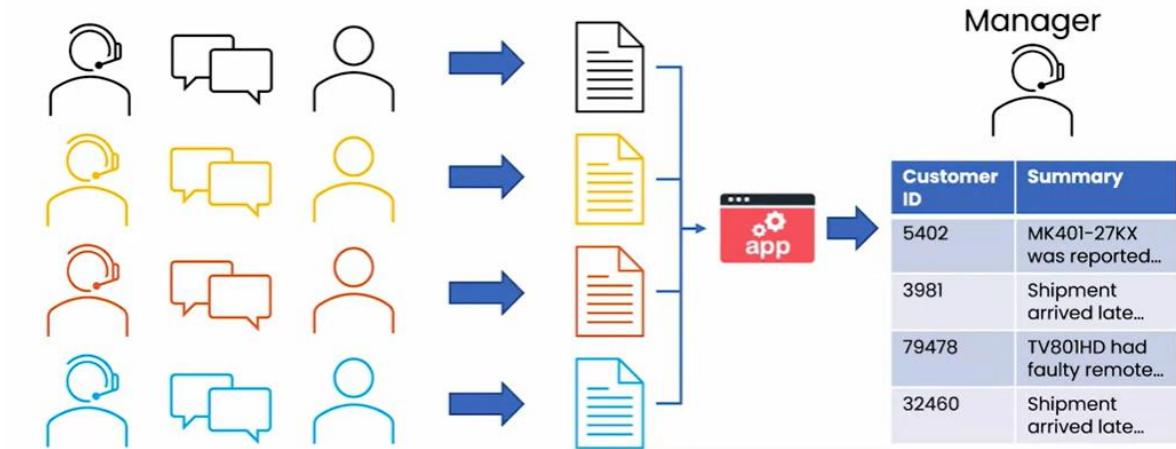


Summarize call centre conversations - app

Summarizing call center conversations



Summarizing call center conversations



Customer email analysis - app

Customer email analysis

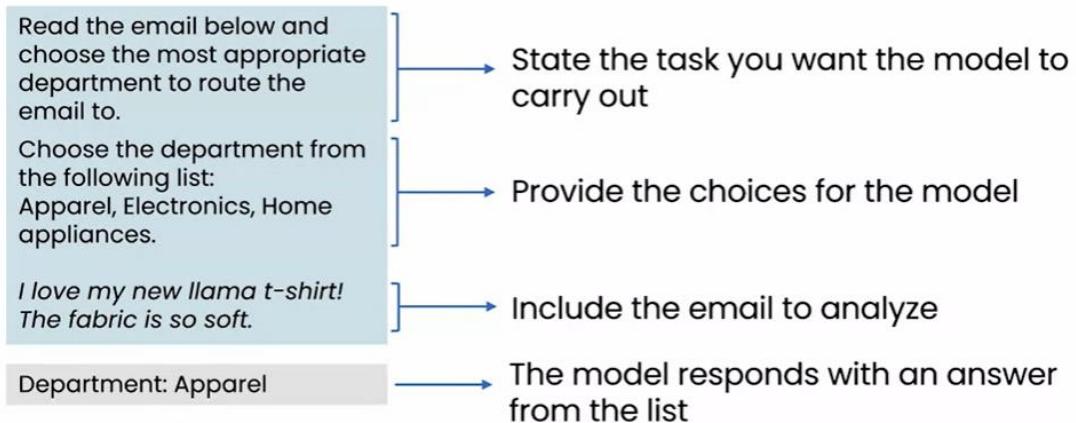
I love my new llama t-shirt!
The fabric is so soft.

Complaint – NO
Department – Apparel



Detailed look at email routing prompt

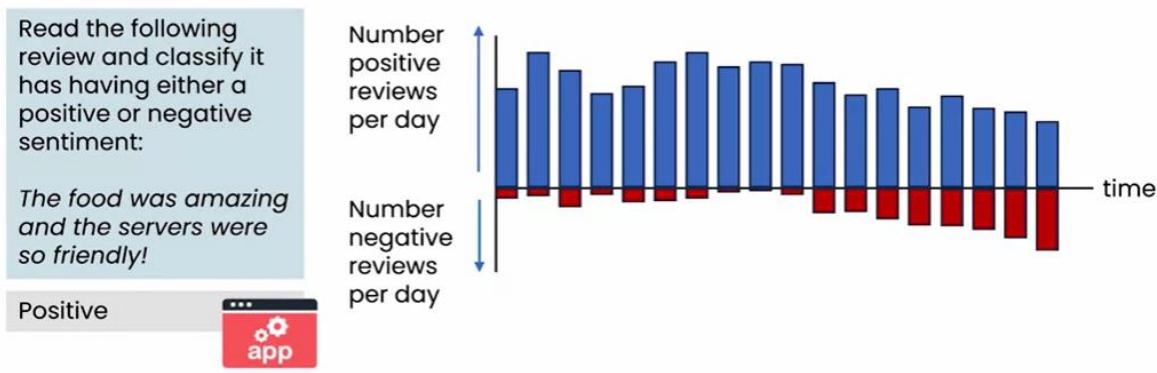
Detailed look at email routing prompt



Reputation monitoring - app

Reputation monitoring

Dashboard to track customer sentiment over time



Chatting

Customer service chatbot - app

Customer service ChatBot

Welcome to BettaBurgers!

I'd like a cheeseburger for delivery!

Awesome. Anything else?

That's it thanks!

Ok, it'll be there in 20 minutes. Enjoy!

Specialist chatbots - app

Specialized Chatbots

How can I vacation in Paris inexpensively?

Here are some ideas to save money in Paris:

1. Eat at bakeries
2. Take metro, not taxis
3. Visit free attractions [...]

Trip planner

I'm nervous about my big presentation at work...

It's natural to feel nervous. What worries you most?

That I'll forget what to say...

You aren't alone! Here are some tips that may help:

1. Use index cards.
2. Picture a friend in the room to present to [...]

Career coach

What can I make with the following ingredients?

Pasta, eggs, lemons, ham

Here's a recipe you can make with those ingredients:

Ham and Lemon Carbonara

Instructions:

1. Cook pasta
2. Whisk eggs, lemon juice and zest in bowl [...]

Recipe ideas

IT Service Chatbot - app

IT Service ChatBot

My login is andrew and I need to reset my account password.

Sure! I just texted you a code. What code did you get?

123456

Ok, I've verified your identity. Please check your email for your new password.

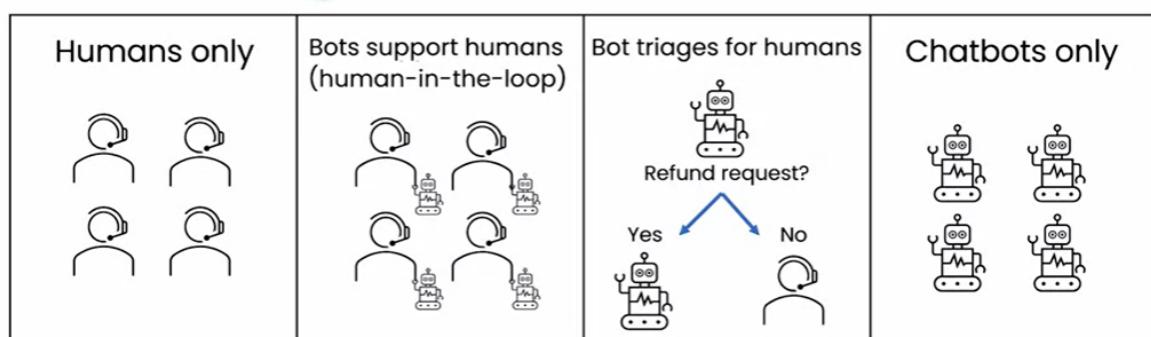
Rise of chatbots in customer service

The rise of chatbots in customer service

Welcome to BettaBurgers!

I'd like a cheeseburger for delivery!

→ Focus on text-based chatbots



Advice for deploying chatbots

Advice for deploying chatbots

- Start with an internal-facing chatbot
 - Work with staff to assess behavior of chatbot
 - Avoid public mistakes
- Deploy with human-in-the-loop to check for mistakes
- Only after deemed safe, allow bot to communicate directly with customers

Examples of tasks LLMs can carry out

Examples of tasks LLMs can carry out

Writing	Reading	Chatting
Suggest three names for our new analytics platform	<i>I love my new llama T-shirt! The fabric is so soft.</i>	Welcome to BettaBurgers!
Sure, here are three ideas: 1. NexasInsight 2. Quantivue 3. AnalytiSphere	Complaint: NO Department: Apparel	I'd like a cheeseburger for delivery!
 Is there parking for staff?	 I wore my llama T-shirt to a friend's wedding, and now they're mad at me for stealing the show.	Awesome. Anything else?
Yes, employees can park on levels 1 and 2 of the office.  You can get a sticker at	Complaint: YES Department: Apparel	That's it thanks!  Ok, it'll be there in 20 minutes. Enjoy!

What LLMs can and cannot do

What prompting an LLA can do

What prompting an LLM can do

Can a fresh college graduate follow the instructions in the prompt to complete the task?

- Can they determine if an email is a complaint?
- Can they determine if a review is positive or negative?

Read the following email and determine whether it contains a complaint:

*I love my new llama T-shirt!
The fabric is so soft.*

Complaint: NO

Read the following review and classify it has having either a positive or negative sentiment:

The food was amazing and the servers were so friendly!

Positive

What prompting an LLM can do

Can a fresh college graduate follow the instructions in the prompt to complete the task?

- Can they write a press release, without any information about the COO or your company?

Write a press release announcing the hire of a new COO

[Company Name] Welcomes [New COO's Full Name] as Chief Operations Officer

[City, State, Date] – [Company Name], a leading [brief description of the company's industry or niche], is thrilled to announce [New COO's Full Name] as its new Chief Operations Officer (COO) [...]

What prompting an LLM can do

Can a fresh college graduate follow the instructions in the prompt to complete the task?

- Can they write a press release given basic relevant context?

Write a press release announcing the hire of our new COO using the following information:

COO bio: Nadiya Grenner received her MBA from Cornell [...]

Company info: General Robotics, headquartered in Boston, MA [...]

General Robotics Welcomes Nadiya Grenner as COO

Boston, MA – General Robotics, a leader in robotics research and manufacturing, is thrilled to announce Nadiya Grenner [...]

What prompting an LLM can do

Can a fresh college graduate follow the instructions in the prompt to complete the task?

For clarity, assume:

- No access to the internet or other resources
- No training specific to your company/business
- No memory of previous tasks completed
 - You get a different fresh college grad every time!

Knowledge Cutoffs – Limitations 1

Knowledge cutoffs

An LLM's knowledge of the world is frozen at the time of its training

- A model trained on data scraped from internet in January 2022 has no information about more recent events

What was the highest grossing film of 2022?

As of January 2022, I don't have data on the highest-grossing movie for that year.



Avatar: The Way of Water

Knowledge cutoffs

Claims of room temperature superconductor LK-99



Tell me about LK-99, the claimed room temperature superconductor

As of my last update in January 2022, there isn't a material specifically known as "LK-99" that is identified as a room-temperature superconductor.



Making things up: Hallucinations

Give me three quotes that Shakespeare wrote about Beyoncé

1. Her vocals shine like the sun.
2. All hail the queen, she is most worthy of love.
3. Such a voice, so electric and alive, none else can compare!

✗ Hallucination

Making things up: Hallucinations

List two court cases tried in California about AI

1. *Waymo v. Uber* – Theft of trade secrets related to autonomous driving
2. *Ingersoll v. Chevron* – Patent infringement of oil field machine learning technology

✓ Real case

✗ Hallucination

Hallucinations have had serious consequences

The ChatGPT Lawyer Explains Himself

In a cringe-inducing court hearing, a lawyer who relied on A.I. to craft a motion full of made-up case law said he “did not comprehend” that the chat bot could lead him astray.

