

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 [...]

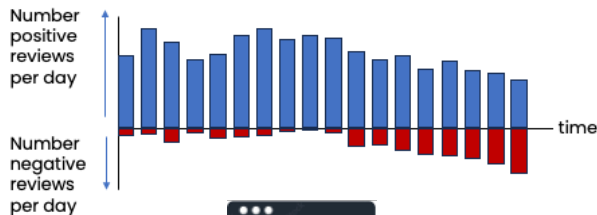


Reading

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!



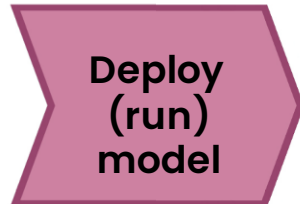
```

1 import torch
2 from torch.utils.data import DataLoader, TensorDataset
3 from torch import nn
4
5 class SentimentLSTM(nn.Module):
6     def __init__(self, vocab_size, output_size, embedding_dim, hidden_dim, n_layers):
7         super(SentimentLSTM, self).__init__()
8         self.output_size = output_size
9         self.n_layers = n_layers
10        self.hidden_dim = hidden_dim
11
12        self.embedding = nn.Embedding(vocab_size, embedding_dim)
13        self.lstm = nn.LSTM(embedding_dim, hidden_dim, n_layers, batch_first=True)
14        self.fc = nn.Linear(hidden_dim, output_size)
15        self.sigmoid = nn.Sigmoid()
16
17    def forward(self, x, hidden):
18        batch_size = x.size(0)
19        x = self.embedding(x)
20        lstm_out, hidden = self.lstm(x, hidden)
21        lstm_out = lstm_out.contiguous().view(-1, self.hidden_dim)
22        out = self.fc(lstm_out)
23        out = self.sigmoid(out)
24        out = out.view(batch_size, -1)
25        out = out[:, -1]
26        return out, hidden
27
28    def init_hidden(self, batch_size):
29        weight = next(self.parameters()).data
