

RAILS: Retrieval-Augmented Intelligence for Learning Software Development

Wali Mohammad Abdullah

*Mathematics & Information Technology
Concordia University of Edmonton
Edmonton, Alberta, Canada
wali.abdullah@concordia.ab.ca*

Md. Morshedul Islam

*Mathematics & Information Technology
Concordia University of Edmonton
Edmonton, Alberta, Canada
mdmorshedul.islam@concordia.ab.ca*

Devraj Parmar

*Mathematics & Information Technology
Concordia University of Edmonton
Edmonton, Alberta, Canada
dparmar@student.concordia.ab.ca*

Happy Hasmukhbhai Patel

*Mathematics & Information Technology
Concordia University of Edmonton
Edmonton, Alberta, Canada
hpatel13@student.concordia.ab.ca*

Sindhuja Prabhakaran

*Mathematics & Information Technology
Concordia University of Edmonton
Edmonton, Alberta, Canada
sprabhak@student.concordia.ab.ca*

Baidya Saha

*Mathematics & Information Technology
Concordia University of Edmonton
Edmonton, Alberta, Canada
baidya.saha@concordia.ab.ca*

Abstract—Large Language Models (LLMs) like GPT-3.5-Turbo are increasingly used to assist software development, yet they often produce incomplete code or incorrect imports, especially when lacking access to external or project-specific documentation. We introduce RAILS (Retrieval-Augmented Intelligence for Learning Software Development), a framework that augments LLM prompts with semantically retrieved context from curated Java resources using FAISS and OpenAI embeddings. RAILS incorporates an iterative validation loop guided by compiler feedback to refine suggestions. We evaluated RAILS on 78 real-world Java import error cases spanning standard libraries, GUI APIs, external tools, and custom utilities. Despite using the same LLM, RAILS outperforms baseline prompting by preserving intent, avoiding hallucinations, and surfacing correct imports even when libraries are unavailable locally. Future work will integrate symbolic filtering via PostgreSQL and extend support to other languages and IDEs.

Index Terms—Retrieval-Augmented Generation, Large Language Models, Java import resolution, code repair automation, compilation feedback

I. INTRODUCTION

Large Language Models (LLMs) such as GPT-3.5 and GPT-4 are increasingly being integrated into software development workflows due to their ability to generate, complete, and repair source code across programming languages [1], [2]. Despite these capabilities, LLMs still exhibit notable limitations, particularly hallucinated outputs, lack of awareness of project-specific context, and the generation of outdated or incorrect API usage [3], [4]. These limitations are especially critical in Java, where missing or incorrect import statements frequently cause compilation errors that are not always easy to debug or resolve automatically [5].

In this work, we define a baseline model as GPT-3.5-Turbo used in a zero-shot setting, without any external retrieval or augmentation—i.e., prompted directly with incomplete or erroneous code. This reflects how LLMs are typically used in practice and serves as our primary baseline for comparing

against RAILS. We refer to this setup as baseline prompting throughout the paper.

We introduce RAILS (Retrieval-Augmented Intelligence for Learning Software Development), a prototype framework that enhances LLM-based code generation using retrieval-augmented generation (RAG) to address these shortcomings. RAILS performs semantic search over a curated corpus of Java documentation and tutorials using a vector database and integrates relevant content directly into the prompt. Additionally, a validation loop re-invokes the model based on compiler diagnostics, enabling multi-turn refinement of broken code. This reduces hallucinations and surfaces context-aware corrections, particularly beneficial for developers who may be unfamiliar with certain imports or third-party libraries.

Our current work targets Java import-related issues, a known high-frequency source of developer frustration and build failure on platforms like Stack Overflow [3], [5]–[7]. While the focus is on Java, RAILS is designed to be language-agnostic and extensible to complex code ecosystems—including those found in high performance computing (HPC) environments. Its retrieval-feedback loop can generalize to scenarios such as missing modules in MPI/C++ workflows, OpenMP pragma misuse, and incompatible dependency chains in scientific software stacks. Future extensions include hybrid symbolic-semantic retrieval (e.g., PostgreSQL with pgvector), runtime validation, and IDE integration (e.g., Visual Studio Code), paving the way for real-time, verifiable developer assistance across academic and enterprise settings.

