ICWR2024
April 24-25, 2024; Tehran, Iran
10th International Conference on Web Research

Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement

Pouria Omrani∗‡, Alireza Hosseini†‡, Kiana Hooshanfar†‡, Zahra Ebrahimian†‡, Ramin Toosi†‡, Mohammad Ali Akhaee†
‡Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran
†School of Electrical and Computer Engineering, College of Engineering, University of Tehran, Tehran, Iran
‡Adak Vira Iranian Rahjoo Company, Tehran, Iran
{pouria.omrani@ieee.org, arhosseini77@ut.ac.ir, k.hooshanfar@ut.ac.ir, z.ebrahimian@ut.ac.ir, r.toosi@ut.ac.ir, akhaee@ut.ac.ir}

***Abstract*— In the domain of Natural Language Pro- cessing (NLP), the integration of Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG) represents a significant advancement towards enhancing the depth and relevance of model-generated responses. This paper introduces a novel hybrid RAG framework that synergizes the Sentence-Window and Parent-Child methodologies with an innovative re- ranking mechanism, aimed at optimizing the query response capabilities of LLMs. By leveraging exter- nal knowledge sources more effectively, the proposed method enriches LLM outputs with greater accuracy, relevance, and information fidelity. We subject our hybrid model to rigorous evaluation against bench- mark datasets and metrics, demonstrating its superior performance over existing state-of-the-art RAG tech- niques. The results highlight our method's enhanced ability to generate responses that are not only contex- tually appropriate but also demonstrate a high degree of faithfulness to the source material, thereby setting a new standard for query response enhancement in LLMs. Our study underscores the potential of hybrid RAG models in refining the interaction between LLMs and external knowledge, paving the way for future research in the field of NLP.**

***Keywords*— LLM, Generative AI, NLP, Retrieval Augmented Generation (RAG).**

JCWR2024
April 24-25, 2024; Tehran, Iran
10th International Conference on Web Research

رویکرد ترکیبی تولید تقویت‌شده با بازیابی برای بهبود پاسخ‌دهی پرسش‌ها در مدل‌های زبانی بزرگ

پوریا عمرانی[۱]، علیرضا حسینی[۱]، کیانا هوشانفر[۱] ، زهرا ابراهیمیان[۱]، رامین طوسی[۱]، محمدعلی اخایی[۲]

[۱] دانشکده مهندسی برق، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران – شرکت آداک ویرا ایرانیان رهجو، تهران، ایران
pouria.omrani@ieee.org, arhosseini77@ut.ac.ir, k.hooshanfar@ut.ac.ir, z.ebrahimian@ut.ac.ir, r.toosi@ut.ac.ir

[۲] گروه مهندسی برق و کامپیوتر، دانشکده مهندسی ، دانشگاه تهران، تهران، ایران
akhaee@ut.ac.ir

چکیده

در حوزه پردازش زبان طبیعی (NLP)، ادغام مدل‌های زبانی بزرگ (LLMs) با تولید تقویت‌شده توسط بازیابی (RAG) نشان‌دهنده پیشرفت قابل توجهی در افزایش عمق و ارتباط پاسخ‌های تولیدشده توسط مدل است. این مقاله چهارچوب RAG ترکیبی نوآورانه‌ای را معرفی می‌کند که روش‌های Sentence-Window و Parent-Child را با مکانیزم بازرتبه‌بندی تلفیق می‌کند، با هدف بهینه‌سازی قابلیت‌های پاسخ‌گویی به پرسش‌ها توسط LLMها. با استفاده مؤثرتر از منابع دانش خارجی، روش پیشنهادی خروجی‌های LLM را با دقت، ارتباط و وفاداری اطلاعاتی بیشتری غنی‌سازی می‌کند. ما مدل ترکیبی خود را با استفاده از مجموعه‌داده‌ها و معیارهای مرجع مورد ارزیابی قرار می‌دهیم و عملکرد برتر آن را نسبت به تکنیک‌های RAG پیشرفته موجود نشان می‌دهیم. نتایج، توانایی بهبود یافته روش ما را در تولید پاسخ‌هایی که نه تنها مناسب با زمینه هستند، بلکه وفاداری بالایی به مواد منبع نیز نشان می‌دهند، برجسته می‌کند و بدین ترتیب یک استاندارد جدید برای بهبود پاسخ‌گویی در LLMها ایجاد می‌کند. مطالعه ما پتانسیل مدل‌های RAG ترکیبی را در پالایش تعامل بین LLMها و دانش خارجی مورد تأکید قرار می‌دهد و راه را برای تحقیقات آینده در زمینه NLP هموار می‌سازد.

کلمات کلیدی: مدل زبانی بزرگ، هوش مصنوعی مولد، پردازش زبان طبیعی، تولید تقویت شده توسط بازیابی