REVIEW

by Prof. DCs. Daniela Ananieva Orozova, Trakia University

of a dissertation for acquiring the educational and scientific degree "Doctor"
in professional direction 4.6. "Informatics and Computer Sciences",
doctoral program "Informatics" (01.01.12)
titled: "Evaluation Framework of Retrieval-Augmented Generation"

by Miroslava Doncheva Dimitrova

By order No. 75/ 27.03.2026 of the Director of IICT–BAS, on the basis of Art. 4, para. 2 of the Act on the Development of the Academic Staff in the Republic of Bulgaria, in connection with the procedure for acquiring the educational and scientific degree "Doctor" in professional direction 4.6. "Informatics and Computer Sciences", doctoral program "Informatics" by Miroslava Doncheva Dimitrova with a dissertation titled "Evaluation Framework of Retrieval-Augmented Generation", I have been appointed a member of the Scientific Jury and reviewer.

In the evaluation of the dissertation, the requirements of the Act on the Development of the Academic Staff in the Republic of Bulgaria (ADASRB), the Regulation on the Implementation of ADASRB (RIDASRB, Council of Ministers Decree No. 26 of 13 February 2019), and the Regulations on the specific conditions of IICT–BAS for the implementation of the law have been applied:

Pursuant to Art. 27, para. 1 of RIDASRB, "the dissertation must contain scientific or scientific-applied results that represent an original contribution to science. The dissertation must demonstrate that the candidate possesses in-depth theoretical knowledge in the respective specialty and the capacity for independent scientific research."

According to Art. 27, para. 2 of RIDASRB, "the dissertation must be presented in a form and volume corresponding to the specific requirements of the primary unit. The dissertation must contain: a title page; an introduction; a presentation; a conclusion – summary of the obtained results with a declaration of originality; a bibliography."

The minimum requirements by groups of indicators for acquiring the educational and scientific degree "Doctor" pursuant to RIDASRB have been met.

The scientific supervisor of the dissertation is Acad. Ivan Popchev.

1. General data on the dissertation

The dissertation is presented in English.

The stated aim (p. 26 of the dissertation) is: "The aim of the dissertation is 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 the aim, four objectives are formulated on p. 27 of the dissertation:

Objective 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 [15] providing reproducibility infrastructure with blockchain-based provenance logging; and (c) a controlled experimental design producing comparative threshold-aware evidence across models and domains.

Objective 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.

Objective 3: "Define metric selection and computation procedures." Select metrics aligned with the evaluation constructs of lexical overlap, semantic similarity, fluency, accuracy, and language modeling, and implement metric computation consistently across models and experimental conditions.

Objective 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.

On p. 26 of the dissertation, three research questions are formulated:

RQ1: "Does varying the similarity threshold produce measurable changes in generation quality?"

RQ2: "Do similarity threshold effects differ across language models?"

RQ3: "Do comparable similarity threshold ranges hold across knowledge domains?"

The dissertation totals 215 pages and includes: a list of tables, a list of figures, a list of equations, a list of algorithms, a glossary of terms and abbreviations (pp. 16–23), an introduction (pp. 24–29), five main chapters (pp. 30–175), a conclusion — summary of the obtained results (pp. 176–178), three appendices (A, B, C, pp. 179–191), a bibliography (pp. 192–204), a list of publications related to the dissertation (p. 205), a citation record (pp. 207–212), a summary of project participation (p. 213), acknowledgements (p. 214), and a declaration of originality (p. 215). Twenty-five figures and 50 tables are presented. The bibliography contains 145 titles and covers foundational and contemporary sources from the fields of information retrieval, natural language processing, generative language models, evaluation frameworks, and research reproducibility. The ratio between the main scientific content (Chapters 1–5) and the supporting sections is balanced.

The structure follows the logic of scientific-applied research: a review of the state of the art, design of the instrumental apparatus, definition of the metric apparatus, systematic experimentation, and discussion. Transitions between parts are smooth, and the reverse traceability from deficiency through research question, objective, chapter, and contribution is summarized in Table I.1 (p. 28).

