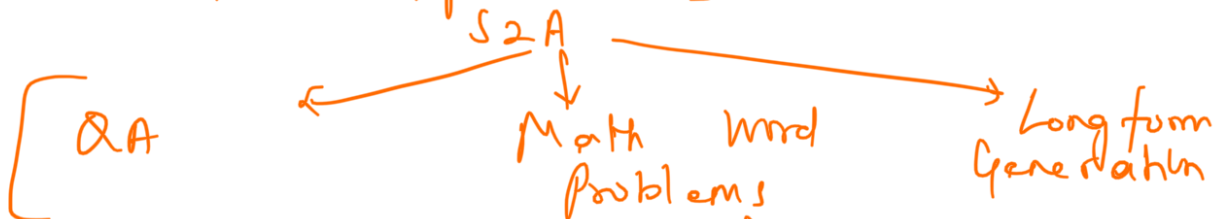


## System 2 Attention

Soft Attention in Transformer Based LLM  
Susceptible to Jor Info  
affects the next token generation

To resolve S2A System 2 Attention

Re generate the input context to  
include only relevant information.  
In Experiments



where S2A increases factuality, and  
objectivity and decreases  
hallucination.

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  $\Rightarrow$  includes irrelevant text  
[focus more on repeated tokens]

QA] S2A - private Dataset  $\Rightarrow$   
Compared to factuality 62.8%  $\rightarrow$  80.3%  
to Clamma 2 13 B Chat

Longform] objectivity  $\uparrow$  57.4%.

Math word problems] GSM-1C S2A accuracy  
from 51.7% - 61.3%.

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

But it ~~not~~ LLMs susceptible to  
be adversely affected by spurious  
correlation in the context.

$\Downarrow$   
Two or more variables are associated  
but not [Causally Related]  
the result of one event is  
the result of another event

probability of Repeated Phrases increases  
When Repetition ] POSITIVE FEEDBACK LOOP

Not just related topics  $\rightarrow$  to resolve that  
[Non trivial Repetitions]  
specific tokens, repeats the  
in the context

---

Implementation of System 2 Attention

Context : (x)  
generated Sequence : (y)

S2A  $y = \text{LLM}(x)$   
(two step process) :

①  $x' = \text{S2A}(x)$   
irrelevant context will be removed

②  $y = \text{LLM}(x')$

① Alternative Implementation..  
No context / question separation

② keep original context  
[append  $x'$  to  $x$ ]

③ Instructed prompting

④ Simplified Relevance / Irrelevance