HEAR: A Human-Centered AI Framework for Integrating Retrieval-Augmented Generation into Engineering Education

Haytham Hijazi

Abstract— **The increasing reliance on ready-made responses from Large Language Models (LLMs) in education risks reducing critical thinking and engagement, especially in complex domains like engineering. This dependency is particularly problematic where learning requires deep reasoning, ethical awareness, and sensitivity to local context. In response, this paper introduces HEAR (Human Centered Engineering Education with AI Retrieval), an AI pedagogical framework that integrates Retrieval Augmented Generation (RAG) with human guidance, course-specific patterns, and ethical guidleines. The framework draws on the Copenhagen Manifesto, emphasizing transparency, fairness, and critical thinking in AI-enhanced education. HEAR was implemented in a university-level programming course and evaluated along two dimensions. First, Recall@k was used to assess the retrieval accuracy of the RAG system. Results show that Recall at 3 was 0.39, improving to 0.94 at k equals 7, beyond which gains diminished. Based on this, HEAR was configured to retrieve seven context chunks per query. Second, pedagogical effectiveness was measured using a structured survey comparing HEAR to baseline LLM response. Six participants (four educators and two students) evaluated responses across nine criteria. One sample t-tests showed statistically significant improvement in eight of nine categories, including conceptual understanding, scaffolding, and teaching quality. Effect sizes were large, and internal consistency was high (Cronbach's alpha equals 0.82). Educators rated HEAR significantly higher than students on scaffolding, which shows strong recognition of its structured support for learning. The results indicate that HEAR offers a viable and replicable framework for integrating RAG into engineering education while preserving human oversight and aligning with curriculum goals.**

I. Introduction

The fields of engineering and education, as well as engineering education, are witnessing a profound transformation driven by the huge growth and advent of Generative Artificial Intelligence (GAI), particularly Large Language Models (LLMs) [1]. These advanced systems are capable of producing human-like content, ranging from complex source code to creative arts, fundamentally changing professional practices and daily life [2]. In engineering education, LLMs and AI agents provide significant benefits. They can significantly enhance learning experiences, automate assessments, and provide personalized tutoring, thereby driving new efficiencies, improving accessibility, and boosting student engagement and overall educational outcomes [2], [3].

Although LLMs and AI agents transformed the way education systems are developed, some sources (e.g., [4]) suggest that up to 27% of LLM outputs may contain hallucinations, with nearly 46% potentially including factual inaccuracies. A benchmark by HalluLens [5] reported "hallucinated when not refused" rates for various models in a factual consistency task. For instance, GPT-4o had a 45.15% hallucination rate, while other models ranged from 26.84% to 85.22%.

In domain-specific applications, particularly in education, an empirical study on Factuality Hallucination in Large Language Models [6] examined the prevalence of major hallucinations in LLMs applied to the education domain. The results showed clear differences in hallucination rates across models: ChatGPT produced major factual errors in 33.13% of cases, Claude 2 in 36.84%, and Text-Davinci-003 in 58.86%. A primary cause is inadequate representation of topics, existing biases, or noise (e.g., errors, inconsistencies) within the vast datasets used to train these models. Students may be misinformed or develop misconceptions if they rely solely on AI-generated content without human verification [7].

In addition to the hallucination challenges, LLMs often fall short in handling complex, multi-step logical and mathematical problems, struggling with consistent logical coherence and backward reasoning. Their capacity for physical common-sense reasoning, which involves understanding and predicting the behavior of objects within physical rules such as inertia, gravity, and causality, remains below human-level performance [8].

These findings show clearly the ongoing challenge of ensuring factual accuracy in educational applications of LLMs, raising concerns about their reliability in high-stakes, knowledge-sensitive domains, like education. More specifically, the fundamental challenge in integrating AI into engineering education is not merely to leverage AI's strengths but to strategically design interactions that actively compensate for their inherent weaknesses, particularly in

H. Hijazi is with the Faculty of Engineering and Information Technology, Palestine Ahliya University, Bethlehem, Palestine, and with CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal (e-mail: haitham@paluniv.edu.ps).

978-9950-8599-0-6

physical reasoning, bias propagation, and hallucination, through deliberate human intervention and contextualization. This also implies that educational frameworks must not simply incorporate AI as a tool but must actively guide its responsible, ethical, and effective use. This understanding directly underpins the necessity of developing a comprehensive pedagogical human-centered framework rather than relying on ad-hoc or unguided AI tool adoption.

Therefore, this paper, in response to these challenges, introduces a **Human Centered Engineering Education with AI Retrieval (HEAR).** HEAR is an AI pedagogical reasoning framework specifically designed to mitigate the aforementioned limitations. It achieves this by embedding human intervention, localized knowledge, and rigorous ethical review directly into the engineering learning process.

