

# Reclaiming Pedagogical Control in DSP Education through Hybrid-RAG-Enhanced Local LLM

Pedro Canavieira, Thiago Nisimura, Mariana Ribeiro,  
Frank B. Morte, Claudio Klautau, Lucas Matni and Aldebaro Klautau

**Abstract**—The growing use of LLMs by students as a substitute for textbooks introduces important pedagogical challenges. Many LLM-generated answers do not follow the notation, methodology, or conceptual approach adopted by instructors, and may even contain errors. This shifts educational influence from professors to Big Tech platforms that define model behavior and collect interaction data. Furthermore, they rely on online services that can lead to increased latency. This paper presents both a quantitative and qualitative evaluation of this problem and proposes an open-source AI chatbot tutor based on locally hosted LLMs and a Hybrid Retrieval-Augmented Generation (Hybrid RAG). The system uses course-specific material about digital signal processing created by the authors, preserving privacy and enabling cost-controlled deployment within institutional infrastructure.

**Keywords**—DSP education, educational chatbot, local LLM, intelligent tutoring system, RAG.

## I. INTRODUCTION

Commercial Large Language Model (LLM) platforms such as DeepSeek, Gemini, and ChatGPT are increasingly being used by students as substitutes for textbooks and traditional study materials. Although these systems provide immediate access to explanations and problem-solving assistance, their responses frequently diverge from the notation, definitions, methodologies, and pedagogical approaches adopted by instructors. In technical disciplines such as Digital Signal Processing (DSP), this inconsistency can generate substantial learning difficulties, since mathematically equivalent formulations may still confuse beginners or conflict with the methodology explained in class. Moreover, LLMs may produce incorrect or hallucinated explanations, especially in analytical domains involving mathematics and signal processing [1], [2].

Another important concern is that most widely used LLMs are proprietary cloud-based systems operated by large technology companies. This creates privacy and sovereignty issues, since students' interactions, academic materials, and institutional data may be processed externally under opaque retention policies, without being shared with instructors. These challenges motivate the development of educational assistants based on locally hosted open-source LLMs, capable of operating inside institutional infrastructure while preserving privacy and enabling cost control [3].

However, simply replacing commercial Application Programming Interfaces (APIs) with local models is insufficient. Open-source LLMs typically exhibit lower reasoning capabilities, particularly in mathematical and multilingual tasks [4]. In addition, the Portuguese language introduces further difficulties due to eventual poorer tokenization efficiency and reduced representation during model training [5]. To mitigate hallucinations and enforce adherence to instructor-provided materials, Retrieval-Augmented Generation (RAG) has emerged as a promising strategy for grounding answers in authoritative references [6].

This paper presents ChatBot-o, an open-source educational chatbot for DSP education based on locally hosted LLMs and the AcademiCK framework [7]. The system extends to support fully local execution using Virtual Large Language Model (vLLM)<sup>1</sup>, eliminating dependence on external APIs while preserving conversational usability.

One goal of this work is to assess the viability of local, open-source systems as a technically competitive, sovereign, and linguistically hybrid alternative for higher education.

Using a DSP course as a case study, we quantitatively evaluate latency and throughput under different combinations of language and RAG activation, and qualitatively analyze how the Hybrid Retrieval-Augmented Generation (Hybrid RAG) implemented in [7] improves adherence to textbook methodology and reduces inconsistencies in LLM-generated responses, including deviations of a specific method to solve a problem suggested by the instructor.

The main contributions of this paper are:

- 1) The development of an open-source chatbot based on the AcademiCK framework, uniquely adapted for local execution on dedicated hardware.
- 2) A quantitative benchmark comparing commercial and local models regarding computational latency across different execution modes combining the interface language (English and Brazilian Portuguese) and Hybrid RAG activation within the context of DSP.
- 3) A qualitative evaluation of LLM usage in the DSP learning process, demonstrating how the proposed Hybrid RAG system aligns the generated responses with the desired instructor's methodology.

