Copyright Notice

These slides are distributed under the Creative Commons License.

DeepLearning.AI makes these slides available for educational purposes. You may not use or distribute these slides for commercial purposes. You may make copies of these slides and use or distribute them for educational purposes as long as you cite DeepLearning.AI as the source of the slides.

For the rest of the details of the license, see https://creativecommons.org/licenses/by-sa/2.0/legalcode

DeepLearning.AI

Software Applications

Using generative AI in software applications

Examples of software applications

Writing

Is there parking for staff?

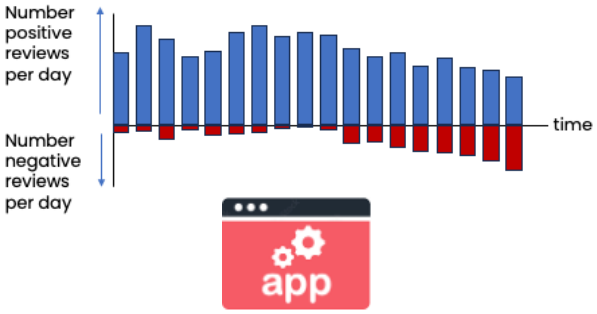
Yes, employees can park on levels 1 and 2 of the office lot. You can get a sticker at [...]

Reading

Classify the sentiment of the following review as either positive or negative:

The mochi is excellent!

Positive



Chatting

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!

```python
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
        hidden = (weight.new(self.n_layers, batch_size, self.hidden_dim).zero_(),
```

Supervised learning for restaurant reputation monitoring

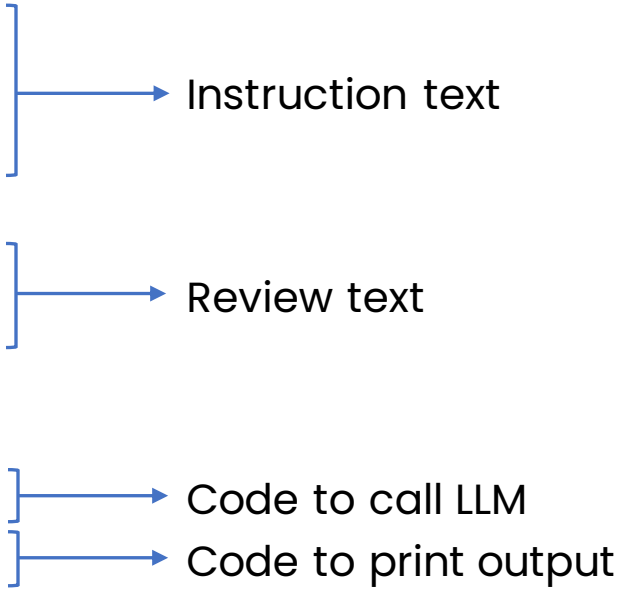
Input A → Output B

Input (A)	Output (B)
Best soup dumplings I've ever eaten.	Positive
The colorful tablecloths made me smile!	Positive
Not worth the 3 month wait for a reservation	Negative

Input: "Best bubble tea I've ever had" Output: Positive

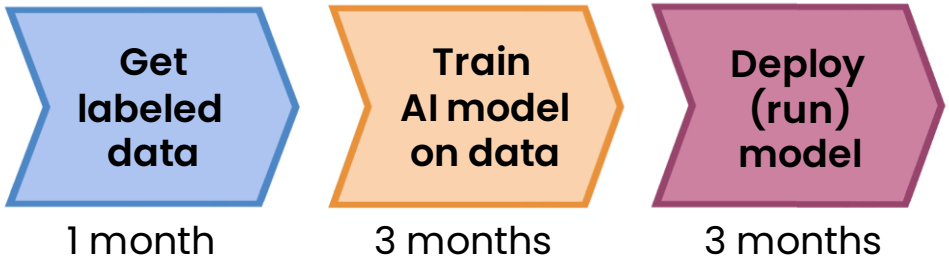
Prompt-based development

```
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)
```