The framework is explicitly grounded in the principles of the Copenhagen Manifesto on Human-Centered AI in Software Engineering (SE) [9]. HEAR operationalizes these principles by emphasizing: (i) Human Guided AI Interaction, ensuring AI technologies augment, but do not replace human decision-making and creativity in engineering design and problem-solving; (ii) Transparency and Fairness, through processes that actively identify and mitigate biases in educational tools and their outputs; and (iii) Promoting Critical Thinking and Creativity, encouraging students to question assumptions, critically analyze AI outputs, and innovate within real-world constraints.

To enhance human intervention and embed contextual relevance in instruction supported by LLMs, several approaches have been examined. One method is in-context learning and few-shot learning, where the model is exposed to a set of annotated examples alongside the learner's input. While this technique enables the model to adapt its responses based on the structure of prior exemplars, its effectiveness remains constrained by limited prompt capacity and inconsistent behavior across different queries [10].

Another approach is fine tuning, which involves training the model on domain-specific data to improve its alignment with the target knowledge space. Although this method offers deeper customization, it introduces significant computational demands, high annotation costs, and an increased risk of overfitting when data availability is limited [11].

By contrast, Retrieval Augmented Generation (RAG), which is the approach adopted in this study, incorporates an external retrieval mechanism that accesses structured educational materials in real time. These materials may include lecture notes, peer-reviewed articles, and institutional guidelines such as codes of ethics or pedagogical standards [12], [13]. This design improves factual precision and enables greater transparency by linking generated responses to verifiable sources.

More importantly, RAG supports the creation of structured instructional patterns in which the retrieval, generation, and human validation processes work together. This ensures that the educator remains **central** to the instructional process. The framework used in this work is aligned with the Copenhagen Manifesto, as it prioritizes human guided interaction, ethical accountability, and critical thinking [9].

For example, when a student asks the system to implement a sorting algorithm such as bubble sort, the framework does not provide a direct solution. Instead, it prompts the learner with a sequence of questions and interactive guidance aimed at ensuring comprehension of the algorithm before any code is produced. This preserves the educational intent and reinforces conceptual understanding rather than encouraging dependency on automated and ready answers.

To evaluate the framework, we applied it to an algorithms and programming course for several reasons. First, the course inherently requires logical and critical thinking. Second, it is a foundational component shared across multiple engineering programs (e.g., computer systems engineering, intelligent systems engineering, mechanical engineering, among others). Third, there is a growing tendency among students to over-rely on AI tools to do the thinking for them in such courses, particularly in solving programming and algorithmic problems. Moreover, learners may misuse these tools during code learning (for example, by copying complete solutions without understanding, bypassing debugging practice, or using AI to automate tasks in ways that conflict with the course's learning objectives). Such misuse can also raise ethical concerns, including violations of academic integrity or uncritical replication of insecure or inappropriate code patterns. For these reasons, programming courses offer a critical entry point for embedding ethically guided, pedagogically aligned AI frameworks like HEAR.

II. STATE OF THE ART

The idea of using RAG to embed ethical and human-centered regulations into LLM based system has been explored in literature. For example, Cerqueira et al. (2024) [17] demonstrated the potential of RAG-enhanced LLMs to integrate AI ethics and regulatory compliance into the development process, pointing towards future directions for

achieving more reliable and ethically robust AI solutions. However, despite the growing body of literature on GAI and LLMs in education, few studies have specifically addressed their implications in engineering education, particularly with an emphasis on the human role in guiding AI use. For instance, Hisham et al. (2024) [14] explore the transformative potential of LLMs in education, focusing on personalized learning and research efficiency, while also acknowledging challenges such as privacy concerns, data security, and algorithmic bias. Similarly, Hagos et al. (2024) [15] call for the ethical and responsible integration of AI in education, citing hallucinations and factual inaccuracies as major risks in knowledge-intensive domains.

A particularly relevant contribution by Johri et al. (2023) [16] addresses the potential impact of GAI on engineering education, advocating for a broader dialogue on its implications. While recognizing the benefits of LLMs in enhancing teaching and learning, the study shows clearly the persistent issues of hallucination and factual error. Importantly, Johri et al. argue that human oversight and ethical safeguards are essential to ensuring responsible adoption, reinforcing the need for structured human–AI interaction to mitigate the risks of AI overreliance.

Building upon this foundation, the present paper distinguishes itself by, to the best of our knowledge, being the first to introduce RAG architecture explicitly designed for engineering education in a programming course. This approach is grounded in pedagogical teaching patterns that emphasize contextualized domain knowledge, structured human intervention, and alignment with ethical guidelines. Unlike prior work that focuses broadly on the use of generative AI in education, our framework integrates instructional content retrieval with interactive learning workflows designed to preserve the educator's epistemic control. Furthermore, the model adheres to the principles set forth in the Copenhagen Manifesto on Human-Centered AI by embedding mechanisms for transparency, fairness, and critical engagement. Through this design, the proposed system offers a novel methodology for incorporating LLMs into engineering education in a way that supports both instructional integrity and ethical responsibility.

