Bsc. Otavio Rodolfo Piske
Application Services - APIs, Events and Integration
Red Hat Czech s.r.o
Purkyňova 71/99, 3080/97b, & 647/111
Tel./email: +420 532 294 111 / opiske@redhat.com

Dear Ing. Viet Anh Phan,

I have been asked by you to prepare an assessment of the graduate thesis of BSc. Nikita Konovalov on the topic *Taxonomy for LLM in the Kafka Component of the Apache Camel Project*. I have studied the final version of the above-mentioned document in the form of a PDF document, source code and dataset sent to me by Nikita on June 2, 2025.

In the thesis, I was satisfied with the explanation about the topics and, specially, with the description of the diverse number of projects involved. I was particularly satisfied with the explanation about Apache Kafka. I was also satisfied to note that the Thesis did cover the security aspects of LLM-based systems, although in some aspects, it could have elaborated more about security risks in pre-training phases. I found the thesis to be well-structured, and the topics, concepts, and overall text were presented effectively. I assess it at a good level.

The project structure was sound, providing a solid foundation. Commendably, the project was functional, and the generated taxonomy was compliant with InstructLab's requirements, demonstrating its practical applicability. The generated taxonomy was also functional, a critical success.

Despite these positives, some issues detract from the project's overall quality. A key architectural misstep was the use of a RUNTIME retention policy instead of SOURCE. Given that Tabaqui analyzes the source code and isn't intended for JVM use, SOURCE retention would have been the appropriate choice. Similarly, while the annotation API is designed for Java 8 and newer, the code analysis being fixed to Java 17 is a rigidity that should ideally be a runtime configuration, allowing for greater flexibility. Additionally, the Annotation API is not available on a public repository, contrary to what had been stated on the thesis.

The decision to incorporate multiple programming languages introduces unnecessary complexity, making the project harder to maintain and diminishing its usability. Furthermore, the missing dependency requirements for the Python code create an immediate barrier to wider adoption and reproducibility.

Although, the thesis did not cover the tradeoffs of using synthetically generated data as part of the training process, nor the application provides the user with support at deciding the appropriate number of snippets to extract, I believe that there was an adequate variety of snippets in the provided solution.

None of the issues presented above, however, are critical enough to significantly penalize the presented work or disprove the practicality of the idea. None of the problems prevent the usage of the work, as I was able to create a taxonomy using Tabaqui with my own code snippets. I also highlight the novelty of the idea, which pursuits a different approach to leverage large language models to the benefits of the development community.

The student generally demonstrated consistent engagement in critical thinking and problem-solving, though there were occasional inconsistencies in interaction, such as proceeding without seeking clarification or misinterpreting requirements. Regular consultation and progress updates were observed consistently during the final three months of the thesis period, with less consistency prior to that.

The student's activity level was not entirely steady, with the bulk of concentrated effort and consultation occurring in the last two to three months before submission, despite foundational work being established earlier.

Lastly, while relevant and high-quality sources were predominantly used, improvements could be made in referencing documentation from open-source projects (e.g., including version information and "date visited") and in the inclusion of relevant industry standards like Java 2 Enterprise Edition or Jakarta EE as references for the annotation-based approach.

In view of the above, I recommend the graduate thesis for defence. Because of the problems with the presented source code, I can only award 76 points (C).

Yours sincerely
Bsc. Otavio R. Piske