HTML हेडर

**HTML HEADERS**

रिट्रीवल-ऑगमेंटेड जेनरेशन (आरएजी) को ज्ञान क्षमताओं में सुधार करने और एलएलएम की मतिभ्रम की समस्या को कम करने के लिए दिखाया गया है। वेब आरएजी प्रणालियों में उपयोग किए जाने वाले बाहरी ज्ञान का एक प्रमुख स्रोत है, और चैटजीपीटी और पेरप्लेक्सिटी जैसी कई वाणिज्यिक प्रणालियों ने वेब खोज इंजनों को अपने प्रमुख पुनर्प्राप्ति प्रणालियों के रूप में उपयोग किया है।

Retrieval-Augmented Generation (RAG) has been shown to improve knowledge capabilities and alleviate the hallucination problem of LLMs. The Web is a major source of external knowledge used in RAG systems, and many commercial systems such as ChatGPT and Perplexity have used Web search engines as their major retrieval systems.

आम तौर पर, ऐसे RAG सिस्टम खोज परिणामों को पुनः प्राप्त करते हैं, परिणामों के HTML स्रोतों को डाउनलोड करते हैं, और फिर HTML स्रोतों से सादे पाठ निकालते हैं। सादे पाठ दस्तावेज़ या खंड LLM में डाले जाते हैं ताकि निर्माण को बढ़ाया जा सके। हालाँकि, HTML में निहित बहुत सी संरचनात्मक और अर्थपूर्ण जानकारी, जैसे शीर्षक और तालिका संरचनाएँ, इस सादे-पाठ-आधारित RAG प्रक्रिया के दौरान खो जाती हैं।

*Typically, such RAG systems retrieve search results, download HTML sources of the results, and then extract plain texts from the HTML sources. Plain text documents or chunks are fed into the LLMs to augment the generation. However, much of the structural and semantic information inherent in HTML, such as headings and table structures, is lost during this plain-text-based RAG process.*

इस समस्या को कम करने के लिए, हम HtmlRAG का प्रस्ताव करते हैं, जो RAG में प्राप्त ज्ञान के प्रारूप के रूप में सादे पाठ के बजाय HTML का उपयोग करता है। हमारा मानना ​​है कि बाहरी दस्तावेज़ों में ज्ञान को मॉडल करने में HTML सादे पाठ से बेहतर है, और अधिकांश LLM में HTML को समझने की मज़बूत क्षमताएँ हैं। हालाँकि, HTML का उपयोग करने से नई चुनौतियाँ सामने आती हैं। HTML में टैग, जावास्क्रिप्ट और CSS विनिर्देशों जैसी अतिरिक्त सामग्री होती है, जो RAG सिस्टम में अतिरिक्त इनपुट टोकन और शोर लाती है। इस समस्या को हल करने के लिए, हम HTML को छोटा करने और सूचना के नुकसान को कम करने के लिए HTML सफाई, संपीड़न और छंटाई रणनीतियों का प्रस्ताव करते हैं। विशेष रूप से, हम एक दो-चरणीय ब्लॉक-ट्री-आधारित छंटाई विधि डिज़ाइन करते हैं जो बेकार HTML ब्लॉकों को छाँटती है और HTML का केवल प्रासंगिक भाग रखती है। छह QA डेटासेट पर किए गए प्रयोग RAG सिस्टम में HTML के उपयोग की श्रेष्ठता की पुष्टि करते हैं।

**To alleviate this problem, we propose HtmlRAG, which uses HTML instead of plain text as the format of retrieved knowledge in RAG. We believe HTML is better than plain text in modeling knowledge in external documents, and most LLMs possess robust capacities to understand HTML. However, utilizing HTML presents new challenges. HTML contains additional content such as tags, JavaScript, and CSS specifications, which bring extra input tokens and noise to the RAG system. To address this issue, we propose HTML cleaning, compression, and pruning strategies, to shorten the HTML while minimizing the loss of information. Specifically, we design a two-step block-tree-based pruning method that prunes useless HTML blocks and keeps only the relevant part of the HTML. Experiments on six QA datasets confirm the superiority of using HTML in RAG systems.**

|  |  |  |
| --- | --- | --- |
| क्रम सं. Sl No | नाम Name | आयु Age |
| 1 | पलाश Palash | 23 |
| 2 | मनष Manash | 26 |
| 3 | आरएजी प्रणालियां. RAG systems. | 59 |
| 4 | दो-चरण two-step | 48 |
| 5 | रात्रि विश्राम Night Stay | 59 |