DEEP LEARNING

Question 1
Which of these is the best definition of “Generative AI”?
1 / 1 point
Any web-based application that generates text.
A form of web search.
Artificial intelligence systems that can map from an input A to an output B.
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.
Question 2
Which of these is the most accurate description of an LLM?
1 / 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.
Correct
A Large Language Model (LLM) has been trained to repeatedly predict the next word using 100 billions - trillions of examples of text from the internet.

3.
Question 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.
Question 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.

5.
Question 5
You hear of a company using an LLM to automatically route emails to the right department. Which of these use cases is it most likely to be?
1 / 1 point
The company has a software-based application that uses an LLM to automatically route emails.
Employees are copy-pasting the emails into a web interface to decide how to route them.
Correct
A software-based application that uses an LLM to automatically read and route emails can process many emails automatically as they are received. So it is most likely the company is using a software-based application to carry out this work.

Question 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?
1 / 1 point
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.
Question 2
Which of the following are tasks that LLMs can do? (Check all that apply)
1 / 1 point
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.
Summarize articles.
Correct
LLMs can take long texts as input and output shorter summaries of those texts.
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).
Earn a university degree (similar to a fresh college graduate).

3.
Question 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?

0 / 1 point

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.

Question 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?

1 / 1 point

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.

Do not use this quote because an LLM can generate toxic output.

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.

Correct

LLMs can generate authoritative sounding, but factually inaccurate text (a behavior known as “hallucinating”). It is important to double check its output when factual accuracy is important to your task.

5.

Question 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?

1 / 1 point

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.

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.

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.

Question 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 / 1 point

True

False

Correct

RAG provides an LLM with additional information and context from external documents that it can reason through to answer a question. So RAG would enable an LLM to answer questions about current news articles.

2.

Question 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?

1 / 1 point

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

Pretraining an LLM on your emails.

Fine-tuning an LLM on your emails, whereby we take a pre-trained LLM and further train it on your emails.

Prompting (without RAG), where we iteratively refine the prompt until the LLM gets the answers right.

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.

Question 3

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

1 / 1 point

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.

Question 4

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 giving the LLM more information only confuses the LLM more and causes it to be more likely to hallucinate.

True, because RAG allows the LLM to reason through accurate information retrieved from a trusted source to arrive at the correct answer.

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.

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.

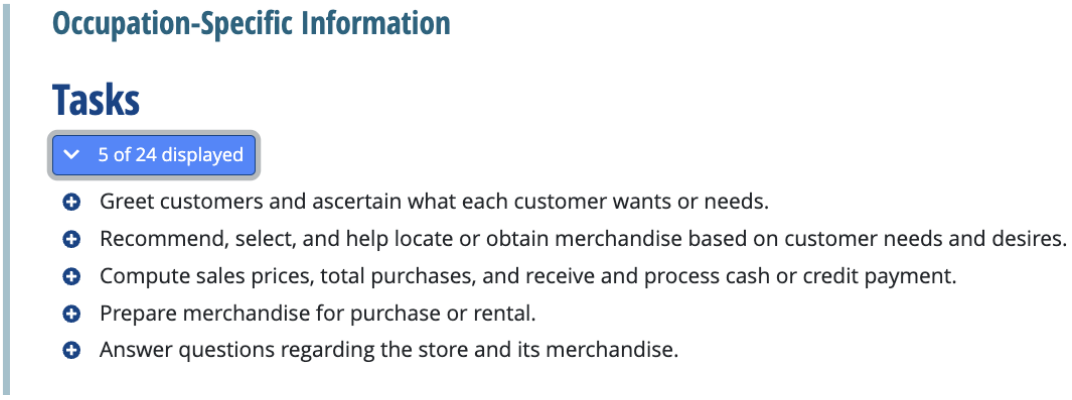
5.
Question 5
An ecommerce company is building a software application to route emails to the right department (Apparel, Electronics, Home Appliances, etc.) It wants to do so with a small, 1 billion parameter model, and needs high accuracy. Which of these is an appropriate technique?
1 / 1 point
Fine-tune 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 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 1,000 examples of emails and the appropriate department.
Correct
Fine-tuning an existing model is an effective way to get it to learn how to correctly route emails to the specific departments in the ecommerce company. Fine-tuning can be done well with 1000 to 10,000 examples.