2. Chapter-by-chapter evaluation of the content

Chapter 1 "Retrieval-Augmented Generation" (pp. 30–58) is a literature-analytical review tracing the development of information retrieval from basic indices and classification schemes to modern modular architectures — Self-RAG, CRAG, FLARE, GraphRAG, and LightRAG. The review covers the architectural components of RAG, innovations in the areas of efficiency, data, iterative retrieval, multimodal extension, domain adaptation, and verification. Section 1.3.8 pays particular attention to similarity thresholds and retrieval selectivity, thereby preparing the theoretical basis for the experimental part. Section 1.4 analyzes the established evaluation frameworks RAGAS, RGB, TREC 2024 RAG Track, and TruLens; in 1.4.6, three research deficiencies are outlined: (D1) the absence of threshold-aware evaluation, (D2) insufficient reproducibility infrastructure, and (D3) limited comparative data for open-source models in the 7–8 billion parameter range. The review is systematic, with clear traceability between the identified deficiencies and the subsequent research questions.

Chapter 2 "Design and Architecture of PaSSER" (pp. 59–78) presents the developed PaSSER platform (Performance Assessment System for Similarity Evaluation and Retrieval) — a browser-based, modular system for configuring and evaluating RAG with open-source LLMs. The three-tier architecture (web interface, backend services, and blockchain subsystem) is justified from the perspective of reproducibility and traceability. An Antelope blockchain network is integrated, deployed as a permissioned private blockchain network, in which the experimental configurations, retrieval parameters, decoding settings, and run identifiers are recorded. This represents an original infrastructure solution that links the evaluation procedure with controlled documentation of conditions. The functionalities for system configuration, data management, retrieval tuning, model interaction, and evaluation are described with the necessary technical precision and are supported by published sources [17], [16] presenting the platform.

Chapter 3 "Model Selection and Evaluation Metrics" (pp. 79–111) defines the analytical apparatus. The selection of seven open-source LLMs in the 7–8 billion parameter range (Mistral 7B v0.1 and v0.3, Llama 2 7B, Orca 2 7B, Granite 3.2 8B, DeepSeek R1 8B, Llama 3.1 8B) is justified by criteria of local deployment feasibility, licensing compatibility, and computational requirements. Twenty-four metrics are defined, grouped into five categories — lexical overlap, semantic similarity, fluency, predictive and answer quality, statistical correlation, and human-readability inspired metrics (B-RT). The original composite indicators Composite Performance Score (CPS), Threshold-aware Composite Performance Score (T-CPS), and Balance Score are formulated; they aggregate heterogeneous metrics into a unified indicator for systematic comparison. The procedure for statistical verification by paired t-tests with effect-size reporting (Cohen's d) is formalized in Algorithm A.3 of Appendix A.

Chapter 4 "Experimental Evaluation and Results" (pp. 112–154) is the most extensive and presents the results of four sequential experimental phases of increasing complexity. Phase I provides system validation and runtime profiling under a fixed top-k regime. Phase II introduces similarity threshold variation in the range 0.50–0.80 and establishes model sensitivity. Phase III extends the range to 0.50–0.95 across four updated models in the agriculture domain, introduces statistical verification, and records CPS improvements of up to +4.58% (Mistral 7B v0.3 at threshold 0.95, Table B.1). Phase IV validates the transferability of the results to a new knowledge domain — biodiversity — where the recorded CPS improvements reach +13.32% (Table 4.14). The shifts in peak-performing thresholds upon transitioning from agriculture to biodiversity fall in the range from −0.05 (Llama 3.1 8B) to −0.35 (DeepSeek R1 8B). The total number of individual evaluations exceeds 38,000, providing a solid empirical basis.

