1. When you increase the *temperature*, what does that do to the completion?

- ○ Makes it more conservative
- ◉ Makes it less conservative
- ○ Makes it longer
- ○ Makes no difference

✓ Correct

2. Which of the following statements accurately describes the *max_tokens* property in the OpenAI API?

- ○ It specifies the maximum number of tokens a user is allowed to use in their prompt.
- ○ It is used to increase a model's processing time.
- ○ It represents the total number of tokens in the model's vocabulary.
- ◉ It sets the maximum number of tokens that can be generated in the completion.

✓ Correct

3. Which of the following is TRUE about the OpenAI API?

- ○ It brings us up to the minute information from the Internet.
- ○ It make our apps more secure.
- ○ It provides human generated content.
- ◉ It allows us to access OpenAI models from within our applications.

✓ Correct

4. Which of the following correctly defines a 'prompt' in the OpenAI API?

- ○ A reminder to do something.
- ○ An error message.
- ◉ The input text provided to guide the AI's response.
- ○ The reply we get from the OpenAI API when we have asked it for a completion.

✓ Correct

5. What is meant by the term 'Zero-shot' when we are creating prompts?

- ○ A mathematical prompt.
- ◉ A prompt consisting of a request but no examples.
- ○ A prompt with one example.
- ○ A prompt with multiple examples.

✓ **Correct**

6. What is meant by the term 'Few-shot' when we are creating prompts?

- ○ A prompt which generates multiple completions.
- ◉ A prompt with one or more examples demonstrating what we are looking for in a completion.
- ○ Using two prompts in one API call.
- ○ A prompt with a maximum of two examples.

✓ **Correct**

7. What information are we REQUIRED to give the OpenAI API to generate an image (do not include parameters which have a default value).

- ○ 1. The size of the image.
  2. A *number* parameter.
  3. A *response_format* such as *url*.
- ○ 1. A description of the image.
  2. The size of the image.
  3. A *number* parameter.
- ◉ 1. A description of the image.
- ○ 1. A description of the image.
  2. The size of the image.
  3. A *number* parameter.
  4. A *response_format* parameter.

✓ Correct

8. Which of the following is true about the *createCompletions* endpoint?

- ○ It only works with the *text-davinci-003* model.
- ○ Setting *temperature* to below 1 will cause an error.
- ○ No parameters are required
- ◉ The *model* property must be used.

✓ Correct

1. Which of these best describes an OpenAI model's behaviour?

- ○ There's no way of making it capable of logically continuing a conversation.
- ○ It can remember the previous prompts it has been given.
- ◉ It has no memory of previous prompts, so all of the context of a conversation must be included in every prompt.
- ○ It can remember just the previous prompt.

✓ Correct

2. What should happen when *presence_penalty* is increased?

- ◉ The model should talk about new topics more often.
- ○ The model will use the same words more frequently.
- ○ The model should repeat itself more.
- ○ The model should supply you with a list of answers.

✓ Correct

3. Which one of the following is true?

- ○ It's important to set a LOW *max_tokens* property when working with the *createChatCompletions* endpoint
- ○ When using the *createChatCompletions* endpoint, you send a prompt to the *GPT-3.5-turbo*/*GPT-4* models in exactly the same way as when using the *text-davinci-003* model on the *createCompletions* endpoint.
- ○ When working with the *createChatCompletions* endpoint, the *messages* property must contain and array of strings.
- ◉ When working with the *createChatCompletions* endpoint, the *messages* property must contain and array of objects.

✓ **Correct**

4. The *frequency_penalty* setting gives us some control over...

- ◉ how likely the model is to use the same words and phrases in a completion.
- ○ how frequently the model discusses new topics.
- ○ how frequently we can make requests to the API.
- ○ how creative and unusual the completions are.

✓ **Correct**

5. Which of the following is true about the array we send to the API in the *messages* property.

- (●) The first object in the array should contain instructions telling the model how we want it to behave.
- ( ) The first object in the array should contain a key/value pair where the key is *content* and the value is the request we are making to the API.
- ( ) The first object in the array should hold the previous completion that came back from the API.
- ( ) Each object in the array should have a key/value pair where the key is *role* and the value is *user*.

