Instructions: Answer all questions. Show all work. Total: 76 points.

Name: ____________________________ Student ID: ____________________________

1. (1 point) The primary function of the attention mechanism in a transformer is to ensure that each word in an input sequence is processed independently, without considering the relationships with other words in the sequence. True or False? A. True B. False

   Solution: False. The attention mechanism allows the model to consider relationships between words in a sequence by calculating attention scores that weigh the importance of different words.

2. (1 point) In a transformer model, does the masked-attention mechanism allow each word in a sentence to directly attend to every other word, thereby capturing relationships within the sentence? A. True B. False

   Solution: False. In masked-attention, a token can only attend to previous tokens, not future ones.

3. (1 point) Consider two methods of representing words in a computer: One-Hot Encoding and TF-IDF. One-Hot Encoding treats each word as a unique, isolated entity, while TF-IDF represents words based on their context within a collection of documents.

   Which statement best describes the fundamental difference in how these methods capture word meaning?

   A. One-Hot Encoding captures semantic relationships by grouping words with similar meanings, while TF-IDF does not.

   B. One-Hot Encoding represents words as dense vectors, while TF-IDF uses sparse vectors.

   C. One-Hot Encoding does not consider the context of a word, while TF-IDF uses the context of words within documents to infer meaning.

   D. Both methods rely on the distributional hypothesis, but TF-IDF is more computationally efficient.

   Solution: C. One-hot encoding treats words as independent entities, while TF-IDF considers word frequency in documents to infer relationships.

4. (2 points) seq2seq models consist of two main parts: the ___**encoder**___ and the ___**decoder**___.

Solution: The encoder processes the input sequence, and the decoder generates the output sequence.

5. (2 points) In GPT models, the pre-training objective is **causal** language modeling, where the model is trained to predict the next token given all the previous tokens in a sequence, while BERT uses **masked** language modeling.

Solution: GPT uses causal language modeling (predicting next token), while BERT uses masked language modeling.

6. (1 point) In an RNN, the **hidden state** is the component responsible for storing and updating information from previous inputs, enabling the network to maintain a 'working memory' of the sequence.

Solution: The hidden state acts as memory that stores information about past inputs.

7. (1 point) In instruction-based prompting, the purpose of an **output indicator** is to specify the required format of the response, ensuring consistency in the output structure and format. It also helps the LLM generate responses in a specific format, rather than in an arbitrary format.

Solution: The output indicator specifies the desired format of the response (e.g., bullet points, JSON).

8. (1 point) T5's key innovation lies in framing most Natural Language Processing (NLP) tasks as a **text-to-text** problem. For example, translation, summarization, and classification can all be addressed using this unified approach.

Solution: T5 frames all NLP tasks as text-to-text problems.

9. (1 point) Which method in natural language processing relies on the principle that words appearing in similar contexts tend to have similar meanings? A. Naive Bayes Classifier B. TF-IDF C. Word2Vec D. Decision Trees

Solution: C. Word2Vec creates word vectors based on context, following the distributional hypothesis.

10. (1 point) True/False: According to the distributional hypothesis, words that frequently occur in similar contexts are likely to have similar meanings. A. True B. False

Solution: True

11. (1 point) Illustrate a concrete data manipulation that highlights a key limitation of the TF-IDF representation. Illustrate the limitation with an example.

Solution: TF-IDF treats words independently and misses semantic relationships. Example: ”The cat sat on the mat” and ”The mat sat on the cat” have different meanings but would have the same TF-IDF vectors.

12. (1 point) True or False: In many cases, a TF-IDF matrix cannot be approximated by a lower-rank matrix because the data is not sparse.

    A. True B. False

Solution: False. TF-IDF matrices are often low-rank because documents naturally cluster by topics.

13. (1 point) When PCA is applied to a TF-IDF matrix, what does the distance between words in the reduced-dimensional space represent?

    A. Similar document frequency but different term usage.
    B. Tendency to co-occur in documents.
    C. Similar grammatical roles, independent of document context.
    D. No meaningful information about their relationship.

