

Design Principles for Systems-meets-AI

Vyas Sekar
Inspired by many students and collaborators

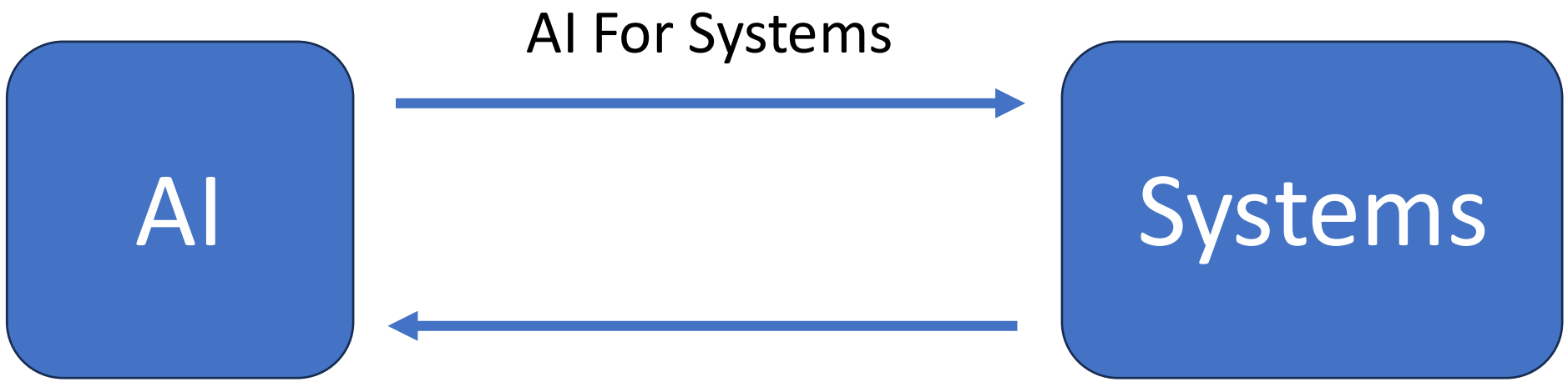
Carnegie Mellon University

CONVIVA

ROCKFISH DATA

AI-meets-Systems

e.g., Coding agents, Auto evolve, Troubleshooting, Design



I know close to nothing about this

e.g., Datacenter fabric, protocols, KV cache, RDMA,

State of the world today

It was the best of times, it was the worst of times,
it was the age of wisdom, it was the age of foolishness,
it was the epoch of belief, it was the epoch of incredulity,
it was the season of Light, it was the season of Darkness,
it was the spring of hope, it was the winter of despair,
we had everything before us, we had nothing before us



Our ability to build and ship today far exceeds our ability to understand, manage, and secure

Two opposing camps in AI-meets-Systems

Camp 1: AI pilled

- RL will solve everything
- Self play or RSI will solve everything
- Models are winning all benchmarks, AGI/ASI is here and there is nothing left to do for us
- Bitter lesson → give up, become data labelers and RL simulators and wait for singularity
- We don't need domain expertise, data is the source of truth

Camp 2: AI Skeptic

- These things are just benchmaxxing
- Intelligence is jagged and overfitted
- It doesn't actually work for any worthwhile real-world task
- We need neurosymbolic systems
- We need classical systems

My position on AI-meets-Systems



"Pilled vs Skeptic" debate is irrelevant ☺

Instead, focus on pragmatic questions:

Q1: Are systems problems different?

Q2: How should we use this new capability?

My position on AI-meets-Systems

Q1: Are systems problems different?

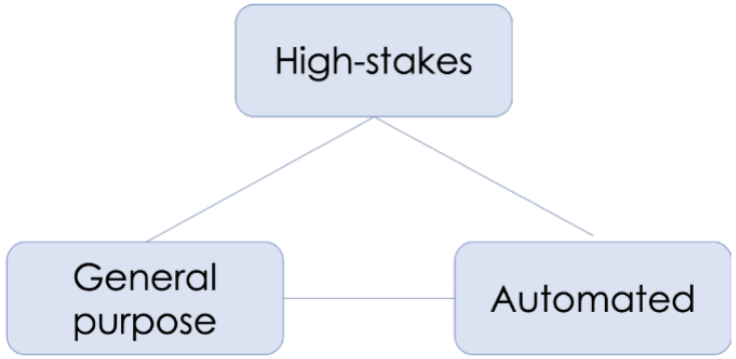
Ans: ?

Q2: How should we use this new capability?

The AI Trilemma

Prediction: collaboration agents will continue to be more successful than automation agents.

