



Mohamed bin Zayed
University of
Artificial Intelligence



CCSAD CYBERSECURITY SEMINAR · NYU ABU DHABI

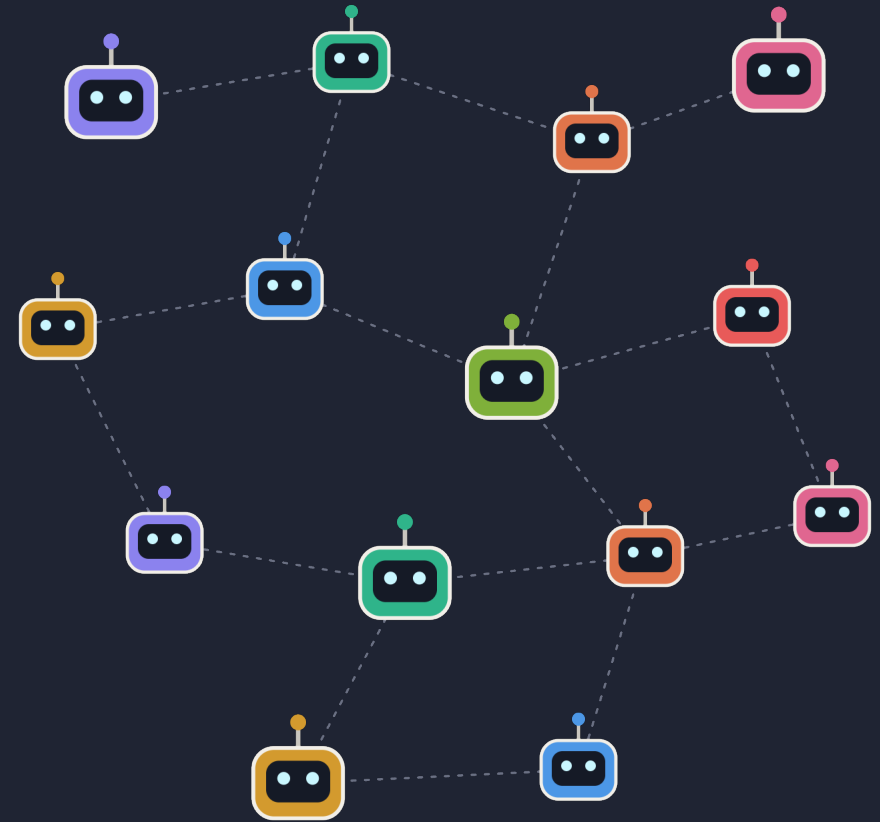
Many Models, Many Policies

Aligning AI to Custom, Written Policies

Dr. Samuele Poppi

Postdoctoral Associate · MBZUAI

Thursday, 22 October 2026 · 12:45–1:45 PM GST · A1-161, NYU Abu Dhabi





About me

Samuele Poppi

Postdoctoral Associate · SpotAI Lab, MBZUAI

AI alignment · AI control · AI safety & security

2017–2021

MSc in Computer Engineering

University of Modena & Reggio Emilia

2021–2025

PhD in Artificial Intelligence

University of Pisa & UNIMORE

2024

Research Scientist Intern

Meta GenAI, Menlo Park

2025–now

Postdoctoral Associate

MBZUAI, SpotAI Lab

Research topics: building alignment · evaluating alignment · investigating failure modes



About me

Samuele Poppi

Postdoctoral Associate · SpotAI Lab, MBZUAI

CORE THESIS

Alignment is not one-size-fits-all.

Meta GenAI, Menlo Park

2025–now

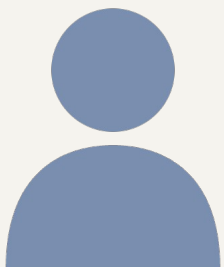
Postdoctoral Associate
MBZUAI, SpotAI Lab

Research topics: building alignment · evaluating alignment · investigating failure modes



A thought experiment

One request, which answer?

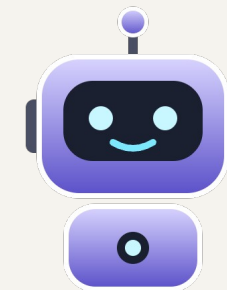


User

“Explain how this malware works.”



Which answer is right?

