1. What are Sequence-to-Sequence Models?

Sequence-to-sequence (seq2seq) models map one sequence to another, typically using an encoder-decoder architecture.

- **The encoder reads the input sequence and encodes it into a context vector.**
- **The decoder uses this vector to generate the output sequence.**

Examples: Machine translation, text summarization, chatbot responses.

---

2. Problems with Vanilla RNNs

- **Vanishing gradients: Hard to learn long-term dependencies.**
- **Fixed-length context: Struggles with variable-length inputs and outputs.**
- **Poor performance on long sequences: All information must be compressed into one vector.**

---

3. What is Gradient Clipping?

Gradient clipping limits the size of gradients during training to prevent exploding gradients, which can destabilize learning.
It's done by capping gradients at a threshold value, keeping updates under control.

---

4. Explain Attention Mechanism

Attention allows the model to focus on different parts of the input sequence when generating each output step.
Instead of relying on a single context vector, attention assigns weights to all input tokens dynamically for each output step.

---

5. Explain Conditional Random Fields (CRFs)

CRFs are used for structured prediction, especially in tasks like Named Entity Recognition.

They model the conditional probability of a label sequence given an input sequence, considering dependencies between output labels.

---

6. Explain Self-Attention

Self-attention lets each position in a sequence attend to all other positions, allowing the model to learn context from the whole sequence.
Used heavily in Transformers to capture dependencies regardless of distance.

---

7. What is Bahdanau Attention?

Also called additive attention, this is a form of attention mechanism introduced by Bahdanau et al.
It learns to align and attend over encoder hidden states using a small neural network to score each input token before generating an output.

---

8. What is a Language Model?

A language model estimates the probability of a sequence of words.
It predicts the next word in a sequence, based on previous words.
Used in text generation, autocomplete, and speech recognition.

---

9. What is Multi-Head Attention?

An extension of self-attention that runs multiple attention mechanisms (heads) in parallel.
Each head learns different aspects of the input.
The outputs are then combined to form a richer representation.

---

10. What is Bilingual Evaluation Understudy (BLEU)?

BLEU is a metric for evaluating machine-translated text against one or more reference translations.
It measures n-gram overlap between the predicted and reference texts.
Higher BLEU scores indicate better quality.