Many hasty deployments are being rolled back.



The AI agent trilemma:

You can only have two

Credit: Arvind Narayanan, ICML Keynote

Examples of AI Fails in Systems



AI/ML for Network Security: The Emperor has no Clothes

https://trusteeml.github.io/

An AI-powered coding tool wiped out a software company's database, then apologized for a 'catastrophic failure on my part'

AI coding governance just got real, our token bill hit six figures and now the CFO cares

AI Trilemma Gets Amplified in Systems

Prediction: collaboration agents will continue to be more successful than automation agents.

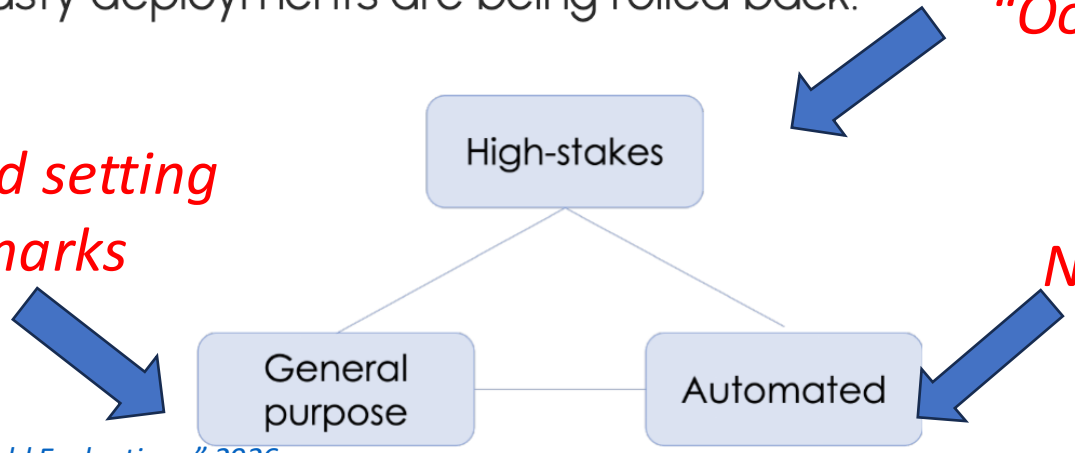
Many hasty deployments are being rolled back.

"Oops" is not an option

Diverse open-world setting

Not closed benchmarks

Needs to work at scale



[Kapoor, Narayanan et al., "Open-World Evaluations," 2026]

The AI agent trilemma:

You can only have two

Credit: Arvind Narayanan, ICML Keynote

AI-for-Systems vs. Other AI Domains

	Chat / Email	Math	AI-for-Systems
Ground truth	Human preference	Formally checkable	**No checker, only invariants**
Reversibility	Fully reversible	Fully reversible	**Often irreversible**
Adversary and Externalities	None Stationary	None Stationary	**Active adversaries** **Constantly shifting environment**

See:
AI for Systems is “AGI-Complete” ACM SIGOPS OSR
Open-World Evaluations for Measuring Frontier AI Capabilities https://arxiv.org/abs/2605.20520

Claim: ***Reliability*** is the real frontier

Not benchmark leaderboards or “agent last exam.”
The open-world + trilemma means scores don’t matter.

“Reliability is defined as the probability that a product, system, or service will perform its intended function adequately for a specified period of time; or will operate in a defined environment without failure”

Intended → Resource usage, Cost, Accuracy, Failure, Scale

See: Towards a Science of AI Agent Reliability https://arxiv.org/pdf/2602.16666v1
https://en.wikipedia.org/wiki/Reliability_engineering

AI-for-Systems vs. Other AI Domains

	Chat / Email	Math	AI-for-Systems
Ground truth	Human preference	Formally checkable	**Some checkers, only invariants**
Reversibility	Fully reversible	Fully reversible	**Often irreversible**
Adversary	None	None	**Active adversaries**
Environment	Stationary	Stationary	**Constantly shifting**
Performance/Cost	Important	Don’t care	**Very important**

My position on AI-meets-System

Q1: Are systems problems different?

AI trilemma gets worse.
Systems is an Open World problem
Reliability is the real frontier

Q2: How should we use this new capability?