30        hidden = (weight.new(self.n_layers, batch_size, self.hidden_dim).zero_(),
31
```

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)
```

Instruction text

Review text

Code to call LLM

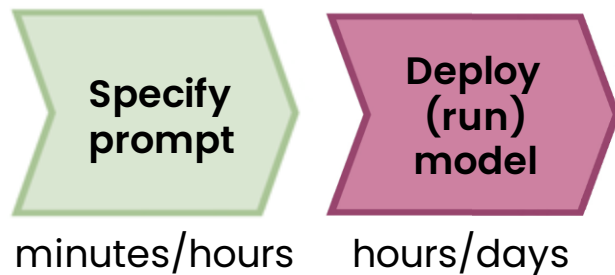
Code to print output

Workflow using Generative AI

Supervised learning



Prompt-based AI





DeepLearning.AI

Software Applications

Trying generative AI code
yourself (optional)

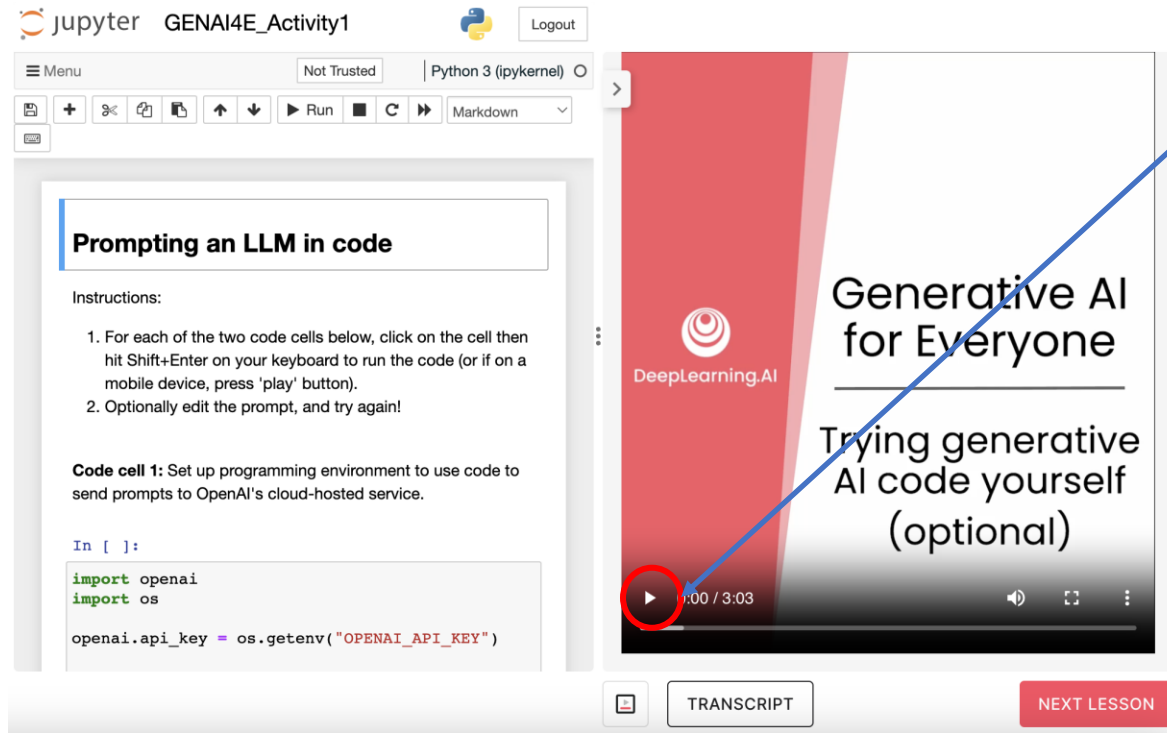
Coding platform (Jupyter notebook)

The one
command
to know:

shift

+

enter



jupyter GENAI4E_Activity1 Logout

Menu Not Trusted Python 3 (ipykernel)

Run

Prompting an LLM in code

Instructions:

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 [ ]:
```

```
import openai
import os

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

Generative AI for Everyone

Trying generative AI code yourself (optional)

0:00 / 3:03

TRANSCRIPT

NEXT LESSON

Play button

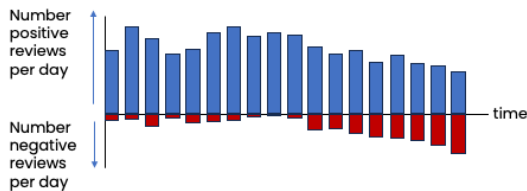
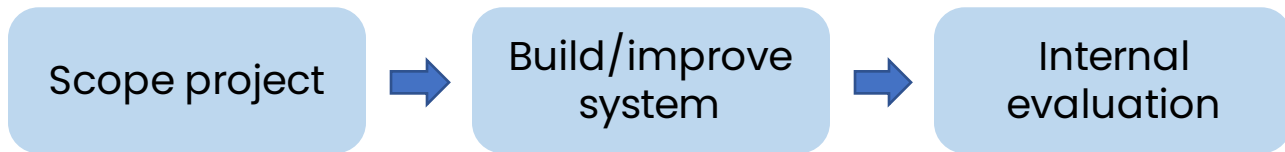


DeepLearning.AI

Software Applications

Lifecycle of a
generative AI project

Lifecycle of a generative AI project



Initially a prototype, that we will improve over time

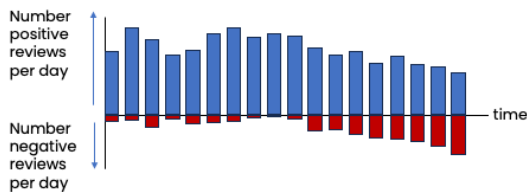
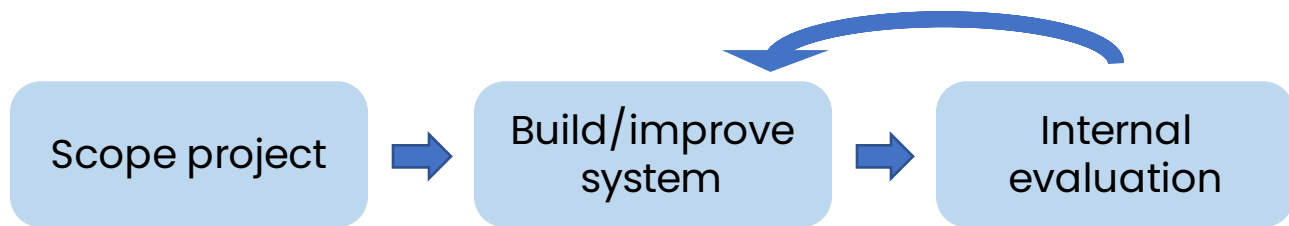


Classify the sentiment of the following review as either positive or negative:

The custard tart was amazing!

