Instructions: Answer all questions. Show all work. Total: 50 points.

Name: _____

1. (1 point) One represents words as vectors of zeros with a single one at the position corresponding to the word. This is known as **one-hot encoding**.

   Solution: One-hot encoding.

2. (1 point) "You shall know a word by the company it keeps" is a famous quote that explains the idea that word meaning can be derived from the context in which a word appears. This idea is known as the **distributional** hypothesis.

   Solution: The distributional hypothesis is a key principle suggesting that words appearing in similar contexts have similar meanings. By analyzing the surrounding words, the model can understand the relationship between words. This is a critical distinction because the distributional hypothesis forms the basis for techniques like TF-IDF and word embeddings, which aim to encode semantic relationships.

3. (1 point) In the context of training RNNs, the process of computing gradients back through time is known as **Backpropagation Through Time (BPTT)**.

   Solution: BPTT is the method used to compute gradients in RNNs, unfolding the network through time and applying backpropagation.

4. (1 point) In the context of training RNNs, the gradients become extremely small as they are backpropagated through many time steps. This is known as the **vanishing gradient** problem.

   Solution: The vanishing gradient problem occurs when the gradients become extremely small as they are backpropagated through many time steps, making it difficult for the network to update its weights effectively and learn long-term dependencies.

5. (3 points) The three main gates in an LSTM cell are the forget gate, the input gate, and the output gate. The ___**forget**___ gate decides what information to remove from the cell state, the ___**input**___ gate determines what new information to store in the cell state, and the ___**output**___ gate controls what parts of the cell state should be revealed as output.

Solution: The forget gate examines the current input and the previous hidden state to decide what information to remove from the cell state. The input gate works together with a candidate memory generator to decide what new information to store. The output gate controls what parts of the cell state should be revealed as output. These three gates work together to control the flow of information within the LSTM cell, allowing it to maintain long-term dependencies.

6. (1 point) __**Dropout**__ is a technique used to mitigate overfitting by randomly deactivating a portion of the neurons during the training process.

Solution: Dropout is a regularization technique used in neural networks, including LSTMs, to prevent overfitting. It randomly deactivates (sets to zero) a fraction of neurons during training. This forces the network to learn more robust features and reduces its reliance on specific neurons. This concept is discussed in the content, and the visual aid clarifies the impact of the regularization.

7. (1 point) In the ELMo architecture, the **bidirectional** LSTM processes text in both forward and backward directions to generate word representations.

Solution: Bidirectional LSTM processes text in both forward and backward directions. The forward LSTM processes the words from the first to the last, while the backward LSTM processes the words from the last to the first. The two LSTM outputs are then concatenated to form a single vector representation for each word.

8. (1 point) In the attention mechanism, the context vector is computed as a **weighted sum** of the encoder hidden states using the attention weights.

Solution: The context vector is calculated as a weighted sum of the encoder hidden states using the attention weights. This is the core functionality of the attention mechanism, allowing the model to focus on the relevant parts of the input sequence. The attention weights determine how much each hidden state contributes to the context vector, effectively enabling the model to 'pay attention' to the most important parts of the input when generating the output. Other options are not appropriate since they don't represent the method the context vector is computed.

9. (2 points) In the decoder transformer block, the **masked attention** mechanism prevents the model from accessing future tokens during training, while the **cross-attention** mechanism allows the decoder to attend to the encoder's output for tasks like translation.

Solution:

- Masked Multi-Head Attention: This mechanism is essential for the parallel training of the decoder. By masking future tokens during training, it ensures the model learns to generate output sequentially, preventing it from "peeking" at the future, and it can also act as regular attention mechanism during the inference. This is critical for ensuring the model learns to generate tokens in the correct order, which can be more efficient and prevent error accumulation problems.
- Cross-Attention: Cross-attention allows the decoder to access information from the encoder's output. It enables the decoder to focus on the relevant parts of the input sequence, which is essential for tasks like translation, where the decoder needs to understand the meaning of the input to generate the correct output. For instance, in translating "I love you" to "Je t'aime", cross-attention allows each French word to focus on related English words (e.g., "Je" attending to "I" and "t'aime" to "love").

10. (3 points) In the self-attention mechanism, the attention score is computed using the ___**query**___ vectors and ___**key**___ vectors, and then normalized using the softmax function. This score is then used to calculate a weighted sum of the ___**value**___ vectors to produce the contextualized vector.

