LLM-assisted Performance Estimation of Embedded Software on RISC-V Processors

Weiyan Zhang[1], Muhammad Hassan[1,2], Rolf Drechsler[1,2]
[1] Institute of Computer Science, Universität Bremen, Germany
[2] Cyber-Physical Systems, DFKI GmbH, Germany
{weiyan, hassan, drechsler}@uni-bremen.de

Motivation

High-level performance estimation using Large Language Models (LLMs) and Machine Learning (ML) offers a promising solution to the growing complexity of embedded systems. We explore the use of LLMs to extract performance-related information from RTL source code and documentation, enabling efficient performance analysis of RISC-V processors.

The proposed method combines LLM-assisted text processing with ML techniques to train Predictive Models (PMs). By extracting data from RTL, we aim to achieve accurate performance estimation without the need for RTL simulations, accelerating the design and optimization process.

Evaluation

- LLM: ChatOpenAI (gpt-3.5-turbo-0125)
- Embedding model: SentenceTransformer
- Testing benchmarks: TACLeBench
- Functional Simulator: Whisper

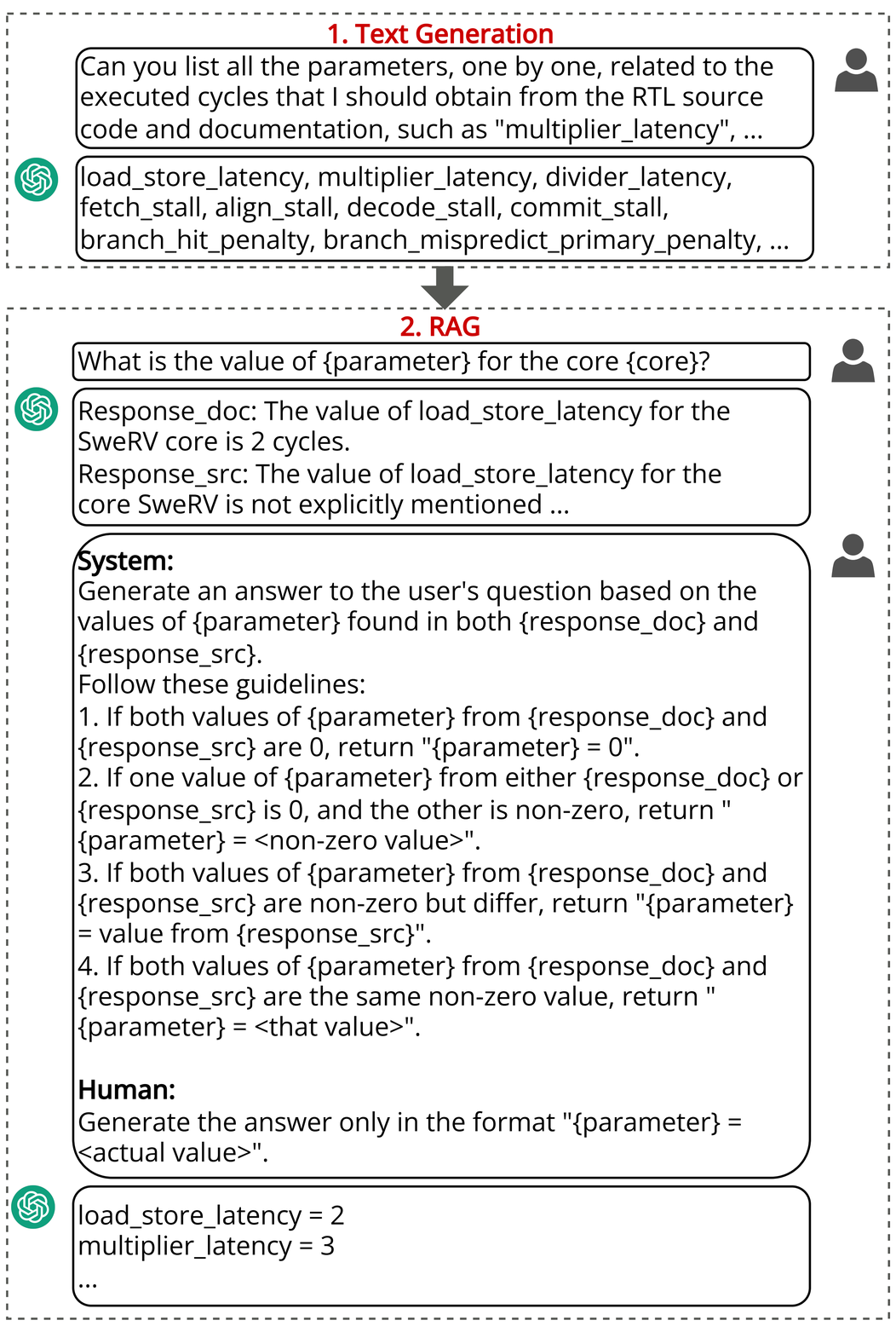

Figure 1: Illustration of user-LLM interactions during phases 1 and 2 for extracting performance-related parameters from RTL documentation and source code.