Positive

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:

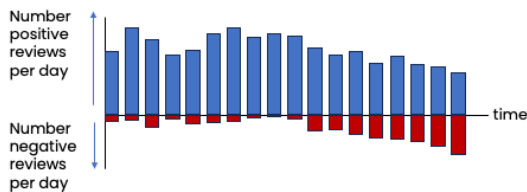
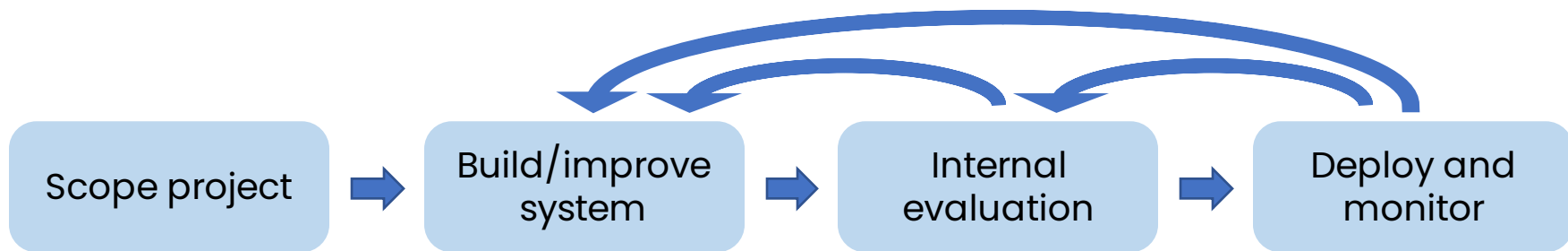
My pasta was cold

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:

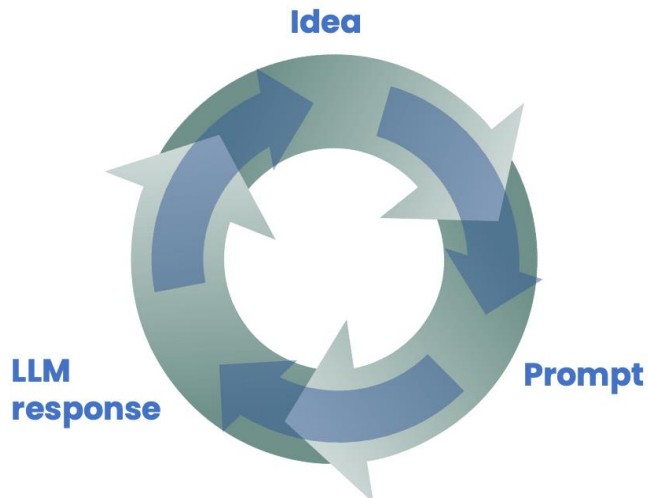
My miso ramen tasted like tonkotsu ramen.

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.

