Assignment - 3

KRISHA GOTHI
240562

Q1 Why encoders got fired ?

Ans
- Decoder-only Transformers use a single autoregressive architecture, avoiding the need for seperate encoder & decoder stacks, which reduces parameter redundancy & simplifies scaling to very large models.
- They are trained using next-token prediction on unpaired text, allowing efficient utilization of massive web-scale corpora, unlike encoder-decoder models that typically rely on paired input-output data.
- During inference, decoder-only models support key-value (KV) caching in causal self-attention, enabling reuse of past computations & significantly reducing latency & ~~atte~~ memory usage for long sequences.
- Encoder-decoder models require cross-attention between encoder and decoder, which introduces additional computational overhead & becomes a bottleneck as context length increases
- Decoder-only architectures scale better to long-context lengths, as attention computation & memory access patterns are more efficient and easier to optimize
- A single decoder-only model can handle multiple tasks via prompting (QA, translation, summarization, dialogue), eliminating the need for task-specific architectural changes

- Overall, decoder-only transformers offer better training efficiency, lower inference cost, and smoother scalability, making them the preferred choice for modern large language models.

Q2 Weaponized Autocomplete

Ans 1) All NLP tasks reduce to continuation

- Translation, summarization, QA, and reasoning can all be framed as "predict the next tokens given a prompt."

2) Single objective learns full languages distribution

- Next token prediction models $P(x_t / x_{<t})$, capturing syntax, semantics, and discourse

3) Task supervision is in the prompt, not the loss

- Instructions like "Translate:" or "Summarize:" guide behavior without changing the objectives.

4) Reasoning exists in training data

- Proofs, explanations, and step-by-step solutions appear as text patterns the model learns to continue

5) Transformers capture long-range dependencies

- Self attention enables variable tracking and multi-step logical consistency.

6) Translation & Summarization are conditional mappings

- Models learn meaning-preserving (translation) and compression (summarization) from data distributions.

1) Scale cause emergence

- Large data + large models → abstract representation and generalisation beyond memorization

8) One loss, many behaviours.

- Minimizing cross-entropy forces efficient internal representations of meaning and structure

9) "Autocomplete" is just the interface

- Internally, the model learns a compressed simulator of language and knowledge

10) Special objectives are not required

- Task specific losses improve efficiency, but next-token prediction is theoretically sufficient.