✓ **Correct**

6. Which of these best describes a firebase 'snapshot'.

- ( ) A backup of data from the database which is created once in a 24 hour period.
- (●) A copy of all of the database data as it exists at that moment.
- ( ) An array of objects holding the current data from the database and a history of changes made to that data.
- ( ) An object which mirrors the database's data. When you mutate the object, you change the database data.

✓ **Correct**

1. Fine-tuning allows us to do which of the following?

- (●) Train an OpenAI model on our own data.
- ( ) Make a model which we can share with other users.
- ( ) Edit OpenAI's models and add our own data.
- ( ) Make a model which never hallucinates.

✓ Correct

2. Which of the following is FALSE about the OpenAI CLI?

- ( ) It provides us with tools to prepare and submit training data.
- (●) It only fine-tunes the GPT-4 model.
- ( ) It can handle large data sets with thousands of data points.
- ( ) It allows us to choose which base model we want to fine-tune.

✓ Correct

3. Which of the following is TRUE about the characters we include in a 'stop sequence'

- ○ They separate the different parts of a prompt, e.g. the instruction and example completion.
- ◉ They will never be included in a completion.
- ○ They can only be alpha-numeric characters.
- ○ A completion will finish after they have been included.

✓ Correct

4. What does the *n_epoch* parameter do in OpenAI?

- ○ It dictates the speed at which the AI model processes and returns results, allowing users to balance performance and accuracy.
- ○ It controls whether the output will be more creative or more conservative. More cycles make it more creative.
- ◉ It sets the number of times the training data will be cycled through when fine-tuning a model. More cycles tends to improve performance.
- ○ It adjusts how long a completion will be relative to prompt length.

✓ Correct

5. We can help a model better understand our prompts by doing which one of the following.

- ○ Keeping our prompts short and general.
- ◉ Using separators in our prompts such as '->' or '###'.
- ○ Using a 'stop sequence' in our prompts.
- ○ Making sure a prompt is always on one line.

✓ Correct

6. Which of the following is TRUE.

- ○ API keys should be returned by a serverless function.
- ○ Secret API keys can be stored on the front end without being compromised.
- ◉ An API key stored in an *env variable* is visible on the front end.
- ○ A serverless function can be accessed from any domain by default.

✓ Correct

7. Which of the following is FALSE about data when fine-tuning.

- ○ All data should be checked by a human before fine-tuning.
- ◉ In terms of accuracy, there is no performance advantage in using a larger data set.
- ○ The data set should consist of example prompt and completion pairs.
- ○ If the data is not formatted correctly in JSONL when sent for fine-tuning, it will cause the fine-tuning to fail.

✓ Correct

1. What is the goal of Natural Language Processing (NLP) in the field of AI?

- ○ To create video games
- ○ To simulate weather conditions
- ◉ To read, decipher, understand, and make sense of human language in a valuable way
- ○ To predict stock prices

Correct
Correct. This is the correct sequence of steps in the ChatGPT process.

2. You're tracking the operational cost of your Chef ChatGPT program, and realize that the number of tokens used has increased. How will this impact the cost of operating Chef ChatGPT?

- ○ More tokens lead to a lower cost.
- ○ The number of tokens doesn't affect the cost.
- ○ The cost is fixed, regardless of the number of tokens.
- ◉ More tokens lead to a higher cost.

Correct
Correct. The cost of using the ChatGPT API is directly related to the number of tokens used in an API call. The more tokens you use, the more it costs.

3. You're tasked with designing a chatbot aimed at providing factual information about scientific topics. How would you adjust the 'temperature' parameter to ensure that the responses are accurate and less likely to be random?

- ○ Set the temperature to a high value
- ○ The temperature parameter is not relevant for this task
- ◉ Set the temperature to a low value
- ○ Base models are less accurate in generating text.

Correct
Correct. A lower temperature value makes the output from the model more deterministic and less random, which would be appropriate for a chatbot providing factual information on scientific topics.

4. You are designing a chatbot that needs to stick closely to the provided conversation context and should avoid introducing unrelated ideas. Which parameter would be helpful to adjust, and how should it be adjusted?

- ○ Increase the temperature.
- ○ Decrease the temperature.
- ◉ Increase the presence penalty.
- ○ Decrease the presence penalty.