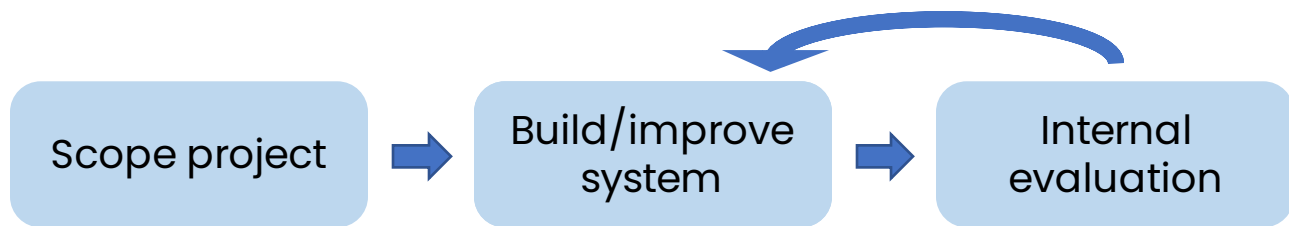
- Prompting
- Retrieval augmented generation (RAG)
 - Give LLM access to external data sources
- Fine-tune models
 - Adapt LLM to your task

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



Take food order



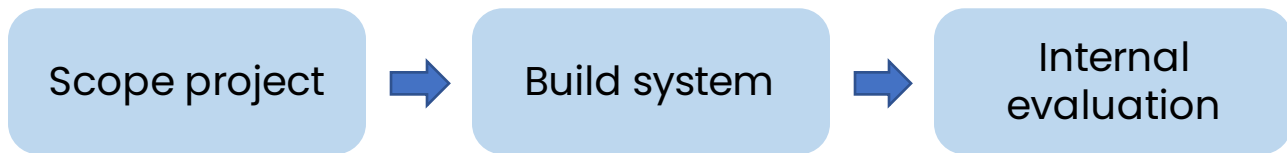
Let team order...

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



Take food
order

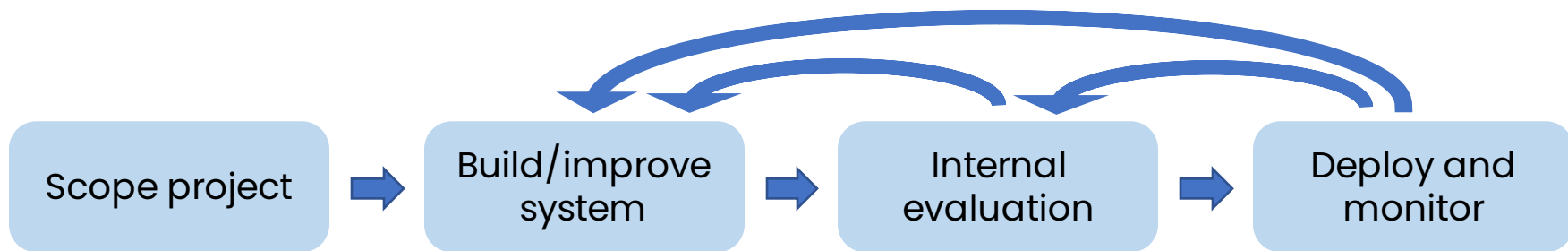


Let team
order...

Do you have pickles?

Yes – would you like to
add some?

Lifecycle of a generative AI project



Take food
order

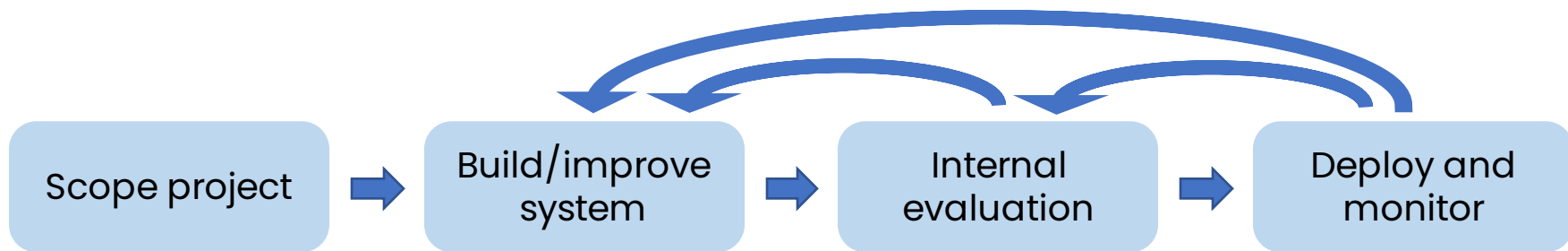


Let customers
order, monitor
LLM responses

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	<u>\$0.002/1K tokens</u>	\$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?