Solution: B) Tendency to co-occur in documents.

14. (1 point) Which of the following best describes why word2vec places semantically related words close to each other in vector space? Options:

    A. To prevent any overlap in word representations
    B. Because similar words often appear in similar contexts and thus learn similar representations
    C. To simplify the interpretation of semantic relationships by users
    D. To reduce the complexity of the model architecture

Solution: B) Similar words appear in similar contexts, so they learn similar vector representations.

15. (1 point) In a word embedding model like word2vec, how would you use vector operations to find that ”king” is to ”queen” as ”man” is to another word?

Solution: vec(king) - vec(man) + vec(woman)   vec(queen)

16. (1 point) Which of the following shows a limitation of using PCA to visualize word embeddings such as countries and capitals?

    A. It increases the dimensionality, making patterns harder to see.

    B. It reduces dimensions but may distort relationships between words.

    C. It groups similar words without reducing dimensions.

    D. It adds color to help distinguish word types.

Solution: B) PCA reduces dimensions but may distort the original relationships between words.

17. (1 point) True or False: In a 2D semantic space defined by two axes (e.g., Sentiment and Strength), each word is represented as a point with coordinates corresponding to its scores along those axes. A. True B. False

Solution: True

18. (1 point) Which of the following best describes why a robust semantic axis is created using expanded pole words?

    A. To increase the computational complexity

    B. To reduce the influence of outliers on the axis

    C. To ensure stability and reliability by averaging similar word vectors

    D. To make the axis more specific to individual words

Solution: C. Using multiple similar words creates more stable axes by averaging vectors.

19. (1 point) A semantic axis is defined by selecting two pole words, such as “happy” and “unhappy”, and computing the direction vector between them. What does this direction vector represent?

    A. A line that connects unrelated words in the embedding space.

B. The average position of the two pole words in the embedding space.

C. A direction that captures the contrast in meaning between the two pole words.

D. A principal component that maximizes variance in the embedding space.

Solution: C. The direction vector captures the semantic contrast between the pole words.

20. (1 point) Consider the semantic axis shown below, constructed from the word embedding space using the direction from “unhappy” to “happy.” Words are projected onto this axis. The origin (0) is shown with a dashed line.

21. (1 point) Which of the following best describes a key advantage of using semantic axes over PCA when analyzing word embeddings?

A. Semantic axes reduce more variance in the data than PCA.

B. Semantic axes automatically discover hidden topics in the embedding space.

C. Semantic axes offer interpretable dimensions aligned with meaningful contrasts.

D. Semantic axes cluster words more accurately than PCA.

Solution: C. Semantic axes use interpretable poles (e.g., ”happy” vs. ”unhappy”), creating meaningful dimensions.

22. (1 point) Which of the following best explains why LSTMs outperform vanilla RNNs on long sequences?

A. LSTMs use more hidden layers

B. LSTMs can operate at a higher learning rate

C. LSTMs explicitly model long-term dependencies via gating mechanisms

D. LSTMs avoid the need for backpropagation through time

Solution: C. LSTMs use gates to control information flow, allowing them to retain information over long sequences.

23. (1 point) During training, a student notices that the validation loss of their LSTM model decreases initially but then starts increasing. Which of the following is the most plausible explanation?

A. The model has learned the data distribution perfectly

B. The model is underfitting the training data

C. The learning rate is decaying too fast

D. The model is overfitting due to lack of regularization

Solution: D. Rising validation loss after initial decline indicates overfitting.

24. (1 point) A model is trained to classify the sentiment of a movie review using an RNN. The review is long, and the classifier uses only the hidden state from the last time step as input. Why might this setup lead to suboptimal performance?

    A. It introduces vanishing gradient issues

    B. It discards information from earlier parts of the sequence

    C. It increases model complexity

    D. It prevents backpropagation through earlier time steps

Solution: B. Using only the final hidden state may lose important information from earlier in the sequence.

25. (1 point) The **forget** gate in an LSTM controls how much of the previous cell state should be retained.

