SysMoBench

Evaluating AI on Formally Modeling Complex Real-World Systems

Q. Cheng, R. Tang, E. Ma, F. Hackett, P. He, Y. Su, I. Beschastnikh, Y. Huang, X. Ma, T. Xu

Why focus on formal methods?

They protect systems.

We use formal methods (TLA+) to verify safety-critical infra.

They're expensive.

It takes months to years of expert effort per system model.

Can AI help?

Are models capable of formal specification generation?

No appropriate benchmarks

Existing work has limited scope.

Pre/post conditions, loop invariants, and logic puzzles.

No eval on *systems*.

Fundamentally different, and harder to automatically evaluate.

(Rego et al., 2025; Xie et al., 2025; Cao et al., 2025; Chakraborty et al., 2025; Ma et al., 2025; Wen et al., 2024)

SysMoBench

11 System Artifacts

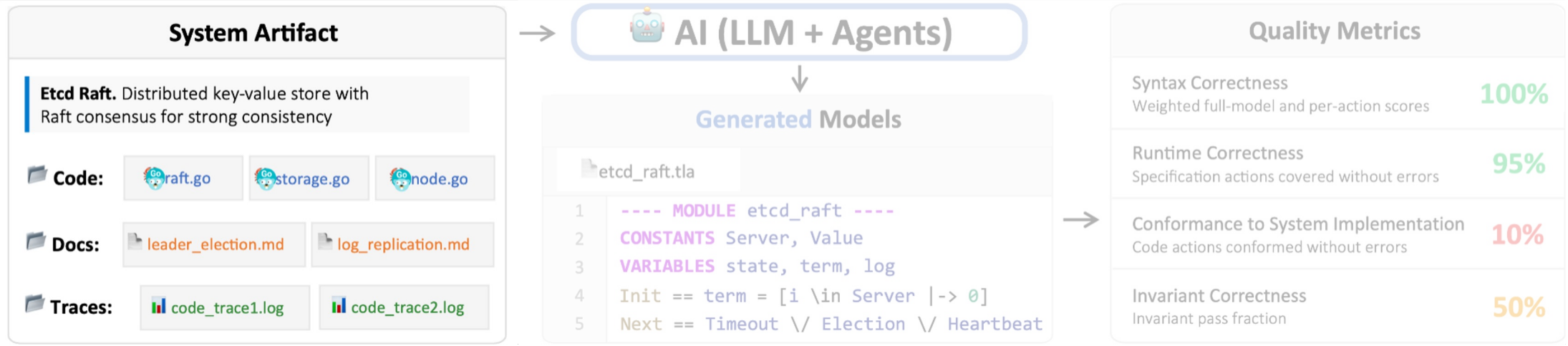
Rust, Go, Java, C

3 Agents

Basic modelling, code translation, and trace learning

4 Frontier LLMs

Claude Sonnet 4, GPT-5, Gemini 2.5 Pro, and DeepSeek-R1



SysMoBench

11 System Artifacts

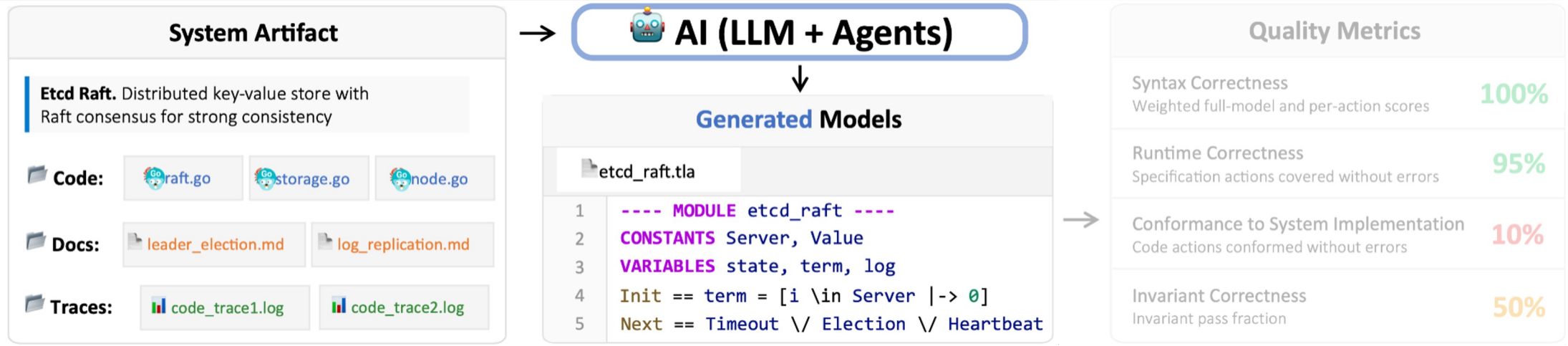
Rust, Go, Java, C

3 Agents

Basic modelling, code translation, and trace learning

4 Frontier LLMs

Claude Sonnet 4, GPT-5, Gemini 2.5 Pro, and DeepSeek-R1



SysMoBench

11 System Artifacts

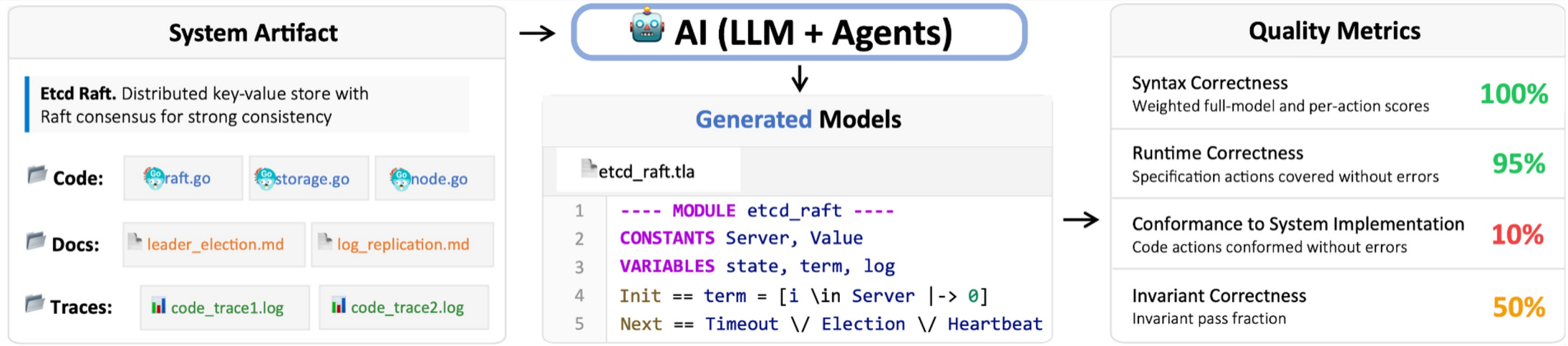
Rust, Go, Java, C

3 Agents

Basic modelling, code translation, and trace learning

4 Frontier LLMs

Claude Sonnet 4, GPT-5, Gemini 2.5 Pro, and DeepSeek-R1



Examples

(a) Asterinas Spinlock

Agent	LLM	Syntax	Runtime	Conformance	Invariant
Basic Modeling	Claude-Sonnet-4	100.00% ✓	100.00% ✓	100.00%	100.00%
	GPT-5	100.00% ✓	100.00% ✓	80.00%	100.00%
	Gemini-2.5-Pro	100.00% ✓	100.00% ✓	80.00%	85.71%
	DeepSeek-R1	100.00% ✓	100.00% ✓	80.00%	100.00%
Code Translation	Claude-Sonnet-4	100.00% ✓	100.00% ✓	100.00%	100.00%
	GPT-5	100.00% ✓	100.00% ✓	100.00%	85.71%
	Gemini-2.5-Pro	100.00% ✓	100.00% ✓	100.00%	100.00%
	DeepSeek-R1	100.00% ✓	100.00% ✓	100.00%	100.00%

(b) Etcd Raft

Agent	LLM	Syntax	Runtime	Conformance	Invariant
Basic Modeling	Claude-Sonnet-4	100.00% ✓	25.00% ✓	7.69%	69.23%
