MCQ

1. What are large language models (LLMs)?
    a. An LLM is an advanced natural language processing framework that uses linguistic algorithms to generate sophisticated conversational agents.
    b. An LLM is a state-of-the-art computer vision system that excels in recognizing and analyzing intricate patterns and features in images and videos.
    c. An LLM is a type of artificial intelligence (AI) that can generate human-quality text. LLMs are trained on massive datasets of text and code, and they can be used for many tasks, such as writing, translating, and coding.
    d. An LLM is an artificial neural network architecture optimized for training large-scale reinforcement learning agents capable of mastering complex tasks in robotics.

2. What are some of the challenges of using LLMs?
    a. They can be used to generate harmful content.
    b. They can be biased.
    c. They can be expensive to train.
    d. All of the above.

3. Which of the following is a common lightweight fine-tuning technique for large language models (LLMs)?
    a. Full fine-tuning.
    b. Low-Rank Adaptation (LoRA).
    c. Randomized Weight Averaging.
    d. Pre-training from scratch.

4. Parameter-Efficient Fine-Tuning (PEFT) includes?
    a. Prompt tuning.
    b. Prefix tuning.
    c. Adapter tuning.
    d. **All of the above.**

5. How does LoRA reduce the number of trainable parameters?
    a. By freezing all model weights and adding small low-rank matrices.
    b. By pruning neurons in the model.
    c. By training only the embedding layers.
    d. By using sparse matrix multiplication.

6. What privacy risks do LMs pose?
    a. Larger models are more susceptible to memorizing.

b. Membership inference attacks.
c. Memorized data can be extracted.
d. All of the above.

7. Why would a model generate a false statement?
    a. The model hasn't learned distribution well enough (non-imitative falsehood).
    b. The model's training objectives incentivizes a false answer (imitative falsehood).
    c. All of the above.
    d. None of the above.

8. What is one of the primary privacy concerns when using LLMs?
    a. The inability to generate human-like text.
    b. The risk of memorizing and reproducing sensitive information from training data.
    c. The lack of multilingual support in models.
    d. Inconsistent grammar in responses.

9. Which method can help mitigate privacy risks in LLMs?
    a. Differential Privacy.
    b. Federated Learning.
    c. Redacting sensitive data from training datasets.
    d. All of the above.

10. Which of the following is an application of multimodal LLMs?
    a. Image captioning
    b. Text-to-image generation
    c. Visual question answering (VQA)
    d. All of the above

True/False

1. Lightweight fine-tuning techniques are particularly effective when the downstream task has limited labeled data. **T**
2. Large language models can inadvertently reinforce biases present in their training data. **T**
3. GPT uses ROPE embeddings similar to Llama. **F**
4. In the attention is all you need paper, the softmax s core is scaled by **1/√dk** because for large values of **dk**, the dot product grow large in magnitude, pushing the softmax function into regions where it has extremely small gradients. **T**
5. Few-shot and zero-shot learning capabilities in LLMs mean they can perform tasks without additional task-specific training data.**T**
6. Knowledge distillation can be used to create smaller versions of LLMs without significant loss of performance. **T**
7. Attention heads in a transformer model allow the LLM to focus on different parts of the input text simultaneously. **T**
8. LLMs are fundamentally incapable of multimodal tasks, such as handling text and images together. **F**
9. In-context to learning completely eliminates the need for any fine-tuning of LLMs for any task. **F**
10. LLMs can be used for many tasks, including ethical decision making. **F**

Subjective - [4*15 Marks=60 Marks]

1. Questions on the Llama series of papers.
    a. Explain how the BPE tokenizer used in Llama1 handles digits and unknown UTF-8 characters. [2 marks]
    b. Explain three changes in architecture proposed in Llama1 compared to the original transformer paper. [3 marks]
    c. Explain the motivation and workings of Ghost Attention proposed in Lama2. [3+2 marks]
    d. Draw and explain an illustration of the compositional approach used by Llama 3 for adding multimodal (text, vision, and audio) capabilities. [5 marks]

    Solution:
    a. Tokenizer. We tokenize the data with the bytepair encoding (BPE) algorithm using the implementation from SentencePiece (Kudo and Richardson, 2018). Notably, we split all numbers into individual digits, and fallback to bytes to decompose unknown UTF-8 characters.
    b. Section 2.2 in Llama1 paper.
    c. Section 3.3 in Llama2 paper.
    d. Figure 28 in Llama3 paper. And Section 2 in Llama3 paper.