Contributions. This paper makes the following key contributions:

- We propose RAILS, a retrieval-augmented framework that addresses a widespread but underexplored challenge in LLM-assisted software development: import-related compilation errors.
- We design a semantic retrieval and prompt construction pipeline that augments LLM responses with contextual

documentation and integrates a validation feedback loop driven by compiler diagnostics.

- We evaluate RAILS across 78 Java import-related scenarios—including standard APIs, third-party libraries, GUIs, and custom utilities—and show that it consistently outperforms baseline prompting in semantic accuracy, compilation success, and intent preservation.

Why Use RAILS Instead of Existing Tools?

While proprietary tools like GitHub Copilot or ChatGPT plugins provide AI-powered code assistance, they often operate as black boxes—offering no control over what context is retrieved or how it is used. In contrast, RAILS offers a transparent and modular retrieval-augmented pipeline where every stage—from semantic search to compiler-informed refinement—is explicitly inspectable, configurable, and reproducible.

RAILS supports model-agnostic deployment (OpenAI or open-weight models), and allows retrieval from private or custom documentation. Its compiler feedback loop mimics real-world debugging behavior and is absent in most commercial tools. This makes RAILS well-suited for research, education, and integration into trusted software development workflows where interpretability, cost control, and fine-grained control are essential.

We organize the rest of this paper as follows: Section II presents the architecture of the RAILS framework, detailing its semantic retrieval, prompt construction, and validation loop. Section III describes the experimental setup, case selection, and benchmarking environments. Section IV summarizes the empirical findings, while Section V discusses the broader implications and potential for future extensions. Finally, Section VII concludes the paper and outlines directions for ongoing and future work.

II. SYSTEM ARCHITECTURE

RAILS is designed to enhance the accuracy of LLM-based code generation by integrating semantic retrieval and iterative validation. As illustrated in Figure 1, the architecture is composed of three key modules: a semantic retriever, a prompt construction engine, and a validation feedback loop.

- **Semantic Retriever:** Uses OpenAI embeddings and FAISS [8] to perform dense vector search over Java documentation and tutorials. This module identifies semantically similar code snippets or explanations relevant to the current coding error.
- **Prompt Construction:** Builds an augmented prompt combining the original code, retrieved context, compiler diagnostics, and previous LLM outputs. Prompting is managed using LangChain [4], enabling multi-turn refinement.
- **Validation Loop:** Compiles the generated code and feeds any resulting errors back into the next prompt iteration.

This loop mimics real-world debugging and helps guide the LLM toward a valid and contextually solution.

A. Vector Database and Retrieval

To enable semantic retrieval, RAILS integrates a vector store powered by FAISS [8], built from a curated corpus of Java documentation and tutorial content, including official Oracle guides [9]. The corpus is segmented using LangChain’s RecursiveCharacterTextSplitter, which splits documents into overlapping chunks of 300 characters with a 50-character overlap—parameters chosen to preserve context while enhancing retrieval precision.

Each chunk is embedded using OpenAI’s text-embedding-ada-002 model via the LangChain OpenAIEMBEDDINGS interface. These embeddings are indexed and persisted using FAISS. The vector index is stored locally using the save_local method and later accessed via LangChain’s retriever API. This setup allows RAILS to query the vector store at runtime and inject the most relevant knowledge chunks into the prompt for error correction.

The end-to-end data preparation and vector store creation workflow is illustrated in Figure 2.