Solution:

- Query, Key, and Value Vectors: These are the fundamental components of the self-attention mechanism. Each word in a sequence is transformed into these three vectors. The query vector is used to represent the word when it is searching for relevant information, the key vector is used to represent the word when it is being searched for, and the value vector contains the information to be retrieved. The attention score measures the relationship between query and key vectors. The value vectors are then weighted by the attention scores to produce the contextualized vector. The query and key vectors are used to compute the attention score by taking a dot product of them. The attention scores are then normalized using the softmax function to provide the weights. The weighted sum of the value vectors is then computed using these softmax weights, which leads to the contextualized vector.
- Contextualization: The purpose of the attention mechanism is to contextualize the word vectors, so that each word vector contains the context of the surrounding words. The concept is explained in the document by saying "The output of the attention mechanism is the *contextualized vector*, meaning that the vector for a word can vary depending on other words input to the attention module."

11. (1 point) Which of the following is a primary limitation of TF-IDF as a word embedding technique, as discussed in the context?

    A. It fails to account for the frequency of words within a document.

B. It does not capture the relationships between words that appear within the same document.

C. It is computationally too expensive for large datasets.

D. It overemphasizes the importance of common words like 'the' and 'a'.

Solution: B. TF-IDF focuses on the importance of a word to a document, not on the relationships between words in that document. This is explicitly mentioned in the text when discussing TF-IDF's limitations.

The other options are incorrect because:

- (A) TF-IDF *does* account for word frequency within a document through the Term Frequency (TF) component.
- (C) TF-IDF is not computationally too expensive; it's a relatively efficient technique.
- (D) TF-IDF *addresses* the overemphasis on common words through the Inverse Document Frequency (IDF) component, which downweights them.

12. (1 point) In the context of Word2Vec, the Skip-gram model focuses on local context windows. Which of the following best describes its primary function and how it relates to the overall goal of word embedding?

A. Skip-gram predicts the center word given the surrounding context words, and it aims to create a sparse matrix of word-document counts for each word.

B. Skip-gram predicts the surrounding context words given a center word, and it aims to learn dense vector representations of words by factorizing a pointwise mutual information matrix.

C. Skip-gram predicts the center word given the surrounding context words, and it is primarily used to categorize the input words, and create a sparse matrix of word-document counts for each word.

D. Skip-gram predicts the surrounding context words given a center word, and it primarily aims to create a vocabulary of words used in the text.

Solution: B. Skip-gram predicts the surrounding context words given a center word, and it aims to learn dense vector representations of words by factorizing a pointwise mutual information matrix.

The core functionality of the Skip-gram model is predicting context words given a center word. It correctly links the model's objective to learning dense vector representations, which is achieved through the factorization of the pointwise mutual information matrix. The other options are incorrect because they either misrepresent the model's prediction task (predicting the center word instead of the context words) or misconstrue its goals (creating a sparse matrix or just a vocabulary). The core idea is to use local context to generate word embeddings.

13. (1 point) In the context of Word2Vec, which statement best describes the primary difference between the CBOW (Continuous Bag of Words) and Skip-gram models?

    A. CBOW predicts the context words given a center word, while Skip-gram predicts the center word given the context words.

    B. Both CBOW and Skip-gram predict the center word given the context words, but they use different algorithms.

    C. CBOW predicts the center word given the context words, while Skip-gram predicts the context words given a center word.

    D. Skip-gram uses a bag-of-words approach, whereas CBOW considers the order of words in the context.

Solution: C. CBOW (Continuous Bag of Words) is designed to predict a target word (the center word) based on its surrounding context words. Skip-gram, conversely, attempts to predict the surrounding context words given a specific target word (the center word). The other options are incorrect. Option 1 reverses the roles of CBOW and Skip-gram. Option 2 is wrong because the prediction tasks are different. Option 4 confuses the general approach with the specifics of each model; both models, in their basic implementations, do not account for word order directly.

14. (1 point) In the forward pass of a standard RNN, what is the correct sequence of operations applied to the input data (x_t) and the previous hidden state (h_{t-1}) to produce the output (o_t) at a given time step?

    A. Concatenation of x_t and h_{t-1}, followed by a linear transformation to get the hidden state (h_t), then a tanh transformation of h_t to get o_t.

    B. Linear transformation of x_t and h_{t-1} separately, followed by a tanh transformation of the concatenation to get the hidden state (h_t), and finally, a linear transformation of h_t to get o_t.

