X





# Week 12 : Assignment 12

**The due date for submitting this assignment has passed.**

**Due on 2025-04-16, 23:59 IST.**

## Assignment submitted on 2025-04-09, 21:56 IST

1) Which statements correctly characterize “bias” in the context of LLMs? ***1 point***

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.

- 1 and 2
- 1 and 3
- 2 and 4
- 1, 3, and 4

Yes, the answer is correct.
Score: 1

Accepted Answers:
*1 and 3*

2) The Stereotype Score (ss) refers to: ***1 point***

- The frequency with which a language model rejects biased associations.
- The measure of how often a model’s predictions are meaningless as opposed to meaningful.
- A ratio of positive sentiment to negative sentiment in model outputs.

- The proportion of examples in which a model chooses a stereotypical association over an anti-stereotypical one.

Yes, the answer is correct.
Score: 1
Accepted Answers:
*The proportion of examples in which a model chooses a stereotypical association over an anti-stereotypical one.*

3) Which of the following are prominent sources of bias in LLMs? ***1 point***

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.”

- 1 and 2 only
- 2 and 3 only
- 1, 2, and 4
- 1, 3, and 4

Yes, the answer is correct.
Score: 1
Accepted Answers:
*1, 2, and 4*

4) In the context of bias mitigation based on adversarial triggers, which best describes the goal of prepending specially chosen tokens to prompts? ***1 point***

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

Yes, the answer is correct.
Score: 1
Accepted Answers:
*To exploit the model’s distributional patterns, thereby neutralizing or flipping biased associations in generated text*

5) Which of the following best describes the “regard” metric? ***1 point***

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

Yes, the answer is correct.
Score: 1
Accepted Answers:
*It is a classification label reflecting the attitude towards a demographic group in the generated text.*

6) Which of the following steps compose the approach for improving response safety via in-context learning? ***1 point***

- [ ] Retrieving safety demonstrations similar to the user query.
- [ ] Fine-tuning the model with additional labeled data after generation.
- [x] Providing retrieved demonstrations as examples in the prompt to guide the model's response generation.
- [ ] Sampling multiple outputs from LLMs and choosing the majority opinion.

Partially Correct.
Score: 0.5
Accepted Answers:
*Retrieving safety demonstrations similar to the user query.*
*Providing retrieved demonstrations as examples in the prompt to guide the model's response generation.*

7) Which statement(s) is/are correct about how high-resource (HRL) vs. low-resource languages (LRL) affect model training? ***1 point***

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

Yes, the answer is correct.
Score: 1
Accepted Answers:
*HRLs get more data, so the model might overfit to HRL cultural perspectives.*
*LRLs are often under-represented, leading to potential underestimation of their cultural nuances.*
*The dominance of HRLs can cause a reinforcing cycle that perpetuates imbalance.*

8) The "Responsible LLM" concept is stated to address: ***1 point***

- ( ) Only the bias in LLMs
- (•) A set of concerns including explainability, fairness, robustness, and security
- ( ) Balancing training costs with carbon footprint
- ( ) Implementation of purely rule-based safety filters

Yes, the answer is correct.
Score: 1
Accepted Answers:
*A set of concerns including explainability, fairness, robustness, and security*

9) Within the StereoSet framework, the icat metric specifically refers to: ***1 point***

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

Yes, the answer is correct.
Score: 1

Accepted Answers:
*A balanced metric capturing both a model's language modelling ability and the tendency to avoid stereotypical bias*

10) Bias due to improper selection of training data typically arises in LLMs when: ***1 point***

- Data are selected exclusively from curated, balanced sources with equal representation
- The language model sees only real-time social media feeds without any historical texts
- The training corpus over-represents some topics or groups, creating a skewed distribution
- All data are automatically filtered to remove any demographic markers

Yes, the answer is correct.
Score: 1

Accepted Answers:
*The training corpus over-represents some topics or groups, creating a skewed distribution*