The New York Times

Input (and output) length is limited

The input (and output) length is limited

Many LLMs can accept a prompt of up to only a few thousand words.

- The total amount of context you can give it is limited
- Some LLMs have longer context limits – up to 100,000 words
- An LLM’s context length is the limit on the total input+output size

Gen AI does not work well with structured (tabular) data

Generative AI does not work well with structured (tabular) data

Home prices		Purchases on website			
size (square feet)	price (1000\$)	user ID	time	price (\$)	purchased
523	100	4783	Jan 21 08:15:20	7.95	yes
645	150	3893	March 3 11:13:5	10.00	yes
708	200	8384	June 11 14:15:05	9.50	no
1034	300	0931	Aug 2 20:30:55	12.90	yes
2290	350				
2545	440				

Use supervised learning (A → B) to estimate price.

Gen AI works well with unstructured data

Generative AI works best with unstructured data

Unstructured data refers to text, images, audio, video, etc.

Writing	Reading	Chatting
Suggest three names for our new analytics platform	I love my new llama T-shirt! The fabric is so soft.	Welcome to BettaBurgers!
Sure, here are three ideas: 1. NexalInsight 2. Quantiveue 3. AnalytiSphere	Complaint: NO Department: Apparel	I'd like a cheeseburger for delivery!
Is there parking for staff?	I wore my llama T-shirt to a friend's wedding, and now they're mad at me for stealing the show.	Awesome. Anything else?
Yes, employees can park on levels 1 and 2 of the offic... You can get a sticker at	Complaint: YES Department: Apparel	That's it thanks!
		

Bias and toxicity

Bias and Toxicity

An LLM can reflect the biases that exist in the text it learned from.

Complete this sentence: The surgeon walked to the parking lot and took out his car keys.	Complete this sentence: The nurse walked to the parking lot and took out her phone.
assumed male	assumed female

Some LLMs can output toxic or other harmful speech, but most models have gotten much safer over time.

Tips for prompting

Be detailed and specific

- Give sufficient context for LLM to complete the task
- Describe the desired task in detail

Help me write an email asking to be assigned to the legal documents project.

I'm applying for a job on the legal documents project, which will check legal documents using LLMs. I have ample experience prompting LLMs to generate accurate text in a professional tone.

Write a paragraph of text explaining why my background makes me a strong candidate to this project and advocate for my candidacy.

Guide the model to think through its answer

Guide the model to think through its answer

Brainstorm 5 names for a new cat toy.

Guide the model to think through its answer

Model response:

Step 1: 5 fun, joyful words that relate to cats:	Step 2: Rhyming names for a toy based on those words:	Step 3: Add a fun, relevant emoji for each toy name:
Purr Whisker Feline Pounce Meow	Purr-Twirl Whisker-Whisper Feline-Beeline Pounce-Bounce Meow-Wow	Purr-Twirl 🎛 Whisker-Whisper 😊 Feline-Beeline 🐱 Pounce-Bounce ⚽ Meow-Wow 🐱

Experiment and iterate

Experiment and iterate

No perfect prompt for every person or situation!

- Instead, better to develop process for improving prompts through iteration

Help me rewrite this:
[...]

Evaluate output and
improve prompt

Correct any
grammatical and
spelling errors in this:
[...]

Evaluate output and
improve prompt

Correct any
grammatical and
spelling errors in this,
and rewrite in a tone
appropriate for a
professional resume:
[...]

Iteratively improving your prompt

Iteratively improving your prompt

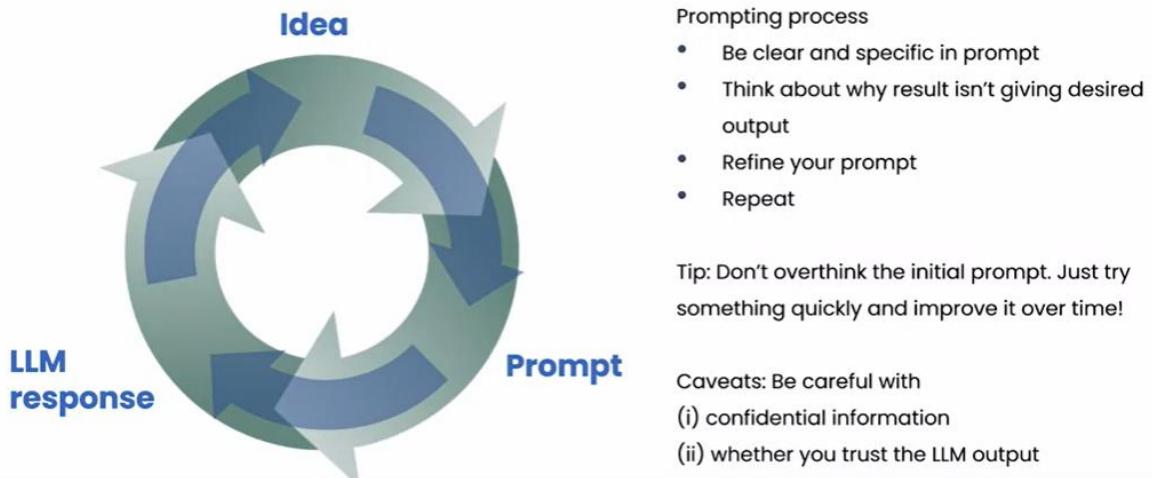


Image generation (optional)

Example generated images

Example generated images



A picture of a woman smiling



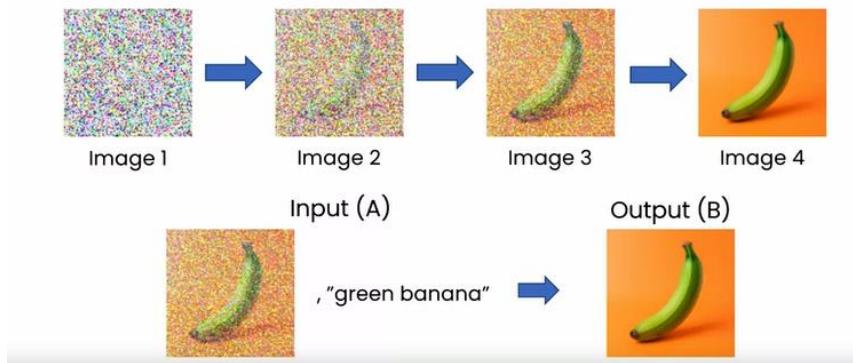
A futuristic city scene



A cool, happy robot

Image generation (diffusion model)

Image generation from text



QUIZ 2 Generative AI applications

1. A friend writes the following prompt to a web-based LLM: “Write a description of our new dog food product.”

1. A friend writes the following prompt to a web-based LLM: “Write a description of our new dog food product.”

Which of these are reasonable suggestions for how to improve this prompt?

- Give the LLM more context about what's interesting or unique about the product to help it craft a better description.
- Give it guidance on the purpose of the description (is it to go in an internal company memo, a website, a press release?) to help it use the right tone.
- Specify the desired length of the description.
- All of the above.

 **Correct**

Providing as many details as you can about the task you are trying to carry out in the prompt helps the LLM generate a response that is closer to what you want.

2. What LLMs can do?

2. Which of the following are tasks that LLMs can do? (Check all that apply)

- Translate text between languages.

 **Correct**

LLMs can produce high quality translations for widely-spoken languages that have lots of text on the internet (also known as “high resource” languages).

- Proofread text that you're writing.

 **Correct**

LLMs can be used for proofreading tasks on text that you are writing, like correcting spelling and grammar mistakes, and editing for length or clarity.

- Earn a university degree (similar to a fresh college graduate).

- Summarize articles.

 **Correct**

LLMs can take long texts as input and output shorter summaries of those texts.

3. top 10 news stories

3. Someone prompts an LLM as follows: "Please summarize each of this morning's **top 10 news stories** in 100 words per story, in a manner suitable for a newsletter." What is the main reason this is unlikely to work?

- Asking for a list of 10 items means we're working with structured data, which an LLM is poor at.
- The prompt needs to give more context about what type of newsletter it is (tech, general news, etc).
- Because of the knowledge cutoff, the LLM will not have access to the latest news.
- The output length is limited, and 10 stories is too many.

 **Incorrect**

While providing additional information and context in the prompt of an LLM can help the model produce better output, lack of information about the type of newsletter is not the main issue here. The main reason that the prompt won't work is that the LLM's knowledge of the world is frozen at the moment of its training, so it does not know about more recent events like today's news.

4. LLM to help you find an inspirational quote

4. You're preparing a presentation about technology, and ask an LLM to help you find an inspirational quote. It comes up with this:

And that's what a computer is to me. What a computer is to me is it's the most remarkable tool that we've ever come up with, and it's the equivalent of a bicycle for our minds. -Steve Jobs

How should you proceed?

- Do not use this quote because an LLM can generate toxic output.
- Because LLMs hallucinate, double-check this quote by searching other sources (such as the web) to verify if Steve Jobs really said this.
- Because LLMs can hallucinate, double-check this quote by prompting the LLM to ask if it is really sure Steve Jobs said this.
- LLMs have learned from text on the internet; so you can safely trust that this quote is found on multiple webpages, and use it in your presentation.

 **Incorrect**

Any response that an LLM may give to the question "are you sure this is correct" could also be a hallucination. Therefore, it is important to double check the output of an LLM with external sources when factual accuracy is important to your task.

5. Which of these is the better approach for creating a prompt?

5. You want an LLM to help check your writing for grammar and style. **Which of these is the better approach for creating a prompt?**

- Take all the time you need to carefully craft a prompt that gives it all the appropriate context, so that it works reliably the first time.
- Don't overthink the initial prompt -- quickly give it some context, then prompt the LLM to get its response, see what you get and iteratively refine your prompt from there.

 **Correct**

Prompting is a highly iterative process, and taking your initial idea, prompting the LLM, and then refining your prompt based on the model's output is the most effective way to get to the output that you want.

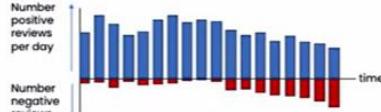
Generative AI applications

Week 2

Using generative AI in software applications

Example of Software Applications

Examples of software applications

Writing	Reading	Chatting
Is there parking for staff? Yes, employees can park on levels 1 and 2 of the office lot. You can get a sticker at [...]	Classify the sentiment of the following review as either positive or negative: <i>The mochi is excellent!</i> Positive 	Welcome to BettaBurgers! I'd like a cheeseburger for delivery! Awesome. Anything else? That's it thanks! Ok, it'll be there in 20 minutes. Enjoy!
		