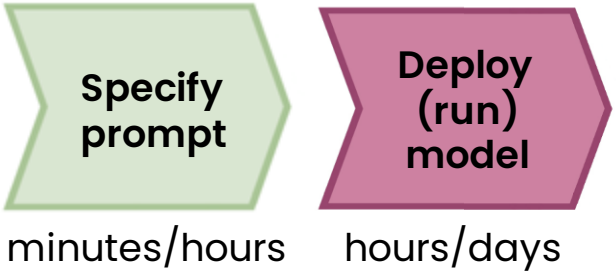


Workflow using Generative AI

Supervised learning



Prompt-based AI



DeepLearning.AI

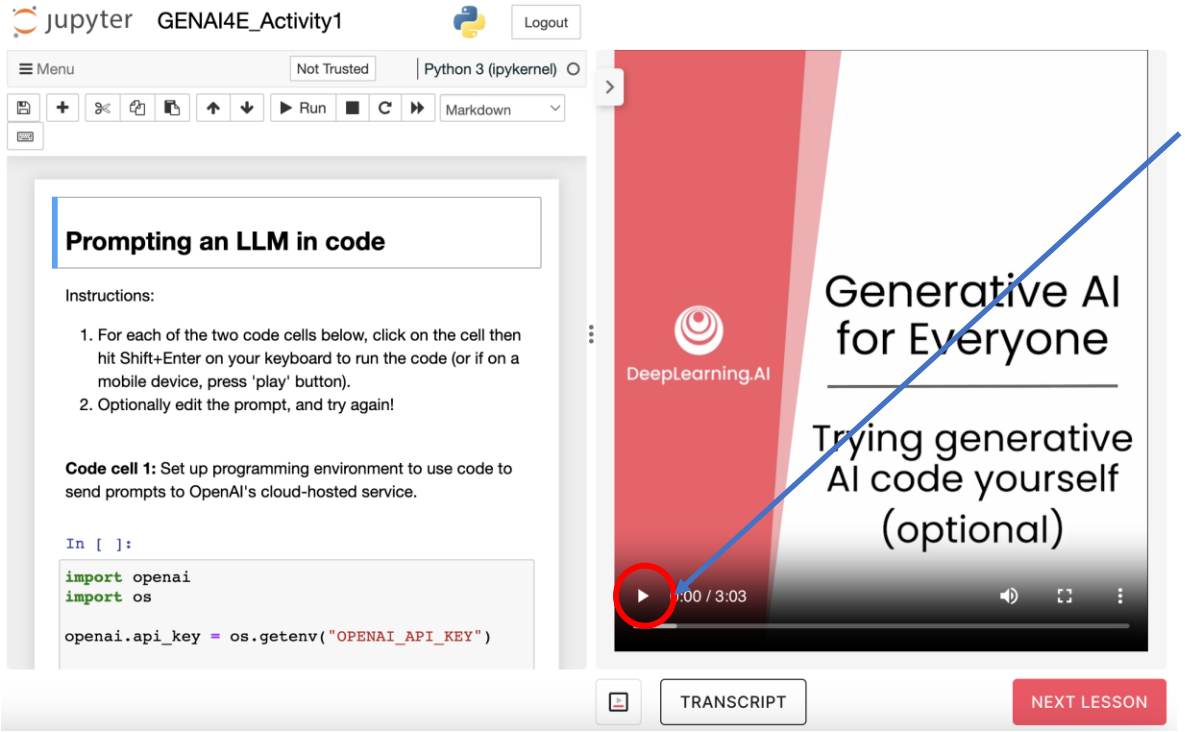
Software Applications

Trying generative AI code yourself (optional)

Coding platform (Jupyter notebook)

The one command to know:

shift + enter



Play button

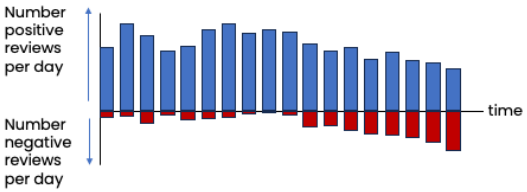
DeepLearning.AI

Software Applications

Lifecycle of a generative AI project

Lifecycle of a generative AI project

Scope project → Build/improve system → Internal evaluation



Initially a prototype, that we will improve over time

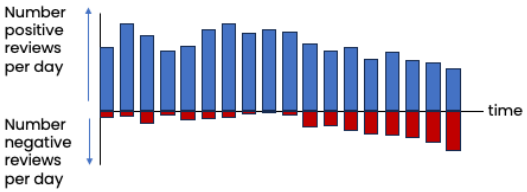
Classify the sentiment of the following review as either positive or negative:

The custard tart was amazing!