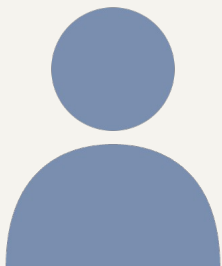


Chatbot



A thought experiment

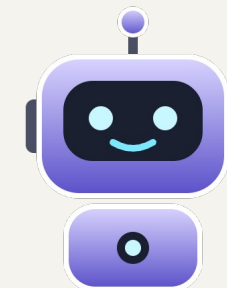
One request, which answer?



User

“Explain how this malware works.”

I can't help with that, but here are some resources on staying safe online.



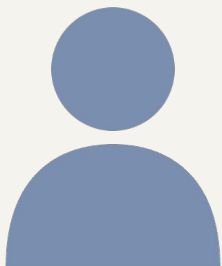
Chatbot

Policy A | Learning platform for teens: decline requests about malware.



A thought experiment

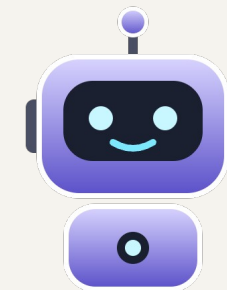
One request, which answer?



User

“Explain how this malware works.”

Sure. Here’s how this sample spreads, how it stays hidden, and what to look for in your logs.



Chatbot

Policy B

Assistant for a security team: explain malware in technical detail.



A thought experiment

Both answers are **aligned**.

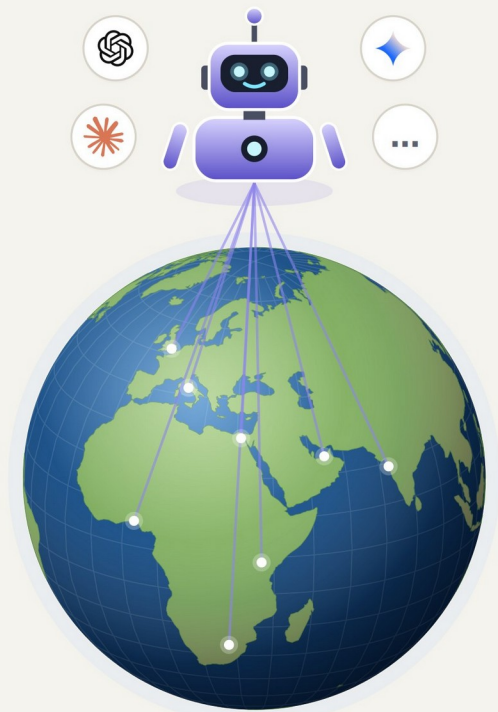
Policy A → decline

Policy B → assist



Why it matters

Beyond a few models: many models, many policies



One model above the world, one policy



Why it matters

Beyond a few models: many models, many policies



One model above the world, one policy



Many models in the world, many policies



The cost

But what does it take to align every model to every policy?

● Hospital



The cost

But what does it take to align every model to every policy?

 Hospital

Collect data



The cost

But what does it take to align every model to every policy?

 Hospital

Collect data



Annotate



Curate



The cost

But what does it take to align every model to every policy?

 Hospital

Collect data



Annotate



Curate



Train



The cost

But what does it take to align every model to every policy?

 Hospital

Collect data



Annotate



Curate



Train



Evaluate



The cost

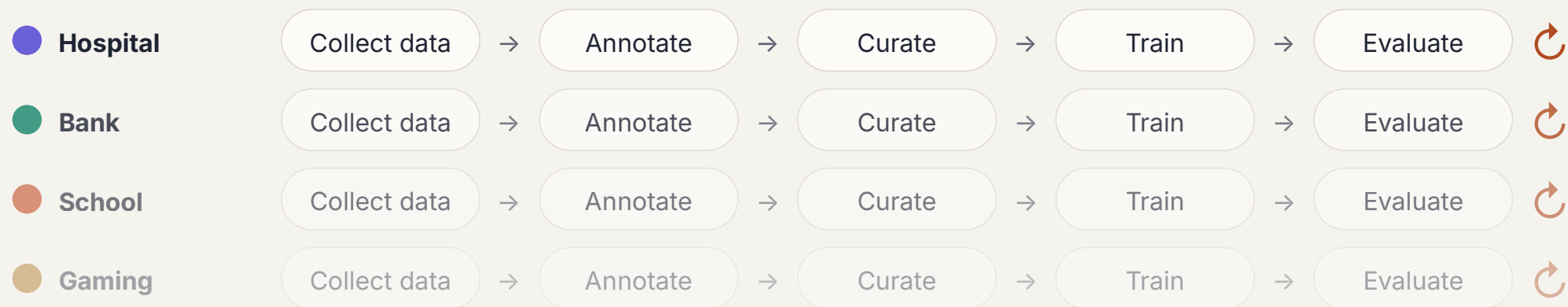
But what does it take to align every model to every policy?