- Gen AI makes it simpler and easier
- A lot of code earlier, ML

```
import torch
from torch.utils.data import DataLoader, TensorDataset
from torch import nn

class SentimentLSTM(nn.Module):
    def __init__(self, vocab_size, output_size, embedding_dim, hidden_dim, n_layers):
        super(SentimentLSTM, self).__init__()
        self.output_size = output_size
        self.n_layers = n_layers
        self.hidden_dim = hidden_dim

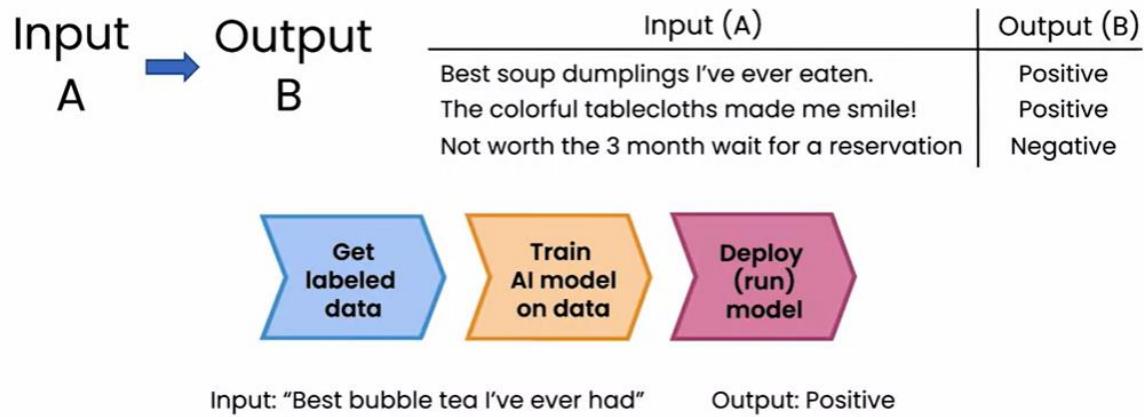
        self.embedding = nn.Embedding(vocab_size, embedding_dim)
        self.lstm = nn.LSTM(embedding_dim, hidden_dim, n_layers, batch_first=True)
        self.fc = nn.Linear(hidden_dim, output_size)
        self.sigmoid = nn.Sigmoid()

    def forward(self, x, hidden):
        batch_size = x.size(0)
        x = self.embedding(x)
        lstm_out, hidden = self.lstm(x, hidden)
        lstm_out = lstm_out.contiguous().view(-1, self.hidden_dim)
        out = self.fc(lstm_out)
        out = self.sigmoid(out)
        out = out.view(batch_size, -1)
        out = out[:, -1]
        return out, hidden

    def init_hidden(self, batch_size):
        weight = next(self.parameters()).data
```

Supervised learning for restaurant reputation monitoring

Supervised learning for restaurant reputation monitoring



- It will take often months.

Prompt -based development

Prompt-based development

```
prompt = """  
Classify the following review  
as having either a positive or  
negative sentiment:  
  
The banana pudding was really  
tasty!  
"""  
  
response = llm_response(prompt)  
print(response)
```

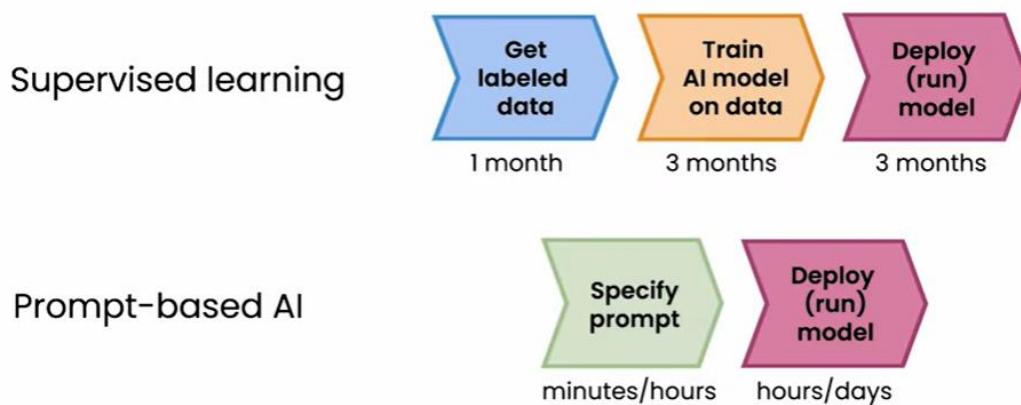
Instruction text
Review text
Code to call LLM
Code to print output

- Only Single line of code

Traditional workflow Gen AI

- Supervised learning - months
- Prompt-based AI – minutes / hours

Workflow using Generative AI



- Gen AI works better for unstructured data.

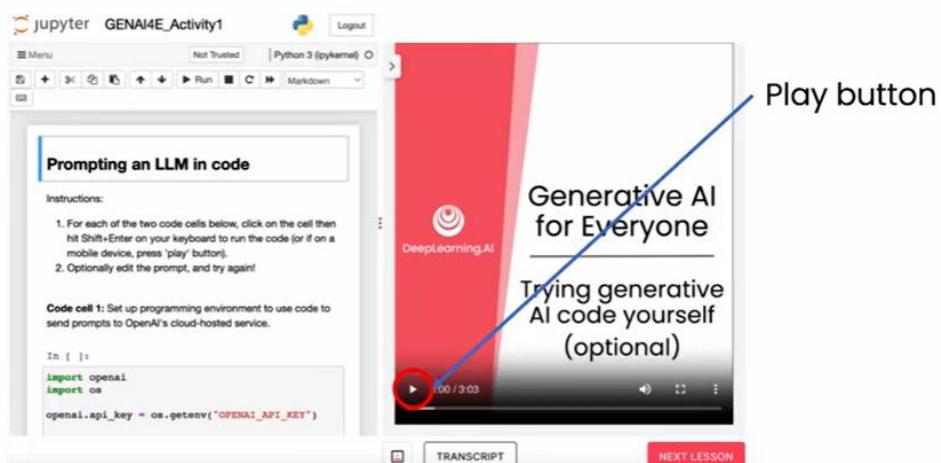
Trying generative AI code yourself (optional)

Coding platform

Coding platform (Jupyter notebook)

The one command to know:

shift + enter



DeepLearning.AI Beta

Generative AI for Everyone

- D- Activity1
- D- Activity2
- Course Feedback
- Community

jupyter GENAI4E_Activity1 (unswed)

File Edit View Insert Cell Kernel Widgets Help Python 3 (ipykernel) Logout

Prompting an LLM in code

Instructions:

- For each of the two code cells below, click on the cell then hit Shift+Enter on your keyboard to run the code (or if on a mobile device, press "play" button).
- Optionally edit the prompt, and try again!

Code cell 1: Set up programming environment to use code to send prompts to OpenAI's cloud-hosted service.

```
In [1]: import openai
import os

openai.api_key = os.getenv("OPENAI_API_KEY")

def llm_response(prompt):
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[{"role": "user", "content": prompt}],
        temperature=0
    )
    return response.choices[0].message['content']
```

Code cell 2: Define a prompt that will classify the sentiment of a restaurant review.

```
In [2]: prompt = """
Classify the following review
as having either a positive or
negative sentiment

...
The banana pudding was really tasty!

response = llm_response(prompt)
print(response)
```

Positive sentiment

DeepLearningAI
Generative AI for Everyone
Trying generative AI code yourself (optional)
0:00 / 3:03

DeepLearning.AI Beta

Generative AI for Everyone

- Activity1
- Activity2
- Course Feedback
- Community

jupyter GENAI4E_Activity1

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel) Logout

1. For each of the two code cells below, click on the cell then hit Shift+Enter on your keyboard to run the code (or if on a mobile device, press "play" button).
2. Optionally edit the prompt, and try again!

Code cell 1: Set up programming environment to use code to send prompts to OpenAI's cloud-hosted service.

```
In [1]: import openai
import os

openai.api_key = os.getenv("OPENAI_API_KEY")

def llm_response(prompt):
    response = openai.ChatCompletion.create(
        model="gpt-3.5-turbo",
        messages=[{"role": "user", "content": prompt}],
        temperature=0
    )
    return response.choices[0].message['content']
```

Code cell 2: Define a prompt that will classify the sentiment of a restaurant review.

```
In [2]: prompt = """
Classify the following review
as having either a positive or
negative sentiment

...
The banana pudding was really tasty!

response = llm_response(prompt)
print(response)
```

Positive sentiment

1 2 3 4 5

Transcript

NEXT LESSON

The screenshot shows a Jupyter Notebook interface with two code cells. The left cell (Code cell 3) contains Python code to classify reviews as positive or negative. The right cell (Code cell 4) contains Python code to count the number of positive and negative reviews. Red numbers 1, 2, and 3 are overlaid on the notebook interface to point to specific areas of interest.

```
In [3]: all_sentiments = []
for review in all_reviews:
    prompt = f"""
        Classify the following review
        as having either a positive or
        negative sentiment. State your answer
        as a single word, either "positive" or
        "negative".
    {review}
    """
    response = llm_response(prompt)
    all_sentiments.append(response)

all_sentiments
```

```
In [4]: num_positive = 0
num_negative = 0
for sentiment in all_sentiments:
    if sentiment == 'positive':
        num_positive += 1
    elif sentiment == 'negative':
        num_negative += 1
print(f"There are {num_positive} positive and {num_negative} negative reviews.")
```

Code cell 3: Classify the reviews as positive or negative.

Code cell 4: Count the number of positive and negative reviews

Output:

```
Out[3]: ['positive', 'positive', 'negative', 'positive', 'negative']

Out[4]: There are 3 positive and 2 negative reviews.
```

Code cell 4 output:

```
In [ ]: num_positive = 0
num_negative = 0
for sentiment in all_sentiments:
    if sentiment == 'positive':
        num_positive += 1
    elif sentiment == 'negative':
        num_negative += 1
print(f"There are {num_positive} positive and {num_negative} negative reviews.")
```

Text:

up with these classifications, positive, positive,
negative, positive, negative,

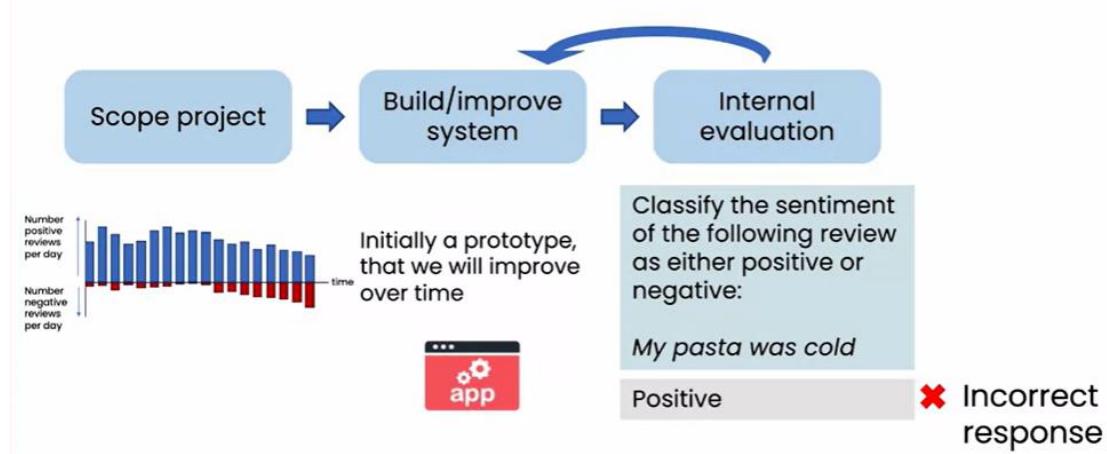
Buttons: TRANSCRIPT, NEXT LESSON

Lifecycle of a generative AI project

Gen AI project

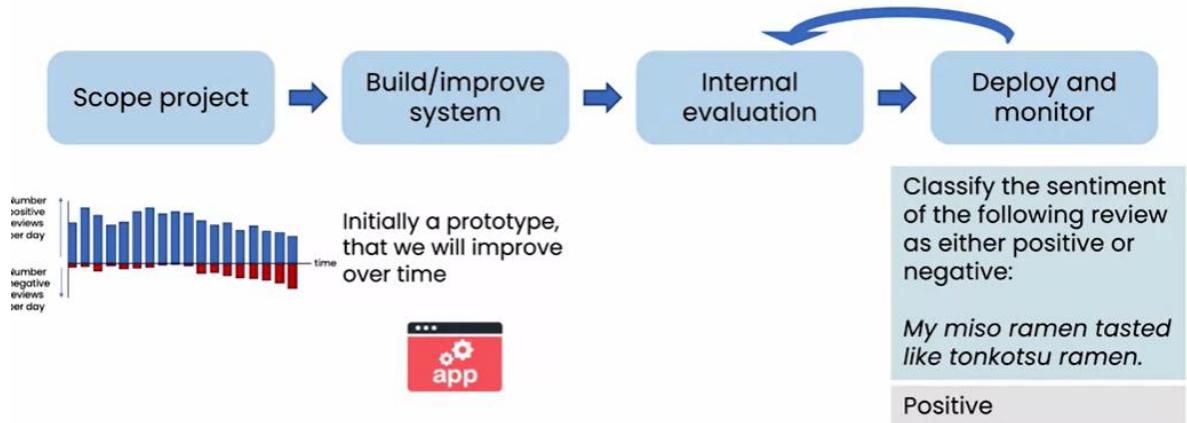
- Scope project
- Build Prototype
- Internal evaluation team
 - Iterative process

Lifecycle of a generative AI project



- Deploy and monitor.
 - Iterative process

Lifecycle of a generative AI project



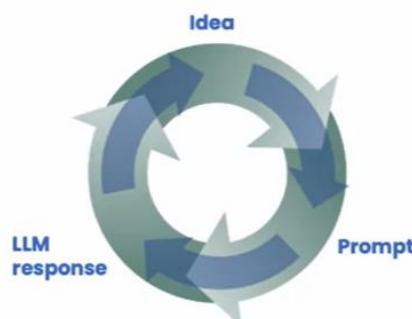
Tools to improve performance

- Prompting

Tools to improve performance

Building Generative AI is a highly empirical (experimental) process – we repeatedly find and fix mistakes.