Positive

Lifecycle of a generative AI project

Scope project → Build/improve system → Internal evaluation



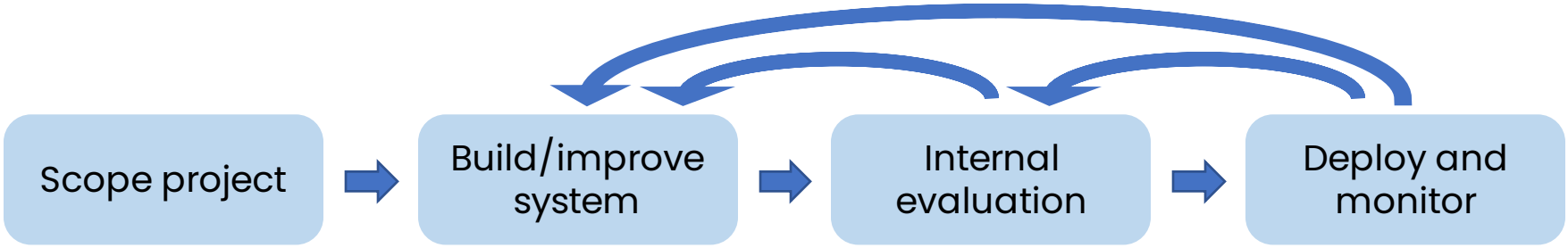
Initially a prototype, that we will improve over time

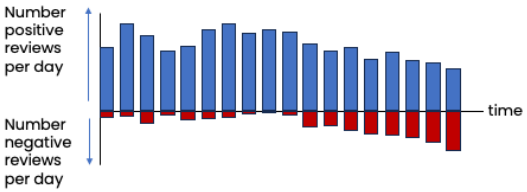
Classify the sentiment of the following review as either positive or negative:

My pasta was cold

Positive ✖ Incorrect response

Lifecycle of a generative AI project





Initially a prototype, that we will improve over time

Classify the sentiment of the following review as either positive or negative:

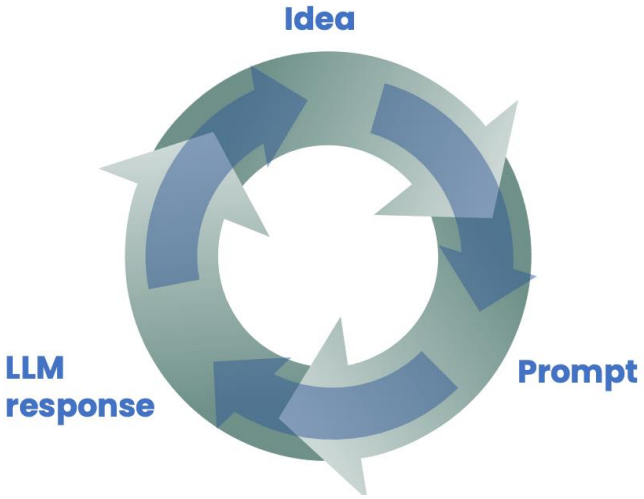
My miso ramen tasted like tonkotsu ramen.

Positive

Tools to improve performance

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

- Prompting



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

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
- Pretrain models
  - Train LLM from scratch

Lifecycle of a generative AI project

Scope project → Build/improve system → Internal evaluation

Take food order

Let team order...

I'd like a mushroom and swiss burger

I'm sorry, we don't have mushrooms, how about onions?