$$\begin{array}{r} 60 \times 250 = 15,000 \leftarrow \text{output} \\ + 15,000 \leftarrow \text{prompt} \\ \hline 30,000 \end{array}$$

→ 40,000 tokens

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

Cost: \$0.002 cents / 1K tokens.
 $\$0.002 \times 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?

Company documents



Benefits



Leave
policy



Facilities



Payroll

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

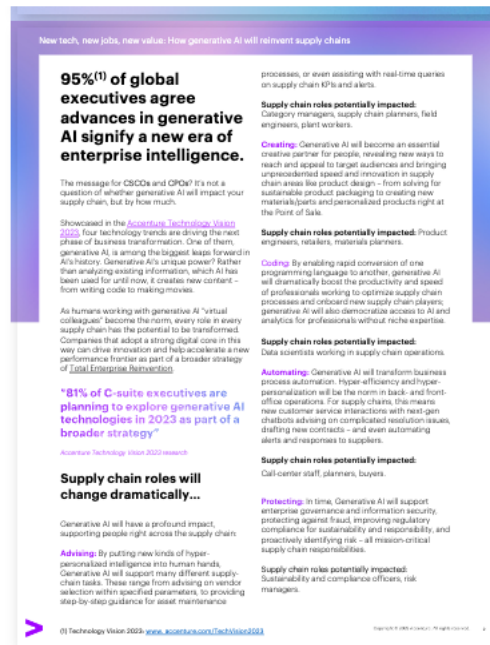
Chat with PDF files



CHATPDF



Accenture-Su... 2 / 4



Chat



What are some roles in supply chain that will be changed by AI?

Generative AI will have a profound impact on supply chain roles. According to 2, some of the roles that will be potentially impacted include data scientists, call-center staff, planners, buyers, sustainability and compliance officers, risk managers, category managers, supply chain planners, field engineers, plant workers, product engineers, retailers, and materials planners. 2

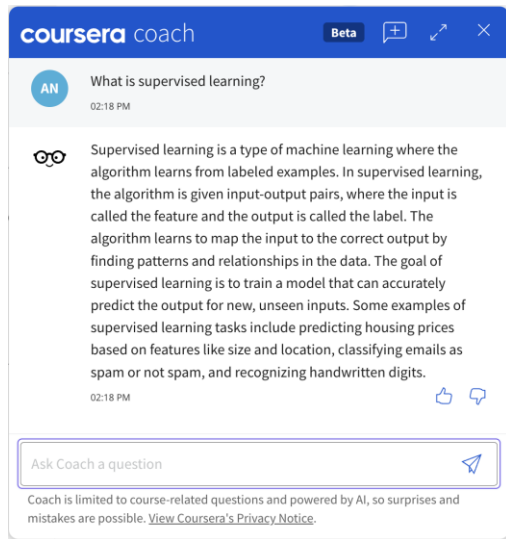
Ask any question...