Design Principles for AI-meets-Systems

Not novel, not rocket science, but worth reiterating
Also not a fixed point, my views and space are evolving

D1. Avoid rigid solutions; exploit the best of all paradigms

D2. Make explainable abstractions a first-class requirement

D3. Explicitly define cost, resource, reliability criteria

D4. Treat evaluations from an adversary point of view

Aside: Inspiration for Design Principles

Hints and Principles for Computer System Design

Butler Lampson

May 14, 2021

Abstract

This new long version of my 1983 paper suggests the goals you might have for your system—Simple, Timely, Efficient, Adaptable, Dependable, Yummy (STEADY)—and techniques for achieving them—Approximate, Incremental, Divide & Conquer (AID). It also gives some principles for system design that are more than just hints, and many examples of how to apply the ideas.

https://arxiv.org/pdf/2011.02455



https://thefrugalarchitect.com/

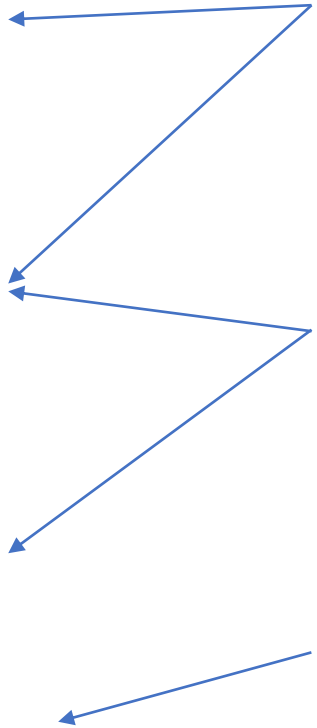
From Principles to Practice

D1. Avoid rigid solutions; exploit the best of all paradigms

D2. Make explainable abstractions a first-class requirement

D3. Explicitly define cost, resource, reliability criteria

D4. Treat evaluations from an adversary point of view



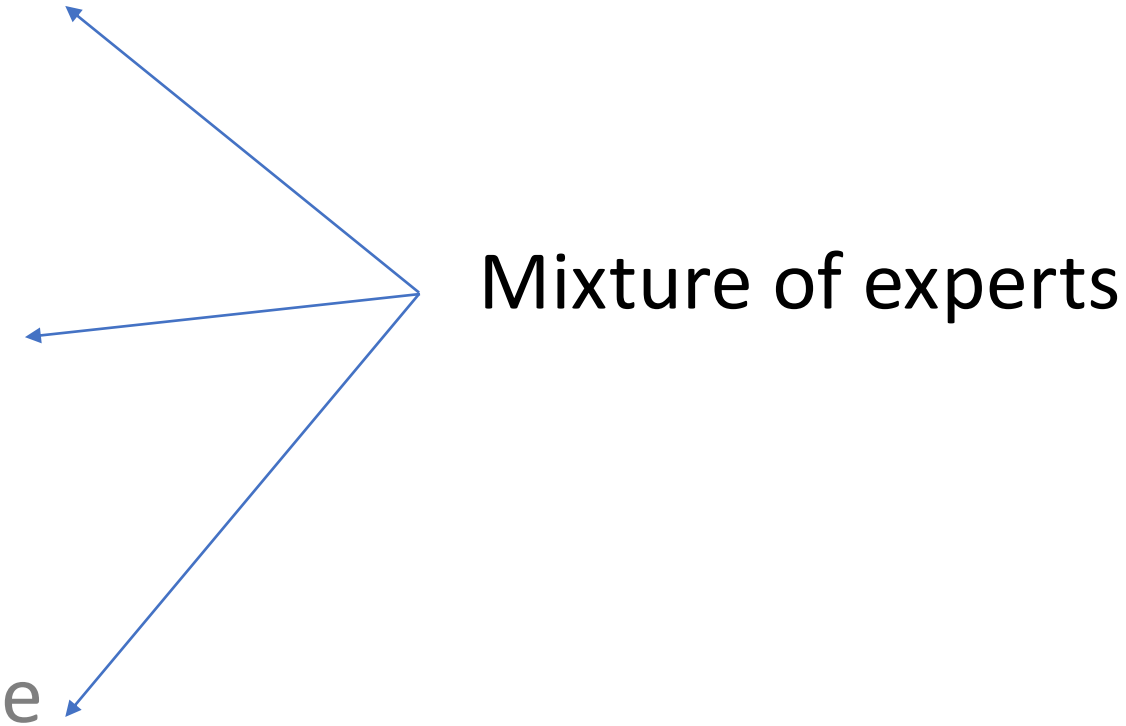
Mixture of Experts

Constrained Creativity

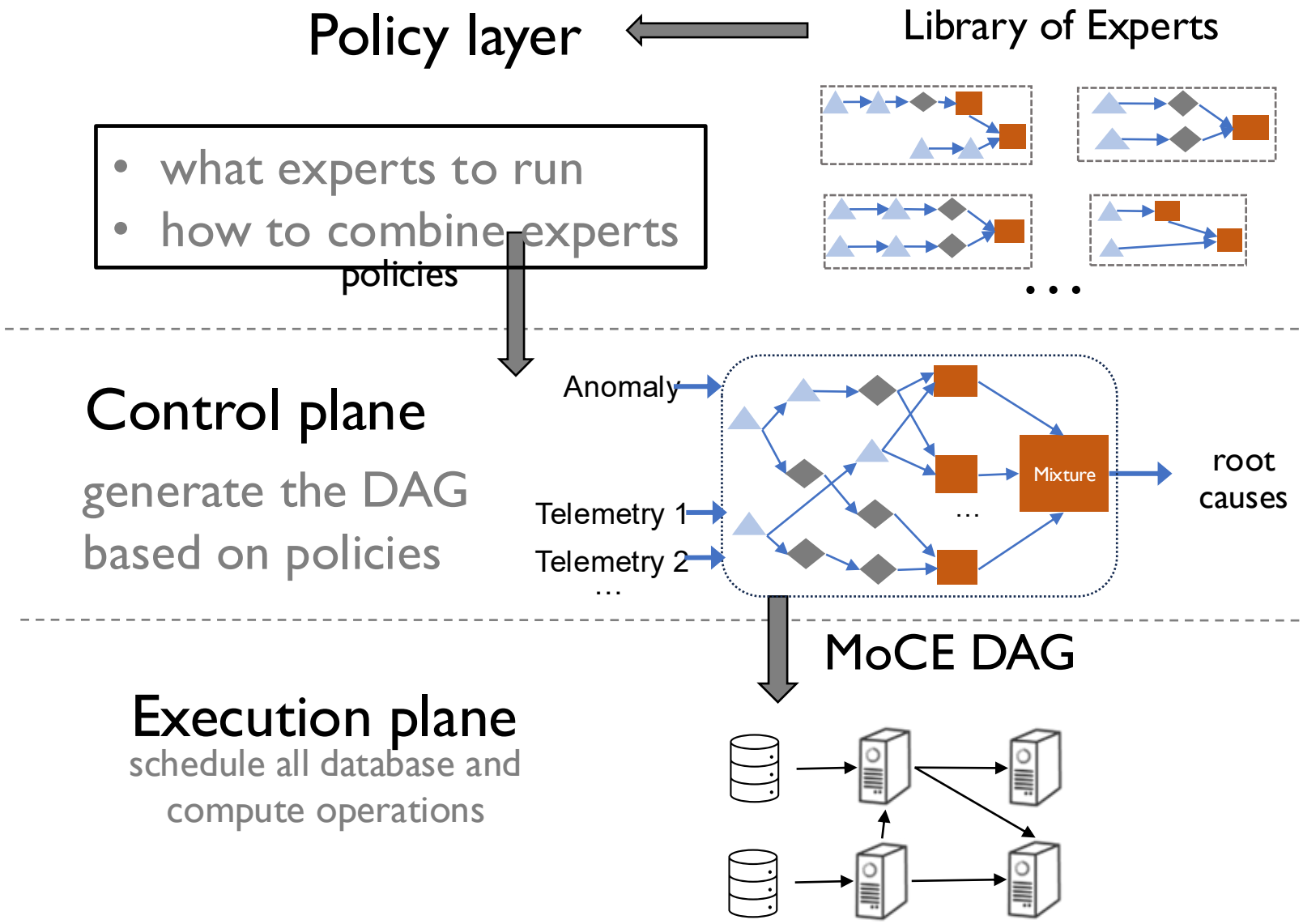
AI-assisted Red Teaming

Idea: Mixture of Experts

1. Expressivity: cover diverse types of failures
2. Extensibility: enable fast development of new algorithms
3. Scalability: run in reasonable time with reasonable resources