    C. A linear transformation of x_t followed by a linear transformation of h_{t-1}, and then summing the results to generate o_t and h_t.

    D. The input (x_t) is directly transformed to produce o_t, and h_{t-1} is ignored.

Solution: A. As described in the content, the forward pass of an RNN involves concatenating the current input (x_t) and the previous hidden state (h_{t-1}) into a combined vector (v_t). This vector then undergoes a linear transformation (using weight matrix W_h and bias b_h) and is passed through the tanh activation function to produce the new hidden state (h_t). Finally, h_t is transformed using the output weight matrix W_o to generate the output (o_t). The other options are incorrect because they misrepresent the order or the specific transformations involved. The key steps, as shown in the Model section, are concatenation, transformation to h_t using tanh, and then output transformation.

15. (1 point) In a sequence-to-sequence model with the attention mechanism, what is the primary function of the context vector when decoding?

A. To store the entire input sequence without any compression.

B. To directly predict the output sequence without any further processing by the decoder.

C. To initialize the decoder's hidden state and provide a weighted summary of the input sequence.

D. To act as a fixed-size bottleneck preventing the decoder from accessing relevant information.

Solution: C. The context vector, in the presence of attention, provides a weighted sum of the encoder's hidden states to the decoder, which allows the decoder to focus on different parts of the input sequence when generating the output. This initialization helps the decoder understand the overall meaning of the input and generate an appropriate output sequence. The other options are incorrect because:

- The context vector does not store the entire sequence without compression.
- The context vector doesn't directly predict the output sequence. The decoder processes the context vector.
- In an attention-based model, the context vector doesn't act as a bottleneck, which is a limitation of the basic seq2seq without attention.

16. (1 point) In the BERT architecture, what is the primary function of the [CLS] token?

A. To mark the end of a sentence.

B. To represent masked words during the Masked Language Modeling pre-training.

C. To classify the input sentence as a whole.

D. To separate sentences in a pair.

Solution: C. The correct answer is 'To classify the input sentence as a whole.' As detailed in the 'Special tokens' section, the [CLS] token's final hidden state is used for sentence-level classification tasks.

- 'To mark the end of a sentence' is incorrect. The [SEP] token marks the end of a sentence.
- 'To represent masked words during the Masked Language Modeling (MLM) pre-training' is also incorrect. The [MASK] token is used during MLM.
- 'To separate sentences in a pair, similar to the [SEP] token' is incorrect. While [SEP] separates sentences, [CLS] is for overall sentence representation.

17. (1 point) Sentence-BERT is primarily designed to improve performance in which of the following NLP tasks, compared to a standard BERT model?

A. Generating human-quality text.

B. Classifying individual words in a sentence.

C. Performing tasks that require sentence-level understanding, such as semantic similarity.

D. Training the BERT model from scratch.

Solution: C. The correct answer is 'Performing tasks that require sentence-level understanding, such as semantic similarity.' Sentence-BERT is specifically designed for tasks that benefit from understanding the meaning of sentences, such as determining semantic similarity, paraphrase identification, and information retrieval. It modifies BERT to produce sentence embeddings that are semantically meaningful, making it ideal for tasks requiring sentence-level context.

Incorrect options:

- 'Generating human-quality text' is more closely associated with language generation models, such as GPT. While BERT can contribute to these tasks, it is not its primary focus.
- 'Classifying individual words in a sentence' is a task typically performed by models designed for word-level understanding, such as token classification tasks, not the primary strength of Sentence-BERT.
- 'Training the BERT model from scratch' is the initial step of BERT models, but it's not a task that Sentence-BERT focuses on. Sentence-BERT builds upon pre-trained BERT models, finetuning them for sentence-level tasks.

18. (1 point) Which of the following best describes the key architectural difference between Generative Pre-trained Transformer (GPT) and Bidirectional Encoder Representations from Transformers (BERT) in the context of language modeling?

A. GPT uses a bidirectional attention mechanism like BERT, allowing it to consider the entire input sequence simultaneously.

B. BERT uses a decoder-only transformer architecture, while GPT utilizes an encoder-decoder architecture.

C. GPT employs a causal (unidirectional) attention mechanism, enabling it to generate text autoregressively, while BERT uses bidirectional attention.

D. Both GPT and BERT use the same transformer architecture, with differences only in the pre-training objectives.

Solution: C C. Incorrect options are incorrect because:

- Option 1: GPT does not use bidirectional attention; it uses causal attention.

