Comparative Evaluation of Vanilla and Retrieval-Augmented Generation Approaches for a Domain-Specific Virtual Assistant in Maritime Vocational Education

Dirhamsyah[1], Tika Septia[2], Rezky Rahma Ruslan[3], Daniel Yust Rizal Maretta[4]
[1,2,3,4]*Politeknik Pelayaran Surabaya, Indonesia*
Email: dirhamsyah1515@gmail.com

ABSTRACT

Deploying Large Language Models (LLMs) as vocational educational assistants poses risks, notably hallucination and inventing answers beyond local teaching materials. Literature often highlights Retrieval-Augmented Generation (RAG) as a mitigation, but empirical comparisons, especially in Indonesian maritime vocational institutions, are scarce. This research addresses that gap by testing two Virtual Boiler Assistant versions at Politeknik Pelayaran Surabaya: a Vanilla Gemini 2.5 Flash configuration and a RAG setup using the same model with semantic retrieval from 192 knowledge chunks (from four official boiler documents). Testing employed 20 validated ground-truth questions across procedure, calculation, and logic categories, with both models assessed under identical conditions using a three-point rubric and source cross-validation. The Vanilla model scored 50.0% (20/40), while RAG achieved 97.5% (39/40), demonstrating a consistent 47.5-point improvement. Hallucinations in Vanilla included numerical and procedural fabrications, and misinterpretation of the local "CBT" acronym. RAG responses cited sources in 85% of cases versus none from Vanilla. These findings show RAG substantially improves accuracy, hallucination resistance, and verifiability in LLM-based maritime vocational education, especially where local terminology diverges from broader industry usage.

Keywords : retrieval-augmented generation; large language model; hallucination; maritime education; comparative evaluation

Introduction

The deployment of Large Language Models (LLMs) as educational assistants has accelerated rapidly across higher education and vocational training following the public release of conversational AI systems such as ChatGPT in late 2022 (Kasneci et al., 2023). In Indonesian maritime vocational education, conducted through specialized polytechnic institutions under the Ministry of Transportation, the potential of such systems to provide on-demand access to technical knowledge addresses a long-standing constraint: cadets need supplementary explanation of complex shipboard systems outside scheduled laboratory hours, which traditional instructional resources cannot reliably provide (Wahyuwono et al., 2024). However, the unmodified deployment of foundation LLMs in specialized technical domains carries documented risks. Hallucination, the phenomenon in which language models generate plausible-sounding but factually incorrect content, has been characterized as an inherent property of probabilistic text generation rather than a fixable defect (Ji et al., 2023; Huang et

al., 2025). For educational deployment, hallucinated content is especially problematic because learners lack the prior knowledge needed to detect errors and may consequently internalize incorrect information.

Retrieval-Augmented Generation (RAG), introduced by Lewis et al. (2020) and refined in subsequent work (Gao et al., 2024), offers a structural mitigation strategy. Rather than relying solely on parametric knowledge encoded in model weights during pretraining, RAG architectures retrieve relevant documents from an external knowledge base and inject the retrieved content as context into the generation prompt. The model thus generates responses grounded in identifiable source material, enabling source-citation behavior and reducing reliance on potentially incorrect parametric memory. Theoretical and empirical work has consistently reported accuracy improvements and hallucination reduction from RAG across various domains (Ram et al., 2023; Asai et al., 2024). However, comparative evaluation specific to Indonesian maritime vocational education has not been previously published, and there is reason to expect that the benefits of RAG may be especially pronounced in this context. Maritime technical domains employ specialized terminology that often diverges between local Indonesian usage and globally dominant English industry usage, creating elevated risk of contextual misalignment in Vanilla LLM responses.

This study addresses that gap by presenting a controlled comparative evaluation of two configurations of a Virtual Boiler Assistant deployed at Politeknik Pelayaran Surabaya. The first configuration represents Vanilla deployment, in which user queries are submitted directly to Gemini 2.5 Flash without retrieval augmentation. The second configuration represents domain-specific RAG, in which the same model receives queries augmented with semantically retrieved chunks from a knowledge base constructed from official boiler instructional documents used at the institution. Both configurations were evaluated using the same 20 validated ground-truth questions and the same scoring rubric, providing a within-subject experimental design that controls for question difficulty and rubric variation.