Illustrative Example: MoCE for RCA



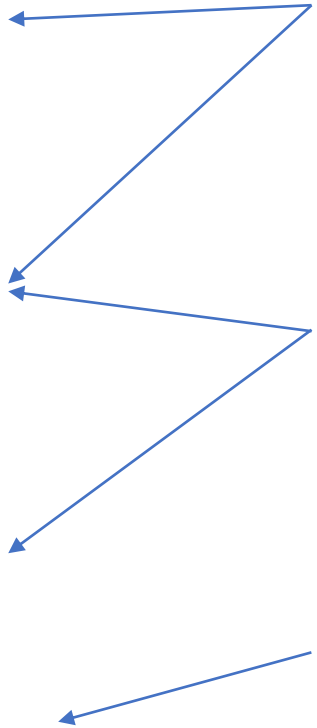
From Principles to Practice

D1. Avoid rigid solutions; exploit the best of all paradigms

D2. Make explainable abstractions a first-class requirement

D3. Explicitly define cost, resource, reliability criteria

D4. Treat evaluations from an adversary point of view



Mixture of Experts

Constrained Creativity

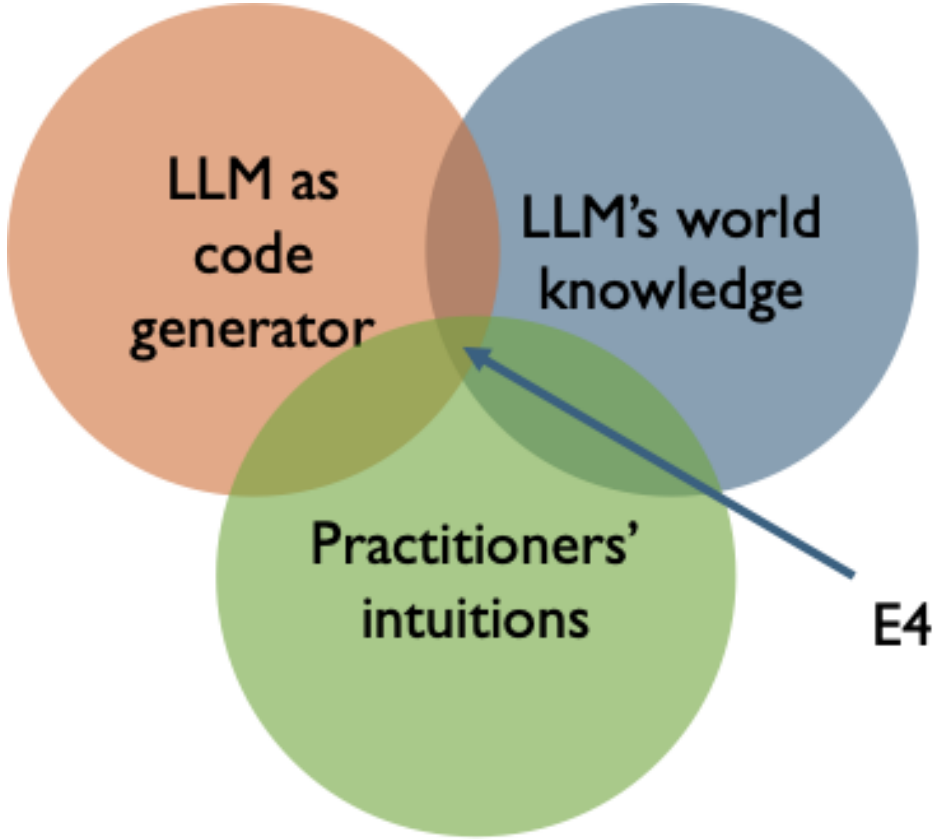
AI-assisted Red Teaming

Idea: Constrained Creativity

Lower barrier to create new hypothesis
Go beyond "human creativity"

