Towards the Reproducible Evaluation of Generative Information Retrieval Systems

Niklas Deckers **Martin Potthast**

Leipzig University and ScaDS.AI
niklas.deckers@uni-leipzig.de, martin.potthast@uni-leipzig.de

Abstract

Search engines based on large language models for text generation can be used to provide SERPs in the form of text instead of the traditional "10 blue links" format. The availability of generative models makes it easy to conceptualize different approaches that utilize them in IR systems. For an evaluation of these systems, evaluation metrics need to be defined, and the lack of reproducibility in available generative models makes controlled and comparable evaluation studies difficult.

By defining a user model that slightly differs from the traditional user models for SERPs, we are able to motivate evaluation metrics for text SERPs. We also explore the system perspective by outlining the spectrum of different pipelines to integrate generative models and retrieval systems.

We introduce a number of helpful tools for the evaluation of generative IR systems: A research prototype for a generative IR system based on the Alpaca model and the ChatNoir search engine allows for demonstrations and easy user experiments. By integrating generative models and the mentioned evaluation methods into the TIRA platform, different approaches for generative IR systems can be defined, evaluated and compared. This helps to address issues like explainability and reproducibility. In this platform, self-hosted LLMs can serve as plug-in replacements for services like ChatGPT, also allowing easier testing and deployment. This is made possible by a technical infrastructure based on Kubernetes, allowing to host different version LLMs on a number of GPU nodes.