

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: September 26, 2026 at 2:30 PM

Speaker Information

Shekh Ahammed Adnan Bashir (Std No. 1018052026) is a part time M.Sc. student in the department of CSE, BUET. He completed his undergraduate studies from CSE, BUET in 2015. His research interest lies in the field(s) of Automated Theorem Provers, Formal Systems, Knowledge Discovery, and Mathematical Logic. He is currently doing his postgraduate thesis under the supervision of Abu Wasif. He will be speaking about his research findings in this talk.



Enhancing Retrieval-Augmented Generation (RAG) Efficiency through Structural Features

Retrieval-Augmented Generation (RAG) supplies Large Language Models (LLMs) with external evidence at inference time. Conventional RAG systems, however, generally treat retrieved passages as context to be interpreted and synthesized by the LLM, even when the answer may be determined by a smaller and more structured subset of the available information. This thesis investigates whether RAG efficiency can be improved by distinguishing information that must be processed by an LLM from information whose consequences can instead be computed explicitly.

The study considers two viewpoints of sentence structure. The first uses surface structural features: relevance, coherence, fluency, grammar, structure, and jargon. These properties are combined into an interpretable weighted score and used to rerank retrieved passages before generation. The second uses logical structural features. In this approach, source text is compiled into facts and Horn-clause rules, allowing queries whose answers are entailed by the compiled knowledge to be answered by symbolic inference rather than neural generation.

The surface-feature approach is evaluated in two stages. First, the proposed feature score is compared with human preference judgments using the LMSYS Chatbot Arena data. The observed agreement is indistinguishable from chance. Second, a controlled context-reduction experiment tests whether the same score can safely remove retrieved passages while holding the retriever, language model, query set, and evaluation procedure fixed. On the 11-Book Corpus dataset, surface filtering reduces average input tokens by 27% for Vanilla RAG and 48% for GraphRAG, but accuracy decreases from 0.30 to 0.10 and from 0.40 to 0.15, respectively, while hallucination increases. On HoVer dataset, the same procedure does not reduce input tokens. These results indicate that properties of evidence text are insufficient to determine whether retrieved evidence can safely be removed.

The thesis therefore develops PrologRAG, a hybrid architecture that separates neural knowledge processing from symbolic reasoning, that illustrates usefulness of logic-feature approach. During an offline compilation phase, an LLM extracts structured knowledge from source text and constructs a predicate-indexed knowledge base of facts and Horn-clause rules. At query time, a natural-language question is translated into one or more candidate logical goals. A symbolic Prolog engine then performs unification and SLD resolution with backtracking. When the query is successfully resolved, the answer is constructed from the resulting proof trace

Postgraduate Seminar Series

Venue: Graduate Seminar Room

Date & Time: September 26, 2026 at 2:30 PM

without neural answer generation. Queries that cannot be resolved follow a structured LLM-based fallback path.

PrologRAG is evaluated on eight datasets spanning narrative prose, philosophical writing, legal texts, encyclopedic question answering, reading comprehension, legal holdings, and claim verification. Resolution rates range from 0% to 90% and vary substantially with the logical explicitness of the source material. Across datasets in which the knowledge-compilation pipeline produces sufficiently compact and useful symbolic representations, PrologRAG reduces average input-token consumption by 14-76%. Prolog-resolved queries use approximately 4.5 times fewer input tokens than fallback queries and, under the evaluation protocol, obtain higher average faithfulness and lower hallucination rates. The principal limitation is not the symbolic inference engine itself but the interface between natural language and logical form: approximately 60% of unresolved primary-corpus queries are associated with predicate mismatch during query translation. The evaluation also reveals a knowledge-granularity constraint. In CaseHOLD, 18 of 20 queries are formally resolved, yet judged accuracy is 0%, because the compiled representation is too coarse to distinguish the answer alternatives required by the benchmark.

Taken together, the results support that a subset of RAG queries, specifically the deductive ones, can be resolved by means of symbolic inference. Context token efficiency can therefore be viewed not only as a retrieval problem, but also as a problem of allocating computation between neural generation and explicit symbolic inference according to the structure of the query and the available knowledge representation.