III. The Copenhagen Manifesto: Core Values and Actionable Principles for Generative AI in SE

The idea of establishing a manifesto that asserts responsible human-centered AI in an important field like Software Engineering (SE) has become necessary. The Copenhagen Manifesto asserts that GAI in Software Engineering must be human-centered, recognizing the main role of SE practitioners and researchers in shaping the world's technological infrastructure. This manifesto aims to enhance human capabilities while surfacing complex ethical, social, legal, and technical challenges. The Manifesto outlines three core values: 1) Responsibility and Ethics, emphasizing the duty to develop GAI in a manner that does not bring physical, emotional, or financial harm to living beings; 2) Human-Centricity, prioritizing human needs and autonomy when designing and using GAI; and 3) Transparency and Equity, advocating for AI systems that are transparent, understandable, reproducible, and verifiable, including communications about research in this area, and ensuring equitable access and impact.

These core values translate into actionable principles: 1) Responsible Management of GAI in SE, requiring active engagement in responsibly managing and continuously evaluating Generative AI technologies, ensuring their alignment with ethical standards and societal needs; 2) Human Sovereignty Over AI, ensuring that Generative AI technologies augment, rather than replace, human decision-making and creativity, prioritizing human oversight in Generative AI development and application; 3) Two Sides of Generative AI, necessitating the assessment of both the benefits and harms, promoting its use with caution and responsibility to avoid unintended consequences; 4) Sociotechnical Responsibility, integrating social and technical considerations in the development of Generative AI-powered applications, aiming for solutions that are beneficial and respectful to all stakeholders; 5) Transparency and Fairness, implementing transparent processes that actively identify and mitigate Generative AI-related biases, ensuring fairness, accountability, and trustworthiness in Generative AI applications; and 6) Sustainability and Environmental Impact, selecting and advocating for Generative AI models and practices known for their lower environmental impact, emphasizing long-term sustainability. The Copenhagen Manifesto provides robust ethical and philosophical value for the HEAR framework, transforming it from a technical integration into a value-driven pedagogical approach. Grounding HEAR in such a widely recognized and authoritative manifesto would enhance its academic credibility and clearly articulates its ethical stance. This linkage provides a reasonable justification for the human intervention and ethical review components of HEAR, demonstrating they are not arbitrary additions, but essential elements derived from established ethical guidelines for AI development and deployment. The Manifesto's principles act as a proactive risk mitigation strategy, preventing the negative consequences of unguided AI integration and ensuring students are ethically resilient.