Correct
Correct. The presence penalty parameter controls how much the model is discouraged from introducing

5. In what way does the 'Chat' feature in GPT Playground differ from the 'Complete' mode when it comes to providing context?

- ○ 'Chat' does not allow the injection of context, while 'Complete' does.
- ◉ 'Chat' allows for the injection of context through user and system roles, while 'Complete' uses a singular prompt.
- ○ 'Chat' provides context automatically, while 'Complete' requires manual input.
- ○ 'Chat' and 'Complete' use the same method of providing context.

Correct
Correct. The 'Chat' feature allows you to provide a system-level role and a user-level message, enabling more specific context to guide the AI's responses. The 'Complete' mode, on the other hand, generates text based on a single prompt without explicitly defined roles.

1. What are the potential benefits and applications of integrating ChatGPT with Excel, and how can it contribute to streamlining data-related tasks, analysis, and decision-making processes in various domains? (Select all that apply)

- [x] Help with writing and validating complex formulas

  Correct
  ChatGPT, with its natural language processing capabilities, can simplify the creation and validation of complex formulas. By interpreting user queries in natural language, it can translate them into the correct Excel formulas, saving time and reducing errors.

- [x] Assistance with standardizing and cleansing data

  Correct
  ChatGPT can indeed assist in data standardization and cleansing by automating these often time-consuming tasks. By identifying inconsistencies and abnormalities in data, it can help maintain a clean and reliable dataset.

- [x] Conducting sentiment analysis in order to classify data

  Correct
  With its text analysis abilities, ChatGPT can be utilized for sentiment analysis tasks. It can help classify data based on sentiments, which is particularly useful in domains such as customer feedback analysis, social media monitoring, and market research.

- [ ] ChatGPT can automatically access and manipulate data within Excel spreadsheets without input from the user

2. In what ways can ChatGPT be utilized to address and resolve issues related to inconsistent date formats in Excel, and how does its functionality contribute to enhancing data organization and accuracy within the spreadsheet?

- ○ Automatically converts dates to a specific format
- ◉ By fixing inconsistent date formats within the ChatGPT console
- ○ Provides pre-built Excel templates for standardizing dates
- ○ Suggests alternative data entry methods for dates

✓ Correct
Inconsistent date formatting can be pasted into the ChatGPT console where it can standardized in a common format. This data can then be outputted in CSV format then copied back into the Excel workbook.

3. How does the use of formulas in ChatGPT contribute to the data cleansing process, and what specific advantages does it offer in terms of ensuring data accuracy, consistency, and improved data-driven decision-making in Excel?

- ○ By generating Excel macros for automating data entry
- ○ By providing visual representations of data in Excel
- ◉ By generating formulas that can standardize, validate, and catogorize data
- ○ By directly editing and modifying Excel files

✓ Correct
ChatGPT can generate formulas for data cleansing tasks, such as standardizing inconsistent data, validating data against certain conditions, and correcting invalid data. These capabilities contribute to ensuring data accuracy and consistency, which are crucial for reliable data analysis and decision-making.

4. In what ways can ChatGPT be utilized to facilitate the extraction of relevant text data from URLs, and what are the advantages of using its capabilities compared to traditional methods or existing web scraping tools when integrating this information into Excel for analysis and processing?

- (●) By extracting text elements from URLs within the ChatGPT console like the root URL or page titles
- ( ) By offering an AI-powered web scraping tool integrated with Excel - Incorrect
- ( ) By suggesting manual methods for extracting text from URLs
- ( ) By generating complex formulas to extract desired text segments

✓ Correct
ChatGPT can extract specific text elements from URLs, such as the root URL or blog titles. These can then be copied from ChatGPT into your Excel worksheet.

5. What are the limitations of ChatGPT when processing URLs, and what specific types of data extraction tasks might present challenges or be beyond the current capabilities of the system when attempting to extract information from webpages associated with the URLs?