Which of these job roles are unlikely to find any use for web UI LLMs?
1 / 1 point
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.
Question 2
What is the relation between AI, tasks, and jobs?
1 / 1 point
Jobs are comprised of many tasks. AI automates tasks, rather than jobs.

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

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



Occupation-Specific Information

Tasks

- Greet customers and ascertain what each customer wants or needs.
- Recommend, select, and help locate or obtain merchandise based on customer needs and desires.
- Compute sales prices, total purchases, and receive and process cash or credit payment.
- Prepare merchandise for purchase or rental.
- 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?
1 / 1 point
This has no business value and should not be done.
Build an AI system to suggest products to the salesperson, who then decides what to recommend to the customer.
Build a chatbot that automatically recommends products that customers can access directly, with no salesperson involved.
Build an AI chatbot that can role-play being a customer to help the salesperson practice having conversations with customers.
Correct
Here the AI is augmenting the work of the salesperson by making suggestions, rather than fully taking over and automating the task.

4.
Question 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.)
0.5 / 1 point
Whether the task is the iconic, defining task for a job role.
This should not be selected

Generative AI has the potential to augment or automate many of the tasks that make up a job role, not just the most iconic, defining task. You should evaluate the business value of all tasks in a role to determine the potential for generative AI augmentation of automation.

Whether to use prompting, RAG or fine-tuning.

Technical feasibility (can AI do it?).

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.

5.

Question 5

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

1 / 1 point

Forming a large team with specialized roles.

Try experimenting and prototyping with a web-based LLM to assess feasibility.

Hiring a dedicated prompt engineer.

Recruiting a large team of data engineers to organize your data.

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.

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 / 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.
Question 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.
Question 3
What are the major steps of the lifecycle of a Generative AI project?
1 / 1 point
Scope project → Build/improve system → Internal evaluation → Deploy and monitor
Scope project → Internal evaluation → Build/improve system → Deploy and monitor
Scope project → Internal evaluation → Deploy and monitor → Build/improve system
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.

4.
Question 4

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

1 / 1 point

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.

Every product should be monitored to track customer satisfaction -- this is good practice for all software.

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.

Question 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).

1 / 1 point

4,500 tokens (6000 * 3/4)

6,000 tokens

8,000 tokens (about 6000 * 1.333)

14,000 tokens (about 6000 * 1.333 + the original 6000 words)

Correct

A token typically represents a single common word or a part of the word. This means that a word can represent anywhere between 1 or more tokens. Therefore, an LLM typically requires more tokens to process the input number of words.

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

1 / 1 point

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.

Question 2

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

1 / 1 point

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.

Question 3

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

1 / 1 point

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.

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.

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.

4.
Question 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 point
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.

5.
Question 5
Now that you've made it to the end of the course, which of these statements are true? (Please check all, because all apply!)
1 / 1 point
You understand how Generative AI technology works, and what it can and cannot do.
Correct
You're well positioned to use Generative AI responsibly to help yourself and others.
Correct
You've achieved the significant accomplishment of finishing this course.
Correct

Andrew is thrilled at your completing, and sends you his warmest thank you and congratulations!
Correct

Fill in the blanks: __________ involves using many prompt-completion examples as the labeled training dataset to continue training the model by updating its weights. This is different from _________ where you provide prompt-completion examples during inference.
1 / 1 point
Prompt engineering, Pre-training
Instruction fine-tuning, In-context learning
In-context learning, Instruction fine-tuning
Pre-training, Instruction fine-tuning
Correct

2.