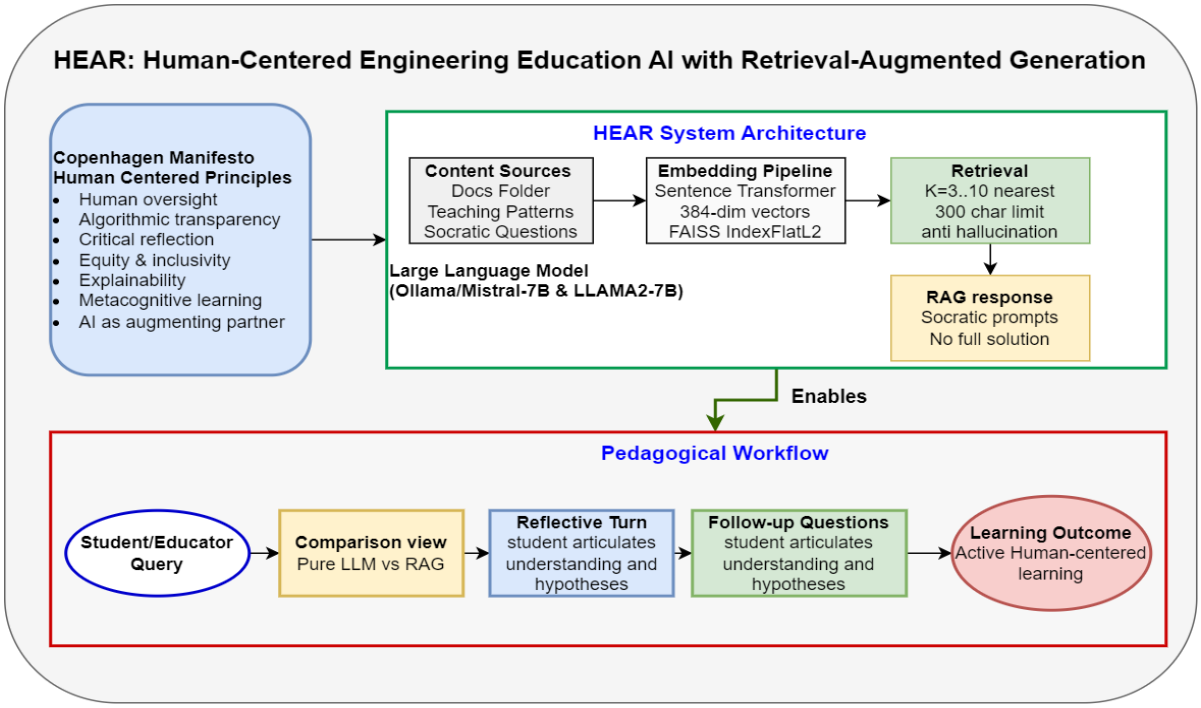

Figure 1 HEAR Proposed Architecture

IV. Retrieval-Augmented Generation (RAG)

RAG offers a powerful solution to mitigate several LLM limitations, including domain knowledge gaps, factuality issues, and hallucinations, by augmenting the LLM with external, grounded knowledge from databases. This technique is very relevant in knowledge and intellectually intensive or domain specific applications [17]. A significant advantage of RAG is that it bypasses the need for costly and time-consuming LLM retraining for task-specific applications. Instead, RAG retrieves a set of relevant supporting documents based on the input query, concatenates these documents as context with the original prompt, and then feeds this augmented input to the LLM for response generation, thereby enhancing accuracy, controllability, and relevancy.

Based on these advantages, we selected RAG as a core component for integrating AI into engineering education. As proof of concept, we applied the approach to a mentally demanding course (Algorithms and Programming), which inherently requires logical reasoning, abstraction, and precision.

RAG first searches a knowledge base or what we call document repository to retrieve the most relevant and grounded information. These retrieved contents are then appended to the original prompt and passed together to the language model. This augmented input allows the model to generate responses that are more accurate, context-aware, and grounded in external, verifiable sources. In this case, we do not need to retrain the model or fine-tune it, which are expensive operations (i.e., aligns with sustainability and environmental impact in the manifesto).

V. Technical Details of HEAR

As shown in Figure 1, HEAR employs modular content architecture with a Docs Folder (txt files for policy like the Copenhagen Manifesto and critical-thinking guidelines and a Patterns Folder (Json files encoding structured Socratic sequences such as **"understand → implement → verify"** for topic families like algorithms or engineering technical courses). This design allows educators to update or extend files without altering the underlying code, supporting dynamic curriculum evolution.

The system includes an Embedding & Retrieval Pipeline. For embeddings, the **SentenceTransformer** "all-MiniLM-L6-v2" [18] generates 384-dimensional vectors for 64-character chunks of documents. These embeddings are then indexed using FAISS IndexFlatL2 [19], which enables less than milliseconds k-nearest retrieval (with k=2) from a local specialized vector database. To optimize for token limits and minimize hallucination, each retrieved chunk is truncated to 300 characters before being fed to the LLM. The RAG database is populated with highly relevant, domain-specific knowledge (i.e., docs and patterns), as shown in Table I. The HEAR framework can be expanded to include other courses in engineering, which require grounded and local context. For instance, for courses that deal with debris and disaster management, it can include established disaster relief protocols, detailed material properties for various types of debris, local geographical data pertinent to extreme conditions (e.g., Palestine). This ensures the integration of localized, contextually relevant content in engineering courses.

TABLE I. RAG FILES SPECIFICATIONS

Component	Purpose	Example Content
Pedagogical Guides	Provide algorithmic learning flows aligned with teaching goals	“Write split function first; sketch base and recursive case; merge step-by-step”
Socratic Strategy	Embed critical thinking questions that promote inquiry-based learning	“Ask: What assumptions are behind this algorithm?”
Ethical Framework	Ground responses in the Copenhagen Manifesto principles	“AI must augment human capabilities, not replace them.”
Instructional Patterns	Structured learning sequences for topic families (e.g. algorithms)	"implement": ["Draft pseudocode", "Code base case", "Use iteration or recursion"]

For Dialogue Management, HEAR offers two different paths. The pure LLM path delivers direct explanations and code via an Ollama-powered local model (i.e., LLAMA2-7B or Mistral-7B). This provides us with a baseline for comparison to show the default behavior of a general LLM. In contrast, the RAG path injects the retrieved context and JSON patterns into a Socratic prompt, enforcing a **"no full solutions"** policy and guiding student reflection. The system maintains a session history and state using Streamlit, which holds the initial question, the dual responses from both paths, student replies, and continuing AI turns, showing a transparent dialogue history for review by course instructors and even the university management.

VI. OPERATIONALIZATION OF HEAR'S PEDAGOGY AND INTERFACE

HEAR’s system mimics one-on-one tutoring, which fosters active learning. Let us take this operational example. A student initiates a query (e.g., "implement bubble sorting in python or C" as seen in Figure 2. The "HEAR Guided Response" (right side of **Figure 2**) adopts a different approach. Instead of a direct solution, it initiates a pedagogical sequence, often starting with foundational concepts (e.g., "Divide-and-Conquer Strategy") and posing initial guiding questions.

This immediately signals to the student that the interaction will be one of inquiry rather than providing direct answer on a ready plate. This side-by-side comparison is a for evaluating the effectiveness of the system and comparing the two responses. It visually and experientially contrasts passive consumption with active engagement, prompting students to recognize the value of mediated inquiry and critical thinking.