Chatting with an Accenture report using ChatPDF

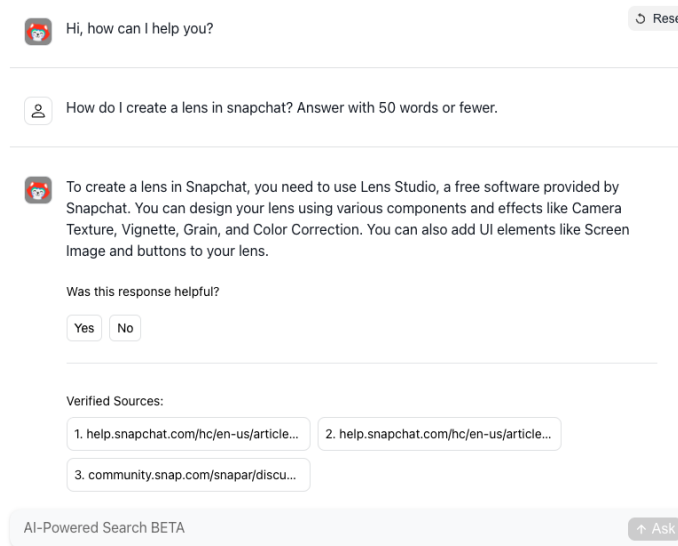
Examples of RAG applications

Answer questions based on a website's articles



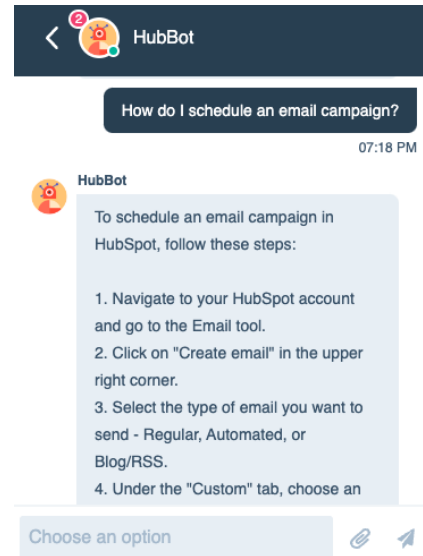
The screenshot shows the Coursera Coach interface. At the top, it says "coursera coach" with a "Beta" tag and icons for chat, share, and close. Below this, a user profile icon labeled "AN" is next to the question "What is supervised learning?" with a timestamp of "02:18 PM". The answer text explains supervised learning as a type of machine learning where an algorithm learns from labeled examples. It mentions that the algorithm is given input-output pairs, where the input is the feature and the output is the label. The goal is to train a model that can accurately predict the output for new, unseen inputs. Examples of supervised learning tasks include predicting housing prices based on features like size and location, classifying emails as spam or not spam, and recognizing handwritten digits. At the bottom, there is a text input field with the placeholder "Ask Coach a question" and a send icon. A disclaimer at the very bottom states: "Coach is limited to course-related questions and powered by AI, so surprises and mistakes are possible. [View Coursera's Privacy Notice.](#)"

Coursera Coach



The screenshot shows the Snapchat AI chat interface. At the top, it says "Hi, how can I help you?" with a "Reset" button. Below this, a user profile icon is next to the question "How do I create a lens in snapchat? Answer with 50 words or fewer." The answer text explains that to create a lens in Snapchat, you need to use Lens Studio, a free software provided by Snapchat. It mentions that you can design your lens using various components and effects like Camera Texture, Vignette, Grain, and Color Correction. You can also add UI elements like Screen Image and buttons to your lens. Below the answer, there is a feedback section with the question "Was this response helpful?" and "Yes" and "No" buttons. At the bottom, there is a "Verified Sources:" section with three links: "1. help.snapchat.com/hc/en-us/article...", "2. help.snapchat.com/hc/en-us/article...", and "3. community.snap.com/snapar/discu...". At the very bottom, it says "AI-Powered Search BETA" and "Ask".