2. a. Describe in detail the 4 steps involved in aligning an LLM to user preferences, starting from training on human-labeled data to ensuring the model remains stable and generalizable. For each step mention **what is the purpose** of the step and **what method or objective** is used to achieve this step?(3+3+3+3)
    b. Why is RL using Human Feedback (RLHF) good ?Give any 3 reasons. (3 Marks)

    Solution: Slide Adaptation 32-53

    2 a. Four Steps in Aligning LLM to User Preferences:

1. Supervised Fine-Tuning (SFT):
    - Purpose: Provide the model with basic task-following capabilities.
    - Method: Fine-tune the model on human-labeled demonstration data to mimic desired behavior.
2. Reinforcement Learning from Human Feedback (RLHF):
    - Purpose: Incorporate user preferences into the model.
    - Method:Collect comparison data and Train a reward model (RM) using ranked outputs annotated by labellers from the SFT model and use it to provide reward signals.

r_θ : The **Reward Model** we are trying to optimize. x : The prompt.
y_w: The better completion. y_r : The worse completion.

$$\text{loss}(\theta) = -\frac{1}{\binom{K}{2}} E_{(x,y_w,y_l)\sim D}\left[\log\left(\sigma\left(r_{\theta^*}(x,y_w) - r_\theta(x,y_l)\right)\right)\right]$$

Reward on better completion

Reward on worse completion

3. Proximal Policy Optimization (PPO):
    - Purpose: Fine-tune the SFT model with reward feedback from the RM while preventing deviations from learned behavior.
    - Method:
        i. SFT model further fine-tuned using RL with the RM providing the reward signal
        ii. A KL-loss is provided to prevent the PPO model from deviating far from SFT

4. PPO-ptx (PPO with Pre Training Auxiliary Objective):
    - Purpose: Ensure the model remains stable and generalizable across tasks.
    - Method: Add auxiliary language modeling objectives on pretraining data during PPO to avoid performance degradation on general NLP tasks.
    - **Solution:** Add a auxiliary LM objective on the pretraining data. Call this variant **PPO-ptx**

$$\text{objective}(\phi) = E_{(x,y)\sim D_{\pi_\phi^{\text{RL}}}}\left[r_\theta(x,y) - \beta\log\left(\pi_\phi^{\text{RL}}(y\mid x)/\pi^{\text{SFT}}(y\mid x)\right)\right] + \gamma E_{x\sim D_{\text{pretrain}}}\left[\log(\pi_\phi^{\text{RL}}(x))\right]$$

2 b. Any 3 of below

- Reward is a more nuanced training signal than autoregressive loss.
- The RM “critiques” actual completions generated from the model itself, whereas SFT training does not use model generations, since it is completely offline.
- The RM more directly captures the notion of “preference”.
- The RM is more data efficient
- SFT may teach the model to lie, while RLHF teaches the model to answer or abstain.

3. Multiple steps are being taken to reduce Harms and Bias in LLM’s, in the same context try to answer the following:
    a. The paper "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models" introduces a dataset to evaluate biases in language models.
        i. What is the primary objective of the CrowS-Pairs dataset, and how does it differ from other bias evaluation datasets like StereoSet? [2 Marks]

ii. Explain the methodology used for constructing the CrowS-Pairs dataset. [2 Marks]
iii. What metric is proposed in the paper to evaluate bias in masked language models? [3 Marks]

b. The paper "StereoSet: Measuring Stereotypical Bias in Pretrained Language Models" addresses the challenge of evaluating social biases in language models.
i. Explain the design of the Context Association Test (CAT) and how it is applied in both intra-sentence and inter-sentence contexts. [3 Marks]
ii. Formulate the Idealized CAT (iCAT) score discussed in paper. [3 Marks]
iii. Based on the experimental results, write the observed relationship between language modeling ability (LMS) and stereotype score (SS) in the evaluated models and what challenge it exhibits. [2 Marks].