## II. PROPOSED SYSTEM: CHATBOT-O

This section describes the Intelligent Tutoring System (ITS) module called ChatBot-o, in reference to the Amazon river

P. Canavieira, T. Nisimura, M. Ribeiro, F. B. Morte, C. Klautau, L. Matni and A. Klautau are with UFPA, Belém-PA, Brazil. E-mails: pedro.canavieira@icen.ufpa.br, thiago.nisimura@icen.ufpa.br, mariana.ribeiro@itec.ufpa.br, frank.morte@itec.ufpa.br, claudio.mello@itec.ufpa.br, lucas.matni@itec.ufpa.br, aldebaro@ufpa.br.

<sup>1</sup>Available at: <https://docs.vllm.ai/en/latest/>, Accessed on May 25, 2026.

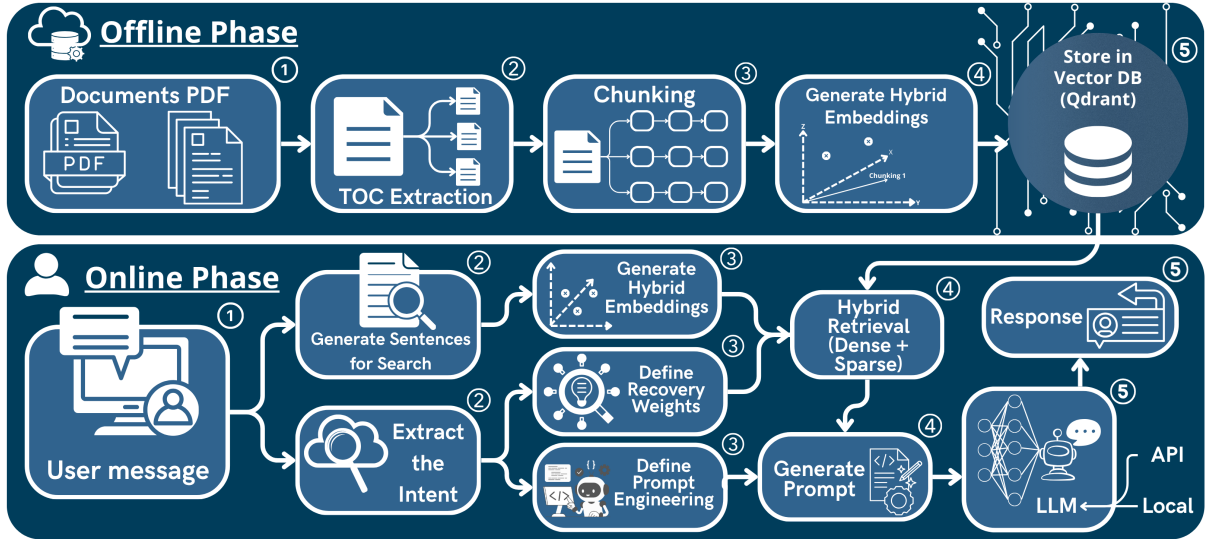


Fig. 1: Proposed Intelligent Tutoring System (ITS) architecture based on AcademiCK.

dolphin (*Inia geoffrensis*). This ITS is based on the open-source assistant AcademiCK, which is detailed in [7]. The next paragraphs provide a short overview of the AcademiCK architecture and how it is used in ChatBot-o.

As illustrated in Fig. 1, AcademiCK operates in two operational stages: an *offline* phase, responsible for the processing and indexing of academic materials, and an *online* phase, dedicated to real-time interaction and response generation. Both phases executes a five-step flow (1–5).

In the *offline* phase, the flow begins with the loading of PDF documents (Step 1), followed by the extraction of the structure and table of contents (*TOC Extraction*) in Step 2 to identify chapters, sections, and semantic divisions.

The content is then segmented via *embeddings* into semantically coherent blocks with an overlap between adjacent segments (Step 3). This strategy preserves contextual continuity and reduces information loss at boundaries [8]. Subsequently, hybrid *embeddings* are generated using the **BGE-M3** model (Step 4), which simultaneously extracts dense (deep meaning) and sparse (acronyms and technical terms) representations. Finally, at Step 5, the *chunks* and their respective vectors are stored in a vector database (**Qdrant**), while metadata, user data, and session logs are managed by a relational database (**PostgreSQL**).