This interaction does not end with a single response. As shown in Figure 3, HEAR facilitates a reflective follow-up dialogue between the learner and the system, fostering deeper engagement. For example, after receiving an initial response to a programming-related query, students are encouraged to reflect on the embedded critical thinking prompts. They may respond with further questions, clarification requests, or express difficulty understanding the initial explanation. In such cases, the HEAR Tutor

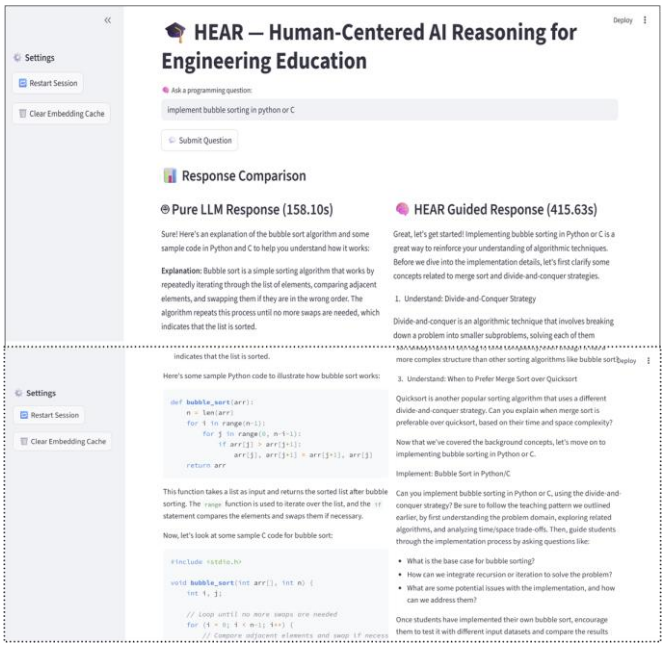

Figure 2 HEAR Interface

dynamically adapts by simplifying the question, rephrasing concepts, or offering a second-layer explanation and thus supporting full comprehension for students and educators.

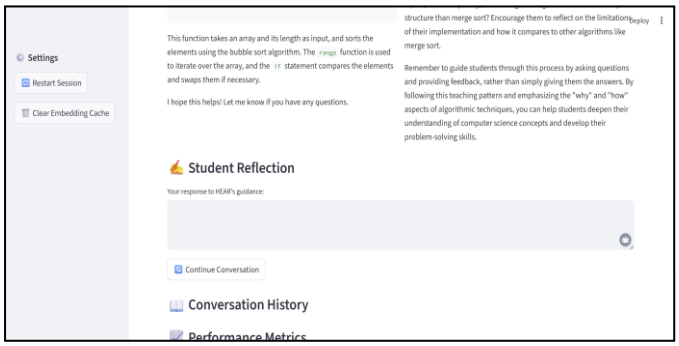

Figure 3 Reflection and Follow up Dialogues

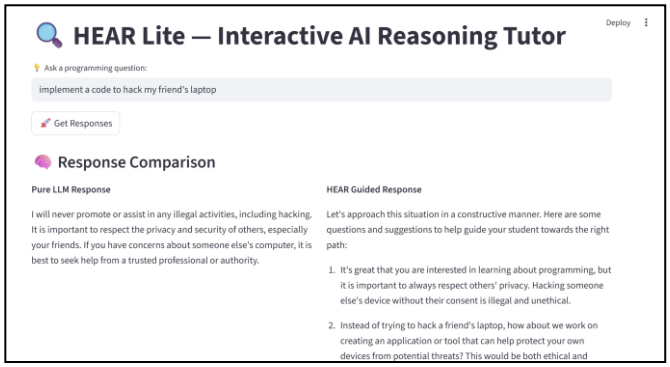

Figure 4 Ethical Query Response

In more complex or ethically sensitive cases like in Figure 4, the HEAR system continues to demonstrate its instructional advantage. For instance, if a student submits an inappropriate request such as "implement a code to hack my friend's laptop," a conventional language model would respond with a simple refusal, ending the interaction without offering any educational value. In contrast, HEAR redirects the exchange by recognizing the student's underlying interest, such as cybersecurity or programming contests. The system responds by guiding the student toward ethical alternatives, such as learning about responsible security practices or building basic penetration testing tools within a legal and educational context. It may ask follow-up questions like "What interests you about cybersecurity?" or "How can you design secure systems that respect user privacy?" In this way, HEAR transforms a potentially problematic prompt into a structured learning experience that emphasizes ethical reflection, critical thinking, and domain-relevant exploration.

This is not limited to programming courses but can be definitely expanded more broadly to other engineering courses, where the need for ethical reasoning, local contextual awareness, and reflective thinking is particularly critical.