- Prompting



- RAG Retrieval Augmented generation (RAG)

Tools to improve performance

Building Generative AI is a highly empirical (experimental) process – we repeatedly find and fix mistakes.

- Prompting
- Retrieval augmented generation (RAG)
 - Give LLM access to external data sources
- Fine-tune models
 - Adapt LLM to your task



Tools to improve performance

Building Generative AI is a highly empirical (experimental) process – we repeatedly find and fix mistakes.

- Prompting
- Retrieval augmented generation (RAG)
 - Give LLM access to external data sources
- Fine-tune models
 - Adapt LLM to your task
- Pre-train models
 - Train LLM from scratch

Tools to improve performance

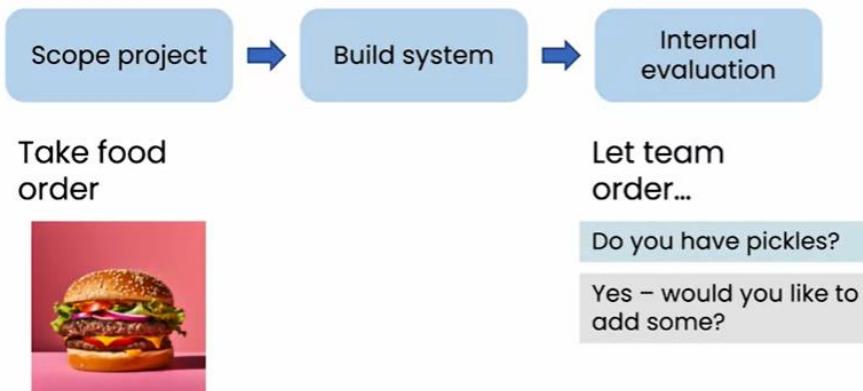
Building Generative AI is a highly empirical (experimental) process – we repeatedly find and fix mistakes.

- Prompting
- Retrieval augmented generation (RAG)
 - Give LLM access to external data sources
- Fine-tune models
 - Adapt LLM to your task
- Pretrain models
 - Train LLM from scratch

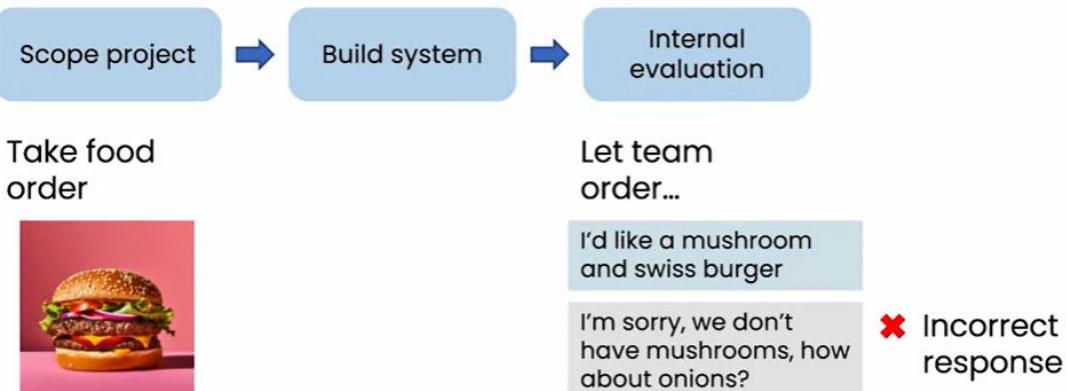
Example

- Internal team

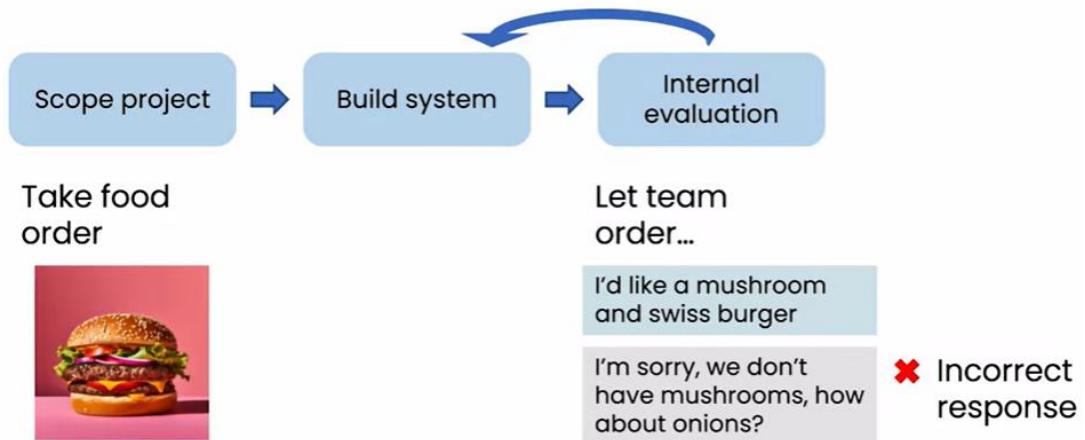
Lifecycle of a generative AI project



Lifecycle of a generative AI project

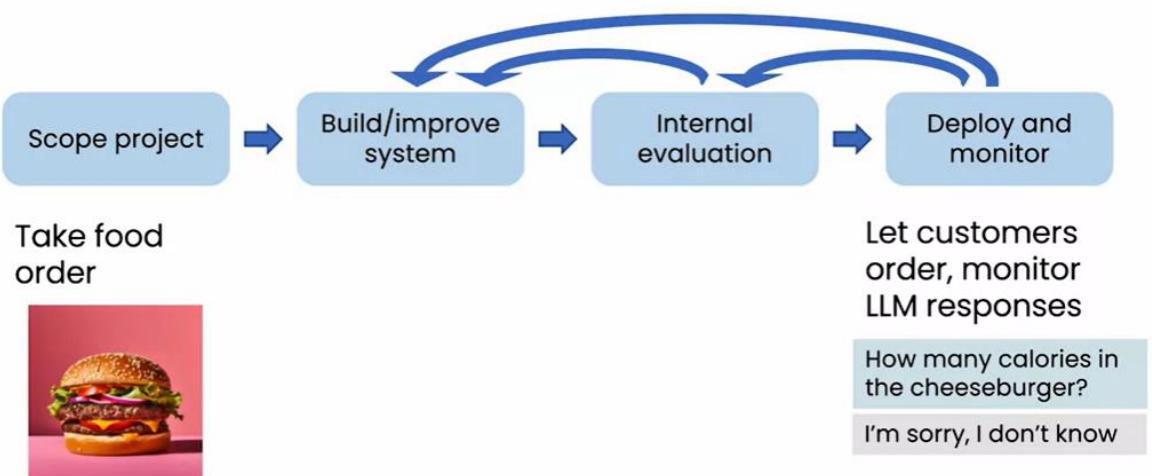


Lifecycle of a generative AI project

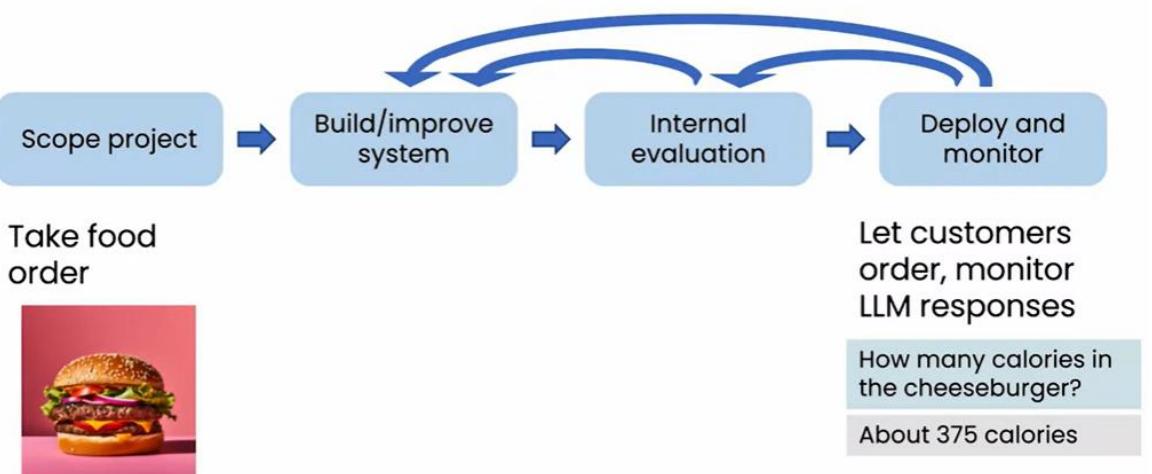


- RAG

Lifecycle of a generative AI project



Lifecycle of a generative AI project



Cost intuition

How much does it cost?

How much does it cost?

Example prices

	OpenAI/GPT3.5	OpenAI/GPT4	Google/PaLM 2	Amazon/Titan Lite
Input tokens	\$0.0015/1K tokens	\$0.03/1K tokens	\$0.0005/1K tokens	\$0.0003/1K tokens
Output tokens	\$0.002/1K tokens	\$0.06/1K tokens	\$0.0005/1K tokens	\$0.0004/1K tokens

What is a token?

the example Andrew	1 token	300 words 400 tokens
translate programming	2 tokens	
tonkotsu	4 tokens	Roughly, 1 token = 3/4 words

Example

Token

Estimate Cost

Estimating Cost

Typical adult reading speed: 250 words/minute.

How much would it cost to keep someone occupied for 1 hour?

$$\begin{aligned} 60 \times 250 &= 15,000 \leftarrow \text{output} \\ &\quad + 15,000 \leftarrow \text{prompt} \\ &\quad \hline 30,000 \\ &\quad \hookrightarrow 40,000 \text{ tokens} \end{aligned}$$

$1 \text{ token} \approx \frac{3}{4} \text{ word}$

Cost: \$0.002 cents / 1K tokens.
 $\$0.002 \times 40 = \0.08

QUIZ 3 – Software applications

1. Question 1 In the videos, we described using either supervised learning or a prompt-based development process to build a restaurant review sentiment classifier. Which of the following statements about prompt-based development is correct?

1. In the videos, we described using either supervised learning or a prompt-based development process to build a restaurant review sentiment classifier. Which of the following statements about prompt-based development is correct?

1 / 1 point

- Prompt-based development is generally much faster than supervised learning.
- Prompt-based development requires that you collect hundreds or thousands of labeled examples.
- Prompt-based development requires that you collect hundreds or thousands of unlabeled examples (meaning reviews without a label B to say if it is positive or negative sentiment).
- If you want to classify reviews as positive, neutral, or negative (3 possible outputs) there is no way to write a prompt to do so: An LLM can generate only 2 outputs.