The *online* phase begins with the input of a user message (Step 1), triggering the parallel execution of sentence generation for semantic search and intent extraction (e.g., summarization, explanation, or code) at Step 2. Thereafter, three complementary processes are executed (Step 3): the generation of hybrid *embeddings* for the query (BGE-M3), the definition of retrieval weights, and the configuration of the tutor *prompt* instructions. This use of dynamic weights aligns with hybrid approaches that improve precision in technical domains[8]. The pipeline, then, advances to the execution of a hybrid search (*Hybrid Retrieval*) in Qdrant (Step 4), combining semantic proximity and keyword matching metrics to reduce ambiguities and mitigate hallucinations.

As the final operation (Step 5), the *prompt* and the retrieved context are sent to the LLM responsible for the final response.

It is at this stage that the main contribution of this work over the original framework [7] is implemented: the adaptation of AcademiCK to support local *open-source* models, eliminating reliance on proprietary APIs and enhancing data privacy and cost-efficiency. In the following sections, we provide a benchmark of different ITS configurations, and a qualitative evaluation of answers to selected questions about DSP.

### III. BENCHMARKING MODELS AND ASSOCIATED STAGES

This benchmark evaluates the feasibility of ITS based on local LLMs compared to commercial LLMs, considering response time [9]. The test utilized 59 DSP questions<sup>2</sup> extracted from [10]. AcademiCK was configured with a Hybrid RAG architecture (BGE-M3), and the augmented information is based exclusively on the textbook material from [10] to avoid licensing issues. All local computations were performed on an NVIDIA RTX A5000 GPU, Ampere architecture, running CUDA 13.2. Ministral-3-8B was selected due to its proven effectiveness in ITS tasks [8]. We also use GPT-5 and DeepSeek-V4-Pro as commercial baselines.

Inference data was collected via streaming<sup>3</sup>, extracting three primary metrics: (1) **total latency** (time between sending the request and generating the last token); (2) **time to first token (TTFT)** (time until the first token is produced); and (3) **tokens out** (total volume generated). To ensure a fair comparison of output volume across different architectures, we standardized the count by applying the same tokenizer (*cl100k\_base*, available via *pip*) to the results of all models.

The goal of this benchmark is to assess ITS configurations regarding their fluidity and usability in real-time conversational interfaces. We also evaluate whether using English (EN) or Brazilian Portuguese (BP) has an impact on aspects such as

<sup>2</sup>Originally 60 questions; one was removed due to convergence failure in the Ministral-3-8b and DeepSeek V4 models (the LLMs generated text endlessly), indicating instability in both local and commercial models.

<sup>3</sup>In streaming mode, the response is sent incrementally via Server-Sent Events (SSE), transmitting each token as soon as it is generated, instead of waiting for the complete sequence. This approach was applied in all scenarios to simulate real-world usage and allows accurate measurement of the adopted metrics.

RAG and response time. In this paper, we did not evaluate quantitatively the accuracy of the ITS response, but in Section IV we qualitatively discuss selected answers.

Hence, the evaluation of three models, with the BP or EN as the interface language, and using or not RAG led to 12 distinct configurations. Descriptive statistics of the results in terms of mean and median are summarized in Table I.

TABLE I: Benchmark results.