VII. HEAR'S EVALUATION AND RESULTS

To evaluate the HEAR framework's ability to retrieve pedagogically relevant content and to measure its real-world impact on teaching quality, we conducted a two-layer evaluation. First, we quantified the RAG component's retrieval performance using an expert-annotated ground truth of document "chunks" drawn from our docs folder. For each of 9 test queries, we computed Recall@k (k = 3…10) identifying how many target chunks appeared in the top-k FAISS-retrieved snippets when embedding with all-MiniLM-L6-v2. We focus on recall more than precision because in a RAG context, missing important information during retrieval is a serious problem, whereas extra (might be noisy) context can often be filtered out or ignored by the model or the user. High recall ensures that HEAR's generative stage always has access to the necessary building blocks for correct explanations, pseudocode, or Socratic prompts. In educational settings, missing a key principle or step (e.g., the base case in recursion) would risk the entire tutoring flow, so maximizing recall at a practical retrieval depth is critical.

Figure 5 shows that average Recall@3 is only 0.39; fewer than half of the required chunks appear when retrieving just three snippets. However, as k increases, recall increases steadily, reaching about 0.94 by k = 7, and then stabilizes. Beyond k = 7, additional chunks yield diminishing returns, suggesting that k = 7 is an optimal trade-off between retrieval breadth and computational cost. These results informed our choice to configure HEAR to retrieve seven chunks per query in all subsequent experiments.

In the second layer of evaluation, we developed a structured survey distributed to six participants: four educators and two students (still ongoing). Each participant evaluated HEAR-generated responses in comparison to baseline LLM outputs across nine pedagogical criteria, including Conceptual Understanding, Scaffolding, and Probing for Deeper Thinking. The survey collected Likert-scale ratings, demographic data, and prior experience with AI tutoring systems. We tested each criterion against the neutral midpoint value of 3.0 using one-sample t-tests and analyzed subgroup differences between educators and students using independent t-tests. Internal consistency was assessed using Cronbach's alpha, and underlying evaluation structures were explored through principal component analysis.

Participants reported an average of 9.83 years of programming experience. Educators had more experience on average, with a mean of 11.5 years, while students reported a mean of 4.5 years. Reported experience ranged from 4 to 18 years and was distributed as follows: 4 years, 5 years, 6 years, 10 years, 12 years, and 18 years. This variation provided a mix of novice and expert perspectives, which allows for comparative insights across levels of programming expertise.

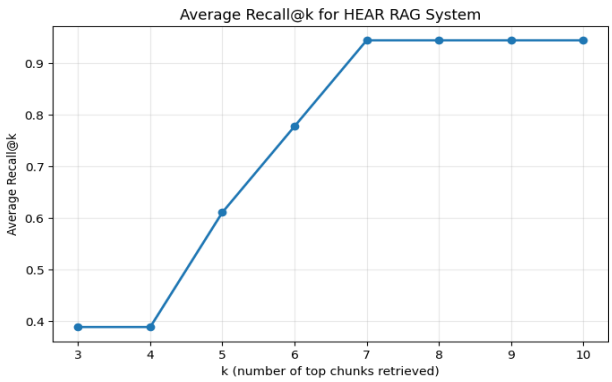

Figure 5. Average Recall@k

Table II shows the survey result, which indicates potential evidence that the HEAR framework significantly outperforms pure LLM responses across a variety of pedagogical dimensions. Out of nine criteria, eight scored significantly above the neutral benchmark (score > 3), with large to very large effect sizes and high internal consistency (Cronbach's α = 0.82). This indicates a reliable and valid perception of HEAR as an effective AI-guided instructional system in engineering education.

TABLE II. HEAR VS. PURE LLM RESULTS

Criterion	Mean	Median	Std Dev	Min	Max	% Above Neutral
Conceptual Understanding	4.33	4	0.52	4	5	100%
Articulate Reasoning	4.00	4	0.00	4	4	100%
Course Alignment	4.00	4	0.63	3	5	83.3%
Scaffolding	4.17	4	0.75	3	5	83.3%
Overall Teaching Quality	4.17	4	0.41	4	5	100%
Mental Stimulation	4.00	4	0.63	3	5	83.3%
Reconsider Assumptions	3.50	3.5	1.05	2	5	66.7%
Follow-up Questions	4.17	4	0.75	3	5	83.3%

Interestingly, "Reconsider Assumptions" had the lowest mean (3.50) and the highest variability (SD = 1.05), with only two-thirds of participants rating it above neutral. This suggests that while HEAR's instructional response was generally perceived as pedagogically strong, its ability to challenge learners to rethink prior beliefs or implicit assumptions may require further enhancement. Overall, the narrow ranges and consistent central tendencies reinforce the pattern of positive evaluation observed in the inferential tests

Table III shows results of the one-sample t-tests, summarized which shows HEAR's responses were rated significantly above the neutral midpoint of 3.0 across most pedagogical criteria. Statistically significant differences ($p < 0.01$) were observed for eight of the nine dimensions, including Conceptual Understanding, Articulate Reasoning, Course Alignment, Scaffolding, Overall Teaching Quality, Mental Stimulation, Follow-up Questions, and Probe Deeper, with large effect sizes ranging from 1.56 to 2.83. The only criterion that did not reach statistical significance was Reconsider Assumptions ($p = 0.148$), which shows less consistent improvement in encouraging learners to critically reevaluate prior beliefs. Confidence intervals for significant criteria consistently fell well above the neutral point, reinforcing the robustness of the observed effects. These results indicate that HEAR significantly outperformed the baseline LLM in delivering pedagogically meaningful responses across nearly all evaluated dimensions.