Chapter 5 "Discussion and Future Work" (pp. 155–175) systematizes the answers to the three research questions, formulates the three scientific-applied contributions, documents the limitations of the research in four groups (scope, experimental design, measurement, and causal interpretation), and outlines directions for future work. The Conclusion — Resume of the Obtained Results (pp. 176–178) consolidates the main results and the three contributions.

The formulated aim and objectives have been achieved: the PaSSER platform (Objectives 1–2), the metric framework CPS/T-CPS/Balance Score (Objective 3), and the four experimental phases (Objective 4) have been completed with concrete results. Research questions RQ1–RQ3 receive empirical answers: RQ1 is confirmed by statistically significant effects of the threshold configuration; RQ2 is confirmed by model-specific sensitivity profiles; RQ3 receives a partial answer — a sensitivity to the knowledge domain is observed, requiring calibration tailored to the specific case.

3. Scientific-applied contributions

In Section 5.2 of the dissertation (pp. 159–162), the author has formulated three scientific-applied contributions that together constitute the integrated evaluation framework. The reviewer accepts the formulations as presented:

Contribution 1 (C1): Evaluation Procedure layer (5.2.1 Threshold-aware Evaluation Procedure, p. 159). "Introduces a threshold-aware evaluation procedure incorporating Composite Performance Score (CPS), Threshold-aware Composite Performance Score (T-CPS), and Balance Score (BS) for characterizing retrieval selectivity across similarity threshold settings." [15], [16]

Contribution 2 (C2): Infrastructure layer (5.2.2 Reproducibility Infrastructure, p. 161). "Implements reproducibility infrastructure through the PaSSER platform, combining blockchain-based provenance logging with complete configuration capture." [15], [17]

Contribution 3 (C3): Evidence layer (5.2.3 Practical Guidance for Open-Source Deployments, p. 162). "Produces practical guidance for open-source RAG deployments, grounded in comparative empirical evidence linking similarity threshold sensitivity, generation quality, and deployment feasibility across seven models in the 7–8 billion parameter range under controlled experimental conditions." [16], [18]

The reviewer accepts the three scientific-applied contributions as correctly formulated and consistent with the research work carried out. The scale of more than 38,000 individual evaluations across two application domains represents one of the more extensive systematic threshold-aware evaluations of open-source LLMs in the reviewed literature.

4. Publications related to the dissertation

Five peer-reviewed publications underpin the dissertation and reflect its essential elements:

[15] I. Radeva, I. Popchev, M. Dimitrova, "Similarity thresholds in retrieval-augmented generation," in Proc. 2024 IEEE 12th Int. Conf. on Intelligent Systems (IS), 2024, pp. 1–7, doi: 10.1109/IS61756.2024.10705214. The publication supports the CPS formulation and the threshold sensitivity analysis of Chapter 4, Phase II.

[16] M. Dimitrova, I. Popchev, I. Radeva, "PaSSER: A platform for evaluating LLMs in RAG," in Proc. 2025 IEEE BdKCSE, 2025, p. 7, doi: 10.1109/BdKCSE67969.2025.11300500. The publication describes the architecture and functionalities of the PaSSER platform presented in Chapter 2.

[17] I. Radeva, I. Popchev, L. Doukovska, M. Dimitrova, "Web application for retrieval-augmented generation: Implementation and testing," Electronics, vol. 13, no. 7, p. 1361, 2024, doi: 10.3390/electronics13071361. The publication presents the PaSSER platform and the initial metric apparatus discussed in Chapter 2.

[18] I. Radeva, I. Popchev, L. Doukovska, M. Dimitrova, "Multi-agent coordination strategies vs. retrieval-augmented generation in LLMs: A comparative evaluation," Electronics, vol. 14, no. 24, p. 4883, 2025, doi: 10.3390/electronics14244883. The publication documents T-CPS and Balance Score presented in Chapter 4, Phase IV.