✖ Incorrect response

Lifecycle of a generative AI project

Scope project → Build system → Internal evaluation

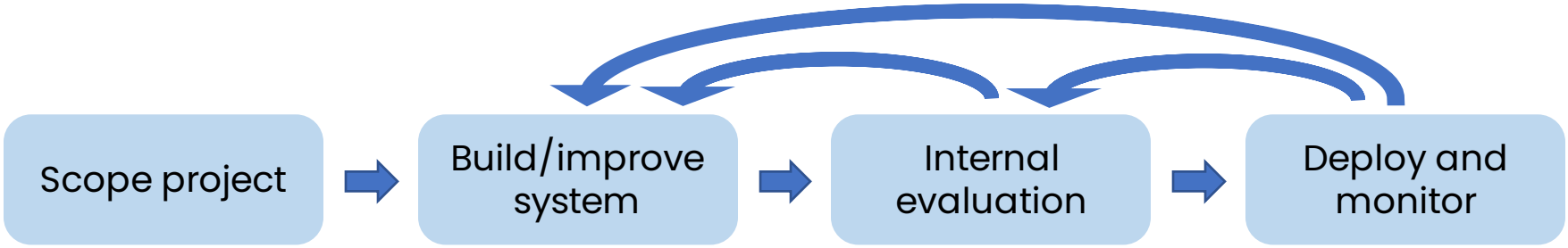
Take food order

Let team order...

Do you have pickles?

Yes – would you like to add some?

Lifecycle of a generative AI project



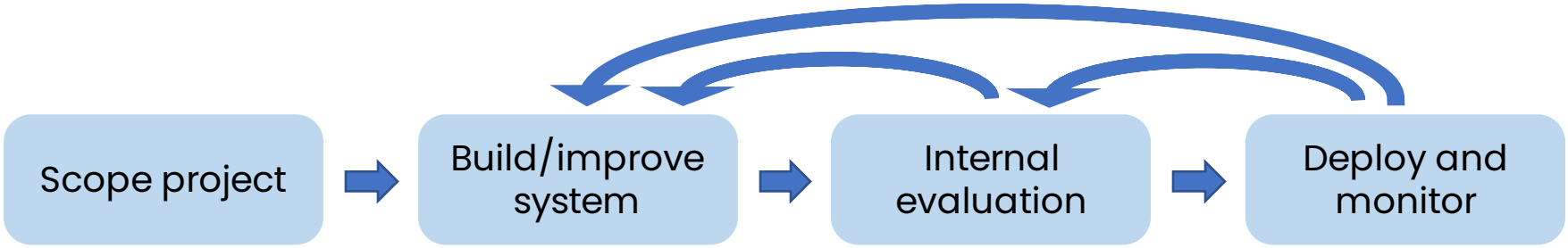
Take food order

Let customers order, monitor LLM responses

How many calories in the cheeseburger?

I'm sorry, I don't know

Lifecycle of a generative AI project



Take food order

Let customers order, monitor LLM responses

How many calories in the cheeseburger?

About 375 calories

DeepLearning.AI

Software Applications

Cost intuition

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

translate programming — 2 tokens

tonkotsu — 4 tokens

300 words
400 tokens

Roughly, 1 token = 3/4 words

Estimating Cost

Typical adult reading speed: 250 words/minute.

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

$60 \times 250 = 15{,}000$ ← output

$+\ 15{,}000$ ← prompt

$30{,}000$

↳ 40,000 tokens

1 token ≈ $\frac{3}{4}$ word

Cost: $0.002 cents / 1K tokens.

$0.002 × 40 = $0.08

DeepLearning.AI

Advanced Technologies: Beyond Prompting

Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) example

General Chatbot

Is there parking for employees?

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

Chatbot with RAG

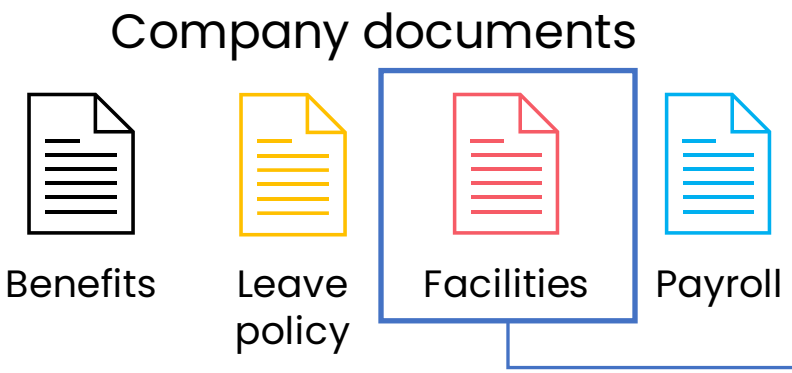
Is there parking for employees?

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

Retrieval Augmented Generation (RAG) example

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 [...]

Is there parking for employees?