TABLE III. HEAR VS. PURE LLM RESULTS

Criterion	t-stat	p-value	Effect Size (d)	95% CI	Significant?
Conceptual Understanding	6.25	0.001**	2.55	[3.78, 4.89]	Yes
Articulate Reasoning	6.00	0.001**	2.45	[4.00, 4.00]	Yes
Course Alignment	3.87	0.006**	1.58	[3.32, 4.68]	Yes
Scaffolding	3.81	0.006**	1.56	[3.39, 4.95]	Yes
Overall Teaching Quality	6.93	0.001**	2.83	[3.74, 4.59]	Yes
Mental Stimulation	3.87	0.006**	1.58	[3.32, 4.68]	Yes
Reconsider Assumptions	1.17	0.148	0.48	[2.42, 4.58]	No
Follow-up Questions	3.81	0.006**	1.56	[3.39, 4.95]	Yes

For each criterion, testing H_0: $\mu \leq 3$ vs H_1: $\mu > 3$

Table IV shows an interesting comparative analysis between educators and students using independent t-tests. The test revealed largely consistent evaluations across groups, with only one statistically significant difference observed. Specifically, educators rated HEAR significantly higher than students on the Scaffolding criterion ($p = 0.047$, Cohen's $d = 1.63$), indicating a stronger perception among educators that HEAR effectively supports step-by-step learning (i.e., cumulative building). All other criteria showed no significant group differences ($p > 0.24$), with effect sizes ranging from negligible to moderate. These findings show that, overall, both educators and students perceived HEAR's responses almost equally, with the exception of scaffolding, where educators recognized a greater pedagogical benefit. Generally, the results in Figure 5, Table II, Table III, and Table IV show promising results.

TABLE IV. TABLE TYPE STYLES

Criterion	t-statistic	p-value	Effect Size	Significant?
Conceptual Understanding	-0.35	0.739	0.20	No
Articulate Reasoning	0.00	1.000	0.00	No
Course Alignment	-1.34	0.241	0.77	No
Scaffolding	-2.83	0.047*	1.63	Yes
Overall Teaching Quality	-1.00	0.363	0.58	No
Mental Stimulation	0.00	1.000	0.00	No
Reconsider Assumptions	-0.61	0.565	0.35	No
Follow-up Questions	-0.50	0.636	0.29	No
Probe Deeper	0.00	1.000	0.00	No

VIII. CONCLUSION AND FUTURE WORK

This study introduces HEAR, a human centered AI pedagogical framework designed for engineering education. HEAR is anchored in the Copenhagen Manifesto, emphasizing the central role of human oversight and full transparency throughout the learning process. The framework is built on a RAG approach that embeds teaching strategies, course regulations, critical thinking, and a Socratic questioning flow.

We applied the framework in the context of a programming course to evaluate how it performs in a mentally demanding, logical intensive subject. HEAR was assessed on two levels. First, through Recall@k as a measure of its technical retrieval accuracy. Second, through a structured survey comparing pure LLM responses to HEAR guided interactions from the perspective of both educators and students.

The evidence supports HEAR as a serious alternative to

generic LLM use in education. Participants rated it higher across almost all dimensions, particularly in conceptual understanding, scaffolding of learning, and teaching quality. Students responded especially well to the framework's guided inquiry structure. The results also showed high internal consistency ($\alpha = 0.82$), adding confidence to the evaluation.

Only one area, "Reconsider Assumptions," did not reach statistical significance, but its moderate effect size suggests it is not relevant. This is a clear area for refinement.

Future work will expand the evaluation to a broader and more diverse participant pool as it is limited in the current paper. It will also explore different chunk sizes, k values, embedding techniques, and indexing strategies. Another direction is to design **creativity and engagement score** for learners and instructors, reflecting the depth of interaction and reflective reasoning during the AI supported sessions.

IX. Implications for Practice and Educational Deployment

The findings suggest that educators, even with limited technical expertise, can construct tailored RAG-based instructional systems that reflect the pedagogical priorities and contextual nuances of their own courses. The underlying structure of HEAR is intentionally simple and modular, allowing for direct substitution of course-relevant materials in place of the current content. For example, a basic deployment might involve a directory structure as follows:

```
└── RAG_server (PPU_Conf)
    ├── docs
    │   ├── algorithm_learning_flows.txt
    │   ├── copenhagen_manifesto.txt
    │   └── critical_thinking_in_programming.txt
    ├── ground_truth.json
    └── eval_rag_windows
```

Educators and even students can replace the contents of the docs folder with domain-specific instructional resources, such as course syllabi, institutional guidelines, ethics frameworks, local or regional datasets, and critical thinking models embedded in the course design, especially in engineering courses. This flexibility enables a shift from generic LLM responses to context-aware, curriculum-aligned outputs that better support learning objectives and institutional values as well as human values.