26. (1 point) The **input** gate in an LSTM determines how much of the new candidate values should be added to the cell state.

27. (1 point) The **output** gate in an LSTM regulates how much of the cell state contributes to the hidden state.

28. (3 points) The **cell state** in an LSTM serves as the memory of the network, and its update is primarily governed by the **forget** and **input** gates.

29. (1 point) A researcher replaces ReLU with tanh as the activation function in an RNN and notices that gradients vanish during training. Explain why this happens. Hint: The derivative of $\tanh(x)$ is $1 - \tanh^2(x)$.

Solution: Tanh saturates for large inputs (outputs near ± 1), making its derivative very small. This causes gradients to vanish during backpropagation.

30. (1 point) Which of the following best defines a contextualized embedding?

    A. It assigns a fixed vector to each word regardless of context.

    B. It generates a vector representation for a word that varies based on its surrounding words.

    C. It relies on manually designed features to encode word meanings.

    D. It uses a lookup table of static vectors without any modification.

Solution: B. It generates a vector representation for a word that varies based on its surrounding words.

31. (1 point) How does ElMo generate contextualized embeddings?

Solution: ELMo runs words through a bidirectional LSTM language model, combining hidden states from both forward and backward directions to create context-aware word representations.

32. (1 point) Which statement best describes the role of the attention mechanism in seq2seq models?

    A. It compresses the entire input sequence into a single fixed-size context vector.

    B. It passes all hidden states from the encoder to the decoder, allowing the decoder to focus on relevant parts of the input during decoding.

    C. It replaces the decoder with a more efficient neural network architecture.

    D. It increases the size of the context vector to accommodate longer sequences.

Solution: B. Attention allows the decoder to focus on relevant parts of the input sequence during each decoding step.

33. (1 point) True or False: The sinusoidal position embedding adds sine waves of the same phase and frequency for all positions to helps maintain unique positional information across different dimensions. A. True B. False

Solution: False

34. Consider the task of using BERT embeddings to resolve polysemy in movie review sentiment classification. For example, the word “dark” might refer to a stylistically appealing film or to a gloomy, negative tone. Explain how the degree of contextualization changes as you move from lower to higher BERT layers, and describe how the self-attention mechanism helps differentiate the meaning of ambiguous words in this context.

Solution: Lower BERT layers capture basic syntax, while higher layers incorporate context-specific meaning. The self-attention mechanism allows each token to weigh all other tokens, helping disambiguate words like ”dark” based on surrounding context in movie reviews.

35. (1 point) True or False: In NLP, using PCA on BERT embeddings not only reduces dimensionality but also ensures that all important information is preserved. A. True B. False

Solution: False. PCA reduces dimensionality while preserving as much variance as possible, but some information is inevitably lost.

36. (1 point) In which of the following scenarios would Sentence-BERT be more advantageous over BERT? Explain your reasoning.

    A. When generating embeddings for individual words in a sentence.

    B. When focusing on tasks that require understanding the context of entire sentences.

    C. When training exclusively with unsupervised learning techniques.

    D. When aiming to achieve state-of-the-art performance across all NLP tasks.

Solution: B. Sentence-BERT is specifically designed for generating sentence-level embeddings.

37. (1 point) Which of the following best describes the difference between deterministic and stochastic sampling methods in text generation?

    A. Deterministic methods always choose the most likely next token, while stochastic methods introduce randomness by sampling from a probability distribution.

    B. Stochastic methods are faster than deterministic methods because they do not consider all possible tokens.

    C. Deterministic methods use temperature control to diversify output, whereas stochastic methods do not.

    D. Stochastic methods always produce higher quality text compared to deterministic methods.

Solution: A. Deterministic methods select the most likely token, while stochastic methods introduce randomness.

38. (1 point) Given a text generation task where maintaining output diversity is crucial, which search method would be more suitable and why?