Retrieval Augmented Generation (RAG) example

3. 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

Examples of RAG applications

Chat with PDF files

panda chat

CHATPDF

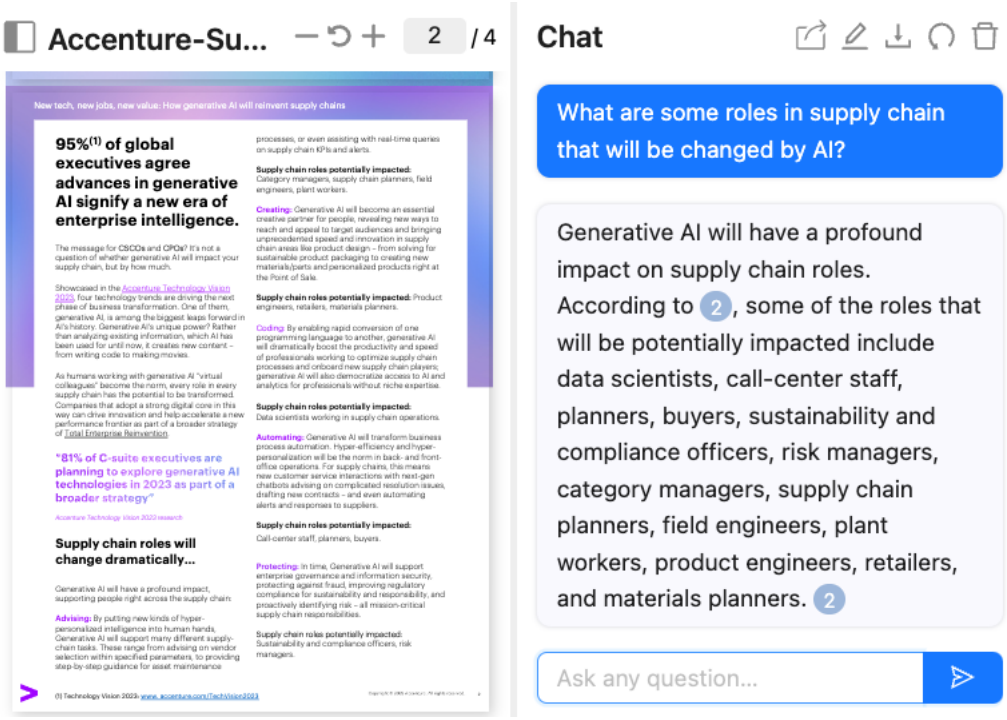
AskYourPDF

PDF.ai

LightPDF

docAnalyzer.AI

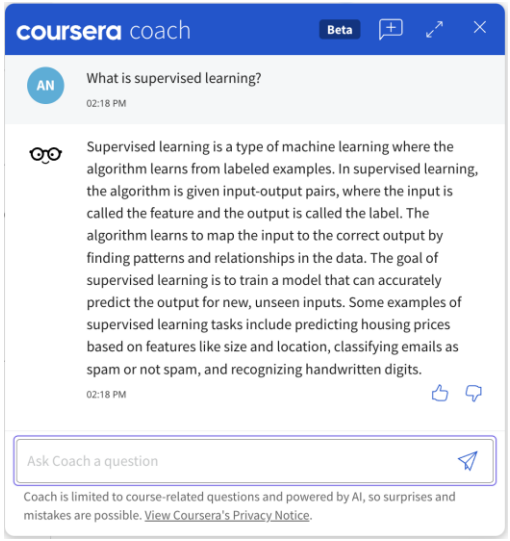
Tactic



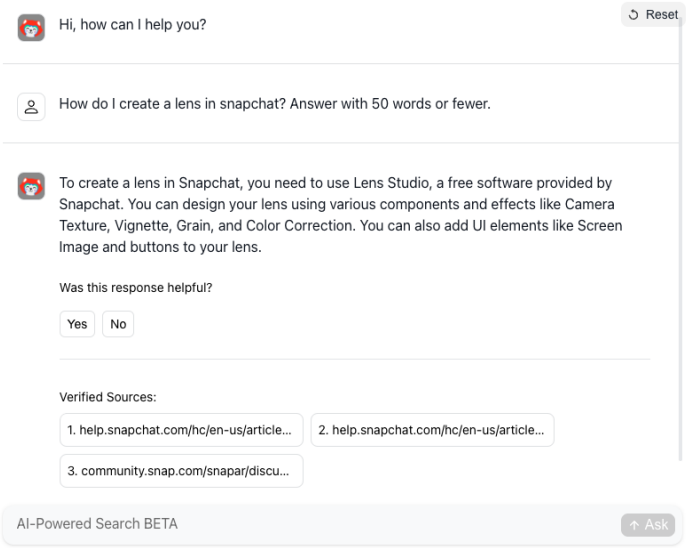
Chatting with an Accenture report using ChatPDF