A a i. The CrowS-Pairs dataset's primary objective is to **measure the degree to which large masked language models (MLMs) exhibit social biases, particularly against historically disadvantaged groups in the United State**s.It differs from StereoSet in that CrowS-Pairs focuses exclusively on stereotypes **concerning two contrasting groups (disadvantaged vs. advantaged) with minimally edited sentences**. CrowS-Pairs achieves higher diversity and validation reliability compared to StereoSet.

a ii.The methodology involves the following steps:

- **Data Collection:** Crowdworkers from the United States, recruited through Amazon Mechanical Turk, write pairs of sentences. The first sentence either demonstrates or violates a stereotype about a disadvantaged group, while the second sentence is minimally edited to reference an advantaged group.
- **Validation:** Each example undergoes validation by five annotators to ensure the sentences represent stereotypes or anti-stereotypes, are minimally distant, and align with one of the nine bias categories. A majority vote decides inclusion in the dataset.

A iii. The proposed metric evaluates bias by measuring the likelihood of a model preferring stereotyping sentences over less stereotyping sentences. This is achieved through **pseudo-log-likelihood MLM scoring**, where the likelihood of unmodified tokens in a sentence is conditioned on the modified tokens.

B i. The Context Association Test (CAT) evaluates bias and language modeling ability by:

- **Intrasentence CAT:** A fill-in-the-blank style task where three attributes (stereotype, anti-stereotype, meaningless) are provided to complete a sentence. The model's choice reflects its bias and language understanding.

- **Intersentence CAT:** A discourse-level task where the model selects an appropriate continuation for a sentence from three options (stereotype, anti-stereotype, meaningless).

B ii.The Idealized CAT (iCAT) score combines the Language Modeling Score (LMS) and the Stereotype Score (SS) into a single metric.

- Formula: iCAT=LMS×min(SS,100−SS)50\text{iCAT} = \text{LMS} \times \frac{\min(\text{SS}, 100 - \text{SS})}{50}iCAT=LMS×50min(SS,100−SS).
- It ensures equal importance is given to minimizing bias and maintaining language modeling performance, where a perfect model has an iCAT of 100.

B iii.Experimental results reveal:

- A strong correlation between LMS and SS.
- Larger models with better language modeling abilities (e.g., GPT-2 Large) tend to exhibit higher stereotypical biases.
- This correlation highlights a challenge: improving language models' performance without exacerbating biases

Q4. Large Language Models (LLMs) are revolutionizing the field of Human-Computer Interaction (HCI) and beyond. Answer the following questions:

a. Define the following concept briefly: [2 Marks]
   i. Design Thinking
   ii. Design Process
b. Compare Design Thinking and the Design Process by highlighting
   i. one key difference [1 Mark]
   ii. an example from the Airbnb case study. [2 Marks]
c. Define the six stages of the Design Thinking framework, providing one example for each stage to illustrate its application. [3 Marks]
d. What is the Double Diamond Model in the Design Process? Briefly explain its key phases. [2 Marks]
e. Highlight one key application of LLMs in the domain of accessibility and explain how it addresses the needs of diverse users. [2 Marks]
f. In context of the emergence of new capabilities in LLMs as model size increases. Provide one example of such a capability that has significant real-world application. [3 Marks]

4. a

- Definitions: (1 mark each)

Design Thinking: 1 mark.

A user-centered, iterative process for problem-solving.

Focuses on empathy, creativity, and experimentation.

Design Process: 1 mark.

It is a series of structured steps followed by designers to identify problems, generate ideas, create solutions, and evaluate results.

4 b

- Key Difference [1 mark for underlined text]

  Design Thinking is a mindset and problem-solving approach that emphasizes empathy, creativity, and iteration. It is a user-centered strategy that involves understanding user needs, defining problems, ideating solutions, prototyping, and testing, often encouraging a non-linear and flexible path to innovation.