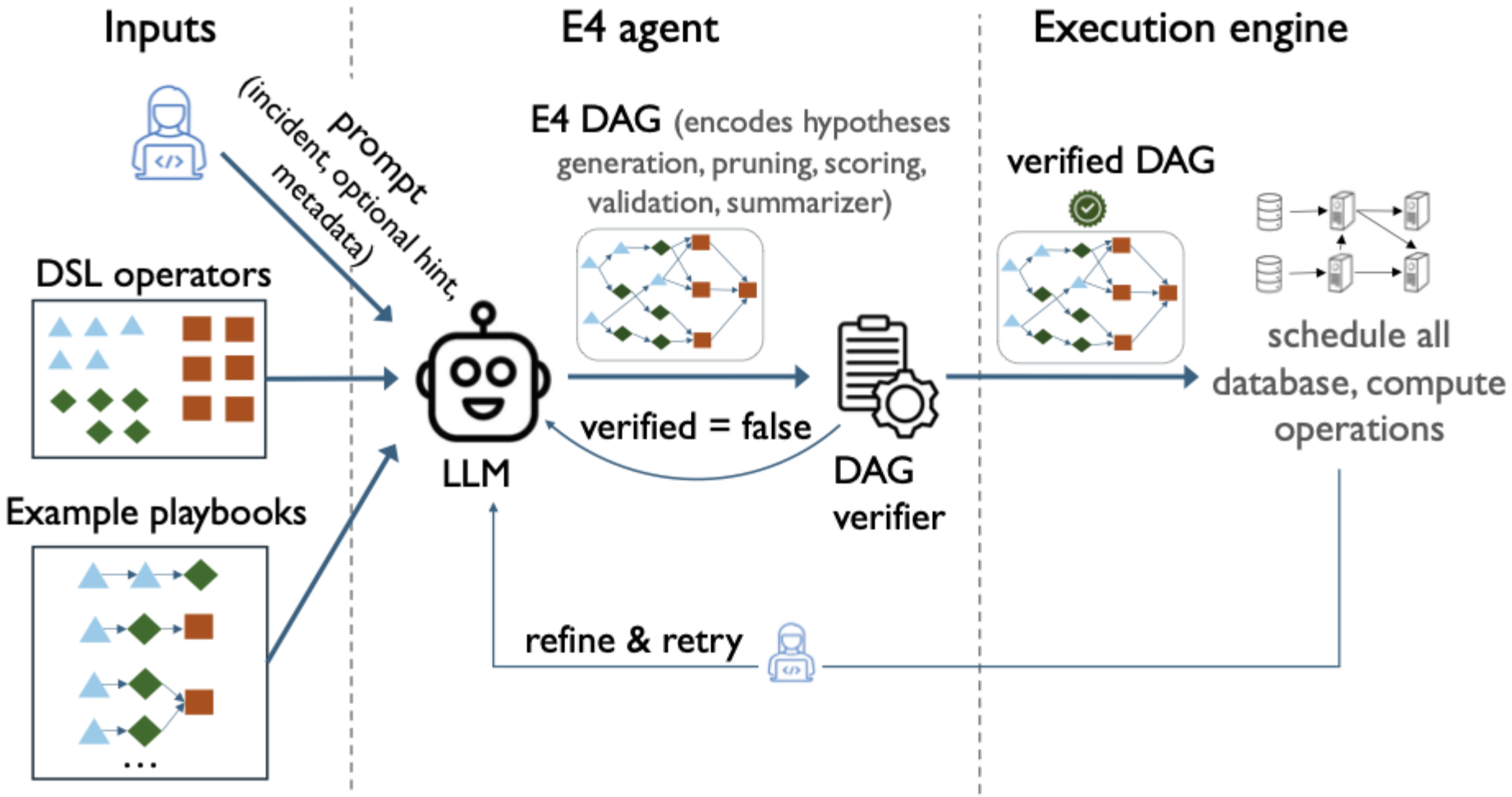
Constrain the LLM into a functional DSL

Start with the simplest building blocks, adding complexity in **explainable**, staggered iterations.



Parallel effort: UW's Self-Defining Systems (Anderson, Mahajan, Peter, Zettlemoyer)

