OPPINION

from

Prof. Olympia Roeva, PhD
Institute of Biophysics and Biomedical Engineering - BAS
Bioinformatics and Mathematical Modeling Department

for awarding of the educational and scientific degree "Doctor of Philosophy"

Professional field:
4.6 Informatics and computer sciences,

with a candidate
Miroslava Doncheva Dimitrova

PhD thesis title
"Evaluation Framework of Retrieval-Augmented Generation"

1. Relevance of the problem developed in the PhD thesis in scientific and scientific-applied terms.

The relevance of the topic is driven by the growing necessity for reliability in automated information retrieval. In fields such as healthcare and law, any inaccuracy can compromise the final outcome, making Retrieval-Augmented Generation (RAG) methods a subject of intense research interest. The text correctly identifies that the web-based functions of certain language models are insufficient for scientific or professional purposes due to their instability and lack of transparency. Furthermore, it highlights the systemic issue of factual variability in parametric models and the difficulties in verifying their responses, which necessitates the development of more precise mechanisms for linking generated text with concrete evidence.

The PhD student aims to develop an evaluation framework for Retrieval-Augmented Generation that supports evidence-based retrieval configuration decisions for RAG systems with open-source LLMs, with particular focus on similarity threshold configuration.

To achieve this objective, the following four tasks have been defined:

1. Define and implement the core components of the evaluation framework by integrating three layers: (a) a threshold-aware evaluation procedure with composite scoring, (b) the Performance Assessment System for Similarity Evaluation and Retrieval (PaSSER) platform providing reproducibility infrastructure with blockchain-based provenance logging, and (c) a controlled experimental design producing comparative threshold-aware evidence across models and domains.
2. Establish model selection criteria. Define selection criteria aligned with local deployment feasibility, licensing constraints, and computational requirements, including profiling of selected models with respect to context window size and decoding settings.
3. Define metric selection and computation procedures. Select metrics aligned with the evaluation constructs of lexical overlap, semantic similarity, fluency, accuracy, and language modelling, and implement metric computation consistently across models and experimental conditions.
4. Conduct controlled testing and analysis. Prepare domain corpora and question-answer datasets with specified preprocessing and retrieval configurations; execute controlled evaluations under systematic parameter variation, including similarity threshold sweeps; and aggregate results to interpret outcomes with respect to retrieval selectivity, generation quality, and reproducibility, producing practical guidance for model and threshold selection.

2. Degree of knowledge of the state of the problem and creative interpretation of the literature

The PhD thesis cites 145 references. The PhD student demonstrates a high level of expertise regarding the state-of-the-art in the field, as well as the appropriate tools and approaches required to address the research tasks and achieve the overall objectives of the PhD thesis.

3. General analytical characteristics of the PhD thesis

The dissertation is well-structured and logically consistent with the research tasks set forth. The thesis consists of 125 pages in English and includes lists of tables, figures, equations, and algorithms; an introduction; Chapter 1 (Retrieval-Augmented Generation); Chapters 2-4 containing the core theoretical frameworks, scientific research results, and applications; and Chapter 5 (Discussion and Future Work), which provides a summary of the results and the contributions of the dissertation. These are followed by Appendices A, B, and C; a bibliography; a list of publications related to the dissertation; a list of citations; participation in research projects; acknowledgments; and a declaration of originality.

In the first chapter, the author systematizes the existing scientific knowledge regarding RAG systems and their evaluation methodologies, justifying the need to address specific deficits (D1–D3). The second chapter is dedicated to the technological implementation, providing a detailed description of the PaSSER platform, including the innovative integration of blockchain to ensure data integrity. The third chapter defines the experimental framework, covering the selection of language models and the algorithms for metric calculation. Empirical validation is presented in the fourth chapter, where the impact of parameters such as the similarity threshold is analyzed through testing in the domains of agriculture and biodiversity. The work concludes with the fifth chapter, which synthesizes the

contributions, analyzes the limitations, and formulates directions for future development, followed by a final summary.