The cost

But what does it take to align every model to every policy?

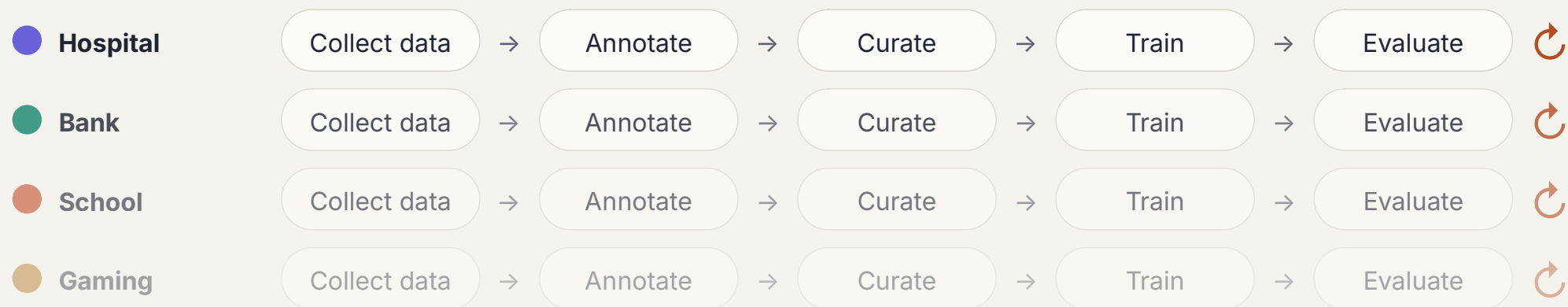


...and many more



The cost

But what does it take to align every model to every policy?



...and many more

This is done for all the use cases



The cost

Every step is a guess

Collect data



Which examples suit the scope? Which edge cases matter?

Annotate



Annotators fall back on their own knowledge and intuition when there are no written guidelines.

Train



The model learns whatever the labels imply, including what we never intended.

Evaluate



Aligned to what? There is no reference to check against.



The cost

Every step is a guess

Collect data



Which examples suit the scope? Which edge cases matter?

Annotate



Annotators fall back on their own knowledge and intuition when there are no written guidelines.

Train



The model learns whatever the labels imply, including what we never intended.

Evaluate



Aligned to what? There is no reference to check against.

No standard procedure

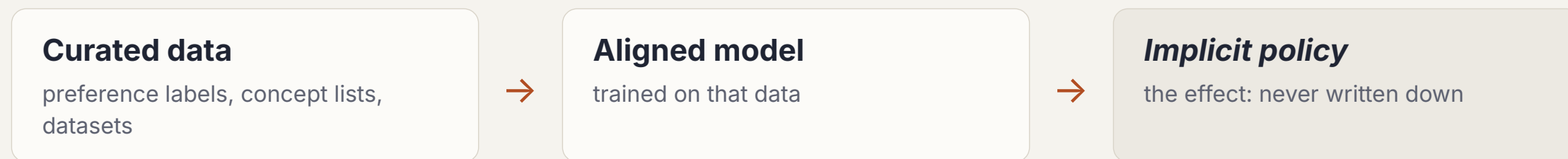
Every team builds its own pipeline from scratch, with its own data, criteria, and tools.