While FAISS enables low-latency vector similarity search, it lacks symbolic querying capabilities and can struggle with large documentation sets due to memory constraints. To address this, we plan to migrate to a hybrid retrieval backend using PostgreSQL and the pgvector extension. This would allow symbolic filters (e.g., keyword or library type) to work alongside semantic similarity, enabling more precise and scalable retrieval pipelines within IDEs. This approach follows recent findings advocating hybrid retrieval for reliable system behavior [10], [11].

B. Prompt Construction

RAILS and the base prompting approach differ significantly in how they construct instructions for the LLM to generate correct code. The baseline model uses a simple prompt that only includes the broken Java code and the associated compiler error, relying entirely on the model’s internal training. This often leads to hallucinated fixes, especially in the absence of contextual knowledge [3], [4]. An example of this baseline prompt is shown in Figure 3.

However, RAILS adopts a retrieval-augmented strategy. It enriches the prompt with semantically retrieved documentation relevant to the faulty code and error messages. This additional context is pulled from a FAISS-based vector database built from Java tutorials and official documentation [8], [9], [12], [13]. The resulting prompt includes four elements: the broken code, the compiler error, the retrieved context, and an instruction to generate compilable output. Figure 4 illustrates an example of the RAILS prompt format.

Furthermore, RAILS supports iterative refinement. If a generated fix still fails to compile, the feedback loop feeds the new error message into the next prompt. This multi-turn repair mechanism follows principles proposed in [14]–[16], and supports better adaptation to specific APIs [17].