- Option 2: BERT uses an encoder, not a decoder-only architecture. GPT uses a decoder-only architecture.
- Option 4: While pre-training objectives differ, the architectural difference is fundamental; causal vs. bidirectional attention.

19. (1 point) Considering the autoregressive nature of GPT and its next-token prediction objective, which of the following statements best describes the primary advantage of nucleus sampling over greedy search in text generation?

    A. Nucleus sampling guarantees more grammatically correct sentences compared to greedy search.

    B. Nucleus sampling ensures that the generated text always adheres to the context window length, unlike greedy search.

    C. Nucleus sampling promotes greater diversity and reduces the likelihood of repetitive text compared to greedy search.

    D. Nucleus sampling is computationally faster than greedy search, enabling quicker text generation.

Solution: C. The core advantage of nucleus sampling, as highlighted in the text, is its ability to balance text quality and diversity. Greedy search, by always picking the most probable token, often leads to repetitive or predictable outputs. Nucleus sampling, by sampling from a dynamic set of tokens based on cumulative probability, introduces randomness and allows the model to explore a broader range of possibilities, thereby generating more varied and less repetitive text. The other options are incorrect because: Nucleus sampling does not guarantee grammatical correctness, ensure the length, or faster speed.

20. (1 point) Which of the following best describes the MAIN advantage of FLAN-T5 over T5 in the context of instruction-following?

    A. FLAN-T5 uses a larger dataset for training, leading to improved performance.

    B. FLAN-T5 is based on a different model architecture compared to T5, enabling it to better interpret instructions.

    C. FLAN-T5 is explicitly fine-tuned to understand and follow diverse, detailed natural language instructions, unlike T5.

    D. FLAN-T5 incorporates Reinforcement Learning with Human Feedback (RLHF), allowing for more nuanced responses.

Solution: C.

- The dataset size is not the main difference; both models are based on the T5 architecture.

- RLHF is a further development, but not the defining characteristic that separates FLAN-T5 from T5.

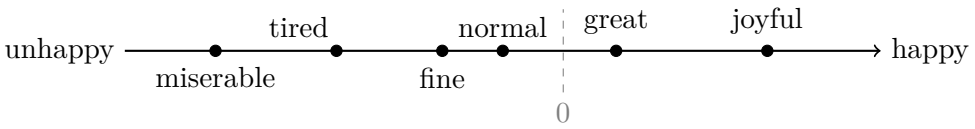
This question avoids common misconceptions by directly addressing the key differences in instruction handling between T5 and FLAN-T5, which is core to the content.

21. (2 points) In TF-IDF, the Inverse Document Frequency (IDF) component helps to downweight words that appear frequently across many documents. IDF for a term t is defined as: $\text{IDF}(t) = \log\left(\frac{N}{n_t}\right)$, where N and n_t are the total number of documents in the corpus and the number of documents containing the term t, respectively. Calculate the IDF value for the term "the" for the following corpus. Show your work.

- "The cat sat on the mat."
- "The dog chased the cat."
- "The mat was comfortable for the cat."
- "The dog barked at the mailman."

Solution: For the term "the" which appears in all four documents: IDF("the") = log(4/4) = log(1) = 0

22. (5 points) 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.



The word "fine" appears on the negative side of the axis, closer to "unhappy" than to "happy." The word "normal" is semantically neutral and equidistant from "happy" and "unhappy." Why is the origin (0) of the embedding space not a reliable indicator of semantic neutrality, even if the pole words ("happy" and "unhappy") were chosen carefully, and the neutral word (e.g., "normal") appears at the middle point of the pole words? Provide an example to illustrate your answer (e.g., draw a 2D scatter plot of word embeddings, semantic axis, and the origin).

Grading Criteria (5 points):

- (2 points) Identifies why the origin is not a reliable indicator of semantic neutrality.
- (3 points) Provide an example to illustrate the answer.

Solution: The origin of the embedding space is arbitrary. Embedding algorithms optimize contextual co-occurrence and do not center semantic meaning around zero. The semantic axis constructed from "unhappy" to "happy" uses these specific words to define a direction, but the geometric origin need not align with semantic neutrality.

23. (5 points) Altinok GeBNLP 2024 studies gender bias in word embeddings trained on Turkish, a language that lacks grammatical gender (e.g., it uses a single pronoun for both "he" and "she"). Yet, the study finds significant gender associations in the resulting embeddings. Explain how gender bias can emerge in word embeddings for a gender-neutral language. Your answer should reference at least one mechanism by which such bias is learned during training, and why this challenges the assumption that grammatical neutrality implies embedding neutrality in terms of stereotypes. You may refer to the following examples from the study:

    - The word "nurse" is more closely associated with female names, while "professor" is more associated with male names.
    - The word "makeup" is grammatically neutral but clusters with culturally feminine terms in the embedding space.