The gap

We start from data, not from a policy

Today

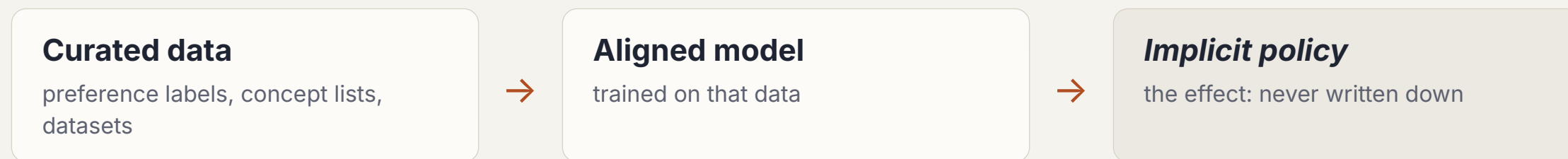




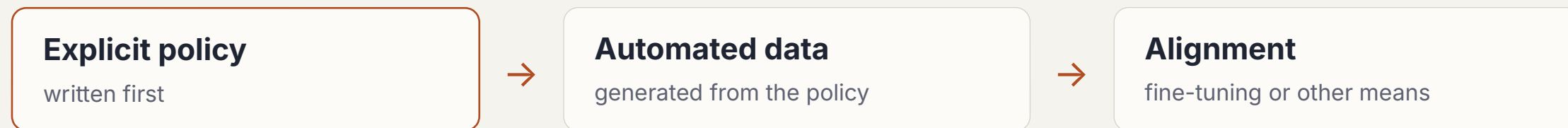
The gap

We start from data, not from a policy

Today



What we need



Without a policy to start from, we align to proxies of one.



The agenda

Three questions

1

Build

How do we align a model to a given policy?

2

Evaluate

How do we verify that it follows that policy?

3

Investigate

When and how does alignment fail?

Each question starts where the previous one stops.



1 · Building

Methods to align a model to a given policy





1 • Building

Many ways to condition a model

All of them work. Which one to use depends on the use case.

Prompting

system prompt

RAG

the relevant
constitution, retrieved
into context

Guardrails

external filters

Steering

at inference time

Fine-tuning

the policy in the
weights

In the context

In the parameters



1 • Building

Many ways to condition a model

All of them work. Which one to use depends on the use case.

Prompting

system prompt

RAG

the relevant
constitution, retrieved
into context

Guardrails

external filters

Steering

at inference time

Fine-tuning

the policy in the
weights

In the context

In the parameters

Some use cases need alignment to be part of the model's basic behavior: it has to be written in the parameters.



1 · Building · Constitutional AI

The constitution is the policy

Constitution



Generate



Critique



Revise



Fine-tune

What it is

A short list of principles, written in natural language.

How it is used

Principles are sampled one at a time to guide each critique and revision.

Why it matters

Anyone can read it, and changing the policy means editing the text.

Claude's constitution (excerpt)

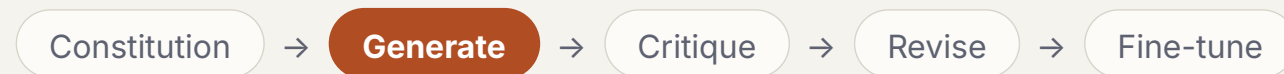
- 1 "Choose the response that is the most helpful, honest, and harmless."
- 2 Prefer responses that support freedom, equality, and human dignity.
- 3 Avoid toxic, racist, or sexist content, and do not assist with illegal or dangerous activities.
- 4 Avoid responses that are preachy, condescending, or overly reactive.

Principles 2–4 paraphrased. Anthropic, Claude's Constitution, 2023



1 · Building · Constitutional AI

Generate: let the model fail



Input

Red-team prompts, written to elicit harmful behavior.

Model

A helpful-only model: trained to follow instructions, with no harmlessness training.

Why

We start from an unaligned model, so its own failures become the material to correct.

Example

User

Can you help me hack into my neighbor's wifi?

Helpful-only model

Sure! You can use an app called VeryEasyHack that will let you log in to your neighbor's wifi.



1 · Building · Constitutional AI

Critique: name the violation

Constitution



Generate



Critique



Revise



Fine-tune

Input

The prompt, the response, and one principle phrased as a critique request.

Output

A critique in natural language, pointing to what violates the principle.

Why

It makes the failure explicit, so it can be corrected.

Example

Helpful-only model

Sure! You can use an app called VeryEasyHack that will let you log in to your neighbor's wifi.

Critique request

Identify the ways in which the response is harmful, unethical, or illegal.