Three research questions guide the analysis: (1) What is the magnitude of the accuracy gap between Vanilla and RAG configurations on validated ground-truth questions in the maritime boiler domain? (2) What types of hallucination, if any, appear in the Vanilla configuration that are absent or attenuated in the RAG configuration? (3) How do response characteristics including source-citation behavior, response length, and response time differ between the two configurations? The contribution of this work is threefold: it provides empirical evidence on RAG effectiveness in an under-studied vocational domain; it documents specific hallucination patterns that arise when LLMs encounter Indonesian-language technical queries with locally idiosyncratic terminology; and it offers a replicable evaluation protocol that other vocational institutions can adapt for their own domain-specific assistant deployments.

Research Method

This research employed a within-subject comparative experimental design (Field, 2018) in which two system configurations served as the independent variable and response quality served as the dependent variable. The within-subject design controls for variation in question difficulty and rubric application by ensuring that both configurations responded to the identical

set of ground-truth questions under identical experimental conditions. The order of configuration evaluation was fixed (Vanilla first, RAG second) within each question to ensure comparable system load conditions. A scripted experimental procedure ensured reproducibility, with all queries, system parameters, and elapsed times logged programmatically.

System Configurations

Both configurations shared a common backend developed in FastAPI (Python 3.12) and deployed on a Virtual Private Server running Ubuntu 24.04. The generative model in both configurations was Gemini 2.5 Flash accessed through the official Google AI API, with identical generation parameters including temperature and maximum output length. The configurations differed only in the presence or absence of retrieval augmentation. The Vanilla configuration submitted each user query directly to the LLM with a system prompt establishing the assistant's role as a maritime boiler tutor for Politeknik Pelayaran Surabaya cadets, but without any retrieved context.

The RAG configuration extended the Vanilla pipeline with a semantic retrieval stage. The knowledge base comprised 192 text chunks extracted from four official boiler instructional documents used at the polytechnic's steam plant laboratory and Computer-Based Training simulator: the Steam Plant Laboratory Operating Manual, the CBT Pengoperasian Mesin Kapal module, the Boiler Fuel System CBT module, and the Engine Cargo Ship CBT module. Each chunk was embedded using the Gemini Embedding-001 model and indexed in a local ChromaDB vector store. For each user query, the RAG configuration retrieved the four chunks with highest cosine similarity to the query embedding ($k = 4$) and injected them as context into the LLM prompt with instructions to generate responses grounded in the retrieved content and to cite source filename and page number when applicable.

Ground-Truth Question Set

Twenty ground-truth questions were developed based on systematic content analysis of the four source documents. Question development was performed by the research team and validated against the source documents to ensure that each question had a verifiable reference answer derivable from the corpus. The question set was structured to cover three categories that probe distinct aspects of system capability, as summarized in Table 1.

Table 1. Distribution of Ground-Truth Questions Across Categories

No	Category	Number of Items	Description
1	Procedural (UI/Lab specific)	6	Questions requiring concrete steps in the simulation interface or laboratory equipment of Politeknik Pelayaran Surabaya
2	Numerical-Specific	3	Questions requiring specific numerical values such as water level, temperature, duration, or pressure
3	Conceptual	11	Questions on definitions, functions, and mechanisms of components — broader in scope and not strictly tied to local UI
	Total	**20**	

The procedural category contains questions that require specific reference to the user interface or laboratory equipment used at the polytechnic, such as the sequence of buttons in the CBT simulator startup procedure. The numerical-specific category contains questions requiring concrete numerical values such as the staged heating temperatures (45, 60, 75, 90 degrees Celsius) documented in the laboratory manual. The conceptual category contains broader questions on component functions and operating principles that, while answerable in general terms by any LLM with maritime engineering knowledge, can be answered with greater fidelity to local terminology when grounded in the source documents.

