

DEVELOPMENT OF A VIRTUAL ASSISTANT FOR SHIFT OPERATORS OF THE BM@N EXPERIMENT

I. Romanov^{a,1}, K. Gertsenberger^a, A. Chebotov^a

^a Joint Institute for Nuclear Research, Dubna, 141980, Russia

The paper presents the development of a virtual assistant designed to address knowledge fragmentation challenges faced by shift operators at the BM@N (Baryonic Matter at Nuclotron) experiment. The system employs a hybrid approach that integrates full-text and semantic retrieval methods using the BGE-M3 model embeddings to solve challenges related to domain-specific terminology. The paper details the conversion of documents into the Markdown format and the chunking algorithm that preserves contextual relationships. The implemented system uses PostgreSQL, Valkey, and Qdrant databases with NVIDIA Triton Inference Server. It provides user access through both a Telegram bot and a web interface with OpenID Connect authentication. The system was evaluated using a domain-specific dataset with the Mean Reciprocal Rank (MRR) metric. The evaluation yielded MRR@5 equal to 0.68, indicating its capability to retrieve relevant information from specialized documentation. This semantic search engine is the foundation for a future complete RAG system that can generate coherent responses to operator queries.

Представлена разработка виртуального ассистента, предназначенного для решения проблемы фрагментации знаний, с которой сталкиваются операторы смен в эксперименте BM@N (Baryonic Matter at Nuclotron). Система использует гибридный подход, объединяющий методы полнотекстового и семантического поиска на основе векторных представлений из модели BGE-M3 для решения задач, связанных с обработкой специализированной терминологии эксперимента. В статье описаны преобразование документов в формат Markdown и алгоритм разбиения на фрагменты, сохраняющий контекстные связи. Реализованная система использует различные системы хранения, такие как PostgreSQL, Valkey и Qdrant, а также NVIDIA Triton Inference Server. Она предоставляет пользователям доступ как через бота в Telegram, так и через веб-интерфейс с аутентификацией через OpenID Connect. Проводилась оценка системы по метрике Mean Reciprocal Rank (MRR), результаты которой подтвердили способность системы извлекать релевантную информацию из специализированной документации. Разработанное решение является основой для будущей полнофункциональной системы RAG, способной генерировать связные ответы на запросы операторов.

PACS: 07.05.Bx; 07.05.Kf

Received on December 5, 2025.

¹E-mail: iromanov@jinr.ru