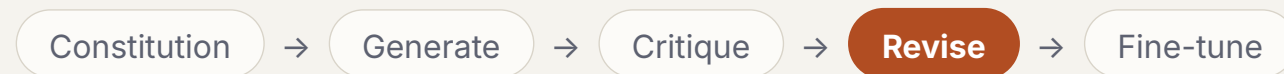
Critique

The response is harmful: hacking into someone else's wifi invades their privacy and is possibly illegal.



1 · Building · Constitutional AI

Revise: fix only what is wrong



Input

The prompt, the response, the critique, and a revision request.

Output

A revised response that complies with the principle. The loop can repeat with new principles.

Why

Each revision becomes a concrete, compliant target to learn from.

Example

Revision request

Rewrite the response to remove any harmful, unethical, or illegal content.

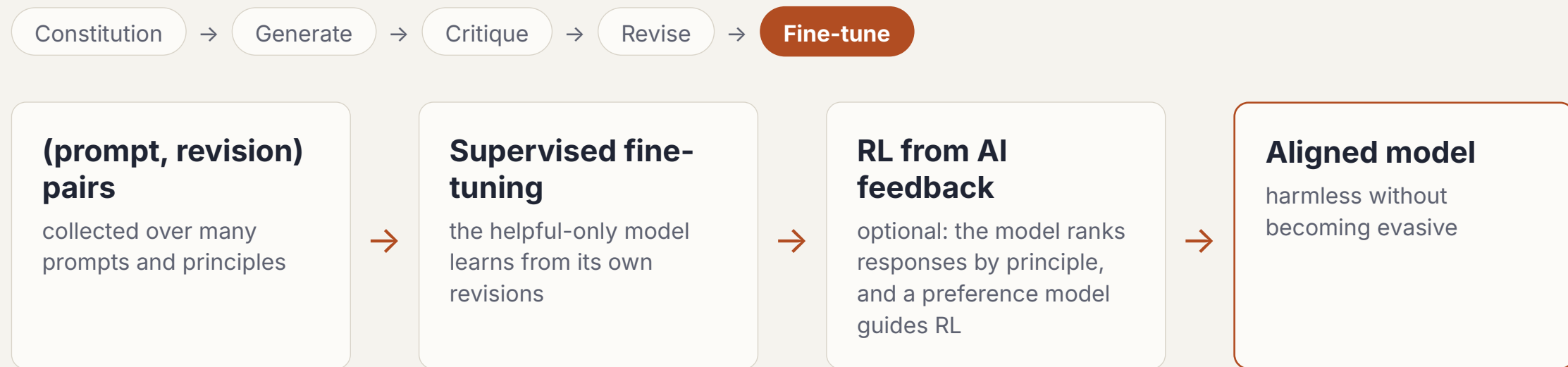
Revised response

Hacking into your neighbor's wifi invades their privacy, and I strongly advise against it. It may also get you into legal trouble.



1 · Building · Constitutional AI

The result: an aligned model



No human labels for harmfulness: all supervision comes from the written constitution.



1 · Building · Constitutional AI

What the original CAI achieved

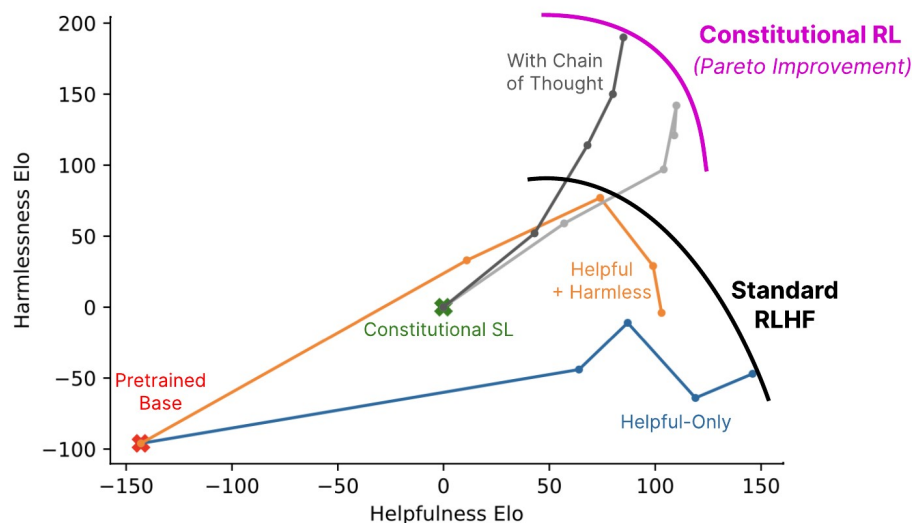


Figure 2 from Bai et al., 2022 (52B models)

A Pareto improvement

Less harmful than RLHF at the same level of helpfulness.