Question 2
Fine-tuning a model on a single task can improve model performance specifically on that task; however, it can also degrade the performance of other tasks as a side effect. This phenomenon is known as:
1 / 1 point
Model toxicity
Catastrophic forgetting
Instruction bias
Catastrophic loss
Correct

3.

Question 3
Which evaluation metric below focuses on precision in matching generated output to the reference text and is used for text translation?
1 / 1 point
BLEU
HELM
ROUGE-2
ROUGE-1
Correct
BLEU focuses on precision and text translation while Rouge focuses on text summarization.

4.

Question 4

Which of the following statements about multi-task finetuning is correct? Select all that apply:

0.5 / 1 point

Multi-task finetuning can help prevent catastrophic forgetting.

Correct

Correct! However, remember that to prevent catastrophic forgetting it is important to fine-tune on multiple tasks with a lot of data.

FLAN-T5 was trained with multi-task finetuning.

Performing multi-task finetuning may lead to slower inference.

This should not be selected

This process only changes the parameters of the network, and therefore does not impact inference speed.

Multi-task finetuning requires separate models for each task being performed.

5.

Question 5

"Smaller LLMs can struggle with one-shot and few-shot inference:" Is this true or false?

1 / 1 point

True

False

Correct

Even when you include a couple of examples, smaller models might still struggle to learn the new task through examples.

6.

Question 6

Which of the following are Parameter Efficient Fine-Tuning (PEFT) methods? Select all that apply.

1 / 1 point

Subtractive

Selective

Correct

Selective methods is a category of PEFT that fine-tunes a subset of the original LLM parameters. It uses different approaches to identify which parameters to update.

Additive

Correct

Additive methods freeze all of the original LLM weights and introduce new model components to fine-tune to a specific task.

Reparameterization

Correct

Reparameterization methods create a new low-rank transformation of the original network weights to train, decreasing the trainable parameter count while still working with high-dimensional matrices. LoRa is a common technique in this category.

7.

Question 7

Which of the following best describes how LoRA works?

0 / 1 point

LoRA decomposes weights into two smaller rank matrices and trains those instead of the full model weights.

LoRA trains a smaller, distilled version of the pre-trained LLM to reduce model size

LoRA continues the original pre-training objective on new data to update the weights of the original model.

LoRA freezes all weights in the original model layers and introduces new components which are trained on new data.

Incorrect

LoRA doesn’t involve training a smaller, distilled version of the pre-trained LLM

8.

Question 8

What is a soft prompt in the context of LLMs (Large Language Models)?

1 / 1 point

A set of trainable tokens that are added to a prompt and whose values are updated during additional training to improve performance on specific tasks.

A strict and explicit input text that serves as a starting point for the model's generation.

A technique to limit the creativity of the model and enforce specific output patterns.

A method to control the model's behavior by adjusting the learning rate during training.

Correct

A soft prompt refers to aa set of trainable tokens that are added to a prompt. Unlike the tokens that represent language, these tokens can take on any value within the embedding space. The token values may not be interpretable by humans, but are located in the embedding space close to words related to the language prompt or task to be completed.

9.

Question 9

"Prompt Tuning is a technique used to adjust all hyperparameters of a language model." Is this true or false?

1 / 1 point

True

False

Correct

Prompt Tuning focuses on optimizing the prompts given to the model using trainable tokens that don’t correspond directly to human language. The number of tokens you choose to train, however, would be a hyperparameter of your training process.

10.

Question 10

"PEFT methods can reduce the memory needed for fine-tuning dramatically, sometimes to just 12-20% of the memory needed for full fine-tuning." Is this true or false?

1 / 1 point

True

False

Correct

By training a smaller number parameters, whether through selecting a subset of model layers to train, adding new, small components to the model architecture, or through the inclusion of soft prompts, the amount of memory needed for training is reduced compared to full fine-tuning.