Illustrative Example: E4 for Agentic RCA



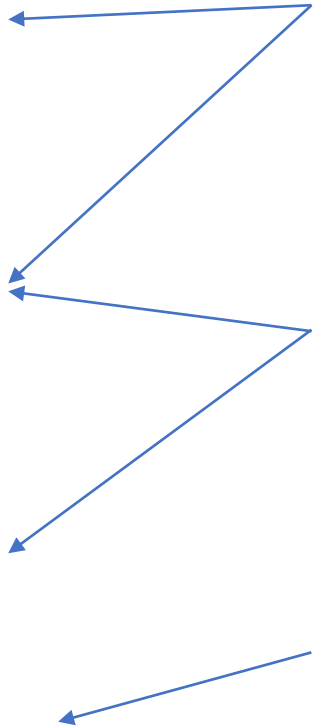
From Principles to Practice

D1. Avoid rigid solutions; exploit the best of all paradigms

D2. Make explainable abstractions a first-class requirement

D3. Explicitly define cost, resource, reliability criteria

D4. Treat evaluations from an adversary point of view

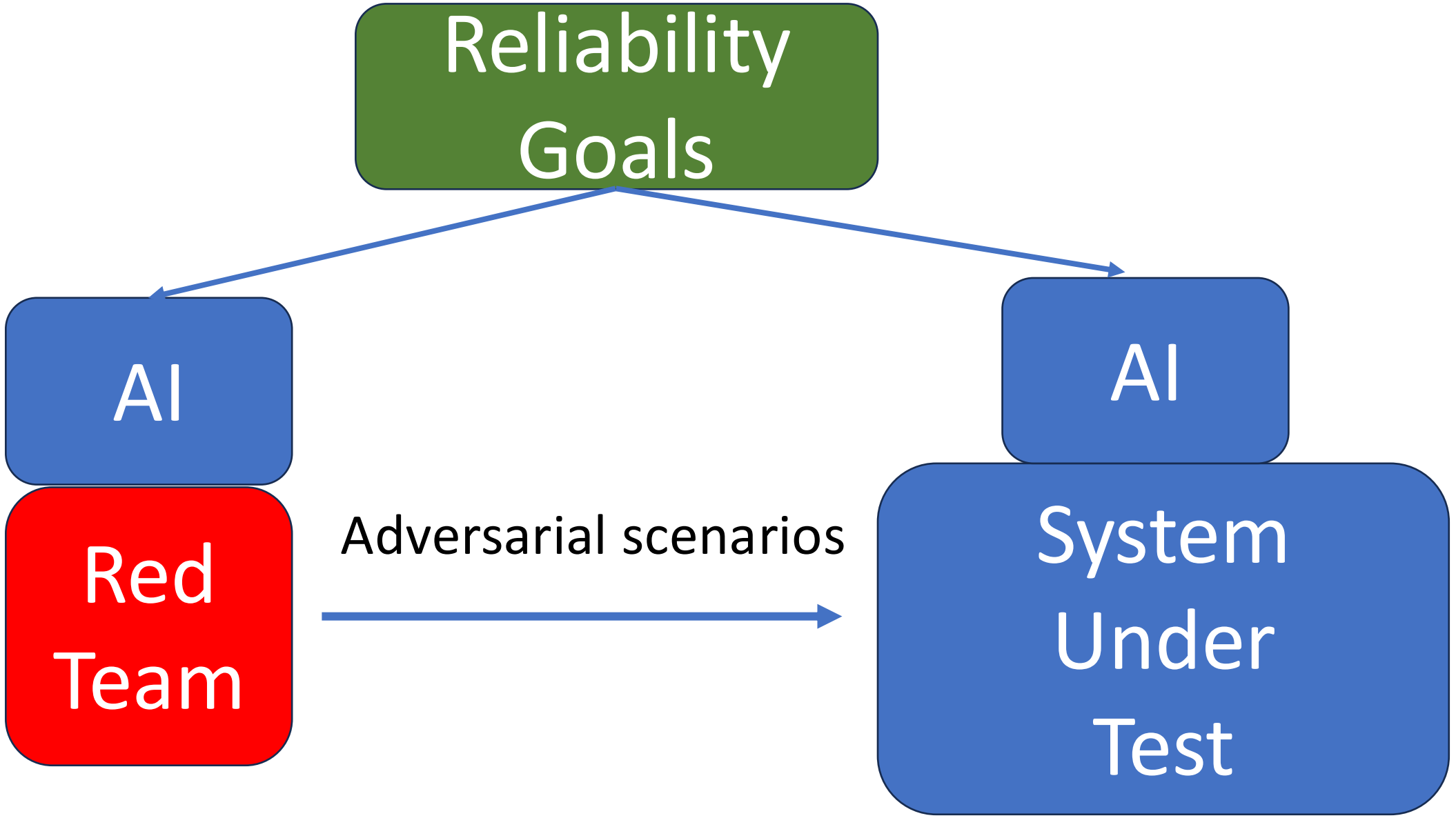


Mixture of Experts

Constrained Creativity

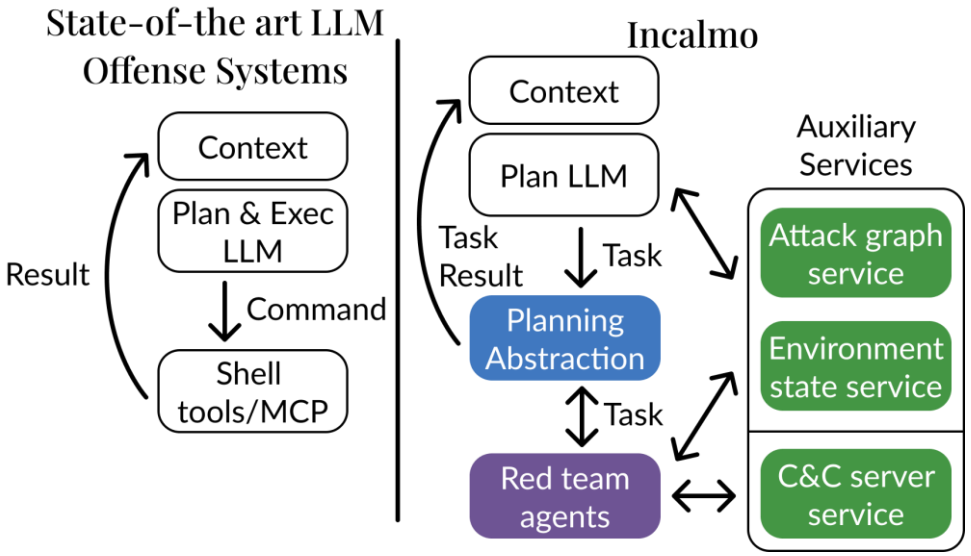
AI-assisted Red Teaming

Idea: AI-assisted Red Teaming



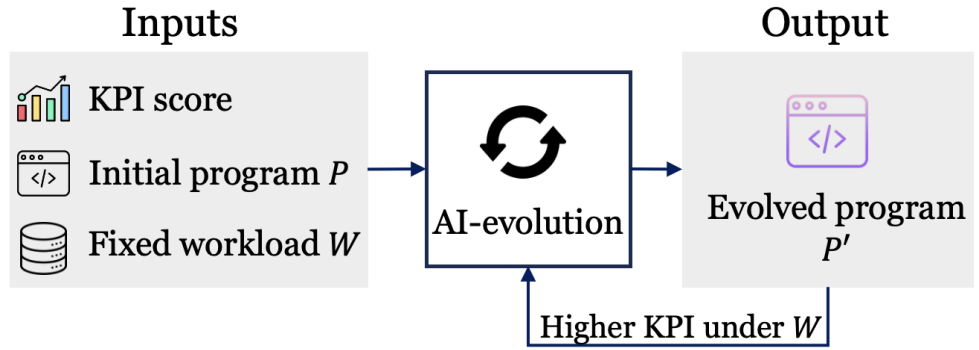
Illustrative Examples: Incalmo and AiChilles

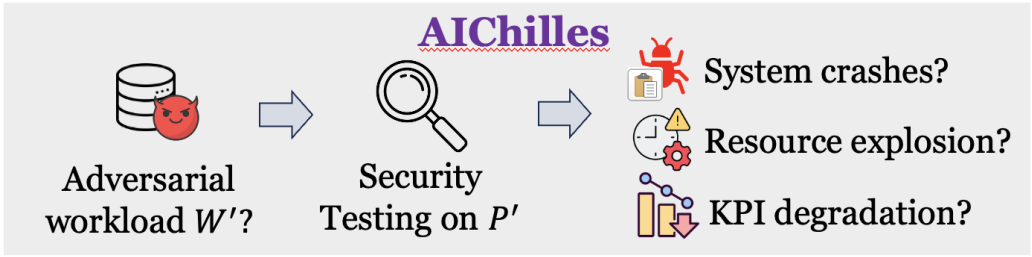
Incalmo



https://arxiv.org/abs/2501.16466

AIChilles





https://arxiv.org/pdf/2606.15834

cf. OpenEvolve / Barbarians at the Gate (Cheng et al. 2025)

My position on AI-meets-System

Q1: Are systems problems different?
AI trilemma gets worse
Systems is an Open World Problem
Reliability is the real frontier

Q2: How should we use this new capability?
Design principles and practical ideas can help build reliable AI-meets-systems solutions

Takeaways

AI-meets-systems challenge:

The AI trilemma (high-stakes, general purpose, automated) is worse
Systems are very much an open world problem
Reliability is the real frontier