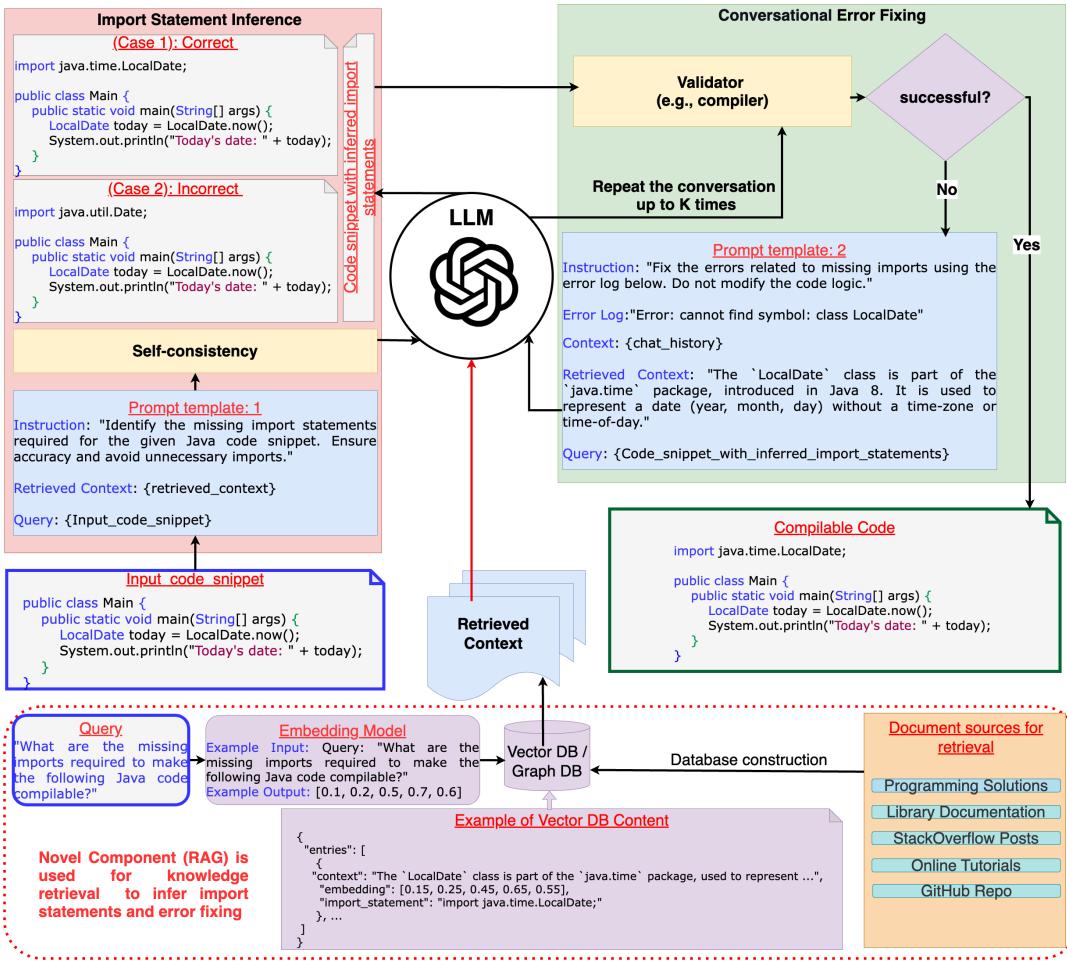


Fig. 1. RAILS methodology: semantic retrieval, prompt generation, and validation feedback loop for context-aware code repair.

```
# build_vectorstore.py
1  from langchain_community.document_loaders import TextLoader
2  from langchain_text_splitter import RecursiveCharacterTextSplitter
3  from langchain_openai import OpenAIEmbeddings
4  from langchain_community.vectorstores import FAISS
5
6  # Load and split
7  doc_loader = TextLoader("test_file.txt")
8  text_splitter = RecursiveCharacterTextSplitter(chunk_size=300, chunk_overlap=50)
9  docs = text_splitter.split_documents(doc_loader.load())
10
11 # Create embeddings and store
12 embeddings = OpenAIEmbeddings()
13 vectorstore = FAISS.from_documents(docs, embeddings)
14
15 # Save locally
16 vectorstore.save_local("faiss_index")
17 print("Vector DB saved to faiss_index")
```

Fig. 2. Code snippet for building the vector store from Java documentation using FAISS and OpenAI embeddings.

While tools like GitHub Copilot also assist with import errors, they offer limited transparency and do not expose retrieval mechanisms or compiler feedback loops. RAILS, in contrast, provides full control over retrieval sources, prompt construction, and validation steps, making it more suitable for IDE integration, debugging, and academic reproducibility.

```
prompt = f"""
You are a Java expert. Here is a code snippet with errors:
{code}

The compiler returned the following error message:
{error_msg}

Fix the code, add correct imports and make it compile.
"""

# Example usage
print(prompt)
```

Fig. 3. Baseline prompt without retrieval context.

III. EMPIRICAL EVALUATION

To assess the effectiveness of RAILS compared to baseline prompting with LLMs, we designed an evaluation framework targeting a recurring class of Java programming challenges: import-related compilation errors. We curated a total of 78 Java code snippets reflecting real-world issues frequently encountered on developer forums such as Stack Overflow. These examples span standard library omissions (e.g., `java.util.ArrayList`), deprecated API us-

```

prompt_template = PromptTemplate.from_template("""
You are a Java expert. Here is the broken code:
{input_code}

Here is the compiler error:
{error_msg}

Here is relevant documentation retrieved from the knowledge base:
{retrieved_docs}

Based on this context, fix the code so that it compiles correctly.
Add all necessary imports and correct usage.
Only return the complete Java code.
""")

```

Fig. 4. RAILS prompt enriched with retrieved documentation and error feedback.

age (e.g., `java.util.Date`), graphical components (e.g., Swing and JavaFX), external dependencies (e.g., Apache Commons, Gson), and references to user-defined utilities.

Each code sample was evaluated under two prompting strategies—baseline prompting and the RAILS framework—with compilation-based validation to assess correctness and semantic alignment. The average end-to-end latency of a RAILS iteration—including retrieval, prompt construction, and LLM inference—was approximately 4.1 seconds per case on our local machine. In contrast, baseline prompting without retrieval averaged 3.3 seconds per case. The additional latency in RAILS is primarily due to the semantic retrieval step, which accounts for roughly 20–30% of total runtime. This trade-off is offset by RAILS’s higher semantic accuracy and is expected to reduce with asynchronous retrieval in future IDE integrations. Portability was confirmed with consistent behavior on Google Colab, supporting reproducibility across environments.

