AI Olympiad, February 2026

1. **Context**

   In Statistical Machine Translation, a common sub-problem is *Word Ordering*. You are given a "bag of words" (a set of scrambled words) and must reconstruct the original sentence order to maximize the probability of the sequence.

   Let $W = \{w_1, w_2, \ldots, w_N\}$ be a set of N unique words. You are given a **Bigram Language Model**, $P(w_j|w_i)$, which provides the probability of word w_j following word w_i. There are two special tokens: $\langle \text{S} \rangle$ (Start) and $\langle \text{E} \rangle$ (End). The sentence must begin with $\langle \text{S} \rangle$ and end with $\langle \text{E} \rangle$.

   Objective: Find a permutation π of the words in W that maximizes the sentence probability:

$$P(\pi) = P(w_{\pi_1}|\langle \text{S} \rangle) \times \left(\prod_{i=1}^{N-1} P(w_{\pi_{i+1}}|w_{\pi_i}) \right) \times P(\langle \text{E} \rangle|w_{\pi_N})$$

   To apply the A* algorithm, we transform this maximization problem into a minimization problem by defining the cost of a transition between word u and word v as the negative log-probability:

$$C(u, v) = -\ln(P(v|u))$$

   a) **Part A: State Space Formalization**

   Formalize this problem as a State Space Graph search suitable for the A* algorithm. Specifically, define:

   i. The structure of a **State** n.
   ii. The **Start State** and **Goal State**.
   iii. The **Successor Function** (how transitions are generated) and the accumulated cost $g(n)$.

   [3 Marks]

   b) **Part B: Heuristic Design**

   The core difficulty lies in the factorial structure of the search space ($N!$), making a trivial heuristic (e.g., $h = 0$) computationally intractable for even modest input sizes (e.g., $N > 15$).

Design a non-trivial heuristic $h(n)$ that estimates the remaining cost to complete the sentence. Your heuristic must meet the following criteria:

i. It must be **admissible**.

ii. It must be **monotonic** (consistent).

Define your heuristic mathematically. You do not need to prove the properties yet, but you must explain the logic behind your design.

Hint: Consider the set of words that have not yet been placed. In a valid completed sentence, every one of these words must have exactly one incoming edge.

[5 Marks]

c) **Part C: Theoretical Proofs**

Using the heuristic $h(n)$ you defined in Part B, provide formal proofs for the following properties:

i. **Admissibility:** Prove that $h(n) \leq h^*(n)$, where $h^*(n)$ is the true minimum cost to the goal.

ii. **Monotonicity:** Prove that $h(n) \leq cost(n, n') + h(n')$, where n' is a successor of n.

[5 Marks]

[Total 13 Marks]