| Mode   | Model           | Total latency in seconds |        | TTFT in seconds |        | Tokens out |        |
|--------|-----------------|--------------------------|--------|-----------------|--------|------------|--------|
|        |                 | Mean                     | Median | Mean            | Median | Mean       | Median |
| BP     | Ministral3-8B   | 27.66                    | 17.19  | 0.04            | 0.04   | 1868.71    | 1182.0 |
|        | GPT-5           | 37.70                    | 25.06  | 33.60           | 22.12  | 362.64     | 275.0  |
|        | DeepSeek-V4-Pro | 66.61                    | 35.17  | 58.07           | 23.52  | 559.39     | 519.0  |
| EN     | Ministral3-8B   | 37.12                    | 16.26  | 0.04            | 0.04   | 2387.02    | 1196.0 |
|        | GPT-5           | 33.30                    | 22.21  | 30.49           | 20.45  | 292.29     | 232.0  |
|        | DeepSeek-V4-Pro | 47.87                    | 31.57  | 40.83           | 23.92  | 476.39     | 435.0  |
| BP-RAG | Ministral3-8B   | 22.32                    | 20.65  | 5.68            | 4.46   | 1102.73    | 1074.0 |
|        | GPT-5           | 79.51                    | 75.78  | 69.26           | 67.66  | 1076.76    | 1034.0 |
|        | DeepSeek-V4-Pro | 70.64                    | 65.67  | 55.44           | 47.95  | 802.14     | 794.0  |
| EN-RAG | Ministral3-8B   | 20.27                    | 17.95  | 4.00            | 3.81   | 1041.68    | 905.0  |
|        | GPT-5           | 86.47                    | 81.76  | 74.91           | 74.79  | 901.29     | 887.0  |
|        | DeepSeek-V4-Pro | 58.21                    | 53.75  | 48.44           | 46.06  | 618.39     | 542.0  |

From the descriptive measures summarized in Table I, it is observed that the TTFT values of the Ministral-3-8B model (for texts without RAG, in both English and Portuguese) are substantially lower than those of the commercial models. This discrepancy is justified by the presence of the *Chain-of-thought* (CoT) mechanism [11] in the evaluated commercial LLMs. Since CoT processing occurs prior to the generation of the first token output, it prolongs the TTFT phase in these models, regardless of the language or the application of RAG. In contrast, Ministral-3-8B begins generation more rapidly as it is not equipped with this iterative reasoning architecture. Another notable aspect is verbosity: Ministral-3-8B proved to be the most verbose model across all scenarios. It is noteworthy that a local model with 8 billion parameters managed to generate a higher volume of text than commercial LLMs (estimated at around 1.5 trillion parameters) for this specific ITS task in DSP.

In the scenario with RAG integration, an increase in TTFT is observed in most configurations. As discussed in Section III, this phenomenon is expected, as the hybrid retrieval step in the database occurs prior to text generation, adding to the TTFT. However, this initial increase does not appear to drastically impact the total latency, since the overall response time is heavily conditioned by a fourth metric: (4) **Inter-Token Latency (ITL)**. ITL represents the accumulated time between the successive emissions of all output tokens; thus, the total system latency can be modeled by the following relationship:

$$\text{Total Latency} = \text{TTFT} + \text{ITL} \quad (1)$$

To facilitate a more in-depth comparison of the computational performance across the architectures, the (5) **Tokens Per Second (TPS)** metric, or *throughput*, was adopted, given by:

$$\text{TPS} = \frac{\text{Tokens Out}}{\text{ITL}}. \quad (2)$$

TPS is one of the primary metrics for defining the model's continuous generation speed, serving as the determining factor

for ensuring fluidity and usability in real-time conversational interfaces. Finally, for the statistical evaluation of these speeds across the models, the Shapiro-Wilk normality test was initially conducted. Due to the presence of deviations from normality ( $p < 0.1$ )<sup>4</sup> in the majority of the analyzed distributions, the non-parametric Mann-Whitney test was employed to infer significant differences in TPS.