    - Greedy Search explores multiple paths simultaneously, whereas Beam Search selects only one path at each step.
    - Both Greedy and Beam Search explore multiple paths simultaneously but differ in how they handle temperature control.
    - Beam Search explores different paths simultaneously, while Greedy Search selects the single most probable token at each step.
    - Greedy Search is deterministic and always produces high-quality outputs, whereas Beam Search introduces randomness for diversity.

Solution: C. Beam Search explores multiple paths simultaneously, while Greedy Search only follows one path.

39. (1 point) Which statement best describes the difference between Top-k Sampling and Nucleus (Top-p) Sampling in text generation?

    A. Top-k Sampling selects from a fixed number of most likely tokens, while Nucleus (Top-p) Sampling dynamically adjusts based on cumulative probability.

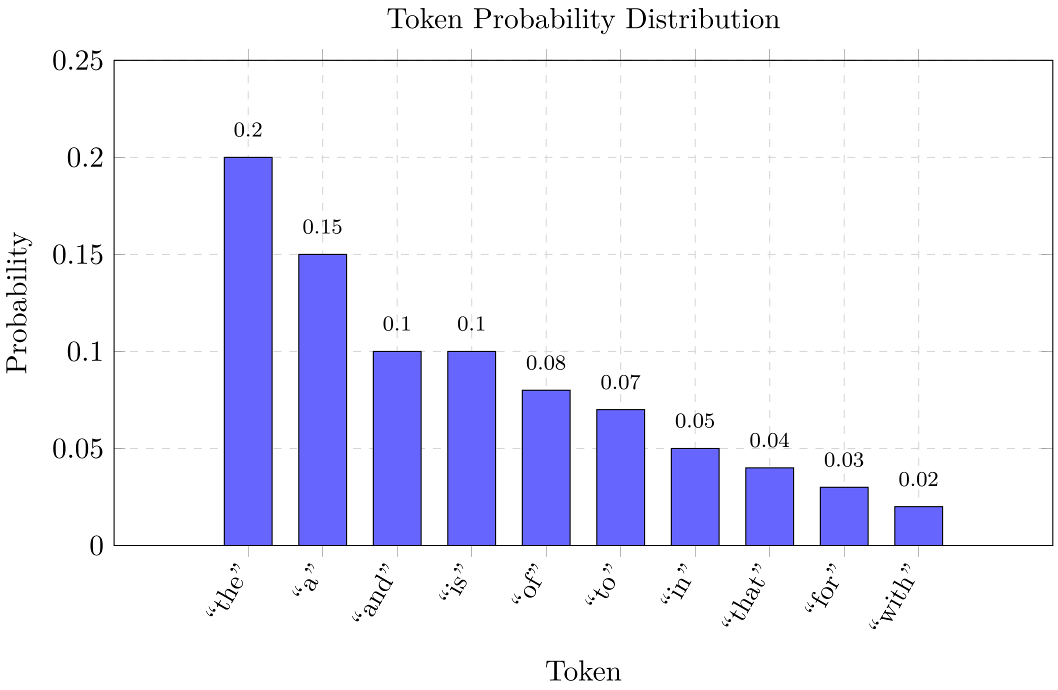
    B. Both methods select from a fixed set of the most probable tokens.

    C. Nucleus (Top-p) Sampling always produces more diverse outputs than Top-k Sampling.

    D. Top-k Sampling is deterministic, whereas Nucleus (Top-p) Sampling introduces randomness.

Solution: A. Top-k uses a fixed number of tokens, while Top-p uses a probability threshold.

40. (1 point) Consider the following hypothetical token set with their respective probabilities:



Identfy the set of tokens that can be sampled with top-p sampling with p = 0.5 and temperature = 1.0.

Solution: ”the”, ”a”, ”and”, ”is” (cumulative probability = 0.55)

41. (1 point) In a scenario where an AI model is struggling with alignment issues due to its reliance on historical data alone, how can Reinforcement Learning with Human Feedback (RLHF) be effectively utilized to improve the model’s performance?

    A. RLHF uses historical data as the primary source for training models but includes occasional human intervention.

    B. RLHF uses human evaluators to rate or rank model-generated responses, guiding iterative improvements through reinforcement learning.

