Assignment - 5

Sravya Mendem
700765926

Part A – Short Answer Questions

1) Positional Encoding concepts :

a) Transformers process tokens in parallel, so they lack an inherent sense of token order. Positional encodings inject information about the sequence order, enabling the model to differentiate "The cat sat on the mat" from "on the mat sat the cat".

b) Two key requirements :

* Unique and differentiable positions :- Each position should have a distinct representation.

* Facilitate relative/absolute position reasoning :- The model should easily compute relationships between positions (e.g:, distance between words).

c) * Unitary : M.M^T = I, meaning no loss of information when transforming vectors.

* Norm-preserving : The length (magnitude) of vectors remains unchanged after adding positional encodings, preventing distortion of embeddings.

2) Attention Mechanism :

a) The attention score is a similarity measure (often a dot product between Query and key vectors) that indicates how much one token should attend to another.

b) Apply the softmax function :

$$\text{Attention\_weights} = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)$$

c) Multiply the attention weights with the value matrix V :

$$\text{Context\_Vector} = \text{attention\_weights} \cdot V$$

The context vector is computed as the weighted sum of the value vectors, Where weights come from the attention distribution.

3) Multi-Head Attention :

a) Multiple attention heads allow the model to learn different types of relationships in parallel (e.g., syntax vs. semantics).

b) Splitting Q, K and V across subspaces lets each head focus on different representation aspects, improving expressiveness and capturing richer dependencies.

c) After attention, concatenation and a linear projection combine all head outputs into one unified representation, restoring dimensionality and mixing information across heads.

4) Ethical Foundations :

a) Ethics are moral principles that guide what is right or wrong based on reasoning and fairness.

* Laws are rules made by governments and may not always be ethical (e.g., unfair laws).

* Feelings are personal emotions that can change and may not reflect what is morally right,

Thus, ethics rely on reasoned judgement rather than legal or emotional responses.

b) Two classical ethical theories and AI example :

* Utilitarianism : focuses on consequences — an action is ethical if it Produces the greatest good for the greatest number.

Example: Approving an AI surveillance system if it increases public safety, even if some privacy is lost.

* Deontology : focuses on duties and moral rules — actions are right

if they follow moral principles, regardless of outcomes.
Example: Rejecting the same AI system if it violates individual privacy rights, even if it helps most people.

c) Why no single ethical theory "wins" in all contexts:
Because real-world situations involve conflicting values (e.g., safety vs. privacy, fairness vs. utility). Each theory emphasizes different moral priorities, so their guidance may differ depending on context. Therefore, ethical decision-making often requires balancing multiple perspectives rather than following one rule absolutely.

5) Types of AI Harms:

a) Definitions:

* Allocational harm: Occurs when AI systems unfairly distribute or deny oppurtunities, resources, or benefits to certain groups.
* Representational harm: Occurs when AI systems reinforce stereotypes, disrespect, or misrepresent certain identities or communities.

b) Examples:

* Allocational harm: A hiring algorithm that rejects qualified women or minority candidates due to biased training data.
* Representational harm: A translation model that always associates "doctor" with "he" and "nurse" with "she", reinforcing gender stereotypes.

c) Why representational harm is harder to measure: Representational harm affects social meaning, dignity, and cultural perception, which are qualitative and context-dependent, not easily expressed in numerical or measurable outcomes - unlike allocational harms such as job offers or credit approvals.

6) Sources of Dataset Bias:

a) Three reasons bias arises during data collection or annotation:

* Sampling bias: Data collected from limited sources that don't represent the full population.
* Annotator Bias: Human labelers apply their own cultural or personal assumptions.
* Historical Bias: Existing societal inequalities or stereotypes are embedded in the data itself.

b) Under-represented data or groups:

* Speakers of minority languages or dialects.
* People from marginalized communities or low-resource regions.
* Non-western cultures and contexts often have less online data available.

c) Bias amplification after preprocessing: Even after preprocessing, model training or fine-tuning can magnify existing imbalances through feedback loops or overfitting dominant patterns.

7) Safety, Security and Privacy:

a) Data poisoning: Data poisoning occurs when malicious or corrupted data is intentionally inserted into a training dataset.

* It manipulates a model's predictions by teaching it false or biased patterns, causing it to make incorrect or attacker-controlled outputs during deployment.

b) Ethical implications of model memorization: When models memorize and reproduce private, sensitive, or copyrighted text, it raises serious ethical issues such as:

* Privacy violations (leaking personal data) * Intellectual property (reusing copyrighted content) infringement
* Loss of trust in AI system's data handling and accountability.

c) Model Stealing: Model Stealing happens when attackers recreate or copy a proprietary model by repeatedly querying it and using the outputs to train their own replica. This threatens: * privacy: If stolen models expose sensitive training data. * Intellectual property: Because original research, designs and data investments are unlawfully copied or misused.