ASE 2026

AcslBench: A Verified C/ACSL Corpus and Benchmark with Compositional Call Chains for Specification Synthesis

Junjie Hu, Cheng Wen, Bin Yu, Jialun Cao, Dugang Liu, Weidi Sun, Haokun Li, Shengchao Qin, Cong Tian

1 Guangzhou Institute of Technology, Xidian University 2 HKUST 3 Shenzhen University 4 Peking University

INTRODUCTION

Formal verification needs verified training data, but for C/ACSL it barely exists: all prior datasets together have fewer than 700 programs, while Lean and Rust/Verus already have millions.

We present AcslBench, built by a verifier-in-the-loop pipeline: an LLM translates verified Rust/Verus seeds into C code with ACSL annotations; Frama-C/WP checks every proof obligation; failed candidates are repaired with verifier diagnostics, up to five iterations. This yields 504,587 fully verified samples (52.9% → 71.5%).

The same components are composed into a 495-task benchmark across function calls. Six LLMs are evaluated: the best reaches only 72.5% pass@5. Dataset is now open-source on huggingface.

PIPELINE OVERVIEW

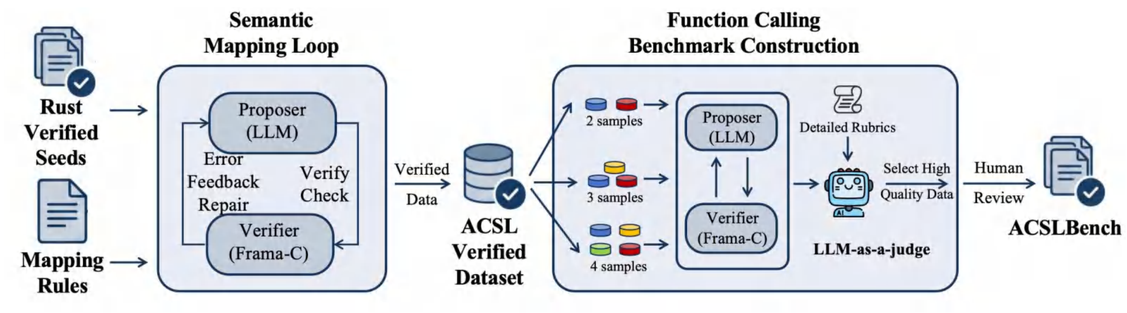


Figure 1: Verifier-in-the-Loop Corpus Construction Pipeline

Seeds — ~0.7M verified Rust/Verus programs from VeruSyn (6.9M)
Translation — LLM maps seeds into C + ACSL
Verification — Frama-C/WP discharges every proof obligation
Repair — failed candidates return with diagnostics, up to 5 iterations

KEY CONTRIBUTIONS

1 Verified Corpus

504,587 Frama-C/WP-verified C/ACSL programs — the first C/ACSL corpus at training scale, three orders of magnitude beyond all prior datasets.

1 Compositional Benchmark

495 curated tasks with 2–10 functions per program, stratified into Simple / Medium / Complex by call count.

1 Synthesis Method

Verifier-in-the-loop translation and composition — **reusable** to synthesize new specification tasks and benchmarks.