 **Correct**

Prompt-based development allows you to take advantage of an LLM's ability to carry out sentiment classification, so you can get up and running very quickly because you don't need to train a model from scratch.

2. What is a token in the context of a large language model (LLM)?

Software Applications
Back Graded Quiz + 10 min

English Due Nov 19, 11:59 PM +08

- Prompt-based development is generally much faster than supervised learning.
- Prompt-based development requires that you collect hundreds or thousands of labeled examples.
- Prompt-based development requires that you collect hundreds or thousands of unlabeled examples (meaning reviews without a label B to say if it is positive or negative sentiment).
- If you want to classify reviews as positive, neutral, or negative (3 possible outputs) there is no way to write a prompt to do so: An LLM can generate only 2 outputs.

 **Correct**

Prompt-based development allows you to take advantage of an LLM's ability to carry out sentiment classification, so you can get up and running very quickly because you don't need to train a model from scratch.

2. What is a token in the context of a large language model (LLM)?

1 / 1 point

- A unit of cryptocurrency (like bitcoin or other "crypto tokens") that you can use to pay for LLM services.
- The part of the LLM output that has primarily symbolic rather than substantive value (as in, "the court issued a token fine", or "the LLM generated a token output").
- A physical device or digital code to authenticate a user's identity.
- A word or part of a word in either the input prompt or LLM output

 **Correct**

Tokens in the context of LLMs refer to a unit of text. Common words are typically represented by a single token, while uncommon words may be broken into two or more tokens.

3. What are the major steps of the lifecycle of a Generative AI project?

0.75 / 1 point

- Scope project → Internal evaluation → Deploy and monitor → Build/improve system
- Scope project → Internal evaluation → Build/improve system → Deploy and monitor

3. What are the major steps of the lifecycle of a Generative AI project? Select 2

3. What are the major steps of the lifecycle of a Generative AI project?

- Scope project → Internal evaluation → Deploy and monitor → Build/improve system
- Scope project → Internal evaluation → Build/improve system → Deploy and monitor
- Scope project → Build/improve system → Internal evaluation → Deploy and monitor

 **Correct**

This sequence accurately represents the recommended steps in the lifecycle of a Generative AI project.

You first scope the project, then build or improve the system, followed by internal evaluation, and finally, deployment and monitoring.

- Scope project → Build/improve system → Deploy and monitor → Internal evaluation

You didn't select all the correct answers

4. You are building a customer service chatbot. Why is it important to monitor the performance of the system after it is deployed?

4. You are building a customer service chatbot. Why is it important to monitor the performance of the system after it is deployed?

- Every product should be monitored to track customer satisfaction -- this is good practice for all software.
- Because of the LLM's knowledge cutoff, we must continuously monitor the knowledge cutoff and update its knowledge frequently.
- In case customers say something that causes the chatbot to respond in an unexpected way, monitoring lets you discover problems and fix them.
- This is false. So long as internal evaluation is done well, further monitoring is not necessary.

 **Correct**

Users can be very creative in the ways they prompt chatbots, so monitoring the system can help you identify any issues with the chatbot's output as they arise and allow you to improve the system in response.

5. You are working on using an LLM to summarize research reports. Suppose an average report contains roughly 6,000 words. Approximately how many tokens would it take an LLM to process 6,000 input words? (Assume 1 token = 3/4 words, or equivalently, 1 word \approx 1.333 tokens).

Advanced Technologies: Beyond Prompting

Retrieval Augmented Generation (RAG)

RAG Example

General Chatbot

Is there parking for employees?

I need more specific information about your workplace to answer that question.

Chatbot with RAG

Is there parking for employees?

Yes, employees can park on levels 1 and 2 of the office lot. You can get a sticker at [...]

Step 1: Given question, search for relevant document for answer.

Retrieval Augmented Generation (RAG) example

1. Given question, search relevant documents for answer

Is there parking for employees?



Step 2: Incorporate retrieved text into an updated prompt.

1. Given question, search relevant documents for answer

Is there parking for employees?



2. Incorporate retrieved text into an updated prompt

Use the following pieces of context to answer the question at the end.

Parking Policy: All employees may park on levels 1 and 2 of the office lot. Use the entrance on Front St [...]

Retrieval Augmented Generation (RAG) example

- Given question, search relevant documents for answer

Is there parking for employees?



- Incorporate retrieved text into an updated prompt

Use the following pieces of context to answer the question at the end.

Parking Policy: All employees may park on levels 1 and 2 of the office lot. Use the entrance on Front St [...]

Is there parking for employees?

Step 3: Generate Answer for new prompt with additional context

Retrieval Augmented Generation (RAG) example

- Generate answer from the new prompt with additional context

Use the following pieces of context to answer the question at the end.

Parking Policy: All employees may park on levels 1 and 2 of the office lot. Use the entrance on Front St [...]

Is there parking for employees?

Yes, employees can park on [...] ① → Link to full onboarding policy

Example RAF Applications

Examples of RAG applications

Chat with PDF files



The screenshot shows a web-based AI interface for ChatPDF. At the top, there's a header with the title 'Accenture-Su...' and a page number '2 / 4'. Below the header is a purple sidebar containing a summary of an Accenture report. The main area is a 'Chat' window with a blue message bubble from the user asking, 'What are some roles in supply chain that will be changed by AI?'. A larger text block on the right side of the window discusses how generative AI will impact supply chain roles like data scientists, call-center staff, planners, buyers, sustainability officers, and compliance officers. At the bottom of the window is a text input field labeled 'Ask any question...'.

Chatting with an Accenture report using ChatPDF

Examples of RAG applications

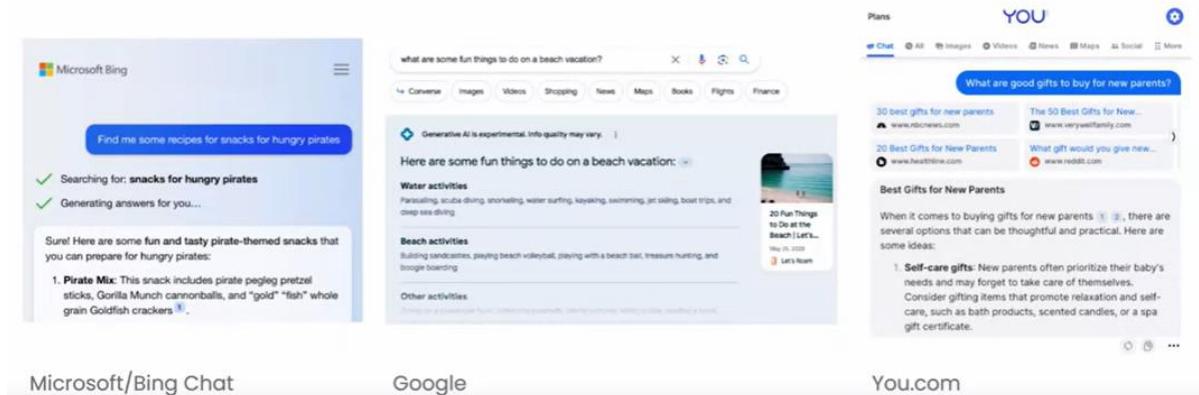
Answer questions based on a website's articles

The image displays three examples of RAG applications:

- Coursera Coach:** A screenshot of the Coursera platform showing a Q&A session. The user asks, "What is supervised learning?" and the AI response provides a detailed explanation of supervised learning, mentioning that it's a type of machine learning where algorithms learn from labeled examples. It includes a link to the original article and a "Rate Coach's answer" button.
- Snapchat:** A screenshot of the Snapchat help center. A user asks, "How do I create a lens in snapchat? Answer with 50 words or fewer." The AI response explains that users need to use Lens Studio to create lenses with various components and effects. It also includes a "Was this response helpful?" poll and a "Verified Sources" section with links to the Snapchat help center and community forums.
- Hubspot:** A screenshot of the HubSpot AI-powered search feature. A user asks, "How do I schedule an email campaign?" The AI response provides step-by-step instructions: 1. Navigate to your HubSpot account and go to the Email tool. 2. Click on "Create email" in the upper right corner. 3. Select the type of email you want to send - Regular, Automated, or Blog/RSS. 4. Under the "Custom" tab, choose an option. It also includes a "Choose an option" dropdown and a "HubBot" profile icon.

Examples of RAG applications

New form of web search



Big Idea: LLM as reasoning engine

Big Idea: LLM as reasoning engine

- LLMs have a lot of general knowledge, but they don't know everything
- By providing relevant context in the prompt, we ask an LLM to read a piece of text, then process it to get an answer
- We're using it as a reasoning engine to process information, rather than using it as a source of information

Fine-tuning

Pretraining and fine tuning

Pretraining and Fine-tuning

Pretraining

My favorite food is a bagel with cream cheese

Input (A)	Output (B)
My favorite food is a	bagel
My favorite food is a bagel	with
My favorite food is a bagel with	cream
My favorite food is a bagel with cream	cheese

Learns from 100Bs of words

Fine-tuning

What a wonderful chocolate cake
The novel was thrilling

Input (A)	Output (B)
What	a
What a	wonderful
What a wonderful	chocolate
What a wonderful chocolate	cake

Learns from 1000s to 10,000s of words

Why Fine-tune?

Example 1: Summarize in certain style or structure

Why fine-tune?

To carry out a task that isn't easy to define in a prompt.

Example 1: Summarize in certain style or structure

Customer: Hi, my monitor won't turn on.
Agent: I'm sorry to hear that.
What model is it?
Customer: It's the 27-inch 4K.
Agent: Can you tell me the model number?
Customer: Hang on one sec...I'm looking on the back...it's the MK401-27X.
Agent: Ok, thank you. Can you try unplugging the HDMI cable

and reinserting?
Customer: Ok, doing that now. Hmm, the image appeared for a few seconds then disappeared.
Agent: Ok, try jiggling the cable while it is in the socket.
Customer: Oh, I see the image coming and going.
Agent: Ok, I think you have a defective cable. We will send you a replacement.

MK401-27KX was reported as broken by customer 5402. Cable identified as problem. Replacement issued.

Example 2: Mimicking a writing or speaking style

Why fine-tune?

To carry out a task that isn't easy to define in a prompt.

Example 2: Mimicking a writing or speaking style

Write a speech about generative AI
in the voice of Andrew Ng

Hello everyone, I'm Andrew Ng, and
I'm here today to talk to you about
generative AI. Generative AI is a type
of artificial intelligence that can
create new content [...]

Pretrained model

Why fine-tune?

To carry out a task that isn't easy to define in a prompt.

Example 2: Mimicking a writing or speaking style

Write a speech about generative AI
in the voice of Andrew Ng

Hello everyone, I'm Andrew Ng, and
I'm here today to talk to you about
generative AI. Generative AI is a type
of artificial intelligence that can
create new content [...]

Pretrained model

Write a speech about generative AI
in the voice of Andrew Ng

Hey everyone, good to see all of you.
So, what I want to do today, is share
some thoughts with you about how
generative AI is going to change the
way that your company [...]

Fine-tuned model

Example 3: To help LLM gain specific knowledge

Medical notes

Why fine-tune?

To help LLM gain specific knowledge.

Medical notes

Pt c/o SOB, DOE. PE: RRR, JVD absent, CTAB.
EKG: NSR. Tx: F/u w/ PCP, STAT CXR, cont. PRN O2.

Extract from admission note for patient with shortness of breath

Legal documents

Legal documents

Licensor grants to Licensee, per Section 2(a)(iii), a non-exclusive right to use the intellectual property, contingent upon compliance with fiduciary duties outlined in Section 8, paragraphs 1-4, and payment as specified in Schedule B, within 15 days hereof.