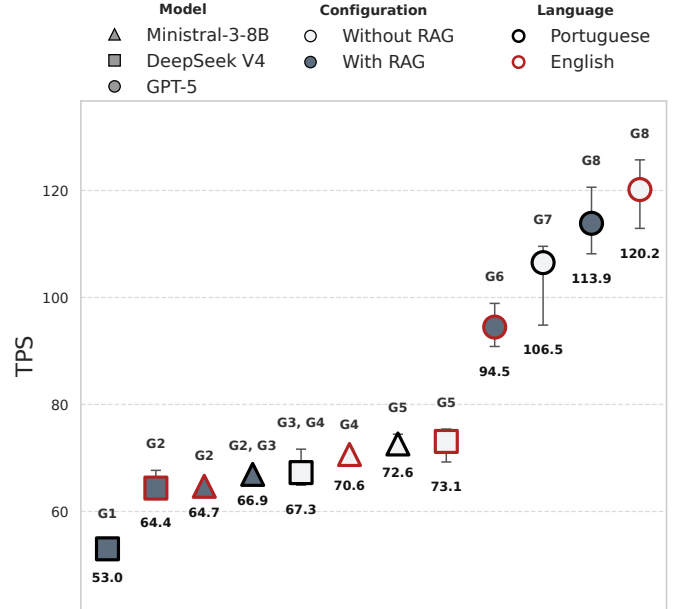


Fig. 2: Median value of TPS for each model and language, with or without RAG. Labels G1 to G8 indicate statistically equivalent groupings based on the Mann-Whitney test.

To understand the performance hierarchy of the configurations, Figure 2 illustrates clear stratifications of TPS, where the error bars represent 90% confidence intervals for the median, estimated via *bootstrap* resampling using the percentile method. The analysis relies on statistical groupings (labeled G1 to G8) derived from the Mann-Whitney test ( $p < 0.1$ ). The rule of interpretation for this labeling posits that configurations sharing at least one common label do not exhibit a statistically significant difference in their median TPS. The data import and experimental analysis codes will be available at an github repository [12].

Analysis of statistical clusters (G1 to G8) reveals three clear stratifications of TPS among the variations in interactions between architecture, RAG usage, and language. At the top of the hierarchy (G6 to G8), GPT-5 dominates exclusively. The computational ceiling (G8) represents a statistical tie between the scenario without RAG in English (120.2 TPS) and, anomalously, with RAG in Portuguese (113.9 TPS). This is an exception to the rule, as the introduction of RAG generally penalizes generation, as occurred with GPT-5's own English output (which dropped from G8 to G6).

In the intermediate range (G2 to G5), Ministral-3-8B and DeepSeek-V4-Pro are concentrated. Without RAG, both mod-

<sup>4</sup>Due to the small sample size ( $n = 59$ ), the significance level was set to 0.1 to reduce the probability of Type II error and preserve statistical power.

els reach their peaks in group G5 (73 TPS). However, the introduction of hybrid retrieval causes a consistent degradation in both, lowering them to groupings G2 and G3 (64-66 TPS). Regarding language, these architectures exhibit opposite affinities: while DeepSeek presents a clear advantage in English, the local Ministral model contradicts the standard dominance and delivers a higher TPS in Portuguese.

Finally, at the base of the distribution (G1), DeepSeek with RAG in Portuguese is isolated at 53.0 TPS. This sharp drop highlights the high computational cost and severe latency penalty that the conjunction of the Portuguese language and the RAG mechanism imposed on this specific architecture.

#### IV. QUALITATIVE ACCURACY ANALYSIS

To evaluate the efficacy of Hybrid RAG in standardizing notations, definitions, and methodologies used by LLMs in the DSP domain, we conducted a qualitative comparative analysis, designing a set of DSP questions specifically chosen for their tendency to create conflicting responses among models. Two proprietary models (GPT-5 and DeepSeek-V4-Pro) and one local model (Ministral3-8B) were evaluated under two conditions: without RAG (direct API access) and with RAG (via the ITS). Due to space constraints, the complete question-answer dataset is available online at [12].

The adopted questions are the following:

**Question 1:** A signal  $x(t)$  is composed by two sines of frequencies  $f_1 = 50$  Hz and  $f_2 = 70$  Hz. What is the minimum sampling frequency according to the sampling theorem?

**Question 2:** Can I choose the sampling frequency arbitrarily when I'm using the bilinear transform for designing a digital filter?