- ( ) ChatGPT cannot understand URLs
- ( ) ChatGPT cannot generate Excel formulas
- (●) ChatGPT cannot extract images or videos from URLs
- ( ) ChatGPT cannot interact with Excel

✓ Correct
ChatGPT is a text-based model and does not handle non-textual data such as images or videos. Therefore, it cannot extract these types of content from URLs. While it can understand and provide information about the URL structure, it cannot access or interpret the specific content hosted at that URL if it's not in

6. To support decision-making processes and data-driven insights, what types of data can be effectively analyzed and classified based on sentiment analysis using ChatGPT?

- ( ) Financial data like stock prices
- ( ) Geographical data like coordinates
- ( ) Numerical data like sales figures
- (●) Textual data like customer reviews

✓ Correct
Customer reviews are textual and often contain subjective opinions and sentiments, making them suitable for sentiment analysis. ChatGPT, with its natural language understanding capabilities, can be utilized to analyze this type of data, extracting sentiments and helping in the classification of reviews based on sentiment.

7. How should data analysts and professionals approach the ethical considerations and transparency when utilizing generated data? (Select all that apply)

- [x] By being transparent about the sources of data and methods of data generation

  Correct
  Transparency is a crucial aspect of ethical data usage. Data analysts should be open about where the data came from and how it was generated or processed. This can help stakeholders understand the data's context and validity, and it can also foster trust.

- [x] By ensuring data privacy and complying with relevant data protection regulations

  Correct
  Respecting data privacy is a key part of ethical data use. This involves protecting personal data from unauthorized access, complying with data protection laws, and implementing appropriate security measures.

- [ ] By creating a false narrative to make the data more compelling
- [x] By respecting the rights of data subjects, especially when dealing with sensitive data

  Correct
  Data subjects' rights must always be respected. This is particularly important when dealing with sensitive data, such as health information, financial details, or personally identifiable information. Consent should be obtained where necessary and data should be anonymized or pseudonymized to protect individuals' identities.

- [ ] By ignoring ethical considerations as long as the data serves its purpose

8. How does the strategic classification of data into groupings contribute to the overall data management and analysis process? (Select all that apply)

- [x] It improves data visualization

  Correct
  Strategic classification of data into groupings indeed enhances data visualization. It allows for the creation of graphs, charts, and other visual aids that help in understanding the patterns, trends, and relationships within the data. This is a fundamental part of the data management and analysis process.

- [ ] It improves data security and privacy
- [ ] It increases the speed of Excel calculations
- [x] It enhances the ability to sort and filter data

  Correct
  When data is classified into strategic groupings, it allows for more efficient sorting and filtering. This can help users to quickly locate and analyze specific subsets of data, enhancing the overall data management and analysis process.

1. Which language controls the styling of a web page?

- ○ HTML
- ○ JavaScript
- ◉ CSS
- ○ XML

Correct

2. What is the below code?

Class

- ○ It's a JavaScript class
- ○ It's an example of invalid code
- ○ It's a CSS class
- ◉ It's a HTML paragraph element

Correct

3. What do we mean by the so-called "variables" in JavaScript?

- ◉ Variables are like containers that we use to store information
- ○ It refers to all the variable ways you can write and run JavaScript code

Correct

4. What do we mean when we say that ChatGPT is "generative AI"?

- ○ It means that its underlying AI model has been generated by another AI model
- ◉ It means that the AI is able to generate new and original content
- ○ It means that it is an innovation we only expect to see once in a generation

Correct

5. What does it mean to deploy a website?

- ○ It's the act of debugging and fixing errors in your website
- ○ It's the act of running your website locally on your computer
- ◉ It's the act of making it available to users via the world wide web

✓ Correct

6. What is GitHub?

- ◉ It's a platform for version control and source code management
- ○ It's the underlying technology that lets you run code on your local machine

✓ Correct

---

1. Interacting with Large Language Models (LLMs) differs from traditional machine learning models. Working with LLMs involves natural language input, known as a _____, resulting in output from the Large Language Model, known as the _____ .

   Choose the answer that correctly fill in the blanks.

   - ○ tunable request, completion
   - ○ prediction request, prediction response
   - ○ prompt, fine-tuned LLM
   - ◉ prompt, completion

   ✓ **Correct**
   The input for working with LLMs is referred to as the prompt and the output from the LLM is referred to as the completion.

2. Large Language Models (LLMs) are capable of performing multiple tasks supporting a variety of use cases. Which of the following tasks supports the use case of converting code comments into executable code?