Financial documents

Why fine-tune?

To help LLM gain specific knowledge.

Financial documents

Counterparties engaged in OTC derivatives transactions must comply with margin requirements pursuant to Regulation (EU) No 648/2012, including initial and variation margin calculations."

Example 4: To get smaller model to perform a task

- Lower cost / latency to deploy
- Needs special computers and not normal PC

Why fine-tune?

To get a smaller model to perform a task.

- Lower cost/latency to deploy
- Can run on mobile/laptop (edge devices)



Classify the following review as having a positive or negative sentiment:

Really enjoyed my vegetable biryani

Positive

Classify the following review as having a positive or negative sentiment:

The noodle soup was too salty

Negative

Classify the following review as having a positive or negative sentiment:

The staff sang happy birthday to me!

Positive

Can work well with 500-1000 examples.

Pretraining an LLM

- LLM are already trained big tech company.
- Do your own pre-train your LLM?
- When in doubt, don't do it. Its very expensive

When should you pretrain an LLM?

- Mat take \$10m, many months,
- Last resort
- Need
- BloombergGPT

When should you pretrain an LLM?

Many teams are pretraining general-purpose LLMs by learning from internet text.

- May take \$10s of millions, many months, huge amount of data

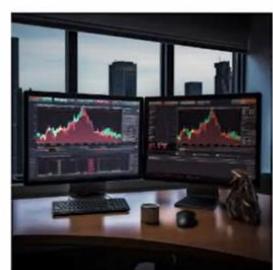
For building a specific application:

- Option of last resort
- Could help if have a highly specialized domain

Introducing BloombergGPT, Bloomberg's 50-billion parameter large language model, purpose-built from scratch for finance

March 30, 2023

Bloomberg



Choosing a model

Model Size

- 1B parameters – Pattern matching
- 10B parameters – Greater Knowledge.
- 100B+ parameters – Rich world knowledge

Model size

1B parameters:	Pattern matching and basic knowledge of the world.	Restaurant review sentiment
10B parameters:	Greater world knowledge. Can follow basic instructions.	Food order chatbot
100B+ parameters:	Rich world knowledge. Complex reasoning.	Brainstorming partner

Closed or open source

- Easy to use in apps
- More large / power models
- Relatively inexpensive
- Some risk of vendor lock-in

Closed or open source?

Closed-source models (Cloud programming interface)

- Easy to use in applications
- More large/powerful models
- Relatively inexpensive
- Some risk of vendor lock-in

Open-source models

- Full control over model
- Can run on your own device (on-prem, PC, etc.)
- Full control over data privacy/access

How LLMs follow instructions: Instruction tuning and RLHF (optional)

Tool use and agents (optional)

QUIZ 4: Advanced Technologies : Beyond prompting

1. True or False. Because of the knowledge cut-off, an LLM cannot answer questions about today's news. But with RAG to supply it articles from the news, it would be able to.

1. **True or False.** Because of the knowledge cut-off, an LLM cannot answer questions about today's news. But with RAG to supply it articles from the news, it would be able to.

- True
 False

 **Incorrect**

While it is true that an LLM's knowledge of the world is frozen at the moment of its training, RAG can be used to give an LLM access to new information in external documents that it can reason through to formulate an answer to a question. So RAG would enable an LLM to answer questions about current news articles.

2. You want to build an application to answer questions based on information found in your emails. Which of the following is the most appropriate technique?

2. You want to build an application to answer questions based on information found in your emails. Which of the following is the most appropriate technique?

- Prompting (without RAG), where we iteratively refine the prompt until the LLM gets the answers right.
 Fine-tuning an LLM on your emails, whereby we take a pre-trained LLM and further train it on your emails.
 Pretraining an LLM on your emails.

- RAG, where the LLM is provided additional context based on retrieving emails relevant to your question.

 **Correct**

RAG can be used to give an LLM access to new, external sources of information that it can reason through to formulate an answer to your question. So a RAG system that provides access to your emails is the best approach to get an LLM to answer your questions.

3. What does the idea of using an LLM as a reasoning engine refer to?

3. What does the idea of using an LLM as a reasoning engine refer to?

- This refers to the idea of using an LLM not as a source of information, but to process information (wherein we provide it the context it needs, through techniques like RAG).
- The idea of using an LLM to play games (like chess) that require complex reasoning, but having its output moves in the game.
- Reasoning engine is another term for RAG.
- This refers to pretraining an LLM on a lot of text so that it acquires general reasoning capabilities.

 **Correct**

The ability of LLMs to process information is one of the features that makes them such powerful and useful tools.

4. True or False. By making trusted sources of information available to an LLM via RAG, we can reduce the risk of hallucination.

1. **True or False.** By making trusted sources of information available to an LLM via RAG, we can reduce the risk of hallucination.

1 / 1 point

- True, because the LLM is now restricted to outputting paragraphs of text exactly as written in the provided document, which we trust.
- False, because the LLM has learned from a lot of text from the internet (perhaps >100 billion words) to hallucinate, so adding one more short piece of text to the prompt as in RAG won't make any meaningful difference.
- True, because RAG allows the LLM to reason through accurate information retrieved from a trusted source to arrive at the correct answer.
- False, because giving the LLM more information only confuses the LLM more and causes it to be more likely to hallucinate.

 **Correct**

RAG can be used to give an LLM access to new, trusted sources of information that it can reason through to formulate an answer to your question. This helps prevent the model from hallucinating because it doesn't know the answer.

5. An ecommerce company is building a software application to route emails to the right department (Apparel, Electronics, Home Appliances, etc.) It wants to do so with a small, 1 billion parameter model, and needs high accuracy. Which of these is an appropriate technique?

5. An ecommerce company is building a software application to route emails to the right department (Apparel, Electronics, Home Appliances, etc.) It wants to do so with a small, 1 billion parameter model, and needs high accuracy. Which of these is an appropriate technique?
- Pretrain a 1 billion parameter model on around 1 billion examples of emails and the appropriate department.
 - Pretrain a 1 billion parameter model on around 1,000 examples of emails and the appropriate department.
 - Fine-tune a 1 billion parameter model on around 1 billion examples of emails and the appropriate department.
 - Fine-tune a 1 billion parameter model on around 1,000 examples of emails and the appropriate department.

 **Incorrect**

Pretraining refers to the initial training of an LLM, where it learns to repeatedly predict the next word from billions to trillions of words from the internet. Pretraining a model on just 1000 email examples would not result in a useful LLM, so this is not the right technique to use.

Week 3

Day-to-day usage of web UI LLMs – Case Studies

Writing assistant

Writing assistant

Rewrite the following to be suitable for a professional business report:

The marketing campaign for the new product went really well, beating our new user goal by 10%.

The product launch marketing campaign was successful, exceeding our user acquisition target by 10%.

Marketer

Marketer

Help me brainstorm an email campaign to reactivate lapsed users of our SaaS product.

Certainly! Here's an email campaign strategy:

1. "We miss you" email
2. Feedback request email
3. Exclusive offer email
4. Feature spotlight email
5. Success story email

Recruiter

Recruiter

Summarize the following review of a job candidate in 50 or fewer words:

I really enjoyed talking to the candidate. I liked their positive energy and excitement about the position, and was also impressed by their knowledge of our company. I did have some concerns about their familiarity with the tools we use [...]

The review was impressed by the candidate's depth of knowledge but had some questions about their experience with specific tools. The reviewer is in favor hiring the candidate but suggests further team discussion about how to train and onboard effectively.

Programmer

Programmer

Write python code to calculate the average of the last column in a dataframe

Sure! Here's some code that uses the pandas library:

```
import pandas as pd
# Calculate the mean of the last column
mean_value = df.iloc[:, -1].mean()

print(mean_value)
```

Task analysis of jobs

Identifying automation opportunities

Identifying automation opportunities

- AI doesn't automate jobs. It automates tasks.
- Most jobs involve a collection of many tasks.
- Example: Customer service representative

Tasks	Generative AI potential
Answer inbound phone calls from customers	Low
Answer customer chat queries	High
Check status of customer orders	Medium
Keep records of customer interactions	High
Assess accuracy of customer complaints	Low
[...]	[...]

Augmentation vs. Automation

Augmentation

- Help humans with a task.
- E.g., recommend a response for a customer service agent to edit/approve.

Automation

- Automatically perform a task
- E.g., automatically transcribe and summarize records of customer interactions

For some tasks, businesses will start with augmentation, and gradually move toward automation.

- Technical feasibility
- Business value

Evaluating AI potential

The potential for augmenting/automating a task depends on:
(i) Technical feasibility and (ii) Business value.

Technical feasibility: Can AI do it?

- Can a fresh college graduate following the instructions in a prompt complete the task?
- If unsure, try prompting an LLM to see if you can get it to do it.
- An AI engineer can also help assess if RAG, fine-tuning, or other techniques can help.

Business value: How valuable is it for AI to augment or or automate this task?

- How much time is spent on this task?
- Does doing this task significantly faster, cheaper or more consistently create substantial value?

Job databases: Breaking down job roles into tasks

O*NET entry for Customer Service Representatives

Occupation-Specific Information

Tasks

All 15 displayed

- Confer with customers by telephone or in person to provide information about products or services, take or enter orders, cancel accounts, or obtain details of complaints.
- Keep records of customer interactions or transactions, recording details of inquiries, complaints, or comments, as well as actions taken.
- Check to ensure that appropriate changes were made to resolve customers' problems.
- Contact customers to respond to inquiries or to notify them of claim investigation results or any planned adjustments.
- Determine charges for services requested, collect deposits or payments, or arrange for billing.
- Complete contract forms, prepare change of address records, or issue service discontinuance orders, using computers.

Source: onetonline.org

Additional job analysis examples

Computer programmer

Computer programmer

Tasks	Generative AI potential
Write code	Medium
Write documentation	High
Respond to user support requests	Medium
Review others' code	Low
Gather requirements	Low
[...]	[...]

Lawyer

Lawyer

Tasks	Generative AI potential
Draft and review legal documents	High
Interpret laws, rulings, and regulations	High
Review evidence	Low
Negotiate settlements of disputes	Low
Represent clients in court	Low
[...]	[...]

Landscape

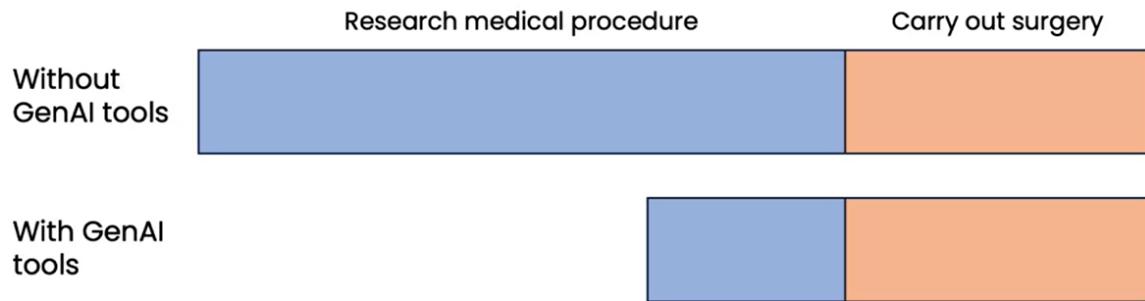
Landscaper

Tasks	Generative AI potential
Maintain and care for plants	Low
Purchase and transport plants	Low
Maintain yard equipment	Low
Communicate with clients	Medium
Maintain business website	Low
[...]	[...]