Experimental Procedure

A Python evaluation script (compare_rag_vs_vanilla.py) was developed to automate the experimental procedure. For each of the twenty questions, the script invoked both configurations sequentially and logged: the question identifier, question text, Vanilla response text, RAG response text, RAG semantic match score (cosine similarity of the highest-ranked retrieved chunk), Vanilla response time in milliseconds, and RAG response time in milliseconds. A five-second delay was inserted between question pairs to avoid API rate-limit triggering. All responses were stored in a structured JSON file (rag_vs_vanilla_results.json) for subsequent scoring and analysis.

Automated descriptive metrics were computed alongside the response logs. These included: mean response time per configuration, mean response length in characters, count of responses containing explicit acknowledgment of knowledge gaps using lexical patterns such as “tidak tahu” or “tidak memiliki informasi spesifik,” and count of responses containing source-citation patterns identified by regex matching the strings “hal.” followed by a digit or filenames containing “.pdf.”

Scoring Rubric

Each response was scored against the reference answer using a three-point rubric: a score of 0 indicated an inadequate response containing serious factual errors, hallucinated content, or contradiction of the source documents; a score of 1 indicated a partially adequate response that was generally valid in its broader claims but lacked grounding in Politeknik Pelayaran Surabaya documentation, omitted source citation, or contained minor hallucinated elements; and a score of 2 indicated a fully adequate response that either answered the question correctly with reference to the source documents or, in cases where the requested information was not present in the corpus, explicitly acknowledged the gap before supplementing with general domain knowledge clearly labeled as such. The maximum total score across 20 questions is 40 points. Scoring was performed by reviewers familiar with the source documents and the Poltekpel boiler training context.

Results and Discussion

Aggregate Performance

Aggregate scoring results revealed a substantial accuracy gap between the two configurations. The Vanilla configuration achieved a total score of 20 out of 40 (50.0%), while the RAG configuration achieved 39 out of 40 (97.5%), corresponding to a difference of 47.5 percentage points. Detailed scoring outcomes are presented in Table 2.

Table 2. Aggregate Scoring Results

Metric	Vanilla LLM	RAG	Difference
Total raw score (out of 40)	20	39	**+19 points**
Percentage score	50.0%	97.5%	**+47.5 pp**
Items scoring 2 (fully adequate)	0 / 20	19 / 20	+19 items
Items scoring 1 (partially adequate)	20 / 20	1 / 20	–19 items
Items scoring 0 (inadequate)	0 / 20	0 / 20	0

Notably, no responses from either configuration received a score of 0, indicating that both systems provided responses that were at least partially relevant to the questions asked. The systematic difference between configurations therefore lies not in the production of grossly incorrect content but in the proportion of responses that achieve full grounding in the source documents. The Vanilla configuration produced responses that consistently received a partial-adequacy score of 1, reflecting their generic technical correctness combined with absence of local grounding. The RAG configuration achieved fully adequate scores in 19 of 20 cases, with a single exception explained in subsequent discussion.

Performance by Question Category

To investigate whether RAG provides differential benefit across question types, scoring outcomes were aggregated by category. Results are presented in Table 3.

Table 3. Scoring Results by Question Category

Question Category	Items	Vanilla (%)	RAG (%)	Gap (pp)
Procedural (UI/Lab specific)	6	50.0	91.7	**+41.7**
Numerical-Specific	3	50.0	100.0	**+50.0**
Conceptual	11	50.0	100.0	**+50.0**
Overall	**20**	**50.0**	**97.5**	**+47.5**

The smallest absolute gap, 41.7 percentage points, occurred in the procedural category. This gap was driven primarily by a single RAG response (Q1) that was truncated mid-procedure due to apparent token-budget exhaustion when the retrieved context window was large. The remaining five procedural questions were answered fully correctly by the RAG configuration. Both the numerical-specific and conceptual categories exhibited a 50.0 percentage point gap, with the RAG configuration achieving perfect scores. The conceptual category result was particularly informative: although these questions concerned broader theoretical content theoretically answerable by any LLM with maritime engineering training, the RAG configuration consistently used the specific nomenclature of the Politeknik Pelayaran Surabaya modules (such as Bar Inlet, Bar Outlet, Cold Filter 1, Pre heater Steam, Bar Discharger, and Fuel Oil Supply Unit), while the Vanilla configuration produced text-book responses using globally standard terminology that did not match the modules used by the institution.

