How Faithful are RAG Models?

This new paper by Wu et al. (2024) aims to quantify the tug-of-war between RAG and LLMs' internal prior.

It focuses on GPT-4 and other LLMs on question answering for the analysis.

It finds that providing correct retrieved information fixes most of the model mistakes (94% accuracy).

GPT-4	Concordance (Prior)	Concordance (w/ RAG)	Slope
Drug Dosage	0.554	0.884	-0.26
Sports Stats	0.240	0.943	-0.18
Latest News	0.133	0.936	-0.10
Wikipedia Dates	0.433	0.995	-0.45
Wikipedia Names	0.350	0.965	-0.13
Wikipedia Locations	0.375	0.920	-0.28
Average	*0.347*	*0.940*	*-0.23*

Source: Wu et al. (2024)

When the documents contain more incorrect values and the LLM's internal prior is weak, the LLM is more likely to recite incorrect information. However, the LLMs are found to be more resistant when they have a stronger prior.

The paper also reports that "the more the modified information deviates from the model's prior, the less likely the model is to prefer it."

So many developers and companies are using RAG systems in production. This work highlights the importance of assessing risks when using LLMs given different kinds of contextual information that may contain supporting, contradicting, or completely incorrection information.

Last updated on September 19, 2024