	GPT-5	47.87% ✗	-	-	-
	Gemini-2.5-Pro	50.00% ✗	-	-	-
	DeepSeek-R1	50.00% ✗	-	-	-
Code Translation	Claude-Sonnet-4	100.00% ✓	66.67% ✓	15.38%	92.31%
	GPT-5	100.00% ✓	20.00% ✗	-	-
	Gemini-2.5-Pro	44.44% ✗	-	-	-
	DeepSeek-R1	100.00% ✓	0.00% ✗	-	-

Spinlock: ~100%

Simpler concurrent lib, ~200 LoC.

Etcd Raft: 67% max

Most models even fail syntax validation.

Examples

(a) Asterinas Spinlock

Agent	LLM	Syntax	Runtime	Conformance	Invariant
Basic Modeling	Claude-Sonnet-4	100.00% ✓	100.00% ✓	100.00%	100.00%
	GPT-5	100.00% ✓	100.00% ✓	80.00%	100.00%
	Gemini-2.5-Pro	100.00% ✓	100.00% ✓	80.00%	85.71%
	DeepSeek-R1	100.00% ✓	100.00% ✓	80.00%	100.00%
Code Translation	Claude-Sonnet-4	100.00% ✓	100.00% ✓	100.00%	100.00%
	GPT-5	100.00% ✓	100.00% ✓	100.00%	85.71%
	Gemini-2.5-Pro	100.00% ✓	100.00% ✓	100.00%	100.00%
	DeepSeek-R1	100.00% ✓	100.00% ✓	100.00%	100.00%

(b) Etcd Raft

Agent	LLM	Syntax	Runtime	Conformance	Invariant
Basic Modeling	Claude-Sonnet-4	100.00% ✓	25.00% ✓	7.69%	69.23%
	GPT-5	47.87% ✗	-	-	-
	Gemini-2.5-Pro	50.00% ✗	-	-	-
	DeepSeek-R1	50.00% ✗	-	-	-
Code Translation	Claude-Sonnet-4	100.00% ✓	66.67% ✓	15.38%	92.31%
	GPT-5	100.00% ✓	20.00% ✗	-	-
	Gemini-2.5-Pro	44.44% ✗	-	-	-
	DeepSeek-R1	100.00% ✓	0.00% ✗	-	-

Spinlock: ~100%

Simpler concurrent lib, ~200 LoC.

Etcd Raft: 67% max

Most models even fail syntax validation.

Key Findings

LLMs struggle with temporal reasoning.

41.9% of timing-related invariants were violated.

LLM-generated specs can find real bugs.

5 real-world critical bugs reproduced.

LLMs remain limited in system comprehension.

Moving forward: we should aim to improve this!

Conclusion

- How do we understand and specify today's complex systems?
- **SysMoBench**: novel benchmark measuring LLM formal methods capabilities on real-world systems (Etcd Raft, Apache Zookeeper)
  - 11 system artifacts, 3 agents, 4 models
- Code open-source, we welcome contributions!

NANJING UNIVERSITY

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

THE UNIVERSITY OF BRITISH COLUMBIA

Microsoft Research 微软亚洲研究院

github.com/specula-org/SysMoBench