**Question 3:** Using partial fractions decomposition, calculate the inverse Z transform of:

$$X(z) = \frac{1 + 2z^{-1} + z^{-2}}{1 - \frac{3}{2}z^{-1} + \frac{1}{2}z^{-2}}, \quad |z| > 1. \quad (3)$$

**Question 4:** Working with frequency  $f$  in Hertz, a PSD  $S(f)$  has a constant value of 5 over the range of frequencies  $Bf = [-20, 20]$  Hz. In this frequency range  $Bf = 40$  Hz, the average power is  $S(f) \times Bf = 5 \times 40 = 200$  watts, and the unit of  $S(f)$  is watts/hertz. For the same example, but now working with angular frequency  $w$  in radians/sec, the PSD  $S(w)$  must still have a constant value over the range of frequencies  $Bw = [-40\pi, 40\pi]$  rad/s and an average power of 200 watts. In this case, what is the value of the PSD  $S(w)$  and its unit?

The LLM models and our categorization of their answers to the four questions are summarized in Table II. The questions are categorized into two subgroups: basic DSP definitions (Q1 and Q2) and method adoption (Q3 and Q4). Questions Q1 and Q2 are simpler and designed to verify whether the LLMs are “Right” or “Wrong”, while Q3 and Q4 verify whether the LLM adopts the same instructor’s methodology or not, and are capable of advanced reasoning. We classify as “Wrong” and “Right” the answers that are strictly incorrect and correct, respectively. The category “Other” is used to indicate that the answer is correct but does not comply with

the methodology outlined by the instructor in the material provided to the Hybrid RAG module (e.g., as a PDF file). The category “Same” indicates the LLM adopted the instructor’s methodology.

TABLE II: Qualitative results by model and RAG mode.

| Mode        | Model           | Questions |          |       |       |
|-------------|-----------------|-----------|----------|-------|-------|
|             |                 | $F_s$     | Bilinear | PFE   | PSD   |
| Without RAG | GPT-5           | Wrong     | Wrong    | Other | Other |
|             | DeepSeek-V4-Pro | Wrong     | Wrong    | Other | Other |
|             | Ministral3-8B   | Wrong     | Wrong    | Wrong | Other |
| With RAG    | GPT-5           | Right     | Right    | Other | Same  |
|             | DeepSeek-V4-Pro | Right     | Right    | Same  | Same  |
|             | Ministral3-8B   | Right     | Right    | Wrong | Wrong |

The LLM answers to each question are analyzed in the following paragraphs.

1) *Q1 -  $F_s$  according to sampling theorem:* This query presents a signal composed of a finite number of sine components, with a maximum frequency  $f_{\max} = 70$  Hz. The sampling theorem requires that:

$$F_s > 2f_{\max},$$

but some references erroneously assume  $F_s \geq 2f_{\max}$ . However, sampling exactly at  $F_s = 2f_{\max}$  does not guarantee a perfect reconstruction [10]. For example, if a signal  $x(t) = \sin(2\pi \times 70t)$  is sampled at  $t = 0, \pm 1/F_s, \pm 2/F_s, \dots$ , with  $F_s = 140$  Hz, all samples will coincide with the zero-crossings of  $x(t)$ , making signal reconstruction impossible.

Without Hybrid RAG, all three models provided incorrect responses, claiming that  $F_s = 2f_{\max} = 140$  Hz. However, augmented with Hybrid RAG, both commercial models correctly applied the strict inequality and emphasized that sampling exactly at  $2f_{\max}$  fails to guarantee perfect reconstruction. The Hybrid RAG-augmented local model also correctly applied the strict inequality but omitted any explanation regarding the mathematical limitation of  $F_s = 2f_{\max}$ .

2) *Q2 - Bilinear:* The role of the sampling frequency  $F_s$  used in the bilinear transform depends on the intended application. For the design of digital filters, when starting with a specification in discrete-time,  $F_s$  is canceled during the calculation and can be chosen arbitrarily. In this case, the designer must only avoid excessively large or excessively small values of  $F_s$ , which may lead to numerical errors. In contrast, in applications where the bilinear is used to find a discrete-time  $G(z)$  that mimics a continuous-time  $G(s)$ , there are requirements governing the choice of  $F_s$ , which cannot be chosen arbitrarily [10].