[20] M. Dimitrova, "Retrieval-augmented generation (RAG): Advances and challenges," Problems of Engineering Cybernetics and Robotics, vol. 83, 2025, doi: 10.7546/PECR.83.25.03. The publication presents the RAG literature review and the analysis of RAG frameworks that form the basis of Chapter 1.

The numbering of the publications follows the numbering used in the bibliography of the dissertation (pp. 192–204).

The analysis of the publication activity shows: two publications in the journal Electronics (MDPI), indexed in Web of Science with an impact factor (JCR-IF) and in Scopus (SJR); two publications in refereed proceedings of IEEE conferences indexed in IEEE Xplore; and one publication in the journal of the Institute of Information and Communication Technologies – BAS, "Problems of Engineering Cybernetics and Robotics". Publication [20] is single-authored. The "CITATION RECORD" section of the dissertation (pp. 207–212) documents 64 recorded citations across four of the five publications. The publications cover all principal aspects of the research and sufficiently test the results before the scientific community.

5. Scientometric profile

The scientometric profile of the doctoral candidate is documented on the basis of reports from Web of Science Core Collection (Clarivate Analytics, actual period 2024–2026) and Scopus/SciVal (Elsevier, actual period 2023–2025; data updated on 15.04.2026; export on 23.04.2026).

Two of the publications supporting the dissertation are indexed in Web of Science Core Collection: [17] Radeva, Popchev, Doukovska, Dimitrova, "Web Application for Retrieval-Augmented Generation: Implementation and Testing", Electronics, vol. 13, no. 7, 2024, with 13 citations and an average annual citation rate of 4.33; and [18] Radeva, Popchev, Doukovska, Dimitrova, "Multi-Agent Coordination Strategies vs. Retrieval-Augmented Generation in LLMs", Electronics, vol. 14, no. 24, 2025, with no recorded citations, which is expected for a publication from December 2025.

The summary indicators from Scopus/SciVal for the actual period 2023–2025 are presented in Table 1.

Table 1. Scientometric indicators from Scopus/SciVal for the period 2023–2025.

Indicator	Value	Reference value
Scholarly Output (total publications)	6	–
Citations (total count)	50	–
Citations per Publication	8.3	field-dependent
h-index	3	field-dependent; career-stage-dependent
h5-index	2	field-dependent
Field-Weighted Citation Impact (FWCI), overall	1.64 (median 1.22)	1.00 (world average)
Field-Weighted Citation Impact for 2024	3.70	1.00 (world average)
Outputs in Top 10% Citation Percentiles (field-weighted)	33.3% (100% for 2024)	10.0% (by definition)
Open Access publications	33.33%	approx. 35% (Scopus 2024; indicative)
Patent-Citations Count	1	0 (typical for a doctoral candidate)

Four of the presented indicators are of decisive weight for the scientometric profile of the candidate: FWCI = 1.64 (median 1.22) for the entire period and FWCI = 3.70 for 2024 document a citation impact above the world average (reference value 1.00). The share of 33.3% of publications in the top 10% most cited (field-weighted) confirms that the citation impact is not attributable to a single publication. The h-index h = 3 across six publications indicates a distribution of citations across more than one work, supporting the stability of the profile. The recorded patent citation (Patent-Citations Count = 1) is an indicator unusual for a doctoral stage, pointing to applied impact beyond the academic sphere.

The publication activity is concentrated in second-quartile journals (Electronics, MDPI, Q2) and in indexed IEEE conference proceedings. The scientometric profile suggests directions for future development: increasing the share of single-authored publications, targeting Q1 journals by CiteScore, and moving from exploratory to confirmatory statistical regimes in subsequent experimental extensions.

6. Participation in research projects

The dissertation has been developed within two research programs of the Ministry of Education and Science: the National Research Program "Smart Crop Production" (Decision of the Council of Ministers No. 866 of 26.11.2020) and the project "BG PLANTNET: Establishment of a National Information Network Genebank — Plant Genetic Resources" (KP-06-N36) of the Scientific Research Fund. Participation in these programs outlines the applied environment in which the domain corpora for agriculture and biodiversity were shaped.