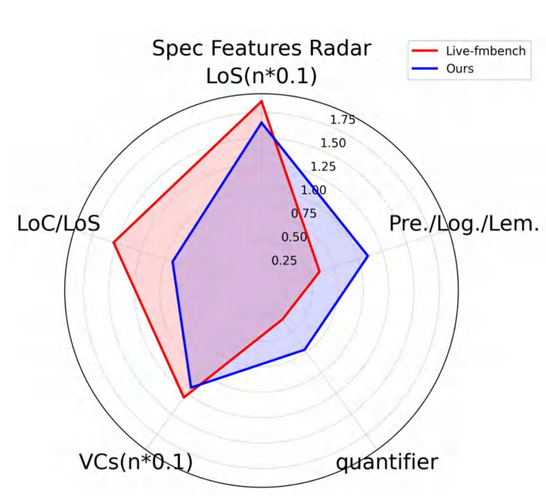


Figure 2: Code Features vs. Live-FMBench

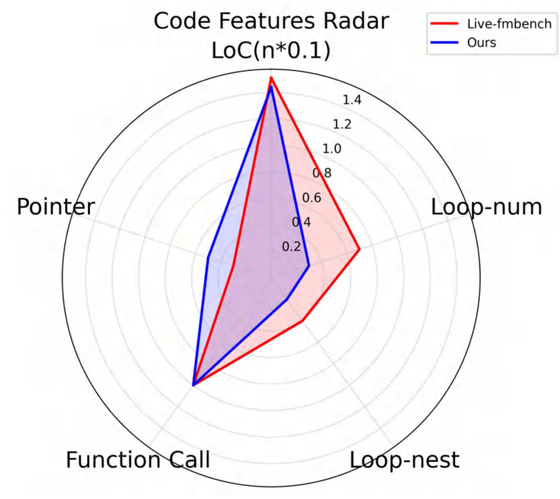


Figure 3: Spec. Features vs. Live-FMBench

EVALUATION & RESULTS ANALYSIS

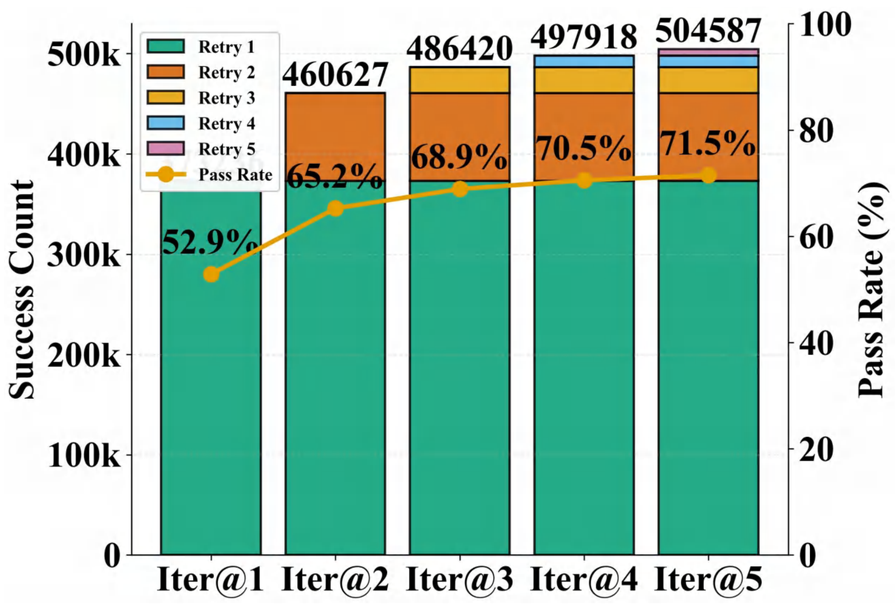


Figure 4: Verified Samples and Cumulative Pass Rate across Repair Iterations (373,236 → 504,587; 52.9% → 71.5%)