Examples of RAG applications

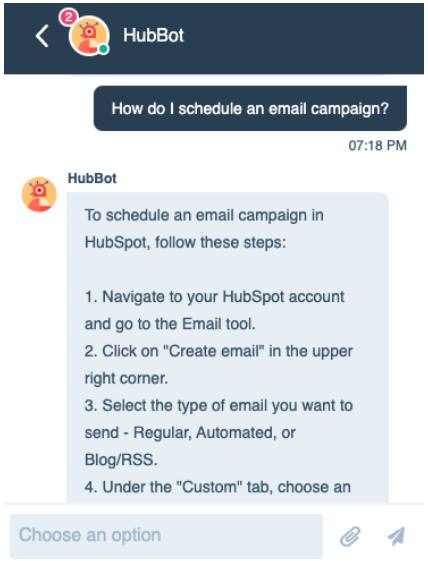
Answer questions based on a website's articles



Coursera Coach



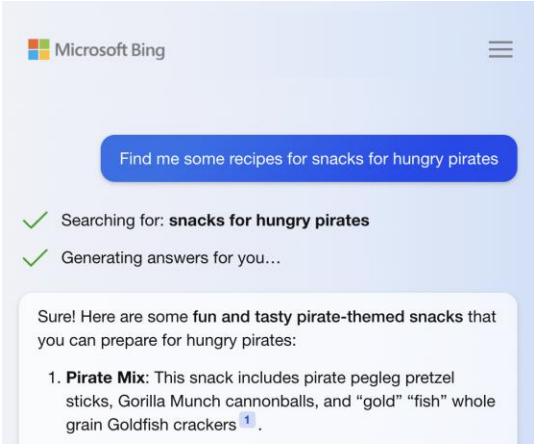
Snapchat



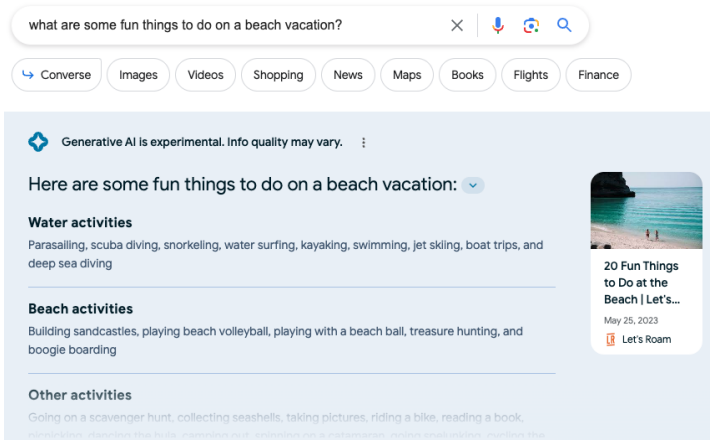
Hubspot

Examples of RAG applications

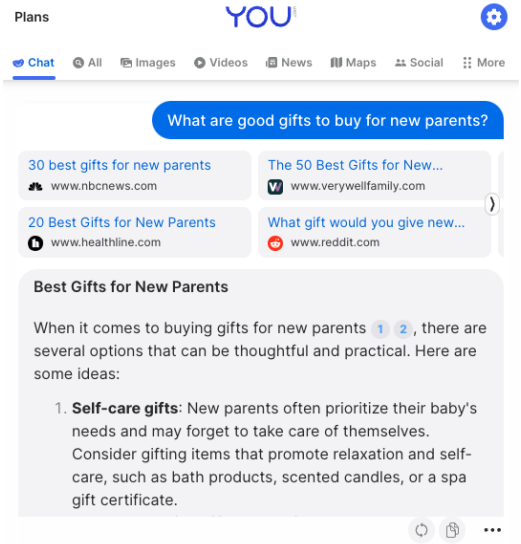
New form of web search



Microsoft/Bing Chat



Google



You.com

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

DeepLearning.AI

Advanced Technologies: Beyond Prompting

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?