   - ○ Text summarization
   - ◉ Translation
   - ○ Information Retrieval
   - ○ Invoke actions from text

   ✓ **Correct**
   Translation focuses on converting languages, including coding languages so in this case the task focuses on translating code comments into executable code

3. What is the *self-attention* that powers the transformer architecture?

- ○ A measure of how well a model can understand and generate human-like language.
- ◉ A mechanism that allows a model to focus on different parts of the input sequence during computation.
- ○ A technique used to improve the generalization capabilities of a model by training it on diverse datasets.
- ○ The ability of the transformer to analyze its own performance and make adjustments accordingly.

Correct
Self-attention is a key component in models like Transformers, where it enables the model to attend to different words in the input sequence to capture their relationships and dependencies.

4. Which of the following stages are part of the generative AI model lifecycle mentioned in the course? (Select all that apply)

- ☐ Performing regularization
- ☑ Deploying the model into the infrastructure and integrating it with the application.

Correct
Once we have a model performing to our needs, we can deploy it into the infrastructure and integrate it with the application.

- ☑ Selecting a candidate model and potentially pre-training a custom model.

Correct
Selecting a candidate model and potentially pre-training a custom model are important stages in the generative AI model lifecycle.

- ☑ Manipulating the model to align with specific project needs.

Correct
It is likely we will have to manipulate the model in some way to align it with the specific needs of the project.

- ☑ Defining the problem and identifying relevant datasets.

Correct
It is crucial to define the problem being solved and identify relevant datasets instrumental to the project.

5. "RNNs are better than Transformers for generative AI Tasks."

Is this true or false?

- ○ True
- ◉ False

Correct
While RNNs can be used for generative AI tasks, they struggle with compute and memory, making it hard to keep context in longer texts. The transformers architecture is more parallelizable and its dynamic attention mechanism helps to capture long-range dependencies in the input.

6. Which transformer-based model architecture has the objective of guessing a masked token based on the previous sequence of tokens by building bidirectional representations of the input sequence.

- ◉ Autoencoder
- ○ Autoregressive
- ○ Sequence-to-sequence

Correct
Autoencoder models are pre-trained using masked language modeling. They use randomly masked tokens in the input sequence and the pretraining objective is to predict the masked tokens to reconstruct the original sentence.

7. Which transformer-based model architecture is well-suited to the task of text translation?

- ◉ Sequence-to-sequence
- ○ Autoencoder
- ○ Autoregressive

Correct
Sequence-to-sequence models use both the encoder and decoders in the transformer-based architecture making them best suited for tasks such as translation, text summarization, and question answering. In the Transformers video, Mike explains it in more detail.

8. Do we always need to increase the model size to improve its performance?

- ○ True
- ◉ False

Correct
Recent trends show that we can build better LLMs without necessarily increasing model size year by year. Models like LLaMa and BloombergGPT have demonstrated the possibility of reducing model size while keeping great performance.

9. Scaling laws for pre-training large language models consider several aspects to maximize performance of a model within a set of constraints and available scaling choices. Select all alternatives that should be considered for scaling when performing model pre-training?

- [ ] Batch size: Number of samples per iteration
- [x] Dataset size: Number of tokens

> **Correct**
> The size of the pre-training data is an important factor to consider when scaling with compute constraints. This is because the size of the dataset directly affects the computational requirements during pre-training, and having a larger dataset generally leads to improved model performance.

- [x] Model size: Number of parameters

> **Correct**
> The size of the model in terms of number of parameters is a key scaling choice to consider with compute constraints because the number of parameters directly impacts the compute needs required during pre-training.

- [x] Compute budget: Compute constraints

> **Correct**
> The compute budget plays a crucial role in scaling during pre-training. When faced with a limited compute budget, we may need to impose restrictions on either the model size or the dataset size.

10. "You can combine data parallelism with model parallelism to train LLMs."

Is this true or false?

- (•) True
- ( ) False

> **Correct**
> Combining data parallelism with pipeline parallelism is known as 2D parallelism.We can achieve 3D parallelism by combining data parallelism with both pipeline parallelism and tensor parallelism simultaneously.

1. 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.