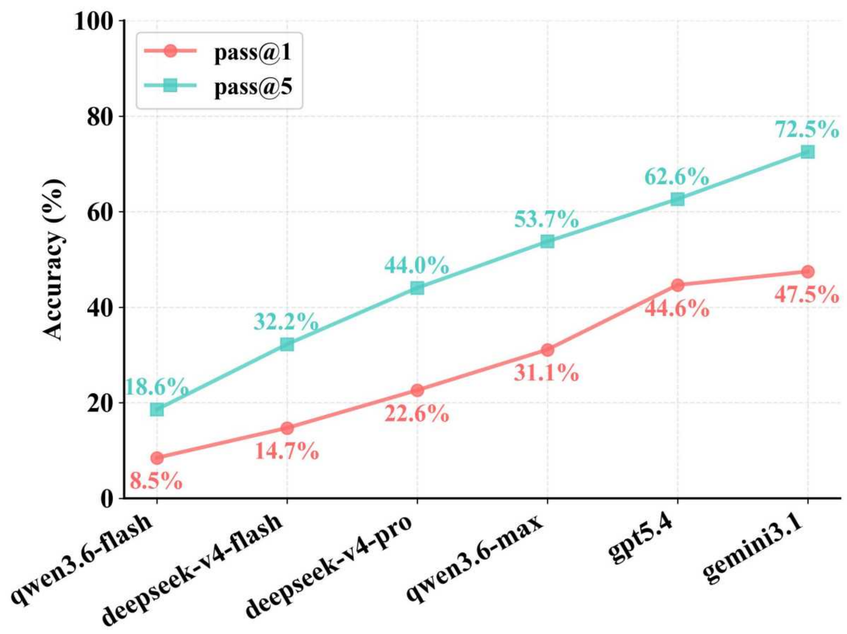


Figure 5: Pass@1 and Pass@5 of Six LLMs on AcslBench

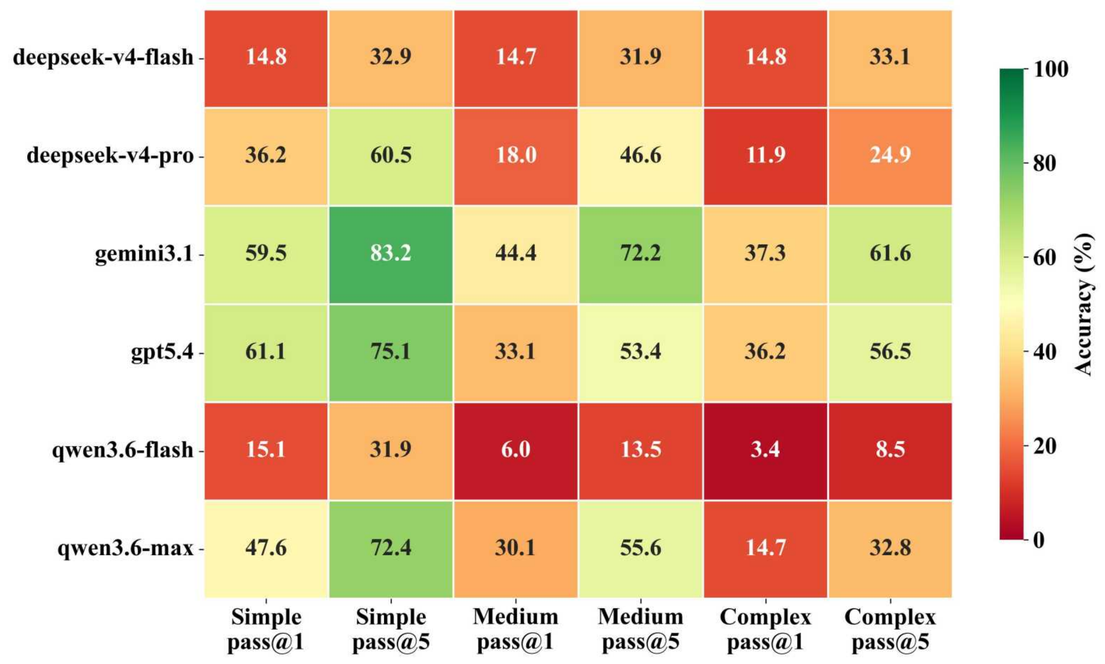


Figure 6: Model Performance by Call-Count Level (Simple / Medium / Complex)

CONCLUSION

AcslBench provides the first large-scale verified C/ACSL corpus (504,587 samples) and a 495-task compositional benchmark. Verifier feedback proved to be a repair signal — five iterations recovered 131,351 samples (52.9% → 71.5%).

However, the corpus is translated from Rust/Verus seeds and may inherit their style rather than native C patterns, and tasks are grouped by call count, not true call-chain depth. Future work includes a multi-language dataset — the same task in multiple specification languages.