Proposed Methodology

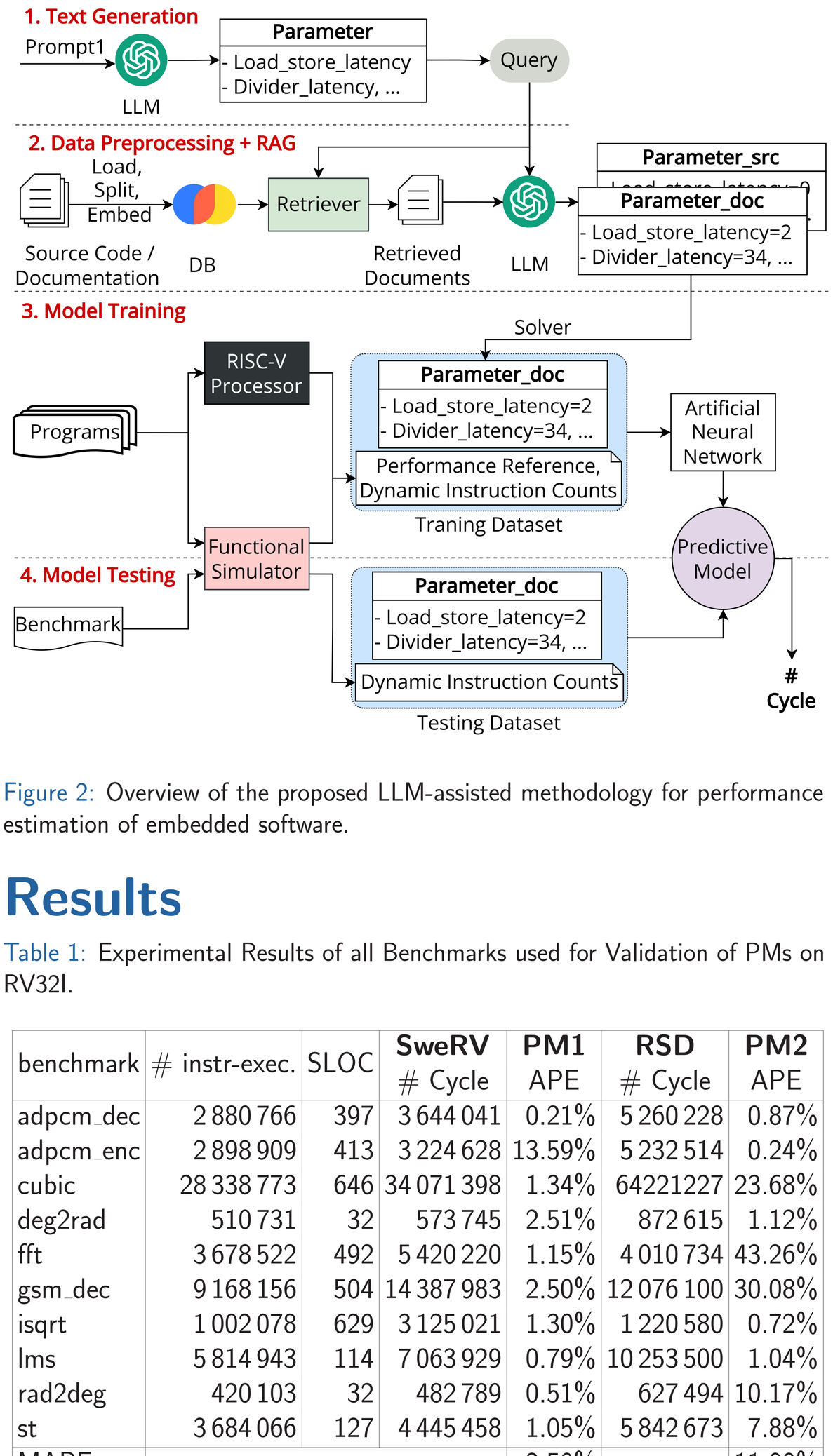

Figure 2: Overview of the proposed LLM-assisted methodology for performance estimation of embedded software.

Results

Table 1: Experimental Results of all Benchmarks used for Validation of PMs on RV32I.

benchmark	# instr-exec.	SLOC	**SweRV** # Cycle	**PM1** APE	**RSD** # Cycle	**PM2** APE
adpcm_dec	2 880 766	397	3 644 041	0.21%	5 260 228	0.87%
adpcm_enc	2 898 909	413	3 224 628	13.59%	5 232 514	0.24%
cubic	28 338 773	646	34 071 398	1.34%	64221227	23.68%
deg2rad	510 731	32	573 745	2.51%	872 615	1.12%
fft	3 678 522	492	5 420 220	1.15%	4 010 734	43.26%
gsm_dec	9 168 156	504	14 387 983	2.50%	12 076 100	30.08%
isqrt	1 002 078	629	3 125 021	1.30%	1 220 580	0.72%
lms	5 814 943	114	7 063 929	0.79%	10 253 500	1.04%
rad2deg	420 103	32	482 789	0.51%	627 494	10.17%
st	3 684 066	127	4 445 458	1.05%	5 842 673	7.88%
MAPE				2.50%		11.90%

Compared to the state-of-the-art method in [1], which reported higher Mean APE (MAPE) for SweRV (6.5%) and RSD (18.6%), our methodology MAPE by 61.54% for SweRV and 36.02% for RSD, demonstrating significant improvements in accuracy.

Selected Publication

[1] W. Zhang, M. Goli, M. Hassan, & R. Drechsler. Efficient ML-based performance estimation approach across different microarchitectures for RISC-V processors, DSD 2023.

Funded by:
Bundesministerium für Bildung und Forschung
Grant number 16ME0127 and 01IW22002
Scale4Edge
ECXL
Universität Bremen
dfki ai Deutsches Forschungszentrum für Künstliche Intelligenz
German Research Center for Artificial Intelligence