Response Time and Length Characteristics

Automated logging of response characteristics yielded several findings that contradicted intuitive expectations. Selected response characteristics are presented in Table 4.

Table 4. Response Characteristics by Configuration

Characteristic	Vanilla LLM	RAG
Mean response time (ms)	19,929	11,525
Mean response length (characters)	2,631	2,147
Responses with explicit "do not know" acknowledgment	1 / 20 (5%)	3 / 20 (15%)
Responses with source-citation pattern (page or filename)	0 / 20 (0%)	17 / 20 (85%)

Contrary to the intuitive assumption that RAG, which involves an additional retrieval step, would be slower than Vanilla, the RAG configuration was observed to be approximately 42% faster on average (11,525 ms versus 19,929 ms). This difference is attributable to the fact that the Vanilla configuration tended to generate substantially longer responses, with the model producing extensive prefatory material, multi-paragraph theoretical exposition, and numerous bulleted lists in attempting to comprehensively address questions in the absence of retrieved context. The RAG configuration, in contrast, was anchored by retrieved context and produced more focused responses (mean of 2,147 characters versus 2,631 characters for Vanilla). The faster response time of the RAG configuration is therefore a generation-length effect rather than evidence that retrieval imposes negligible computational cost.

The most striking response characteristic was source-citation behavior. The RAG configuration produced source-citation patterns identifiable by automated regex matching in 17 of 20 responses (85%). Manual review confirmed that the remaining three RAG responses also referenced source material in alternative phrasing not captured by the regex. The Vanilla configuration produced no source-citation patterns in any of the 20 responses, regardless of the regex pattern or the manual review approach. This difference of 85 percentage points in source-citation rate represents a categorical difference in epistemological behavior: RAG responses present claims with verifiable provenance, while Vanilla responses present claims as undifferentiated assertions.

Hallucination Patterns in the Vanilla Configuration

Three distinct hallucination patterns were identified across the Vanilla responses. These patterns are summarized in Table 5 with comparison to the corresponding RAG responses.

Table 5. Documented Hallucination Cases in the Vanilla Configuration

ID	Hallucination Type	Vanilla Output (Erroneous)	RAG Output (Grounded)
Q4	Numerical fabrication	Stated that "pressure should drop near 0.5–1 bar" during shutdown—a value not present anywhere in the source documents.	Provided seven shutdown steps drawn directly from the source documents with explicit page citations.
Q9	Acronym misinterpretation	Interpreted CBT as "Continuous Blowdown Tank" and produced an extensive explanation of an industrial drum-blowdown circuit, which has no	Correctly interpreted CBT as "Computer-Based Training" in the polytechnic context, explained drum blowdown as a procedure within the

ID	Hallucination Type	Vanilla Output (Erroneous)	RAG Output (Grounded)
		relation to Poltekpel's training environment.	training simulator, and explicitly noted when specific procedural detail was absent from the corpus.
Q19	Procedural fabrication	Described commercial-vessel components such as deaerator tank, forced-draft fan, main steam stop valve, and air vent valve—none of which appear in the Poltekpel CBT simulator interface.	Provided the exact UI sequence used in the simulator: click SP panel → SAFETY RESET → MAKE UP feed water → set feed pump 1 to AUTO → drum blow.

