

基于大语言模型的 RISC-V 软件生态使用示例生成

崔星 吴敬征 罗天悦 凌祥 王旭 芮志清

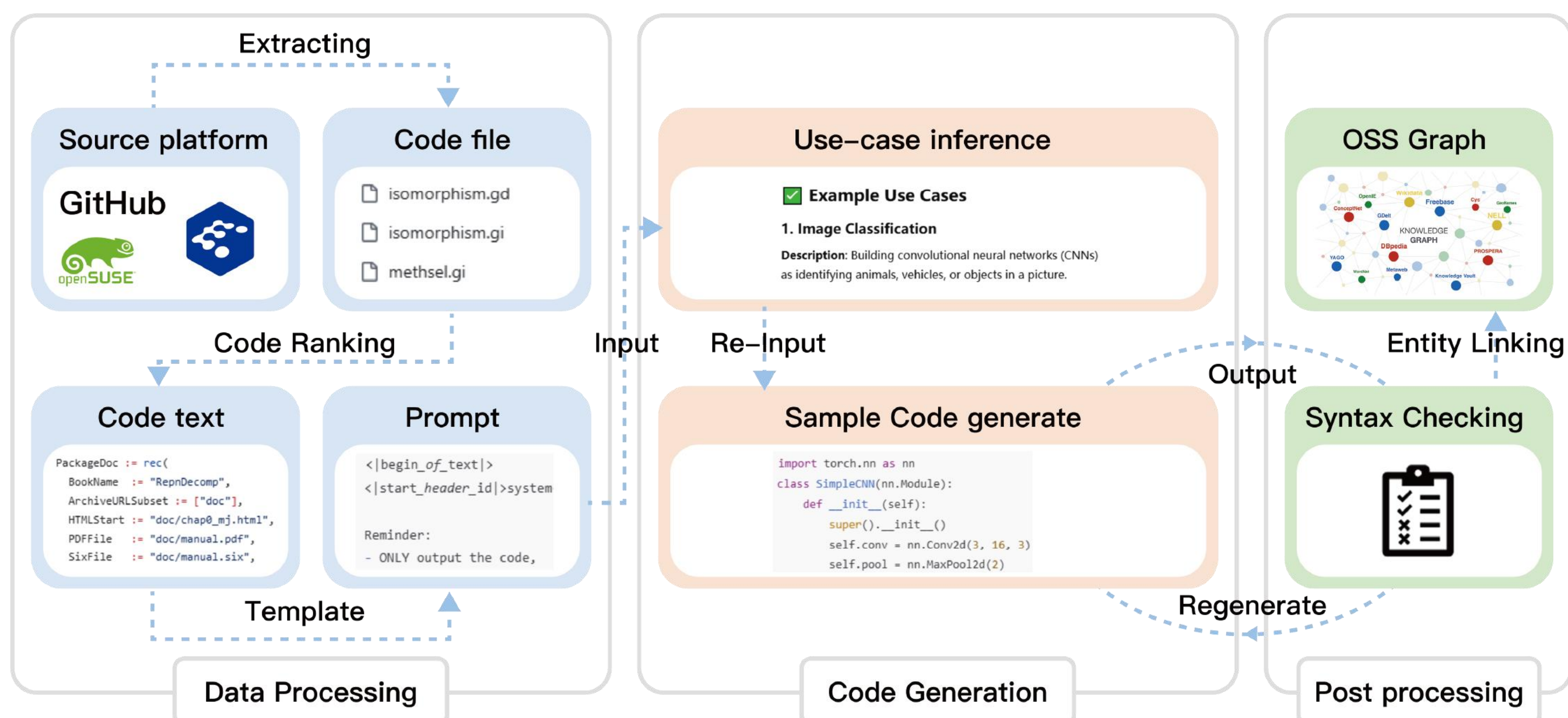
第五届 RISC-V 中国峰会

联系人：崔星，13051316652，cuixing@iscas.ac.cn

Background

As the RISC-V instruction set architecture evolves rapidly, its open-source software (OSS) ecosystem is steadily expanding to include mainstream Linux distributions (e.g., Ubuntu, Debian, openSUSE) and emerging platforms (e.g., openEuler, openKylin). Despite improved compatibility, the ecosystem still lacks high-quality, architecture-specific usage examples—resources that are essential for software adaptation, debugging, and cross-platform migration. These examples help reduce the learning curve and enhance development efficiency. Large language models (LLMs) have recently shown strong capabilities in code generation and semantic understanding, offering new opportunities for automation. However, most existing LLMs are optimized for x86 and ARM architectures, with limited support for RISC-V-specific calling patterns, usage contexts, and system interfaces, limiting their utility in this domain.

Methodology



We present a Llama-3.1-based framework for usage scenario inference and example code generation, consisting of four core components:

- OSS collection and database construction:** We collect RISC-V-compatible projects from platforms like openEuler and openSUSE to build a unified, multi-language repository.
- Semantic modeling and scenario extraction:** We use LLM-based vector encoding to model code structure and automatically extract typical usage scenarios such as initialization, API calls, and error handling.
- Example generation and refinement:** The system generates scenario-driven example code and applies heuristic post-processing to improve syntax, readability, and contextual accuracy.
- Entity linking and version control:** Generated examples are mapped to entities in a software knowledge graph to support consistent cross-version and cross-project management.

Results

Key results of this framework are as follows:

- Multi-Language Support:** The system encompasses 53 programming languages, with a primary focus on Python, C, and C++, ensuring compatibility with a wide array of software packages.
- Large-Scale Code Generation:** It has successfully generated over 9 million lines of example code, illustrating common use cases such as initialization, API utilization, and parameter configuration.
- Comprehensive Software Coverage:** The framework supports 59,707 distinct software packages within the RISC-V ecosystem, thereby facilitating comparative analysis and cross-architecture migration.
- Continuous Integration and Updates:** To maintain currency and relevance, the system automatically synchronizes with nine data sources daily to refresh its source code base and update the generated examples.

Contributions

- We apply LLMs to RISC-V-specific example generation, addressing a key gap in cross-architecture development.
- We build a heterogeneous knowledge base across operating systems and code sources to support large-scale scenario generation.
- We introduce a knowledge graph-based entity linking mechanism to enhance contextual coherence and maintainability.

Conclusions

This work demonstrates the potential of LLMs in supporting the RISC-V software ecosystem by addressing the critical gap in usage examples. Leveraging cross-platform OSS analysis and automated example generation, we build a scalable, structured, and semantically aware code generation framework. This lays a solid foundation for future efforts in automated software adaptation, cross-architecture debugging, and developer assistance.



This work is supported by
Open Source Map