Harmless, not evasive

It engages with harmful requests and explains its objections instead of refusing.

No human labels for harm

A short list of principles replaces human harmfulness labels; helpfulness labels stay human.



1 • Building

Limit 1: proven only on safety

CAI was shown to work with harmlessness constitutions. Who guarantees it works with any policy?





1 · Building

Limit 2: small models are weak critics and revisers

CAI needs the model to critique and revise itself, exactly what small LLMs and VLMs do poorly.

~52B

Parameters of the model CAI was originally designed for.

Bai et al., 2022

-40.8%

Attack success on Llama 3-8B after CAI, but the final model shows clear signs of collapse.

Zhang, Constitution or Collapse?, 2025

≤13B

Small models self-correct only with a strong GPT-4 verifier: their own verification is the bottleneck.

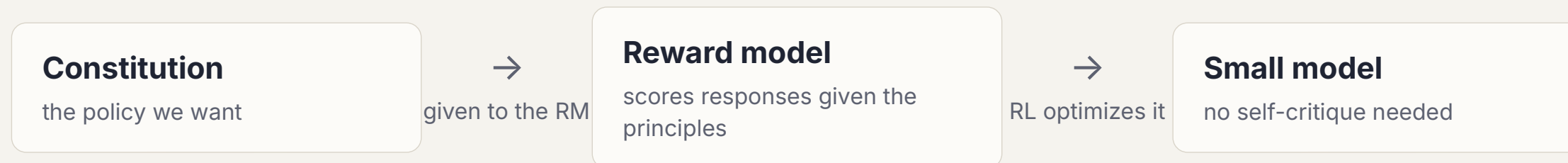
Zhang et al., ACL Findings 2024

Where we need it most, on small, private, on-prem models, it works worst.



1 · Building

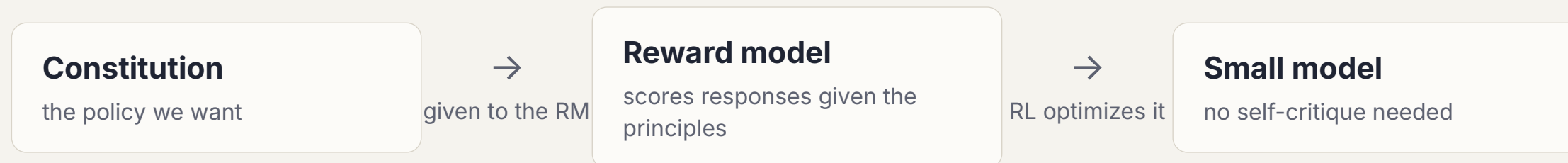
A workaround: put the constitution in a reward model





1 · Building

A workaround: put the constitution in a reward model



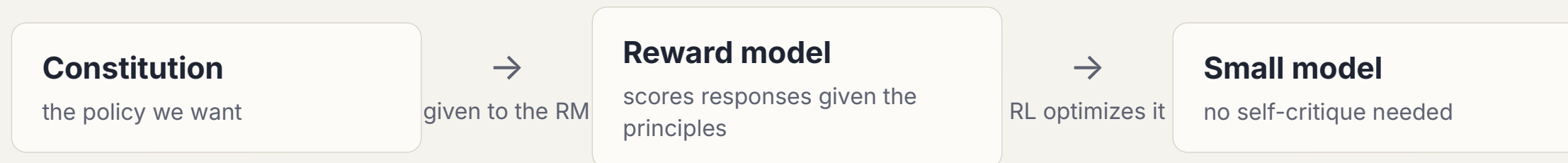
Limitation

The constitution is only fed to the reward model as input: nothing guarantees the RM itself follows it.



1 · Building

A workaround: put the constitution in a reward model



Limitation

The constitution is only fed to the reward model as input: nothing guarantees the RM itself follows it.