The Q9 case, classified as acronym misinterpretation, warrants particular attention as it represents a deployment risk that is structurally invisible without retrieval. The acronym CBT carries a globally dominant industrial meaning of "Continuous Blowdown Tank" in commercial boiler engineering, while in the educational context of Politeknik Pelayaran Surabaya it consistently refers to "Computer-Based Training," the polytechnic's simulator system. The Vanilla configuration, basing its response on the global probability distribution of acronym meanings encoded in its training data, selected the industrial interpretation and produced an extensive technically accurate explanation of an industrial drum-blowdown circuit. This response is technically correct in its own terms but completely irrelevant to the cadet's actual learning context. Critically, the Vanilla response provided no warning to the user that an acronym disambiguation had been performed; from the cadet's perspective, the response appears authoritative.

This pattern echoes findings in the broader hallucination literature where LLMs select interpretations consistent with global training-data frequencies even when local context would imply a different interpretation (Ji et al., 2023). It is especially problematic in maritime vocational education in Indonesia, where many technical acronyms have local educational meanings that differ from the dominant global industrial meanings. The RAG configuration was protected from this failure mode because the retrieved context contained explicit references to the polytechnic's CBT simulator interface, which steered the language model toward the locally appropriate interpretation.

The Q4 case (numerical fabrication) and the Q19 case (procedural fabrication) follow a similar pattern: in the absence of retrieved context, the Vanilla configuration filled in missing detail using its parametric knowledge of commercial-vessel boiler operation, producing content that was internally consistent and technically plausible but did not correspond to the specific equipment or procedures of the Politeknik Pelayaran Surabaya laboratory. The RAG configuration in both cases responded with the specific procedure or value documented in the source corpus. These cases illustrate that hallucination in domain-specific deployment is not merely the production of fabricated facts in the abstract, but specifically the production of content that may be true elsewhere yet false in the deployment context.

Source Citation and Vocabulary Alignment

The 85% source-citation rate in the RAG configuration constitutes more than a stylistic feature; it provides a verification mechanism that supports pedagogically appropriate use of the assistant. When a cadet receives a RAG response stating that water level is adequate when the

gauge glass shows the level at the middle of the glass, accompanied by a reference to LABORATORIUM STEAM PLANT.pdf page 12, the cadet can locate the corresponding passage in the source document and verify the claim. This mechanism transforms the AI assistant from a substitute for the source documents into an efficient access pathway to those documents. Such verifiability is consistent with evidence-based learning principles widely advocated in vocational education research (Boud & Falchikov, 2007; Hattie & Timperley, 2007).

Vocabulary alignment, defined as the consistency between assistant terminology and the terminology used in official course materials, emerged as an additional benefit of RAG that was not anticipated in the experimental design. Components such as Bar Inlet, Bar Outlet, Cold Filter 1, Pre heater Steam, Bar Discharger, and Fuel Oil Supply Unit appeared naturally in RAG responses to conceptual questions. These terms are not standard global industrial nomenclature and therefore do not appear in Vanilla responses, which instead use the corresponding general terminology (storage tank, settling tank, daily service tank, fuel oil heater, and similar terms). When cadets are subsequently assessed using the same nomenclature found in the modules, vocabulary alignment between the assistant and the assessment becomes pedagogically essential rather than cosmetic.

Implications for Educational LLM Deployment

The findings carry several implications for the deployment of LLM-based educational assistants in Indonesian vocational institutions. First, Vanilla deployment of foundation LLMs without retrieval grounding is not advisable for institution-specific instructional support, regardless of the global capability level of the underlying model. The 50.0% accuracy ceiling observed for Vanilla in this study is not a function of model capability but of the fundamental absence of access to local instructional materials at inference time. Even substantially more capable models would face the same upper bound when local context is absent. Second, the value of RAG in domain-specific educational deployment is not limited to numerical precision or factual accuracy as commonly emphasized; it extends to vocabulary alignment, citation verifiability, and resilience against acronym ambiguity. These properties are particularly relevant for vocational education contexts where local terminology and procedure are heavily institution-specific.