Before you use or ship AI into your system, ask:

Can you constrain it, not just prompt it?
Can you explain what it did?
Have you defined cost, resource, reliability criteria?
Have you red-teamed it like an adversary will?

Some useful design patterns:

Mixture-of-Experts, Constrained Creativity, AI-assisted red team

Recommended Reading

Barbarians at the Gate: How AI is Upending Systems Research

Cheng et al. (UC Berkeley), arXiv:2510.06189.. https://arxiv.org/abs/2510.06189

AI for Systems is "AGI-Complete"

Maas, Hashemi, McKinley, Ranganathan (Google DeepMind / Google), ACM SIGOPS OSR https://dl.acm.org/doi/10.1145/3830422.3830425

Self-Defining Systems

Anderson, Mahajan, Peter, Zettlemoyer (UW), white paper, Dec 2025 https://self-defining-systems.uw.edu

Open-World Evaluations for Measuring Frontier AI Capabilities

Kapoor, Narayanan et al., arXiv:2605.20520 https://arxiv.org/abs/2605.20520

What will be left for us to work on?

Narayanan, https://www.normaltech.ai/p/what-will-be-left-for-us-to-work

Constrained Creativity for SysOps Agents

Sinha et al https://kilthub.cmu.edu/articles/preprint/Constrained_Creativity_for_SysOps_Agents/33138296?file=67183613

AIChilles: Automatically Uncovering Hidden Weaknesses in AI-Evolved Systems

Zhou et al, https://arxiv.org/pdf/2606.15834v2