Snapchat

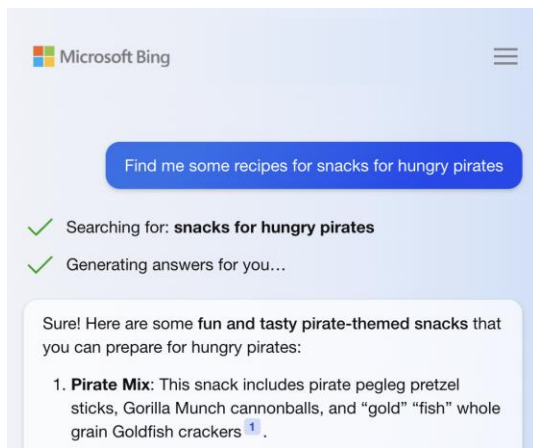


The screenshot shows the HubSpot HubBot interface. At the top, it says "HubBot" with a "Reset" button. Below this, a user profile icon is next to the question "How do I schedule an email campaign?". The answer text provides a step-by-step guide: "1. Navigate to your HubSpot account and go to the Email 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". At the bottom, there is a text input field with the placeholder "Choose an option" and a send icon.

Hubspot

Examples of RAG applications

New form of web search



Microsoft Bing

Find me some recipes for snacks for hungry pirates

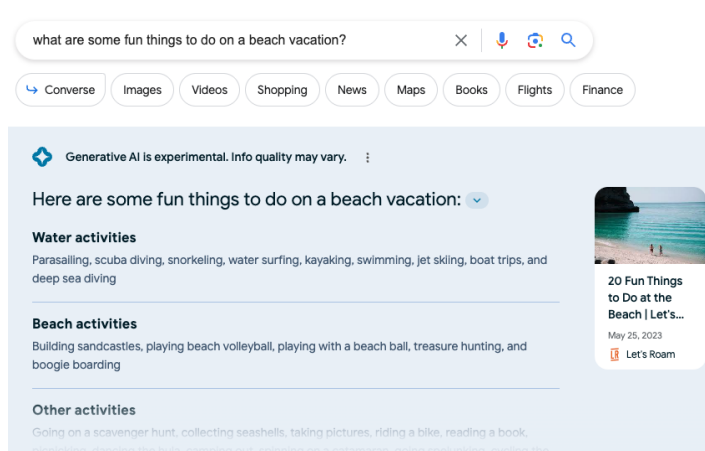
✓ Searching for: **snacks for hungry pirates**

✓ 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 ¹.

Microsoft/Bing Chat



what are some fun things to do on a beach vacation?

Converse Images Videos Shopping News Maps Books Flights Finance

Generative AI is experimental. Info quality may vary.

Here are some fun things to do on a beach vacation:

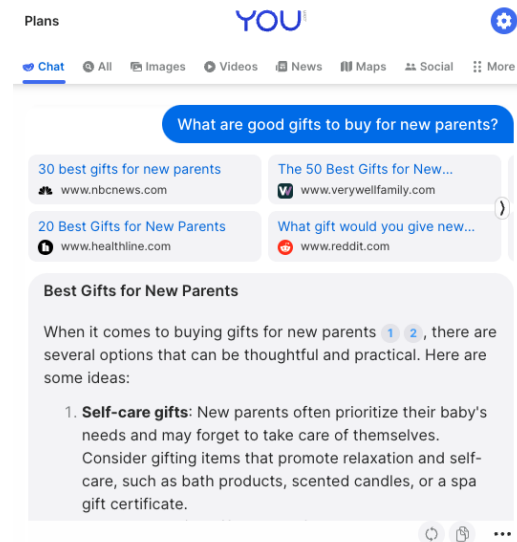
Water activities
Parasailing, scuba diving, snorkeling, water surfing, kayaking, swimming, jet skiing, boat trips, and deep sea diving

Beach activities
Building sandcastles, playing beach volleyball, playing with a beach ball, treasure hunting, and boogie boarding

Other activities
Going on a scavenger hunt, collecting seashells, taking pictures, riding a bike, reading a book.

20 Fun Things to Do at the Beach | Let's...
May 25, 2023
Let's Roam

Google



Plans YOU

Chat All Images Videos News Maps Social More

What are good gifts to buy for new parents?