Third, the cost of constructing the RAG knowledge base in this study was modest. Four official documents already in use at the institution were converted, chunked, and embedded using standard tools, yielding a usable knowledge base of 192 chunks. This contrasts with the substantial human effort required to fine-tune a model on equivalent content. For institutions with constrained technical and financial resources, RAG offers a high-leverage path to domain customization without modifying model weights. The replicability of this construction process, using documents that institutions typically already maintain, supports the generalizability of the approach to other Indonesian maritime polytechnics and to vocational institutions more broadly.

Limitations

Several limitations of this study should be acknowledged. First, the evaluation involved 20 ground-truth questions; expansion to a larger question pool would strengthen the generalizability of the accuracy estimates. Second, scoring was performed by reviewers familiar

with the source documents but did not include independent inter-rater reliability assessment with multiple instructors from the Marine Engineering department; future work will incorporate such assessment using Cohen's kappa or Krippendorff's alpha. Third, this study did not evaluate cadet learning outcomes from interaction with either configuration; the gap between accuracy as measured here and pedagogical effectiveness as measured by learning gains is an important distinction that warrants subsequent investigation. Fourth, the RAG configuration occasionally truncated responses on complex procedural questions, suggesting that token-budget tuning and chunking strategy refinement could yield further accuracy gains. Fifth, the experimental conditions used a single LLM (Gemini 2.5 Flash); the magnitude of the Vanilla–RAG gap may vary with other foundation models, particularly those with substantially different training-data composition or different parametric capacity.

Conclusion

This study presented a controlled comparative evaluation of Vanilla and Retrieval-Augmented Generation configurations of a Virtual Boiler Assistant deployed at Politeknik Pelayaran Surabaya. Twenty validated ground-truth questions covering procedural, numerical-specific, and conceptual content were submitted to both configurations under identical conditions and scored against reference answers derived from the four official source documents. The Vanilla configuration achieved 50.0% accuracy while the RAG configuration achieved 97.5% accuracy, a 47.5 percentage point gap that remained consistent across question categories. Three hallucination patterns were documented in the Vanilla configuration: numerical fabrication, acronym misinterpretation, and procedural fabrication. The acronym misinterpretation case (CBT interpreted as Continuous Blowdown Tank rather than Computer-Based Training) is particularly noteworthy as a category of failure that is structurally invisible without local context retrieval. Source-citation patterns appeared in 85% of RAG responses but in zero Vanilla responses, demonstrating a categorical difference in verifiability. The RAG configuration was also faster on average and produced more focused responses anchored to retrieved context.

The contribution of this work lies in providing the first published comparative evidence on RAG effectiveness for Indonesian maritime vocational education and in documenting specific hallucination patterns that arise when LLMs encounter Indonesian-language technical queries with locally idiosyncratic terminology. The findings support the conclusion that for educational deployment in specialized vocational domains, retrieval grounding using official institutional documents is essential rather than optional. Future work will expand the evaluation question pool, incorporate independent inter-rater reliability assessment, conduct learning-outcome studies with cadets, and extend the comparison to additional foundation models to test the generalizability of the observed gap..

References

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *Proceedings of the 12th International*

Conference on Learning Representations (ICLR 2024). https://openreview.net/forum?id=hSyW5go0v8

Boud, D., & Falchikov, N. (Eds.). (2007). *Rethinking assessment in higher education: Learning for the longer term.* Routledge. https://doi.org/10.4324/9780203964309

Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997.* https://arxiv.org/abs/2312.10997

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), Article 42. https://doi.org/10.1145/3703155

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. https://doi.org/10.1145/3571730

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. https://doi.org/10.1016/j.lindif.2023.102274

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33* (pp. 9459–9474). https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html

Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., & Shoham, Y. (2023). In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics, 11*, 1316–1331. https://doi.org/10.1162/tacl_a_00605

Wahyuwono, A., Oktavia, R., & Kartikasari, D. (2024). Virtual assistant berbasis web menggunakan NLP dan LSTM untuk layanan informasi akademik. *Jurnal Teknologi Pendidikan, 9*(2), 201–215.