7. Abstract

The abstracts are prepared in Bulgarian and English, 48 pages in volume, and present the main results, contributions, and conclusions of the dissertation. The content of the abstracts corresponds to the content of the dissertation.

8. Critical remarks and recommendations

With respect to the presentation and layout of the results, the following remarks may be made:

1. Research questions RQ1–RQ3 are empirical in nature, whereas Objectives 1–3 build the research infrastructure. The logical connection between the two groups is reflected in Table I.1 (p. 28) and may be clarified orally during the defense.
2. In several places the text contains formulations that may be read as "suggesting" validation against human evaluations (e.g., validated framework), which is not within the scope of the present work. During the oral presentation the author may clarify that the work is in a "developed and demonstrated" mode.
3. The weights of the Composite Performance Score (Table 4.3: lexical overlap 30%, semantic similarity 25%, fluency and quality 25%, language modeling 20%) are set by expert judgment, and their robustness is supported by the sensitivity analysis in publication [18]. A brief oral summary of the weight-selection rationale during the defense would assist readers not familiar with [18].
4. In Phase IV, several concurrent changes are introduced relative to Phase III — hardware configuration, question-generation approach, and knowledge domain — which complicates the isolation of the pure domain effect. This is acknowledged in Section 5.3 (Experimental Design Limitations). A brief oral clarification at the defense would strengthen the reading of the cross-domain transferability results.
5. The structure of the multiple comparisons (four models × ten threshold levels per phase) requires the p-values to be interpreted as exploratory rather than confirmatory. This distinction is drawn in Section 5.3.3 (Measurement and Analysis Limitations). A brief oral restatement of this emphasis when presenting the results from Chapter 4 would be helpful.

6. Phase I is designated in the dissertation as system validation and runtime profiling (p. 26), with no direct contribution to the answers of research questions RQ1–RQ3. This distinction is present in the text. A brief restatement during the oral presentation would help the reader correctly position the four experimental phases.
7. The three presented contributions are formulated as scientific-applied. However, pursuant to Art. 27, para. 1 of RIDASRB, the dissertation must contain results that represent an "original contribution to science." During the oral presentation, the author may clarify which of the three contributions fall into this category in her view, and on what grounds.

8. **Questions to the doctoral candidate**
   1. What are the main criteria for selecting the seven open-source language models in the 7–8 billion parameter range?
   2. On what grounds was the Antelope blockchain chosen for recording the provenance of the experimental data?
   3. What is the number of questions used in the evaluation for the agriculture and biodiversity domains, and how were they generated?
   4. What is the reason for extending the investigated threshold range from 0.50–0.80 in Phase II to 0.50–0.95 in Phases III and IV?
   5. Which of the presented scientific-applied contributions can be identified as an original contribution to science?
   6. Which of the directions for future work formulated in Section 5.4 (pp. 170–175) are considered the most immediate priorities for the continuation of the research?

CONCLUSION

The dissertation by Miroslava Doncheva Dimitrova contains scientific-applied results that represent an original contribution to science and meets the requirements of the Act on the Development of the Academic Staff in the Republic of Bulgaria, its implementing regulations, and the Regulations on the specific conditions of IICT–BAS.

Miroslava Doncheva Dimitrova possesses in-depth theoretical knowledge in the fields of natural language processing, information retrieval, and evaluation methodologies for language models, as well as the capacity for independently conducting large-scale scientific research.

I give a decidedly positive assessment of the presented dissertation, the abstract, and the results obtained.

I propose to the honorable Scientific Jury to award the educational and scientific degree "Doctor" to Miroslava Doncheva Dimitrova in professional direction 4.6. "Informatics and Computer Sciences", doctoral program "Informatics" (01.01.12).

Date: 27.04. 2026

На основание
ЗЗЛД

Reviewer:
(Prof. DC