Introduction to Large Language Models
Assignment- 12

Number of questions: 10 **Total mark: 10 X 1 = 10**

QUESTION 1: [1 mark]

Which statements correctly characterize “bias” in the context of LLMs?

1. Bias can generate objectionable or stereotypical views in model outputs.
2. Bias is always intentionally introduced by malicious data curators.
3. Bias can cause harmful real-world impacts such as reinforcing discrimination.
4. Bias only affects low-resource languages; high-resource languages are unaffected.

a. 1 and 2

b. 1 and 3

c. 2 and 4

d. 1, 3, and 4

Correct Answer: b

Explanation:

- **(1) True:** Model outputs can reflect harmful stereotypes if training data or modelling procedures contain biases.
- **(3) True:** Biased outputs may perpetuate discrimination or unfair treatment in real-world contexts.

- Statements (2) and (4) are not necessarily correct:

    - **(2) False:** Bias in data is often unintentional, reflecting existing societal or historical imbalances.
    - **(4) False:** Bias can affect any language; high-resource languages are not inherently immune.

QUESTION 2: [1 mark]

The Stereotype Score (ss) refers to:

a. The frequency with which a language model rejects biased associations.
b. The measure of how often a model’s predictions are meaningless as opposed to meaningful.
c. A ratio of positive sentiment to negative sentiment in model outputs.
d. The proportion of examples in which a model chooses a stereotypical association over an anti-stereotypical one.

Correct Answer: d

Explanation:

- **Stereotype Score (ss)** is a metric that measures how frequently the model picks a stereotypical continuation or association **instead of** a non-stereotypical or anti-stereotypical one.

- Essentially, it's a proportion (or fraction) of test items for which the model output aligns with the stereotype.

QUESTION 3: [1 mark]

Which of the following are prominent sources of bias in LLMs?

1. Improper selection of training data leading to skewed distributions.
2. Reliance on older datasets causing "temporal bias."
3. Overemphasis on low-resource languages causing "linguistic inversion."
4. Unequal focus on high-resource languages resulting in "cultural bias."

a. 1 and 2 only

b. 2 and 3 only

c. 1, 2, and 4

d. 1, 3, and 4

Correct Answer: c

Explanation:

1. **Improper selection of training data** (true) can lead to some groups or topics being over-represented, causing bias.
2. **Reliance on older datasets** (true) can introduce out-of-date or "temporal bias" that doesn't reflect current social norms or language usage.
3. "Overemphasis on low-resource languages" is not commonly described as "linguistic inversion"; typically the bias is the opposite — under-representation of low-resource languages.
4. **Unequal focus on high-resource languages** (true) can lead to cultural biases and poor performance or misrepresentations of underrepresented cultures.

QUESTION 4: [1 mark]

In the context of bias mitigation based on adversarial triggers, which best describes the goal of prepending specially chosen tokens to prompts?

a. To directly fine-tune the model parameters to remove bias
b. To override all prior knowledge in a model, effectively "resetting" it
c. To exploit the model's distributional patterns, thereby neutralizing or flipping biased associations in generated text
d. To randomly shuffle the tokens so that the model becomes more robust

Correct Answer: c

Explanation:

- **Adversarial triggers** are carefully crafted token sequences that, when prepended to the prompt, steer the model's output in a certain direction (e.g., reducing bias or toxicity). They work within the model's learned distribution rather than overriding its knowledge.
- They do *not* retrain the model; they exploit patterns in the existing parameters to mitigate biased outcomes.

QUESTION 5: [1 mark]

Which of the following best describes the "*regard"* metric?

a. It is a measure of how well a model can explain its internal decision process.
b. It is a measurement of a model's perplexity on demographically sensitive text.
c. It is the proportion of times a model self-corrects discriminatory language.
d. It is a classification label reflecting the attitude towards a demographic group in the generated text.

Correct Answer: d

Explanation:

- **Regard** is typically measured by classifying the *tone* of text toward a demographic group (e.g., "positive," "negative," or "neutral" regard).
- It's used to assess whether certain demographics consistently receive negative or disrespectful language.

QUESTION 6: [1 mark]
Which of the following steps compose the approach for improving response safety via in-context learning?

a. Retrieving safety demonstrations *similar* to the user query.

b. Fine-tuning the model with additional labeled data after generation.
c. Providing retrieved demonstrations as examples in the prompt to guide the model's response generation.
d. Sampling multiple outputs from LLMs and choosing the majority opinion.

Correct Answer: a, c

Explanation:

- One strategy for safe or polite generation with large language models is to retrieve "safety demonstrations" from a database of safe examples. Then you include these examples in the prompt to the LLM, showing it how to respond safely.

- Fine-tuning (b) is a different technique, not part of the described in-context learning approach.

- Majority vote (d) is also not typically a method described under "improving response safety via in-context learning."

QUESTION 7: [1 mark]

Which statement(s) is/are correct about how high-resource (HRL) vs. low-resource languages (LRL) affect model training?

a. LRLs typically have higher performance metrics due to smaller population sizes.
b. HRLs get more data, so the model might overfit to HRL cultural perspectives.
c. LRLs are often under-represented, leading to potential underestimation of their cultural nuances.
d. The dominance of HRLs can cause a reinforcing cycle that perpetuates imbalance.

Correct Answers: b, c, d

Explanation:

- (b) **True:** If the model sees far more data in certain HRLs, it might be overly biased or "overfit" to those languages' norms and perspectives.
- (c) **True:** LRLs often lack extensive corpora, so the model learns fewer details about these languages, risking lower performance and cultural misrepresentations.
- (d) **True:** The more a model focuses on HRLs, the more beneficial it appears to be for those languages, attracting further data, thus perpetuating imbalance.
- (a) is not correct: LRLs typically have lower performance metrics due to insufficient training data, not higher.

QUESTION 8: [1 mark]

The "Responsible LLM" concept is stated to address:

a. Only the bias in LLMs
b. A set of concerns including explainability, fairness, robustness, and security
c. Balancing training costs with carbon footprint
d. Implementation of purely rule-based safety filters

Correct Answer: b

Explanation:

- **Responsible LLM** research focuses on a broad range of ethical, social, and technical concerns:
    - Fairness & bias mitigation
    - Explainability & transparency
    - Robustness to adversarial inputs
    - Security & safe deployment

QUESTION 9: [1 mark]

Within the StereoSet framework, the ***icat*** metric specifically refers to:

a. The ratio of anti-stereotypical associations to neutral associations
b. The percentage of times a model refuses to generate content deemed hateful
c. A measure of domain coverage across different demographic groups
d. A balanced metric capturing both a model's language modelling ability and the tendency to avoid stereotypical bias

Correct Answer: d

Explanation:

- In the **StereoSet** framework, *icat* is designed to measure how well the model balances contextual accuracy (i.e., good language modelling) and reduced stereotyping.
- It's a combined metric that looks at correctness in typical language modelling tasks while also penalizing stereotypical responses.

QUESTION 10: [1 mark]

Bias due to improper selection of training data typically arises in LLMs when:

a. Data are selected exclusively from curated, balanced sources with equal representation
b. The language model sees only real-time social media feeds without any historical texts
c. The training corpus over-represents some topics or groups, creating a skewed distribution

d. All data are automatically filtered to remove any demographic markers

Correct Answer: c

Explanation:

- **Improper data selection** leads to over-representation of certain domains, topics, or demographic groups, causing the learned model to be skewed.

- Balanced data curation and filtering are actually methods to *reduce* bias. If data come only from certain communities or perspectives, the model lacks balanced coverage, and biases surface.