- ○ In-context learning, Instruction fine-tuning
- ○ Prompt engineering, Pre-training
- ○ Pre-training, Instruction fine-tuning
- ◉ Instruction fine-tuning, In-context learning

✓ **Correct**

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:

- ○ Model toxicity
- ○ Catastrophic loss
- ◉ Catastrophic forgetting
- ○ Instruction bias

✓ **Correct**

3. Which evaluation metric below focuses on precision in matching generated output to the reference text and is used for text translation?

- (●) BLEU
- ( ) ROUGE-1
- ( ) ROUGE-2
- ( ) HELM

Correct
BLEU focuses on precision and text translation while Rouge focuses on text summarization.

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

- [x] 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.

- [ ] Multi-task finetuning requires separate models for each task being performed.
- [x] FLAN-T5 was trained with multi-task finetuning.

Correct
The FLAN family of models have been trained with multi-task instruction finetuning.

- [ ] Performing multi-task finetuning may lead to slower inference.

5. "Smaller LLMs can struggle with one-shot and few-shot inference:"

Is this true or false?

- (●) True
- ( ) False

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

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

- [ ] Subtractive
- [x] 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.

- [x] 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.

- [x] Additive

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

. Which of the following best describes how LoRA works?

- ○ LoRA freezes all weights in the original model layers and introduces new components which are trained on new data.
- ○ LoRA continues the original pre-training objective on new data to update the weights of the original model.
- ◉ 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

Correct
LoRA represents large weight matrices as two smaller, rank decomposition matrices, and trains those instead of the full weights. The product of these smaller matrices is then added to the original weights for inference.

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

- ◉ 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

9. "Prompt Tuning is a technique used to adjust all hyperparameters of a language model."

Is this true or false?

- ○ 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. "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?

- ◉ 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.

1. Which of the following are true in regards to Constitutional AI? Select all that apply.

- ☑ Red Teaming is the process of eliciting undesirable responses by interacting with a model.

Correct
Red Teaming is the process of eliciting undesirable responses, and it is necessary for the first stage of Constitutional AI, as we need to fine-tune the model with those "red team" prompts and revised answers.

- ☑ In Constitutional AI, we train a model to choose between different responses.

Correct
This is the role of the preference model, that will learn what responses are preferred following the constitutional principles.

- ☐ For constitutional AI, it is necessary to provide human feedback to guide the revisions.
- ☑ To obtain revised answers for possible harmful prompts, we need to go through a Critique and Revision process.

Correct
This process is necessary for Constitutional AI, and its done by asking the model to critique and revise the elicited harmful answers.

2. What does the "Proximal" in Proximal Policy Optimization refer to?

- ○ The algorithm's ability to handle proximal policies.
- ○ The algorithm's proximity to the optimal policy
- ○ The use of a proximal gradient descent algorithm
- ◉ The constraint that limits the distance between the new and old policy

Correct
The "Proximal" in Proximal Policy Optimization refers to the constraint that limits the distance between the new and old policy, which prevents the agent from taking large steps in the policy space that could lead to catastrophic changes in behavior.

3. "You can use an algorithm other than Proximal Policy Optimization to update the model weights during RLHF."

Is this true or false?

- ◉ True
- ○ False

Correct
For instance, you can use an algorithm called Q-Learning. PPO is the most popular for RLHF because it balances complexity and performance, but RLHF is an ongoing field of research and this preference may change in the future as new techniques are developed.

4. In reinforcement learning, particularly with the Proximal Policy Optimization (PPO) algorithm, what is the role of KL-Divergence? Select all that apply.

- [x] KL divergence measures the difference between two probability distributions.

  Correct
  KL-Divergence is a mathematical measure of the difference between two probability distributions.

- [ ] KL divergence is used to train the reward model by scoring the difference of the new completions from the original human-labeled ones.

- [x] KL divergence is used to enforce a constraint that limits the extent of LLM weight updates.

  Correct
  PPO used KL divergence to introduce a constraint that limits the changes to the LLM weights to prevent dramatic changes from the original model.

- [ ] KL divergence encourages large updates to the LLM weights to increase differences from the original model.

5. Fill in the blanks: When fine-tuning a large language model with human feedback, the action that the agent (in this case the LLM) carries out is _____ and the action space is the _____.