    Grading Criteria (5 points):

    - (2 points) Clearly explains a process by which gender bias is learned during training.
    - (3 points) Explains why grammatical neutrality does not guarantee embedding neutrality.

Solution: This happens because embeddings learn from patterns in how words appear together in large amounts of text. If certain words frequently appear in contexts that reflect real-world gender roles or stereotypes, the embeddings will capture those associations. For example, if the word "nurse" often appears near female names, and "professor" near male names, embeddings will reflect that pattern by associating each word with a specific gender. Similarly, a word like "makeup," which is grammatically neutral, may still appear in contexts involving femininity and therefore be embedded closer to female-associated terms. Thus, grammatical neutrality doesn't prevent bias—models pick up on how language is used in society.

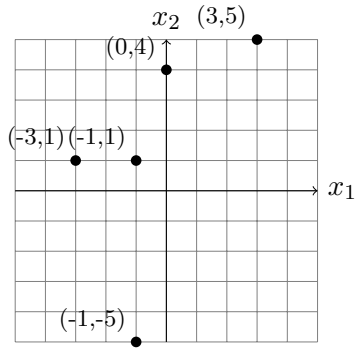
24. (5 points) In Transformer architectures, layer normalization plays a critical role in stabilizing training. The layer normalization is defined by:

$$\hat{x}_i = \gamma \frac{x_i - \mu_i}{\sqrt{\sigma_i^2 + \epsilon}} + \beta_i,$$

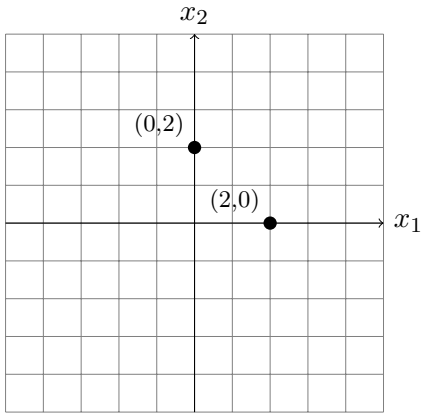
    where μ_i is the mean and σ_i is the standard deviation of the input x_i, ϵ is a small constant to prevent division by zero, and γ and β are learnable parameters. Apply the layer normalization with $(\epsilon, \gamma, \beta) = (0, 1, 1)$ to the following data points and draw the normalized data points as 2D scatter plot. Show your work.

    Grading Criteria (5 points):

- (3 points) Correctly applies the layer normalization formula.
- (2 points) Correctly plots the normalized data points.



Solution:



25. (8 points) Imagine you're tasked with creating an instruction-based prompt to help an LLM deliver an insightful movie review for a specific film. The intended audience is general moviegoers seeking a balanced analysis of the film's strengths and weaknesses. Construct a prompt that incorporates the following key components: Instruction, Data, Output Indicator, Persona, Context, and Audience. In your answer, outline how each component contributes to a clear and effective review. For the prompt, explicitly specify the context, persona, output indicator, and audience.

    Grading Criteria (8 points):

    1. (1 point for each component) Explain how each component impacts the LLM's output.
    2. (2 points) Create a complete prompt that explicitly includes all six components.

Solution:

- **Instruction** defines what the LLM should do.
- **Data** provides essential background information needed to complete the task effectively.
- **Output Indicator** specifies the expected format and structure of the response.
- **Persona** establishes the tone, perspective, and voice the LLM should adopt.
- **Context** sets the situational background and relevant circumstances for the task.
- **Audience** identifies who will be reading the output, helping tailor the content appropriately.

Prompt:

> **Persona:** You are a movie critic who is known for your clear, balanced, and engaging commentary. **Instruction:** Write a detailed movie review for the film The Midnight Escape. **Data:** Include information from the provided synopsis, key scenes, and highlighted themes. **Output Indicator:** The review must end with an overall rating out of 10 and a bulleted list summarizing the film's major strengths and weaknesses. **Context:** This review is intended for a modern audience that appreciates thoughtful analysis of both narrative and technical aspects without requiring specialized knowledge. **Audience:** General moviegoers who seek comprehensive yet accessible film critiques.