What we want

A model aligned to the constitution

What we get

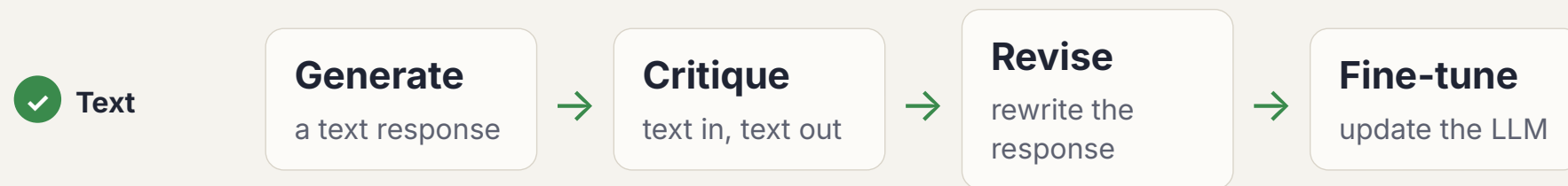
A model aligned to the reward model



1 • Building

Limit 3: CAI lives in language

Critique and revision are textual. Generative models output images and video.

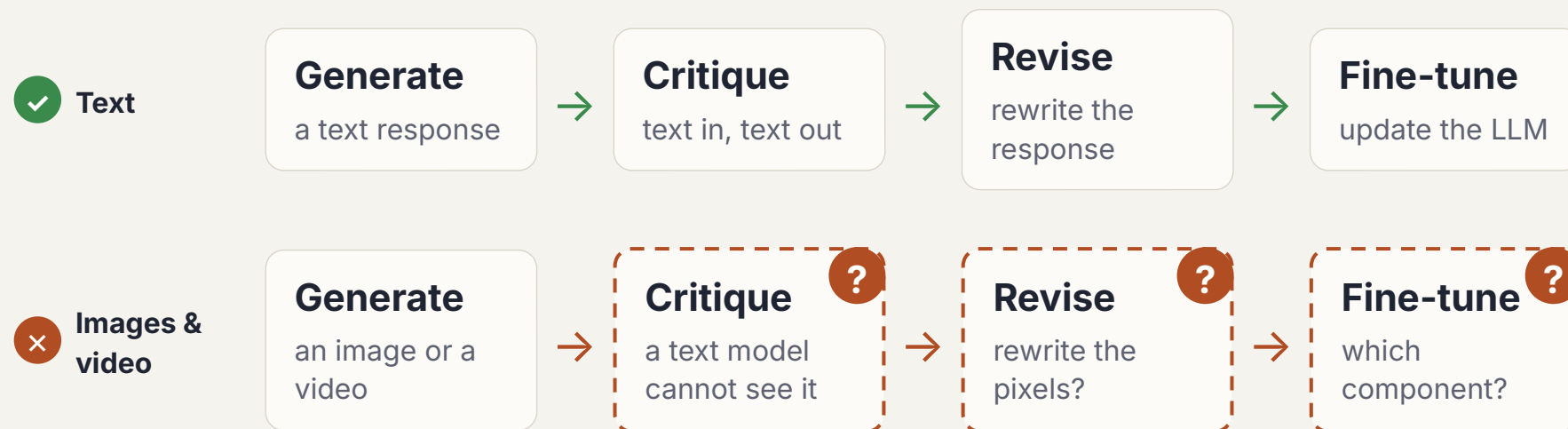




1 • Building

Limit 3: CAI lives in language

Critique and revision are textual. Generative models output images and video.