- ( ) Generating the next token, the context window
- ( ) Calculating the probability distribution, the LLM model weights.
- ( ) Processing the prompt, context window.
- (•) Generating the next token, vocabulary of all tokens.

  Correct
  The LLM generates tokens based on the text in the context window, and the probability of all tokens in the

6. How does Retrieval Augmented Generation (RAG) enhance generation-based models?

- ○ By optimizing model architecture to generate factual completions.
- ○ By increasing the training data size.
- ○ By applying reinforcement learning techniques to augment completions.
- ◉ By making external knowledge available to the model

Correct
The retriever component retrieves relevant information from an external corpus or knowledge base, which is then used by the model to generate more informed and contextually relevant responses. This incorporation of external knowledge enhances the quality and relevance of the generated content.

7. How can incorporating information retrieval techniques improve your LLM application? Select all that apply.

- ☑ Overcome Knowledge Cut-offs

Correct
Retrieving data from external sources enables the model to incorporate information it did not see during training when generating text.

- ☐ Faster training speed when compared to traditional models
- ☑ Improve relevance and accuracy of responses

Correct
By retrieving from curated, verified datasets you can improve the relevance and accuracy of the model's completions.

---

8. What are correct definitions of Program-aided Language (PAL) models? Select all that apply. 1

- ☐ Models that enable automatic translation of programming languages to human languages.
- ☑ Models that offload computational tasks to other programs.

Correct
It offloads these tasks to a runtime symbolic interpreter such as a python function, which reduces the workload for the LLM and improves accuracy as symbolic interpreters tend to be more precise with computational tasks.

- ☐ Models that integrate language translation and coding functionalities.
- ☑ Models that assist programmers in writing code through natural language interfaces.

Correct
Program-aided Language (PAL) models are designed to assist programmers in writing code using natural language interfaces. They aim to facilitate the coding process by providing support and guidance through human-like interactions.

9. Which of the following best describes the primary focus of ReAct?

- ○ Exploring action plan generation in LLMs.
- ○ Studying the separate topics of reasoning and acting in LLMs.
- ○ Investigating reasoning abilities in LLMs through chain-of-thought prompting.
- ◉ Enhancing language understanding and decision making in LLMs.

Correct
The ReAct framework aims to enhance both language understanding and decision-making capabilities in LLMs by combining reasoning and acting components.

10. What is the main purpose of the LangChain framework?

- ○ To evaluate the LLM's completions and provide fast prototyping and deployment capabilities.
- ○ To connect with external APIs and datasets and offload computational tasks.
- ○ To provide prompt templates, agents, and memory components for working with LLMs.
- ◉ To chain together different components and create advanced use cases around LLMs, such as chatbots, Generative Question-Answering (GQA), and summarization.

Correct
The LangChain framework is built around LLMs and allows the chaining of various components to create more advanced applications for LLMs. It supports use cases like chatbots, Generative Question-Answering (GQA), and summarization.

1. Setting the temperature property to a higher number makes the model…

- (●) provide answers which are more creative and daring.
- ( ) use common words more frequently.
- ( ) provide answers with more context.
- ( ) use more tokens and give longer answers.

✓ Correct

2. When working with the OpenAI API, the objects that are held in the messages array each have a role property, but what values can the role property hold?

- ( ) assistant, AI, human
- ( ) user, admin, human
- ( ) system, assistant, admin
- (●) system, user, assistant

✓ Correct

3. Which of the following is true about tokens and credit when using the OpenAI API?

- ○ Words provided by the API use more tokens than words included in the prompt.
- ○ Creating more complex text requires more tokens regardless of text length.
- ◉ Both the words in the prompt you send to the API and the data you get back count towards the total number of tokens used.
- ○ Only the data you get back from the API counts towards the total number of tokens used.

✓ **Correct**

4. What does the frequency_penalty setting do?

- ○ It makes sure the model will only use a given word once.
- ○ It penalizes a word which has already appeared in the provided text.
- ◉ It gives the developer some control over the likelihood of words and phrases being repeated.
- ○ It makes the model use the same phrases as much as possible.

✓ **Correct**

5. When asking the OpenAI API to create images, which property or properties are required.

- ○ prompt