Read & Re-derive "Attention Is All You Need" (Individual HW 4)
Due: 11/15/2025. Send the PDF to yxw1259@miami.edu

Task (do all):

1. Read the paper *Attention Is All You Need* .
2. Re-derive the key formulas in your own words/steps.
3. Write what was **hard for you and why**.

Deliverable:

- PDF, 2–4 pages max. No code required. Use your own wording.

Reflection (5–8 sentences):

- List 2–3 parts you found difficult and how you tried to understand them.

Requirements:

- For CS students, please understand all the concepts and technical equations of this paper.
  You may need to understand:
  Scaled dot-product attention
  - What Q/K/V represent conceptually (queries, keys, values) and why we combine them.
  - What the softmax weights mean (importance of each token to the current token).
  - Why do we stabilize the scores (to avoid over-confident, noisy weights).
  - Where the mask is applied (to the scores before softmax) and what it does.

  Multi-head attention
  - Why do we use multiple heads (capture different patterns/relations in parallel).
  - The idea of project → split into heads → attend per head → concat → project back.
  - Basic shape thinking only (batch, length, model size, heads)—no numbers needed.
    The role of the final output projection (mix information from all heads).

  Positional encoding (sin/cos)
  - Why do we need positions (order isn't in tokens by default).
  - High-level idea of adding a position signal to each token embedding.
    Intuition: lets the model reason about relative and absolute positions and generalize to longer texts.

  Masks: causal vs. padding
  - Causal mask: blocks looking into the future (used in GPT-style decoding).
  - Padding mask: ignores pad tokens so they don't affect attention.
  - Where masks act: applied to the attention scores before softmax.

  One diagram (encoder or decoder layer)
  - Boxes you should include: (Self-Attention) → (Feed-Forward) with residual + LayerNorm around each.
  - For decoder: show masked self-attention, then cross-attention to the encoder, then FFN.
  - Arrows: show token embeddings + positions going in, and same length sequence coming out.
  - Label where masks are used and where skip connections (residuals) wrap sublayers.

- For non-CS students, please write down a 2-4 page summary of this paper, what you learned, why it is important, etc. You can consider including the following questions.
  - Attention in EMRs: In one sentence, explain what "attention" does when a model reads a clinical note (e.g., HPI, meds, allergies).
  - Multi-head value: Why is it helpful for a model to read the same chart with several "heads" (different angles)? Name two angles (e.g., meds/allergies, recent labs, problem list, timelines).
  - Masks in drafting notes: Give one scenario where a causal mask is appropriate while generating a discharge summary and one scenario where a padding mask matters when comparing notes of different lengths.

- Why positions matter: Describe a healthcare example where word/order changes the meaning (e.g., “hold warfarin before surgery” vs. “restart after”). What does positional information give the model here?
- Safety & limits: Name one benefit and one risk of using attention-based LLMs for clinical summaries (e.g., catching key problems vs. missing context or hallucinating) and give a one-sentence mitigation for the risk.