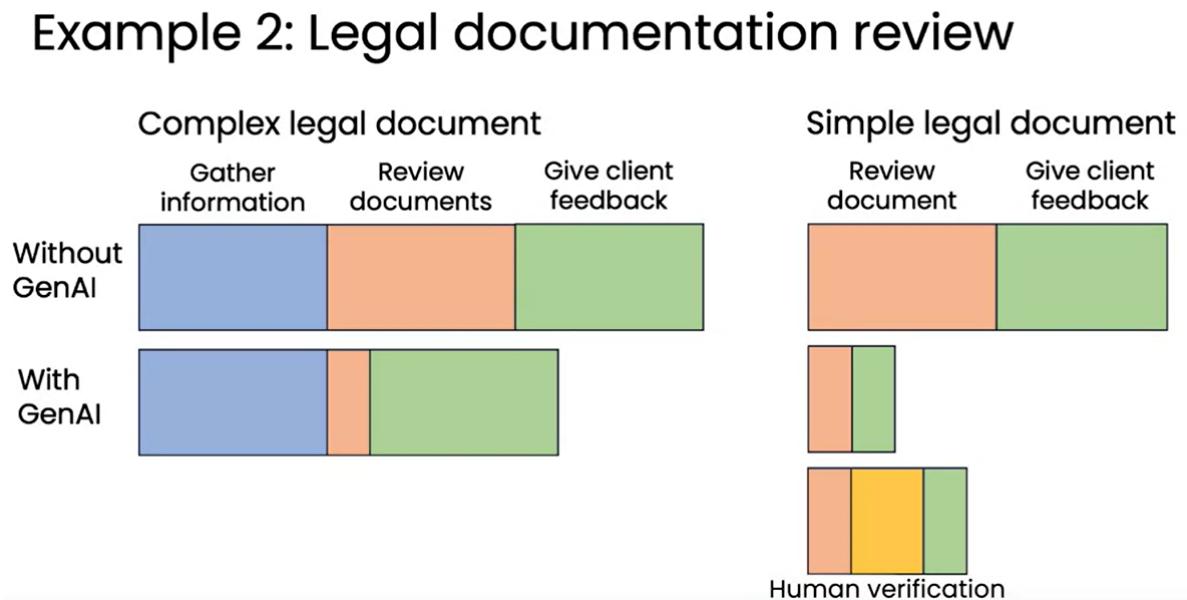
New workflows and new opportunities – Case Studies

Example 1: Surgeon

Example 1: Surgeon

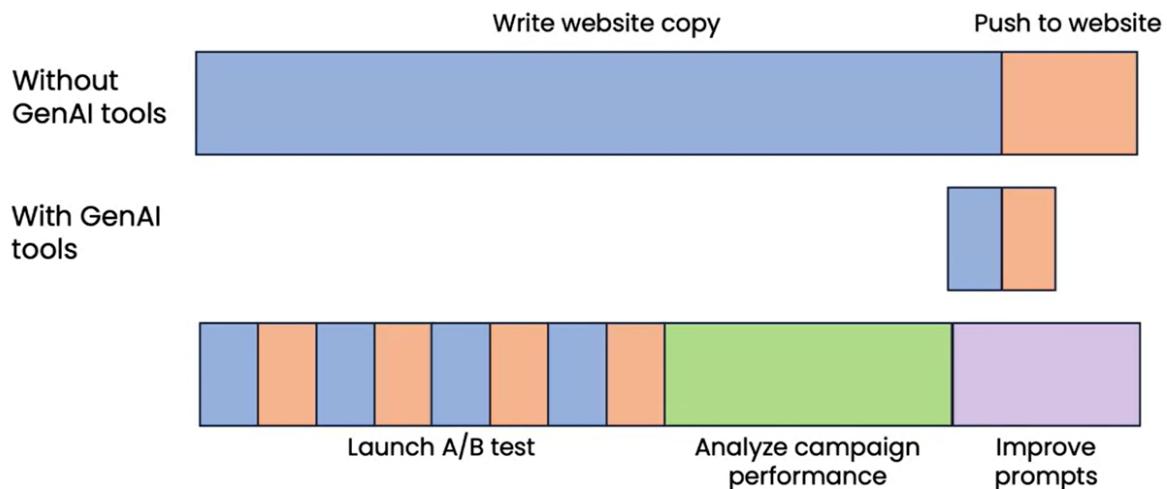


Example 2: Legal documentation review



Example 3: Marketing documentation

Example 3: Marketing automation



Analysing Customer Tasks

Analyzing Customers' Tasks

Example: Helping customers build websites

Tasks	Generative AI potential
Select website template	Low
Write title	High
Select images	Medium
Write copy for homepage	Medium
Optimize copy for SEO	High
[...]	[...]

Teams to build generative AI software

Common Roles

- Software engineer
- ML Engineer
- Product manager

- Prompt engineer

Common roles

- Software engineer
 - Responsible for writing software application
 - Ideally someone who has learned basics of LLMs/prompting
- Machine learning engineer
 - Responsible for implementing AI system
 - Ideally familiar with LLMs/prompting, RAG, fine-tuning
- Product manager
 - Responsible for identifying and scoping the project
- Prompt engineer?
 - Usually not hired as a dedicated role

Getting started with a small team

One Person

- Software
- ML Engineer

Two person

- Can be ML + Software

Getting started with a small team

One person team:

- Software engineer (who has learned some prompting), or
- Machine learning engineer, or
- Just yourself!

Two person team:

- Can be machine learning engineer + software engineer.
- But many other configurations are fine.

Additional roles

Additional roles

Data engineer: Responsible for organizing data and ensuring data quality

Data scientist: Responsible for analyzing data to make recommendations to guide project or business decisions

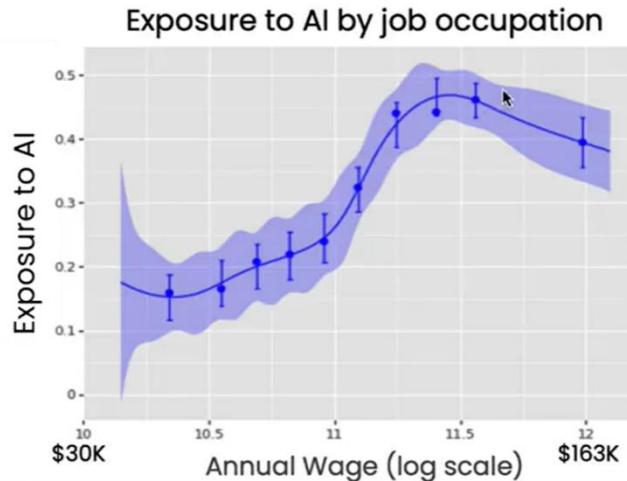
Project Manager: Responsible for coordinating project execution

Machine learning researcher: Responsible for developing advanced AI technologies

Automation potential across sectors

Gen AI will impact higher-paid jobs more

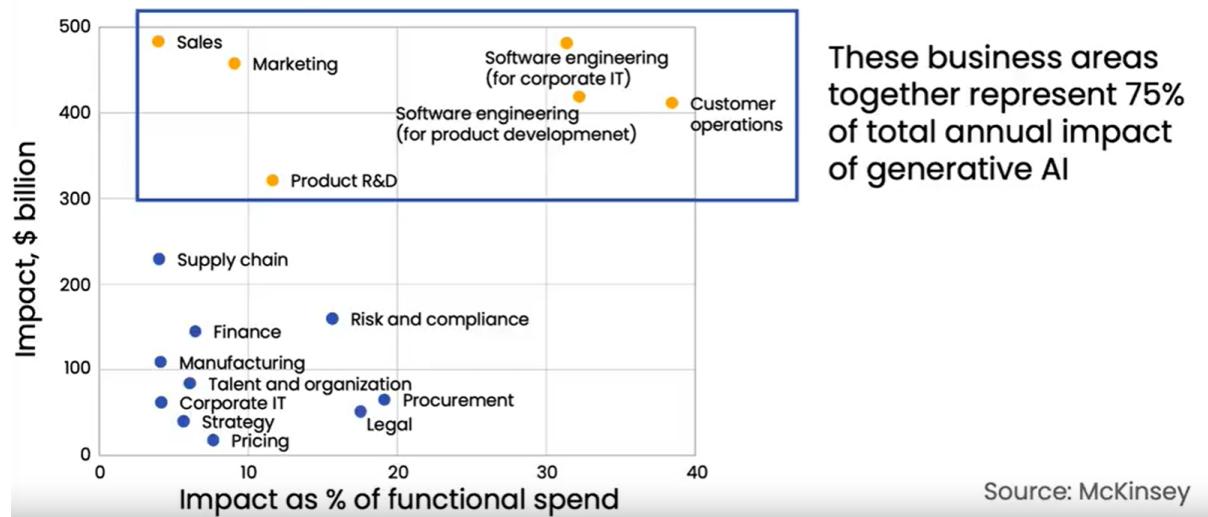
Generative AI will impact higher-paid jobs more



Credit: Eloundou et al., 2023

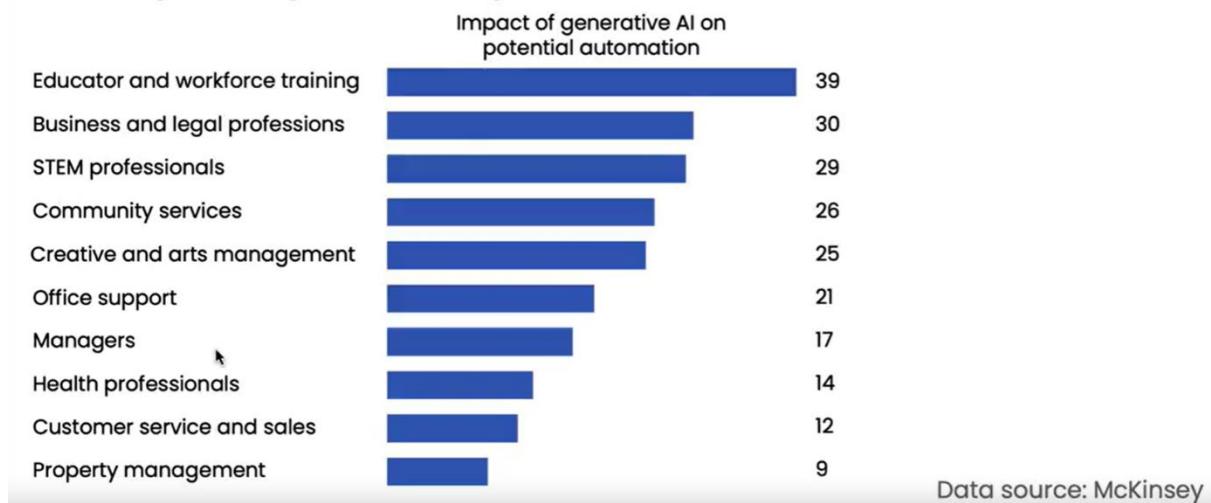
Analysis by functional role

Analysis by functional role



Analysis by industry sector

Analysis by industry sector



Quiz: Generative AI and business

1. Which of these job roles are unlikely to find any use for web UI LLMs?

1. Which of these job roles are unlikely to find any use for web UI LLMs?

- Marketer
- Recruiter
- Programmer
- None of the above

 **Correct**

All of the above roles carry out one or more tasks that could be augmented with a web UI LLM.

2. What is the relation between AI, tasks, and jobs?

2. What is the relation between AI, tasks, and jobs?

1 / 1 point

- Tasks are comprised of many jobs. AI automates jobs, rather than tasks.
- Jobs are comprised of many tasks. AI automates jobs, rather than tasks.
- Jobs are comprised of many tasks. AI automates tasks, rather than jobs.

 **Correct**

- Tasks are comprised of many jobs. AI automates tasks, rather than jobs.

3. Here are some of the tasks of a retail salesperson from O*NET (We encourage you to check out the page yourself.)

3. Here are some of the [tasks of a retail salesperson from O*NET](#). (We encourage you to check out the page yourself.)

