Shinkai Tests

Nicolas Arqueros

July 2024

1 Tests

Task	Dataset	Avg #Tokens	Max #Tokens	#Samples
Multi-hop QA	HotpotQA	9.4k	15.9k	300
	2WikiMultihopQA	8.8k	15.9k	300
	MuSiQue	15.5k	16.0k	200
	HotpotWikiQA-mixup	142.4k	370.8k	250
Single-hop QA	NarrativeQA	29.7k	63.7k	200

Table 1: The statistics of benchmarks employed in our evaluation. The token number is calculated using the GPT-4 tokenizer from the TikToken. #Samples denote the total number of benchmarks.