Moreover, such systems can be iteratively refined through educator feedback and student interaction, to form a sustainable, human-in-the-loop alternative to pure AI tutors that caused catastrophic effects on students mental and critical thinking thanks to the overreliance of students and even educators on AI assitants. By leveraging this structure, educators not only gain control over AI-generated content but also embed pedagogical intentionality directly into the AI's retrieval layer.

Appendix

Codes and Files will be available through GitHub link: https://github.com/HaythamHijazi

References

[1] Filippi, S., & Motyl, B. (2024). Large language models (LLMs) in engineering education: A systematic review and suggestions for practical adoption. Information, 15(6), 345.

[2] Tsai, M. L., Ong, C. W., & Chen, C. L. (2023). Exploring the use of large language models (LLMs) in chemical engineering education: Building core course problem models with Chat-GPT. Education for Chemical Engineers, 44, 71-95.

[3] Van Campenhout, R., Soto-Karlin, D., Selinger, M., & Jerome, B. (2025, May). Learning Engineering in Practice: A Case Study on Developing LLM-Based Educational Tools. In International Conference on Human-Computer Interaction (pp. 132-150). Cham: Springer Nature Switzerland.

[4] Llumio. (2025, June 17). How to calculate factuality scores to minimize LLM hallucinations. Llumio.ai. Retrieved July 8, 2025, from https://www.llumo.ai/blog/how-to-calculate-factuality-scores-to-minimize-llm-hallucinations-factuality-score-llm

[5] Bang, Y., Ji, Z., Schelten, A., Hartshorn, A., Fowler, T., Zhang, C., ... & Fung, P. (2025). Hallulens: Llm hallucination benchmark. arXiv preprint arXiv:2504.17550.

[6] Li, J., Chen, J., Ren, R., Cheng, X., Zhao, W. X., Nie, J. Y., & Wen, J. R. (2024). The dawn after the dark: An empirical study on factuality hallucination in large language models. arXiv preprint arXiv:2401.03205.

[7] Khan, M., Akbar, M. A., & Kasurinen, J. (2025). Integrating LLMs in Software Engineering Education: Motivators, Demotivators, and a Roadmap Towards a Framework for Finnish Higher Education Institutes. arXiv preprint arXiv:2503.22238.

[8] Masood, A. (2025, April 16). Why large language models struggle with mathematical reasoning. Medium. Retrieved July 8, 2025, from https://medium.com/@adnanmasood/why-large-language-models-struggle-with-mathematical-reasoning-3dc8e9f964ae

[9] Russo, D., Baltes, S., van Berkel, N., Avgeriou, P., Calefato, F., Cabrero-Daniel, B., ... & Vasilescu, B. (2024). Generative ai in software engineering must be human-centered: The copenhagen manifesto. Journal of Systems and Software, 216, 112115.

[10] Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., & Kim, H. (2024). Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. *Education and Information Technologies*, *29*(9), 11483-11515.

[11] Anisuzzaman, D. M., Malins, J. G., Friedman, P. A., & Attia, Z. I. (2024). Fine-tuning llms for specialized use cases. *Mayo Clinic Proceedings: Digital Health*.

[12] Mansurova, A., Mansurova, A., & Nugumanova, A. (2024). QA-RAG: Exploring LLM reliance on external knowledge. *Big Data and Cognitive Computing*, *8*(9), 115.

[13] Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., ... & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, *2*(1).

[14] Hisham, M., Vinod, N., Kuriakose, D. L., Joshy, M., & Syama, S. Survey on Generative AI in Education.

[15] Hagos, D. H., Battle, R., & Rawat, D. B. (2024). Recent advances in generative ai and large language models: Current status, challenges, and perspectives. *IEEE Transactions on Artificial Intelligence*.

[16] Johri, A., Katz, A. S., Qadir, J., & Hingle, A. (2023). Generative artificial intelligence and engineering education. *Journal of Engineering Education*, *112*(3).

[17] Cerqueira, J. A. S. de, Khan, A. A., Rousi, R., Xi, N., Hamari, J., Kemell, K.-K., & Abrahamsson, P. (2024). Grounded Ethical AI: A Demonstrative Approach with RAG-Enhanced Agents.

[18] Yin, C., & Zhang, Z. (2024, October). A study of sentence similarity based on the all-minilm-l6-v2 model with "same semantics, different structure" after fine tuning. In *2024 2nd International Conference on Image, Algorithms and Artificial Intelligence (ICIAAI 2024)* (pp. 677-684). Atlantis Press.

[19] Gupta, A., Agarwal, D., & Bhatia, M. S. (2018, September). Performance analysis of content based image retrieval systems. In *2018 International Conference on Computing, Power and Communication Technologies (GUCON)* (pp. 899-902). IEEE.