Although non-RAG LLMs provided valid filter design methodologies, they unanimously stated that  $F_s$  is not arbitrary, presenting explanations that are unnecessarily complex. Among the non-RAG systems, only DeepSeek noted that  $F_s$  cancels out, but contradictorily asserted that the choice of  $F_s$  is not arbitrary. In contrast, all RAG-augmented models correctly specified that  $F_s$  can be chosen arbitrarily and appropriately warned against using extremely small values to prevent numerical instability.

3) *Q3 - PFE for inverse transforms*: Partial Fractions Expansion (PFE) is a useful method for inverting both Laplace and Z-transforms. Unlike rational functions in  $s$ , which use positive powers,  $X(z)$  is typically expressed in negative powers of  $z$ . Since students may be confused when dealing with negative powers, it is assumed here that the instructor's method is to convert  $X(z)$  to positive powers. This conversion often returns slightly different results, but mathematically equivalent expressions for  $x[n] = \mathcal{Z}^{-1}\{X(z)\}$ .

For the example of Eq. (3), using negative powers of  $z$ , the conventional procedure (see, e.g., Section 3.2 of [13]) leads to

$$x[n] = 2\delta[n] - 9\left(\frac{1}{2}\right)^n u[n] + 8u[n]. \quad (4)$$

In contrast, the methodology that uses positive powers of  $z$  leads to (an expression with the same amplitude values):

$$x[n] = \delta[n] + 8u[n-1] - 4.5\left(\frac{1}{2}\right)^{n-1} u[n-1], \quad (5)$$

which is assumed here to be the result desired by the instructor.

Without RAG, the commercial models (GPT and DeepSeek) avoided positive powers of  $z$ , deriving Eq. (4) via alternative methodologies. On the other hand, the non-RAG Ministral-3 model utilized positive powers of  $z$  but did not recognize that Eq. (3) is an improper rational function, producing an incorrect PFE. With RAG, GPT maintained its initial approach, again reaching Eq. (4). In contrast, the RAG-augmented DeepSeek model derived Eq. (5), making it the only evaluated model to align with the preferred methodology. Lastly, the RAG-augmented Ministral-3 model repeated its previous error, failing to produce a mathematically valid expression.

4) *Q4 - PSD  $S(\omega)$* : While the unit of a continuous-time Power Spectral Density (PSD)  $S(f)$ , with  $f$  in Hz, is clearly W/Hz, the unit of  $S(\omega)$ , with  $\omega$  in rad/s requires a proper definition. In this paper, it is assumed that  $S(\omega)$  has the property of

$$\mathcal{P} = \frac{1}{2\pi} \int_{-\infty}^{\infty} S(\omega) d\omega, \quad (6)$$

where  $\mathcal{P}$  is the average power in W. Hence, the expected answer to Q4 is that  $S(\omega) = 5$ , which after normalization by  $2\pi$  has the unit of W/(rad/s).

Without RAG, the LLMs answered that  $S(\omega) = 5/(2\pi)$  W/(rad/s), while the RAG led the commercial models to follow the instructor's methodology. The Ministral3-8B with RAG derived a wrong answer of  $S(\omega) = 10$  W/rad.

## V. CONCLUSIONS

This work demonstrated that locally hosted, open-source LLMs combined with Hybrid RAG constitute a viable alternative for intelligent tutoring systems in DSP education. The proposed ChatBot-o architecture extends the AcademiCK framework to support private institutional deployment, enabling greater control over privacy, infrastructure costs, and pedagogical alignment.

The benchmark analysis revealed that response latency and throughput depend on the interaction between architecture, language, and RAG activation. While commercial models

achieved the highest throughput, the local Ministral-3-8B model demonstrated competitive responsiveness and particularly favorable behavior in BP. The qualitative analysis of the LLM answers highlighted the importance of Hybrid RAG, which significantly improved accuracy for the adopted questions about DSP. In fact, according to our criteria, the non-RAG mode led always to inadequate LLM responses. This analysis also indicated that the adopted open source model is not competitive with the baseline commercial models when the questions required advanced reasoning (Q3 and Q4). Future work includes developing larger and public datasets, broader DSP topics and adaptive tutoring strategies.