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.*

Customer tells agent about a problem with monitor.

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.

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	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 [...]	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 [...]
Pretrained model	Fine-tuned model

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

Why fine-tune?

To help LLM gain specific knowledge.

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.

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."

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	Classify the following review as having a positive or negative sentiment: The noodle soup was too salty	Classify the following review as having a positive or negative sentiment: The staff sang happy birthday to me!
Positive	Negative	Positive

Can work well with 500-1000 examples.

DeepLearning.AI

Advanced Technologies: Beyond Prompting

Pretraining an LLM

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

Bloomberg

March 30, 2023

DeepLearning.AI

Advanced Technologies: Beyond Prompting

Choosing a model

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?

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

DeepLearning.AI

Advanced Technologies: Beyond Prompting

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

How do chat systems learn to follow instructions?

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

What is the capital of France?

What is the capital of Germany?
Where is Mumbai?
Is Mt Fuji or Mt Kilimanjaro taller?
[...]

How do chat systems learn to follow instructions?

What is the capital of South Korea?

The capital of South Korea is Seoul.

Help me brainstorm some fun museums to visit in Bogatá.

Sure, here are some suggestions:
[...]

Write a Haiku poem about Japan's cherry blossoms.

Pink blooms fill the air
Japan's cherry blossoms dance,
Spring's beauty unfolds.

Tell me how to break into Fort Knox.

I can't assist with that.

How do chat systems learn to follow instructions?

Fine-tuning

Help me brainstorm some fun museums to visit in Bogatá.

Sure, here are some suggestions: […]

Input (A)	Output (B)
Help me brainstorm some fun museums to visit in Bogatá.	Sure,
Help me brainstorm some fun museums to visit in Bogatá. Sure,	here
Help me brainstorm some fun museums to visit in Bogatá. Sure, here	are
Help me brainstorm some fun museums to visit in Bogatá. Sure, here are	some
Help me brainstorm some fun museums to visit in Bogatá. Sure, here are some	suggestions:
•••	•••

Reinforcement learning from human feedback (RLHF)

Helpful, Honest, Harmless

Step 1: Train an answer quality (reward) model

Prompt:
Advise me on how to apply for a job.

Input (A)	Output (B)
Response	Score (reward)
I'm happy to help! Here are some steps to follow […]	5
Just try your best!	3
It's hopeless – why bother?	1

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

DeepLearning.AI

Advanced Technologies: Beyond Prompting

Tool use and agents (optional)

Tool use for food order taking

Send me a burger!

Ok, it's on the way!

Tool use for food order taking

Send me a burger!

ORDER(Burger, 9876, 1234 My Street)

User message:
Ok, it's on the way

Ok, it's on the way!

Passed to restaurant ordering system:

y Street

as response

You have ordered:

1 burger

for delivery to:

1234 My Street

Is this correct? (If yes, your credit card will be charged)

Yes

No

Tools for reasoning

How much would I have after 8 years if I deposit $100 into a bank account that pays 5% interest?

After compounding $100 at 5% interest for 8 years, you would have $147.04 dollars.

✖ Incorrect value

LLMs are not great at precise math

Tools for reasoning

How much would I have after 8 years if I deposit $100 into a bank account that pays 5% interest?

After compounding $100 at 5% interest for 8 years, you would have CALCULATOR(100 * 1.05^8) dollars.

External calculator program

100 * 1.05^8 = 147.74

After compounding $100 at 5% interest for 8 years, you would have 147.74 dollars.

Agents

- Use LLM to choose and carry out complex sequences of actions
- Cutting edge area of AI research

Help me research BetterBurgers top competitors

Here are steps I need to carry out to research competitors:
1. Search top competitors
2. Visit web site of each competitor
3. For each competitor, write summary based on homepage content

SEARCH("BetterBurgers Competitors")

VISIT(http://www.fastburger.com)

VISIT(http://www.burgerworld.com)

...

Summarize the following text:

At Fast Burger, we pride ourselves on the speed of our delivery ...