4. Evaluation of contributions of the PhD thesis and their significance

I accept the contributions formulated in the dissertation. The scientific and applied contributions are systematized in the following areas:

1. Infrastructural Implementation: The PaSSER platform has been developed – an open-source, web-based environment that integrates multi-metric evaluation mechanisms and innovative data tracking through the Antelope blockchain technology.
2. Methodological Contribution: New metrics for comprehensive evaluation have been proposed: the Composite Score (CPS) for unified comparison, the Threshold Composite Score (T-CPS), which accounts for stability via a "reward-penalty" mechanism based on the Coefficient of Variation (CV), and the Balance Score, which measures the trade-off between quality and consistency.
3. Empirical Verification and Domain-Specific Analysis: A large-scale study (exceeding 38,000 evaluations) was conducted on seven open-source LLMs. The results prove that sensitivity to the similarity threshold is not a static property of the model but is highly domain-dependent. Significantly more pronounced effects were observed in the "Biodiversity" domain (up to a 13.32% improvement in CPS) compared to "Agriculture." It was established that the transition between domains leads to a noticeable shift in optimal threshold configurations and a sharp increase in output instability (up to 105%), highlighting the necessity for context-specific tuning of RAG systems.

The scale of the experimental work summarized in the third contribution is particularly impressive. Through systematic testing of models in the 7-8B parameter range, the author successfully subjects the notion that retrieval parameters are universal to critical analysis. The proven variability of the optimal threshold when switching domains (with shifts of up to -0.35 for models like DeepSeek R1) and the parallel tracking of output instability (CV) are valuable contributions with direct practical application. The comparative analysis between CPS and T-CPS underscores the importance of consistency: it is demonstrated that a model with high average performance may prove unreliable when its instability is taken into account, making the proposed methodology a robust filter for professional RAG applications.

5. Assessment of PhD thesis publications

The results of the dissertation have achieved wide dissemination within the scientific community. The doctoral candidate presents five publications in international conference proceedings and the journals Problems of Engineering Cybernetics and Robotics and Electronics. Two of these publications are indexed in the IEEE Xplore Digital Library. Miroslava Dimitrova is the sole author of one of the publications. The publications have a total of 64 citations, with one of them alone reaching 51 citations. All of this attests to the high level of scientific research and its findings.

6. **Assessment of the compliance of the autoreferate with the requirements for its preparation, as well as the adequacy of reflecting the main points and contributions of the PhD thesis**

The autoreferate correctly reflects the content of the PhD thesis and gives an idea of the problems under consideration, the results obtained, and the thesis's contributions.

7. **Critical notes on the PhD thesis**

I have no critical remarks regarding the PhD thesis. I would like to briefly express my impressions of the work.

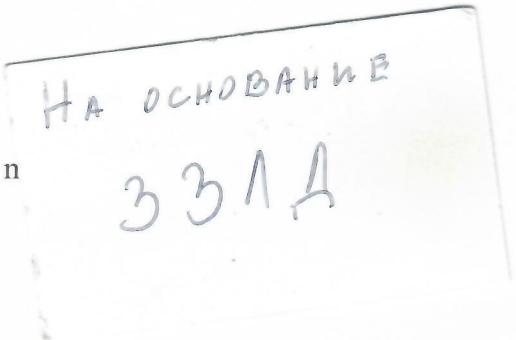
The development of the open-source PaSSER platform is a significant applied contribution, providing the research community with a ready-to-use tool for conducting controlled and transparent experiments. The scale of the presented experimental work, encompassing over 38,000 individual evaluations, is impressive and lends a high degree of credibility to the derived statistical dependencies. The doctoral candidate demonstrates exceptional precision in defining the new composite metrics (CPS and T-CPS), successfully transforming the calibration of the similarity threshold from a heuristic decision into a rigorous scientific and measurable process.

The thesis represents a mature and methodologically sound study, offering original solutions for improving the stability and quality of contemporary language models.

8. **Conclusion with a clear positive or negative assessment of the PhD thesis**

Given the proven scientific value of the research conducted and the contributory nature of the results achieved, I provide a positive evaluation of the PhD thesis of Miroslava Doncheva Dimitrova.

The PhD thesis meets the requirements of the Law on the Development of the Academic Staff in the Republic of Bulgaria, the Internal Regulations for its application, as well as the Regulations for the terms and conditions for acquiring scientific degrees and occupying academic positions at the IICT – BAS. The achieved scientific and scientific-applied results give me reason to propose to the respected Scientific Jury to award the educational and scientific degree "Doctor of Philosophy" to Miroslava Doncheva Dimitrova in the professional field 4.6 Informatics and Computer Sciences, PhD programme Informatics.



17.04.2026 Scientific Jury n

Sofia