    C. RLHF employs machine learning algorithms to generate feedback based on historical performance data without direct human input.

    D. RLHF relies solely on historical data for training without any current human input.

Solution: B. RLHF uses human evaluators to rate model outputs, guiding improvements through reinforcement learning.

42. (1 point) Identify two specific differences of GPT-2 and T5 that affect their ability to follow explicit task instructions, and provide a concrete example for each.

Solution: 1. T5 formats tasks as text-to-text with embedded instructions (e.g., ”translate English to French: How are you?”), while GPT-2 treats instructions as context for prediction.

2. T5 is trained on multiple instruction-based tasks, so it can switch between operations like summarization when prompted with ”summarize: [text]”, while GPT-2 may not consistently follow such directives.

43. (1 point) Identify and explain how natural language instruction following differs between T5 and the Flan T5. Provide a concrete example demonstrating the difference.

Solution: Flan T5 is fine-tuned to handle diverse phrasings of instructions. For example, whether given ”translate to French: How are you?” or ”please translate this to French: How are you?”, Flan T5 consistently produces the correct translation.

44. (4 points) Consider the use of a simple recurrent neural network (RNN) for sentiment analysis of movie reviews. Answer the following:

    (a) (1 point) Explain how the RNN updates its hidden state with each incoming token. Describe the role of this hidden state in retaining information across the sequence.

    (b) (1 point) Describe how the RNN processes a movie review to determine its overall sentiment. Specify how individual tokens contribute to the final classification.

(c) (2 points) Discuss the limitation of the RNN. Explain its effect on the model's ability to capture dependencies from earlier parts of a review and how that impacts sentiment classification accuracy.

Solution:

- Each word combines with the previous hidden state to create an updated hidden state, forming a memory of the sequence.
- The RNN processes each token sequentially, updating the hidden state. The final hidden state captures sentiment features from all tokens and is used for classification.
- The vanishing gradient problem makes it difficult for RNNs to learn from tokens early in the sequence. For long reviews, important sentiment cues at the beginning might be forgotten.

45. (4 points) Consider an LSTM cell applied to a time-series forecasting task. Answer the following:

(a) (1 point) Explain the role of the forget gate within the LSTM cell. How does it modify the cell state based on the current input and previous hidden state, and how does this action help preserve gradient stability?

(b) (1 point) Outline the mechanism of the output gate. How does it determine which information from the cell state is transferred to the hidden state for subsequent processing?

(c) (2 points) Summarize how the coordination of these three gates enables the LSTM to capture long-term dependencies and effectively reduce the vanishing gradient problem compared to a simple RNN.

Solution:

1. The forget gate uses a sigmoid function to decide which parts of the previous cell state to keep or discard, helping maintain stable gradients over long sequences.
2. The output gate controls which parts of the cell state become the hidden state, using a sigmoid function to filter the cell state after tanh activation.
3. The gates work together to create additive updates rather than multiplicative ones, helping prevent vanishing gradients and allowing LSTMs to capture long-range dependencies better than simple RNNs.

46. (10 points) Explain how TF-IDF addresses the limitations of a simple word-document count matrix when representing words for text analysis. In your explanation, address the following points:

(a) (1 point) How does TF-IDF account for words like "the" and "in" that appear frequently across many documents?

(b) (1 point) How does TF-IDF give more weight to words that are unique to specific documents?

(c) (2 points) Provide an example scenario where TF-IDF would be particularly useful for analyzing a collection of documents, and explain why.

Solution: TF-IDF downweights common words like ”the” and ”in” by using the inverse document frequency (IDF) component. Words appearing in many documents get a low IDF score.

TF-IDF gives higher weight to unique, document-specific words by assigning them high IDF scores. Words appearing in few documents receive higher weights.

Example: In a collection of scientific papers, TF-IDF would help identify the key technical terms that distinguish each paper's topic. Common words like ”study” or ”analysis” would be downweighted, while specialized terms unique to specific papers would be highlighted, making it easier to categorize papers by subject matter.