## ACKNOWLEDGEMENTS

The authors would like to thank the National Council for Scientific and Technological Development (CNPq) (process: 171382/2025-0).

## REFERENCES

- [1] L. Marquez-Carpintero, A. Lopez-Sellers, and M. Cazorla, "Simulation of teaching behaviours in intelligent tutoring systems: a review using large language models," *Artificial Intelligence Review*, 2025, DOI: 10.1007/s10462-025-11464-8.
- [2] J. Haupt, Q. Lu, Y. Shen, J. Chen, Y. Dong, D. McCreary, M. Akçakaya, and G. B. Giannakis, "Deploying AI for signal processing education: selected challenges and intriguing opportunities," *IEEE Signal Processing Magazine*, vol. 43, no. 1, pp. 32–46, 2026, DOI: 10.1109/MSP.2025.3600846.
- [3] Y. Jayaram, "Private LLMs for higher education: secure GenAI for academic & administrative content," *American International Journal of Computer Science and Technology*, vol. 6, no. 4, pp. 28–38, 2024, DOI: 10.63282/3117-5481/AIJCSCT-V6I4P103.
- [4] C. Merlos, F. Hussain, S. Kar, L. Shri, O. Olugbenle, M. Banavar, and A. A. AlMomani, "WIP: comparison of large language models for applied mathematics questions in engineering courses," in *2024 IEEE Frontiers in Education Conference (FIE)*, 2024, pp. 1–5, DOI: 10.1109/FIE61694.2024.10893476.
- [5] L. Shen, W. Tan, S. Chen, Y. Chen, J. Zhang, H. Xu, B. Zheng, P. Koehn, and D. Khashabi, "The language barrier: dissecting safety challenges of LLMs in multilingual contexts," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 2668–2680, DOI: 10.18653/v1/2024.findings-acl.156.
- [6] Z. Li, Z. Wang, W. Wang, K. Hung, H. Xie, and F. L. Wang, "Retrieval-augmented generation for educational application: a systematic survey," *Computers and Education: Artificial Intelligence*, vol. 8, p. 100417, 2025, DOI: 10.1016/j.caeai.2025.100417.
- [7] C. K. Mello, "AcademiCK: Assistente de estudos para o ambiente acadêmico," 2025, [Online]. Available: <https://bdm.ufpa.br/handle/prefix/8945> (in Portuguese).
- [8] C. Y. Yang, G. J. Liu, and C. J. Yu, "Optimizing RAG in programming education: a comparative study of models and strategies," *IEEE Access*, vol. 13, pp. 190 890–190 903, 2025, DOI: 10.1109/ACCESS.2025.3629843.
- [9] Y. Zhong, S. Liu, J. Chen, J. Hu, Y. Zhu, X. Liu, X. Jin, and H. Zhang, "DistServe: Disaggregating prefill and decoding for goodput-optimized large language model serving," *18th USENIX Symposium on Operating Systems Design and Implementation (OSDI 24)*, pp. 193–210, 2024, [Online]. Available: <https://arxiv.org/pdf/2401.09670>.
- [10] A. Klautau, *Digital Signal Processing Using Python, Matlab, and Octave*. <https://www.amazon.com/dp/6501609135>, 2026. [Online]. Available: [https://ai6g.org/books/dsp/ak\\_dsp\\_book.html](https://ai6g.org/books/dsp/ak_dsp_book.html)
- [11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou *et al.*, "Chain-of-thought prompting elicits reasoning in large language models," *Advances in neural information processing systems*, vol. 35, pp. 24 824–24 837, 2022, [Online]. Available: <https://arxiv.org/abs/2201.11903>.
- [12] P. Canavieira, T. Nisimura, and M. Ribeiro, "LLM for teaching DSP," 2026, [Online]. Available: [https://github.com/CAIO-LIMA-37/Intelligent\\_Tutor\\_System\\_DSP](https://github.com/CAIO-LIMA-37/Intelligent_Tutor_System_DSP).
- [13] A. V. Oppenheim and R. W. Schaffer, *Discrete-Time Signal Processing*. Pearson, 2014.