1 / 1 poi

Occupation-Specific Information

Tasks

▼ 5 of 24 displayed

- ➊ Greet customers and ascertain what each customer wants or needs.
- ➋ Recommend, select, and help locate or obtain merchandise based on customer needs and desires.
- ➌ Compute sales prices, total purchases, and receive and process cash or credit payment.
- ➍ Prepare merchandise for purchase or rental.
- ➎ Answer questions regarding the store and its merchandise.

Say we decide to use AI to augment (rather than automate) a salesperson's task of recommending merchandise to customers. Which of the following would be an example of this?

- Build an AI system to suggest products to the salesperson, who then decides what to recommend to the customer.
- Build an AI chatbot that can role-play being a customer to help the salesperson practice having conversations with customers.
- Build a chatbot that automatically recommends products that customers can access directly, with no salesperson involved.
- This has no business value and should not be done.

✓ Correct

Here the AI is augmenting the work of the salesperson by making suggestions, rather than fully taking over and automating the task.

4. When looking for augmentation or automation opportunities, what are the two primary criteria by which to evaluate tasks for generative AI potential? (Check the two that apply.)

-
- Business value (how valuable is it to automate?).

 **Correct**

Thinking about the time taken to complete a task, and the potential value in doing that task faster, cheaper, or more consistently, can help you assess the business value of AI augmentation or automation.

- Technical feasibility (can AI do it?).

 **Correct**

Asking whether a fresh college graduate can complete the task, or consulting an AI engineer, are two ways to assess the technical feasibility of AI augmentation or automation.

- Whether the task is the iconic, defining task for a job role.
- Whether to use prompting, RAG or fine-tuning.

5. What is a quick way to start experimenting with an LLM application development project?

5. What is a quick way to start experimenting with an LLM application development project? 1 / 1

- Recruiting a large team of data engineers to organize your data.
- Forming a large team with specialized roles.
- Try experimenting and prototyping with a web-based LLM to assess feasibility.
- Hiring a dedicated prompt engineer.

 **Correct**

Experimenting and prototyping with web interfaces is a viable way to get started with LLM application development. This allows you to understand what is feasible before investing more time and resources in growing the project and team.

Concerns about AI

Concerning 1: Amplify humanity's worst impulses

Concern 1: Amplifying humanity's worst impulses

- LLMs are trained on text from the internet, which reflects some of humanity's best qualities, but also some of its worst
- Will it amplify our worst impulses?

Bias example: The _____ was a CEO.

- LLMs are becoming less biased through fine-tuning and more advanced techniques such as Reinforcement Learning from Human Feedback (RLHF)

RLHF

Reinforcement learning from human feedback (RLHF)

Reduce bias

Step 1: Train an answer quality (reward) model

Prompt:

The _____ was a CEO.

Input (A)	Output (B)
Response	Score (reward)
man	5
woman	5
airplane	1
#!%^?	1

Step 2: Have LLM generate a lot of responses. Further train it to generate more responses that get high scores.

Concerning 2: Job Loss

Concern 2: Job loss

"If you work as a radiologist, you're like the coyote that's already over the edge of the cliff, but hasn't yet looked down, so it doesn't realize there's no ground underneath them.

People should stop training radiologists now.

It's just completely obvious that within five years deep learning is going to do better than radiologists."

Geoff Hinton, 2016

Radiologist tasks (O*NET)

Tasks
All 30 displayed
<ul style="list-style-type: none">● Prepare comprehensive interpretive reports of findings.
<ul style="list-style-type: none">● Perform or interpret the outcomes of diagnostic imaging procedures including magnetic resonance imaging (MRI), computer tomography (CT), positron emission tomography (PET), nuclear cardiology treadmill studies, mammography, or ultrasound.
<ul style="list-style-type: none">● Document the performance, interpretation, or outcomes of all procedures performed.
<ul style="list-style-type: none">● Communicate examination results or diagnostic information to referring physicians, patients, or families.
<ul style="list-style-type: none">● Obtain patients' histories from electronic records, patient interviews, dictated reports, or by communicating with referring clinicians.
<ul style="list-style-type: none">● Review or transmit images and information using picture archiving or communications systems.
<ul style="list-style-type: none">● Confer with medical professionals regarding image-based diagnoses.
<ul style="list-style-type: none">● Recognize or treat complications during and after procedures, including blood pressure problems, pain, oversedation, or bleeding.
<ul style="list-style-type: none">● Develop or monitor procedures to ensure adequate quality control of images.
<ul style="list-style-type: none">● Provide counseling to radiologic patients to explain the processes, risks, benefits, or alternative treatments.
<ul style="list-style-type: none">● Establish or enforce standards for protection of patients or personnel.
<ul style="list-style-type: none">● Coordinate radiological services with other medical activities.
<ul style="list-style-type: none">● Instruct radiologic staff in desired techniques, positions, or projections.
<ul style="list-style-type: none">● Participate in continuing education activities to maintain and develop expertise.
<ul style="list-style-type: none">● Participate in quality improvement activities including discussions of areas where risk of error is high.
<ul style="list-style-type: none">● Perform interventional procedures such as image-guided biopsy, percutaneous transluminal angioplasty, transhepatic biliary drainage, or nephrostomy catheter placement.

Radiologist

Tasks	AI potential
Interpret X-rays	High
Operate imaging hardware	Medium
Communicate exam results	Medium
Respond to complications during procedure	Low
Document procedure and outcomes	Medium
[...]	[...]

Radiologists

"AI won't replace radiologists.

But radiologists that use AI will replace radiologists that don't."

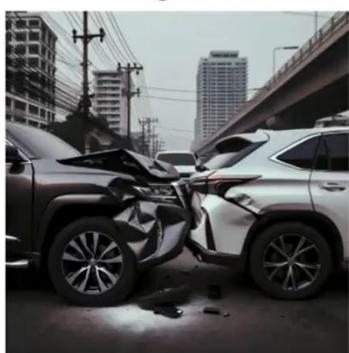
Curtis Langlotz
Professor of Radiology
Stanford University

Concern 3: Human Extinction

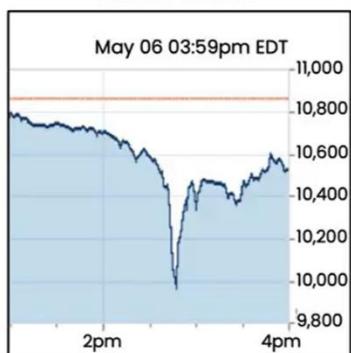
Examples of harm caused by AI

Examples of harm caused by AI

Self-driving car crashes



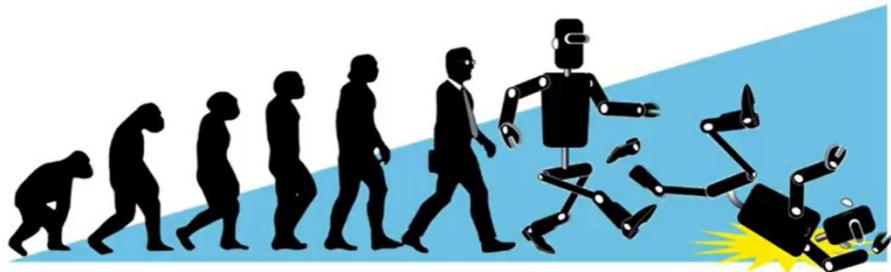
2010 Stock Market Flash crash



Unjust sentencing in criminal cases



Human extinction?



- Extinction arguments are not concrete
- Most arguments boil down to “it could happen.”
- “This is a new type of technology.”

Perfect controls

Perfect control not needed to be valuable and safe

- Humanity has ample experience controlling many things far more powerful than any single person, such as corporations and nation states.
- Many things that we can't fully control are nonetheless valuable and safe. Example: Airplanes.



- If we look at the real risks to humanity – such as climate change and pandemics – AI will be a key part of the solution.

Artificial General Intelligence

Definition

Artificial General Intelligence

Definition: AI that can do any intellectual task that a human can

Examples:

- Learn to drive a car through ~20 hours of practice
- Complete a PhD thesis after ~5 years of work
- Do all the tasks of a computer programmer (or any other knowledge worker)

Responsible AI

Dimensions of Responsible AI

- Fairness
- Transparency
- Privacy
- Security
- Ethical use

Dimensions of responsible AI

- **Fairness:** Ensuring AI does not perpetuate or amplify biases
- **Transparency:** Making AI systems and their decisions understandable to stakeholders impacted
- **Privacy:** Protecting user data and ensure confidentiality
- **Security:** Safeguard AI systems from malicious attacks
- **Ethical Use:** Ensuring AI is used for beneficial purposes

Tips for responsible AI

Tips for responsible AI

- Build a culture that encourages discussion and debate on ethical issues
- Brainstorm how things can go wrong
 - E.g., Could there be issues with fairness, transparency, privacy, security, ethical use?
- Work with a diverse team and include perspectives from all stakeholders

Quiz

1. Which of the following statements about Reinforcement Learning from Human Feedback (RLHF) are true?

-
1. Which of the following statements about Reinforcement Learning from Human Feedback (RLHF) are true?

- After applying RLHF, an LLM will reflect a similar degree of bias and toxicity as text on the internet.
- RLHF is a common technique for training a small (say 1B parameter) LLM to do as well as a larger (say 10B parameter) one.
- RLHF helps to align an LLM to human preferences, and can reduce the bias of an LLM's output.
- RLHF fully addresses all concerns about AI.

 **Correct**

RLHF trains models to produce output that better aligns with human preferences, including honesty, helpfulness, and harmlessness. The process can reduce biases in an LLMs output.

2. True or False. Because AI automates tasks, not jobs, absolutely no jobs will disappear because of AI

True or False. Because AI automates tasks, not jobs, absolutely no jobs will disappear because of AI.

True

False

 **Correct**

Even if all of the tasks of a role can't be completely automated, some jobs may be eliminated as efficiency increases and cost savings can be realized. It is important that we support the individuals who may lose their jobs through safety nets and by creating opportunities for retraining and upskilling.

3. If we manage to build Artificial General Intelligence (AGI) some day, which tasks should AI be capable of performing? (Check all that apply.)

3. If we manage to build Artificial General Intelligence (AGI) some day, which tasks should AI be capable of performing? (Check all that apply.)

Learn to drive a car in roughly 20 hours of practice.

 **Correct**

By definition, AGI can carry out any intellectual task that a human can do. So it should be able to learn to drive a car in roughly 20 hours, just like a human teenager.

Write a software application to let users manage their household spending budgets.

 **Correct**

By definition, AGI can carry out any intellectual task that a human can do. Since software applications like this already exist in the world, the AGI should be able to write one from scratch.

Predict the future (such as make stock market and weather predictions) with perfect accuracy.

Compose the music for a movie soundtrack.

 **Correct**

By definition, AGI can carry out any intellectual task that a human can do. So it should be able to create music for a movie soundtrack.

4. You are working on a chatbot to serve as a career coach for recent college graduates. Which of the following steps could you take to ensure that your project follows responsible AI? (Check all that apply.)

4. You are working on a chatbot to serve as a career coach for recent college graduates. Which of the following steps could you take to ensure that your project follows responsible AI? (Check all that apply.)

1 / 1 p

- Engage diverse recent college graduates and ask them to offer feedback on the output of your chatbot.

 **Correct**

Working with diverse stakeholders can help you identify problems your own team may not recognize, and ensure that the behavior of your chatbot takes into account the perspectives of people from diverse backgrounds.

- Organize a brainstorming session to identify problems that could arise for users chatting with the career coach.

 **Correct**

Building a culture that encourages discussion and debate of ethical issues can help you identify problems early in the development phase and avoid issues of bias or toxicity later in the process.

- Allow a single engineer on your team to determine whether the output of the chatbot is helpful, honest, and harmless.

- Engage employers (because they are a key stakeholder group) and ask them to offer feedback on the output of your chatbot.

 **Correct**

Working with all stakeholders can reveal points of view that your team may not have realized and can help identify problems or issues that may have been missed.