A. Evaluation Setup

Experiments were conducted in two computational environments to validate performance across different hardware configurations:

- **Local Machine:** A Windows Subsystem for Linux (WSL) environment on a laptop equipped with an Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz, 16 GB RAM, and Ubuntu 24.04.
- **Google Colab (Cloud Environment):** A cloud-hosted runtime with 12.67 GB RAM and Python 3 running on a Google Compute Engine backend. This setup validated reproducibility and generalizability of RAILS in an online, server-based environment.

Each case was processed using two pipelines:

- 1) **Baseline Prompting:** GPT-3.5-turbo was prompted with the broken Java snippet and corresponding compilation error, without access to any external documentation.
- 2) **RAILS (RAG-Augmented Prompting):** GPT-3.5-turbo received the same inputs along with semantically retrieved context drawn from a FAISS vector store built from Java tutorials, API documentation, and community-contributed examples [9].

B. Experiment Pipeline

For each of the 78 code examples, both prompting approaches followed a consistent three-step evaluation loop:

- 1) The original code was compiled and error messages were logged.
- 2) The LLM (with or without RAG) generated a fix based on the code and error.
- 3) The proposed fix was recompiled, and the outcome—success, partial fix, or failure—was recorded.

Each trial produced logs capturing the LLM output, compiler responses, and (for RAILS) the retrieved documents. These logs were stored in separate files for both approaches, `log_baseline.txt` and `log_rails.txt`, to ensure consistency and traceability. We validated semantic correctness for non-compiling cases by resolving missing dependencies and confirming expected runtime behavior. Although most RAILS fixes succeeded in the first attempt, the feedback loop remains critical for future multi-turn or cross-file repair scenarios.

C. Reproducibility Details

The RAILS vector store was built using content from the official Java documentation [9]. The documents were preprocessed and segmented using LangChain’s `RecursiveCharacterTextSplitter`, producing overlapping text chunks of 300 characters with a 50-character overlap. We embedded 1,456 chunks using OpenAI’s `text-embedding-ada-002` model (API v2.1, temperature=0), resulting in a FAISS index of approximately 22MB.

During inference, the retriever used cosine similarity to return the top $k = 4$ most relevant chunks per query. These chunks were injected into the LLM prompt for context-aware code generation. To facilitate replication and transparency, all code, data, and logs associated with RAILS are available online at [18].

IV. RESULTS

We observed notable performance differences between RAILS and baseline prompting, particularly in cases involving external libraries, GUI frameworks, and project-specific utilities.

As shown in Table I, both approaches performed reliably on standard Java SE constructs such as collections, date/time handling, and file path utilities. However, a performance gap emerges in domains that require contextual or external knowledge, including external libraries (e.g., Apache Commons), JavaFX, and custom utilities.

In all 78 cases, RAILS maintained 100% semantic correctness in its proposed solutions. While 18 examples failed to compile due to missing runtime dependencies (e.g., unavailable JARs), the retrieved context and generated logic were correct. In contrast, the baseline model failed to compile 14 cases and produced semantically incorrect outputs in four others, often rewriting user-defined methods or substituting custom code with standard library approximations.

TABLE I
SEMANTIC CORRECTNESS OF CODE GENERATED BY RAILS vs.
BASELINE PROMPTING

Case Type	RAILS	Baseline Prompting
Standard JDK Import (e.g., <code>ArrayList</code> , <code>HashMap</code>)	✓	✓
Deprecated API Fix (<code>java.util.Date</code> → <code>LocalDate</code>)	✓	✓
Swing UI Component (<code>JFrame</code>)	✓	✓
Java.nio.file Usage (<code>Paths</code> , <code>Files</code>)	✓	✓
External Lib (Apache Commons I/O)	✓	✗ (Missing import)
External Lib (Gson, Text)	✓	✗ (No retrieval or awareness)
GUI with JavaFX (Alert)	✓	✗ (Incorrect or failed usage)
Custom Utility Class (e.g., <code>MySpecialUtils</code>)	✓	✗ (Hallucinated or substituted logic)