  Design Process is a structured sequence of steps followed to create a product or solution, often including stages like research, ideation, prototyping, testing, and implementation. It is a linear or iterative methodology aimed at achieving a specific design outcome.

- Airbnb case [1 mark for underlined text]

  In its early days, Airbnb struggled to gain traction. The founders applied design thinking by staying in their own listings to empathize with users and understand their pain points. They discovered that poor-quality photos of rental spaces were a major issue.

  Outcome: They redefined the problem by focusing on how better visuals could increase bookings. This led to an idea: they went door-to-door in New York City, photographing hosts' homes themselves. This empathetic, iterative, and creative approach was a pivotal moment that helped Airbnb take off.

  Design Thinking drove Airbnb's initial problem discovery and innovative idea by emphasizing empathy and understanding the user's experience.

  [1 mark for underlined text]

  After understanding user needs and validating the initial idea (high-quality photos), Airbnb moved to a more structured **design process**. They researched their target market, designed a scalable system to support hosts, built prototypes of the platform, gathered feedback, and iteratively improved features like booking, payments, and host management in a systematic manner.

Outcome: This structured approach allowed Airbnb to develop a reliable platform, scaling it into a successful, global marketplace.

In summary, Design Thinking is about discovering and defining the problem creatively, while the Design Process is about systematically developing and executing the solution.

4 c. (0.5 for each stage)

a. Empathize: close to purpose/goal (0.25 marks). No marks for methods/steps/tools/types. Example (0.25 marks)
b. Define: close to purpose/goal (0.25 marks). No marks for methods/steps/tools/types. Example (0.25 marks)
c. Ideate: close to purpose/goal (0.25 marks). No marks for methods/steps/tools/types. Example (0.25 marks)
d. Prototype: close to purpose/goal (0.25 marks). No marks for methods/steps/tools/types. Example (0.25 marks)
e. Test: close to purpose/goal (0.25 marks). No marks for methods/steps/tools/types. Example (0.25 marks)
f. Implement: close to purpose/goal (0.25 marks). No marks for methods/steps/tools/types. Example (0.25 marks)

4d. Double Diamond Design Process:

[1 mark for underlined text or making a labelled diagram]

Developed by the UK Design Council, it's a visual representation of the design process,

[0.5 mark for explaining why it is called Double Diamond]

In all creative processes a number of possible ideas are created ('divergent thinking') before refining and narrowing down to the best idea ('convergent thinking'), and this can be represented by a diamond shape. But the Double Diamond indicates that this happens twice – once to confirm the problem definition and once to create the solution. source.

[0.25 mark for writing phase names, 0.25 mark for correctly explaining all of them]

divided into two key phases:

1. Discover & Define (Problem Space): Marks only for purpose
2. Develop & Deliver (Solution Space): Marks only for purpose

Goal: Balance divergent (exploring) and convergent (focusing) thinking. No marks for goal.

Divided into four distinct phases (source)

1. Discover: Designers try to look at the world in a fresh way, notice new things and gather insights.
2. Define: Designers try to make sense of all the possibilities identified in the Discover phase. The goal here is to develop a clear creative brief that frames the fundamental design challenge.
3. Develop: Solutions or concepts are created, prototyped, tested and iterated.
4. Delivery: the resulting project is finalised, produced and launched.

4 e. How LLMs Contribute to Accessibility:

[if any one of the following is written then 1 mark]

● LLMs enable applications to understand and respond to voice commands, offering hands-free navigation for users with physical disabilities.

● LLMs can generate alternative text descriptions for images, videos, and other media, making content accessible to visually impaired users.

[else if any one of the following is written then 2 mark]

● Personalized Accessibility Features: LLMs can adapt the content, offering alternative formats (text, audio, visuals) based on individual needs.

● Multilingual Support: LLMs enable real-time translation and localized content, enhancing social inclusion for users from different linguistic backgrounds.

[else if anything sensible 0.5-1.0 marks]

[else zero mark]

4 f. Refer to this GIF

a) If the mentioned capability is either in GIF or has evidence in literature. 1 mark
b) Mention of LLM name which has performed well on it. 1 mark.
c) Real world application of that capability. 1 mark.