System 2 Attention

Soft Attention in transformer Based LLM
⇓
Susceptible to irr info
affects the next token generation

to Resolve → S2A System 2 Attention
↓
Regenerates the input context to include only relevant information.

In Experiments
S2A
- QA
- Math word Problems
- Long form Generation

where S2A increases factuality, and objectivity and decreases sycophancy.

Before S2A, Several approaches tried to mitigate these issues
- By adding More supervised training data
- Reinforcement learning Strategies.

But the Researchers of S2A thought that there is an issue in the transformer itself mainly in attention Mechanism

Soft Attention

Assigns probability to the large context of the text ⇒ includes irrelevant text
[focuses more on repeated tokens]

QA] S2A - trivia QA Dataset ⇒ factuality 62.8% → 80.3%.
Compared to Llama 2 13B that

Long form] objectivity ↑ 57.4%.

Math word problems] GSM-IC S2A accuracy from 51.7% - 61.3%.

[Correlation in Next word Prediction Improves accuracy of the Next word]

But it ~~notes~~ LLMs susceptible to be adversely affected by spurious correlation in the context.

Two or more variables are associated but not [Causally Related] → One event is the result of another event

probability Repeated Phrases increases When Repetition] POSITIVE FEEDBACK LOOP → to resolve that

[Non trivial Repetitions]

Not just specific tokens, repeats the related topics in the context

---

Implementation of System 2 Attention

Context : (x)

generated Sequence : (y)

$y = LLM(x)$

S2A (two step process) :

① $x' = S2A(x)$
Irrelevant context will be removed

② $y = LLM(x')$

Alternative Implementations..

① No context / Question separation

② keep original context
[append x' to x]

③ Instructed prompting

④ Emphasize Relevance | Irrelevance