Observations:

- Both pipelines succeeded for standard Java SE examples including core classes like `ArrayList`, `HashMap`, and `Paths`.
- RAILS retrieved and integrated context for external dependencies (e.g., Commons I/O, Gson) even without having access to the libraries in the local Java environment.
- In GUI (JavaFX) and custom utility cases, RAILS preserved semantic intent; the baseline approach frequently substituted logic or fabricated fallback code.
- Hallucinations in the baseline pipeline were observed in four cases (e.g., replacing `MySpecialUtils.normalize()` with unrelated built-ins), highlighting a lack of retrieval-grounding.

V. DISCUSSION

The expanded evaluation in Section IV demonstrates that RAILS consistently produces semantically correct and intent-preserving code fixes, even in scenarios involving deprecated APIs, third-party libraries, JavaFX GUI components, and user-defined utilities. These improvements stem from RAILS’s retrieval-augmented prompting strategy, which supplements LLM reasoning with documentation-grounded context retrieved at runtime.

In contrast, baseline prompting frequently struggled to resolve import-related issues that fell outside the standard Java SE domain. Notably, the baseline model hallucinated replacement logic in four cases and failed to compile in fourteen more—typically due to missing external packages or incorrect import suggestions. RAILS, by comparison, failed to compile in 18 cases but maintained semantic correctness in all of them by issuing logically valid and well-structured code. This highlights the value of retrieval grounding in reducing semantic drift [4], [16], [17].

To further illustrate these trends, Figure 5 presents a radar chart comparing semantic correctness across five representative import-related error categories. RAILS (blue) maintains

near-perfect coverage in all categories, including external libraries and custom project code. Baseline prompting (red), however, underperforms in the JavaFX and external library domains and demonstrates zero correctness in custom utility scenarios—primarily due to hallucinations or logic substitutions.

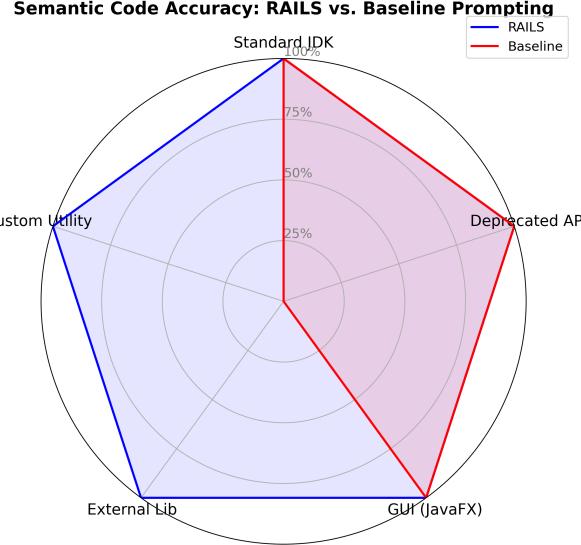


Fig. 5. Radar plot showing semantic code accuracy across categories. RAILS (blue) sustains high correctness, while Baseline (red) underperforms in complex import scenarios.

This consistency makes RAILS a strong candidate for real-world adoption. As modern software increasingly integrates external APIs, community libraries, and modular internal codebases, developer tools must handle ambiguous or context-dependent imports. Retrieval-augmented repair frameworks like RAILS can bridge this gap by offering grounded, verifiable, and reproducible solutions. In team-based environments or during onboarding, such systems improve developer confidence and reduce debugging overhead.