30 best gifts for new parents
www.nbcnews.com

The 50 Best Gifts for New...
www.verywellfamily.com

20 Best Gifts for New Parents
www.healthline.com

What gift would you give new...
www.reddit.com

Best Gifts for New Parents

When it comes to buying gifts for new parents ¹ ², there are several options that can be thoughtful and practical. Here are some ideas:

1. **Self-care gifts:** New parents often prioritize their baby's needs and may forget to take care of themselves. Consider gifting items that promote relaxation and self-care, such as bath products, scented candles, or a spa gift certificate.

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

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

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)

Large model: 100B+ parameters  Small model: 1B parameters

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.



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 Bogotá.

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 Bogotá.

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

Input (A)

Output (B)

Help me brainstorm some fun museums to visit in Bogotá.

Sure,

Help me brainstorm some fun museums to visit in Bogotá. Sure,

here

Help me brainstorm some fun museums to visit in Bogotá. Sure, here

are

Help me brainstorm some fun museums to visit in Bogotá. Sure, here are

some

Help me brainstorm some fun museums to visit in Bogotá. 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

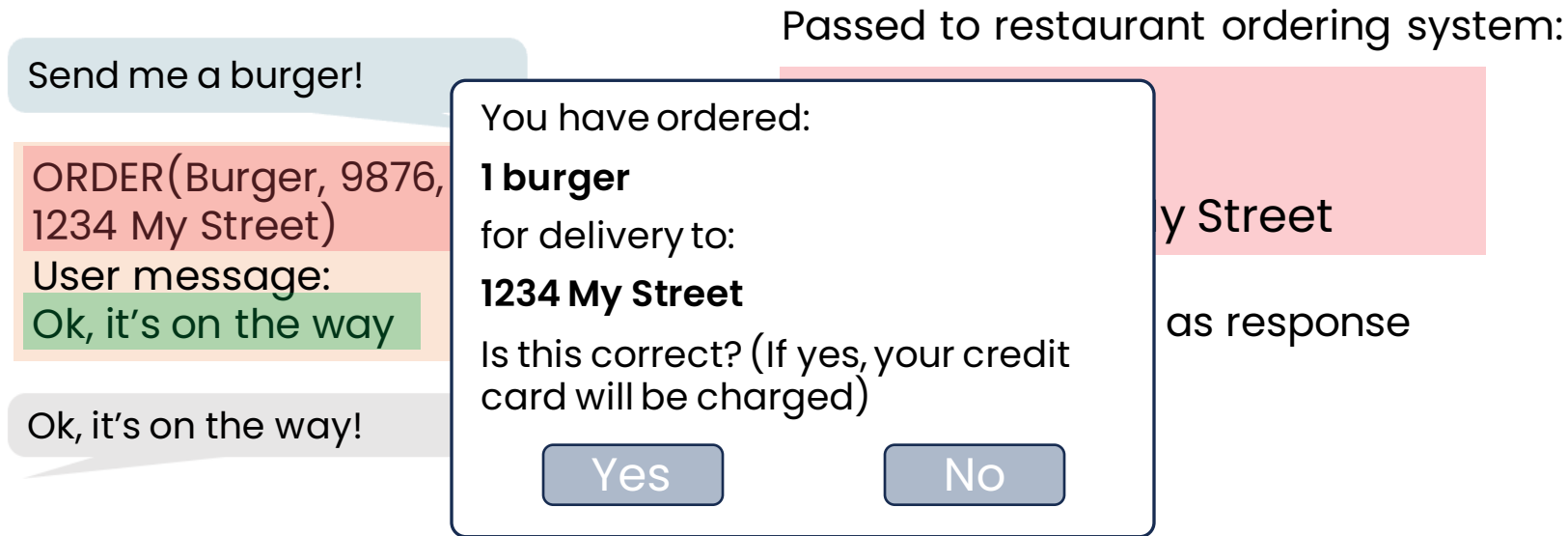
Tool use and agents
(optional)

Tool use for food order taking

Send me a burger!

Ok, it's on the way!

Tool use for food order taking



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.

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

External calculator program

$$100 * 1.05^8 = 147.74$$

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 ...