VI. DESIGN SCOPE AND FUTURE WORK

A. Current Scope and Model Constraints

RAILS is currently scoped to Java import repair, where semantic evaluation and compilation validation can be reliably tested. While our prototype uses GPT-3.5-turbo, RAILS supports interchangeable LLMs, including open-weight and cost-effective alternatives. Token window constraints may affect long or multi-file inputs, motivating future integration of summarization and hierarchical prompting.

B. Retrieval Precision

Our current dense retrieval setup with FAISS enables low-latency matching but may miss certain syntactic cues. We plan to adopt hybrid retrieval via PostgreSQL and pgvector, enabling symbolic filters (e.g., by library or API) to complement semantic similarity for greater precision.

C. Extensibility Across Languages and Environments

RAILS's modular design is language-agnostic and intended for broader repair tasks. Planned extensions include support for C++ and Python (e.g., OpenMP, pip imports), runtime bug repair, and plugin integration with IDEs such as VS Code. These directions aim to support real-time, context-aware developer assistance across diverse software ecosystems.

D. Evaluation Comparisons

Future evaluations will include comparisons against retrieval-only pipelines (e.g., BM25), static analysis tools, and fine-tuned small models. Our present setup isolates the contribution of retrieval-grounded prompting while preserving extensibility to multi-strategy systems.

RAILS demonstrates a practical design trade-off, prioritizing semantic fidelity over end-to-end automation, making it a reproducible and verifiable baseline for advancing LLM-powered code repair.

VII. CONCLUSION

This paper presented RAILS, a retrieval-augmented framework that enhances Java code synthesis by combining semantic search with a validation feedback loop to improve the reliability of LLM-generated code. Evaluated across 78 Java import error scenarios, including standard APIs, deprecated libraries, GUI toolkits, and custom utilities, RAILS consistently produced accurate and context-based fixes. It succeeded in 100% of semantically verifiable cases, even when compilation failed due to missing dependencies. In contrast, baseline prompting often hallucinated fixes or altered logic, especially in cases requiring third-party or project-specific knowledge.

The modular architecture of RAILS enables reproducibility, extensibility, and plug-and-play model integration, making it well-suited for modern IDEs. The planned extensions include hybrid symbolic-semantic retrieval (via pgvector), cross-language repair capabilities (e.g., Python, C++), and expanded benchmarks using automated pipelines. In the long term, our goal is to deploy RAILS as a real-time plugin for educational and industrial IDEs, providing verifiable context-aware LLM support for robust developer assistance [19], [20].

A. Ethical Considerations

RAILS sources information from publicly available documentation and community forums. As such, the system may inherit outdated practices, biased examples, or incorrect assumptions from its retrieval corpus. Future work will explore filtering and explainability strategies to align retrieved content with modern coding standards and ensure responsible LLM-assisted development.

ACKNOWLEDGMENT

This research is funded by the generous support of Concordia University of Edmonton under the Seed Grant Program. A portion of our experiments were executed on Google Colab, and we thank its compute infrastructure for enabling reproducible, cloud-based testing. We also plan to scale experiments

on the Digital Research Alliance (<https://www.alliancecan.ca>) of Canada's Graham cluster in future iterations.

REFERENCES

- [1] A. Kabir, S. Wang, Y. Tian, T.-H. Chen, M. Asaduzzaman, and W. Zhang, “Zs4c: Zero-shot synthesis of compilable code for incomplete code snippets using llms,” *ACM Transactions on Software Engineering and Methodology*, vol. 34, no. 4, pp. 1–30, 2025.
- [2] Y. Wang, W. Wang, S. Joty, and S. C. Hoi, “Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation,” *arXiv preprint arXiv:2109.00859*, 2021.
- [3] T. Zhang, D. Yang, C. Lopes, and M. Kim, “Analyzing and supporting adaptation of online code examples,” in *2019 IEEE/ACM 41st International Conference on Software Engineering (ICSE)*. IEEE, 2019, pp. 316–327.
- [4] S. Zhou, U. Alon, F. F. Xu, Z. Wang, Z. Jiang, and G. Neubig, “Docprompting: Generating code by retrieving the docs,” *arXiv preprint arXiv: 2207.05987*, 2022.
- [5] H. Zhang, S. Wang, H. Li, T.-H. Chen, and A. E. Hassan, “A study of c/c++ code weaknesses on stack overflow,” *IEEE Transactions on Software Engineering*, vol. 48, no. 7, pp. 2359–2375, 2021.
- [6] X. Xia, L. Bao, D. Lo, P. S. Kochhar, A. E. Hassan, and Z. Xing, “What do developers search for on the web?” *Empirical Software Engineering*, vol. 22, pp. 3149–3185, 2017.
- [7] H. Phan, H. A. Nguyen, N. M. Tran, L. H. Truong, A. T. Nguyen, and T. N. Nguyen, “Statistical learning of api fully qualified names in code snippets of online forums. in 2018 ieee/acm 40th international conference on software engineering (icse),” 2018.
- [8] J. Johnson, M. Douze, and H. Jegou, “Faiss: A library for efficient similarity search and clustering of dense vectors,” *arXiv*, 2017.
- [9] Oracle. (2023) The java™ tutorials: Packages and import statements. Accessed: 2025-05-23. [Online]. Available: <https://docs.oracle.com/javase/tutorial/java/package/import.html>
- [10] A. Goyal, A. Friesen, A. Banino, T. Weber, N. R. Ke, A. P. Badia, A. Guez, M. Mirza, P. C. Humphreys, K. Konyushova *et al.*, “Retrieval-augmented reinforcement learning,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 7740–7765.
- [11] A. E. Hassan, D. Lin, G. K. Rajbahadur, K. Gallaba, F. R. Cogo, B. Chen, H. Zhang, K. Thangarajah, G. Oliva, J. Lin, W. M. Abdullah, and Z. Ming (Jack) Jiang, “Rethinking software engineering in the era of foundation models: A curated catalogue of challenges in the development of trustworthy fmware,” in *Companion Proceedings of the 32nd ACM International Conference on the Foundations of Software Engineering*, 2024, pp. 294–305.
- [12] X. Gu, H. Zhang, and S. Kim, “Deep code search,” in *Proceedings of the 2018 IEEE/ACM 40th International Conference on Software Engineering (ICSE)*. IEEE, 2018, pp. 933–944.
- [13] H. Husain, H.-H. Wu, M. A. Gazit, and M. Brockschmidt, “Codesearchnet challenge: Evaluating the state of semantic code search,” *arXiv preprint arXiv:1909.09436*, 2019.
- [14] Y. Lu, M. Bartolo, A. Moore, S. Riedel, and P. Stenetorp, “Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity,” *arXiv preprint arXiv:2104.08786*, 2021.
- [15] Y. Zhang, Z. Jin, Y. Xing, and G. Li, “Steam: simulating the interactive behavior of programmers for automatic bug fixing,” *arXiv preprint arXiv:2308.14460*, 2023.
- [16] Y. Zhou, A. I. Muresanu, Z. Han, K. Paster, S. Pitis, H. Chan, and J. Ba, “Large language models are human-level prompt engineers,” in *The Eleventh International Conference on Learning Representations*, 2022.
- [17] C. K. Saifullah, M. Asaduzzaman, and C. K. Roy, “Learning from examples to find fully qualified names of api elements in code snippets,” in *2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE, 2019, pp. 243–254.
- [18] <https://anonymous.4open.science/r/rails/>, 2025.
- [19] Pinecone, “Pinecone and weaviate as tools for vector retrieval,” <https://www.pinecone.io/>, 2024, accessed: Dec. 25, 2024.
- [20] Z. Yuan, Y. Shang, Y. Zhou, Z. Dong, Z. Zhou, C. Xue, B. Wu, Z. Li, Q. Gu, Y. J. Lee *et al.*, “Llm inference unveiled: Survey and roofline model insights, 2024b,” URL <https://arxiv.org/abs/2402.16363>.