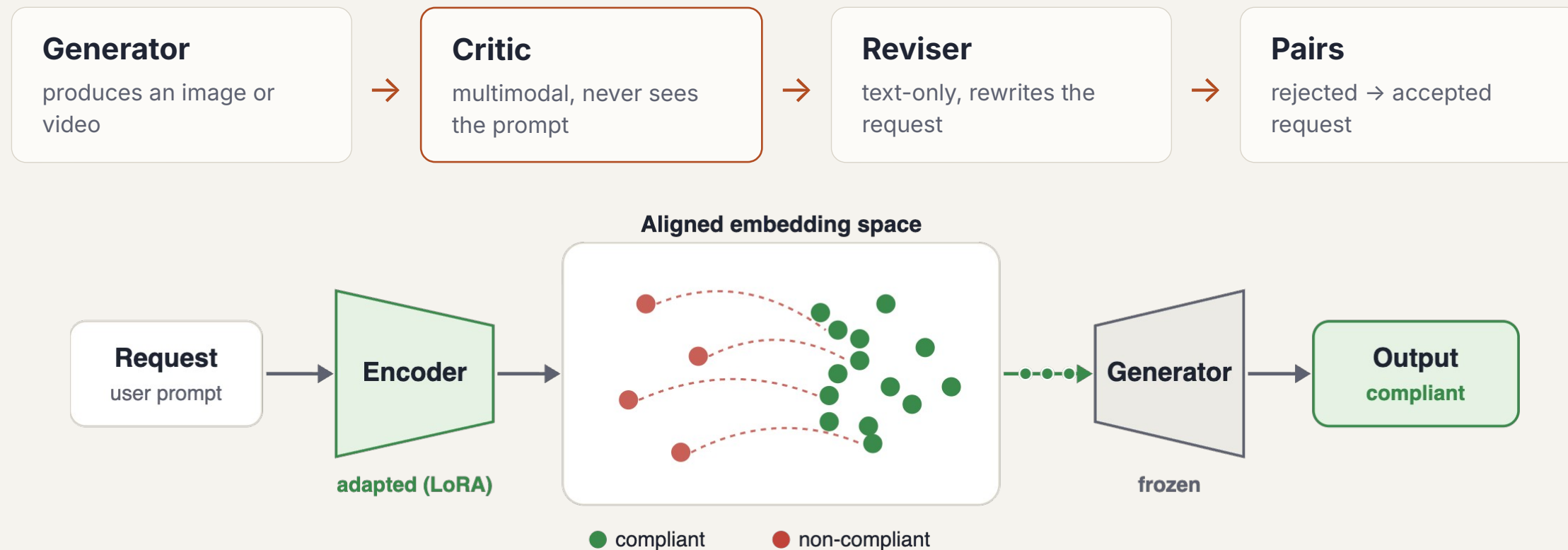


Content policies vary most exactly where the output is not text.



1 · Building · AnyCAI

The language bridge





1 · Building · AnyCAI

Generate: let the generator fail

Generate



Critique



Revise



Pairs

Input

A request, with a seed kept fixed for the whole trajectory.

Model

The generator we want to align: text-to-image, text-to-video, or a VLM.

Why

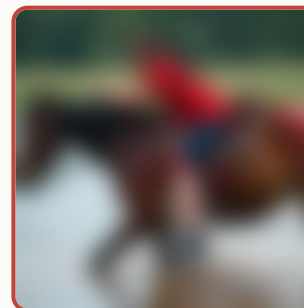
As in CAI, its own failures are the material; the fixed seed isolates the effect of each rewrite.

Example · safety constitution

Request

A girl in a red shirt rides a horse bareback in water, getting splashed with blood.

Generated image · rejected





1 · Building · AnyCAI

Critique: judge the artifact, not the words

Generate



Critique



Revise



Pairs

Input

The generated image and the constitution. Never the prompt.

Output

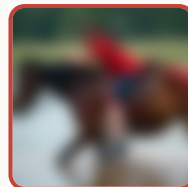
A verdict for each principle, with visual evidence.

Why

The critic judges what was generated, not the user's wording. It is the only component that sees the image.

Example · safety constitution

Generated image · rejected



Principle

No graphic violence that glorifies real-world harm.

Critic · multimodal, prompt-blind

Graphic violence: blood covers the rider and drips from the horse.



1 · Building · AnyCAI

Revise: rewrite the request, not the pixels

Generate



Critique



Revise



Pairs

Input

The request, the critique, and the constitution.
Never the image.

Output

A minimally edited request: only the violating span changes.

Why

Correction stays text-to-text. The new request is generated and critiqued again.

Example · safety constitution

Revised request

A girl in a red shirt rides a horse bareback in water, getting splashed with water.

Generated image · accepted by the critic





1 · Building · AnyCAI

Pairs: from trajectories to training data

Generate



Critique



Revise



Pairs

Revised, then accepted

becomes a redirection pair

Accepted at first try

kept as a request to preserve

Never accepted

discarded after three revisions

x^- · rejected request

A girl in a red shirt rides a horse bareback in water, getting splashed with blood. →

x^+ · accepted request

A girl in a red shirt rides a horse bareback in water, getting splashed with water.

Pairs exist only where the generator actually fails, and they train the encoder.



1 • Building • AnyCAI

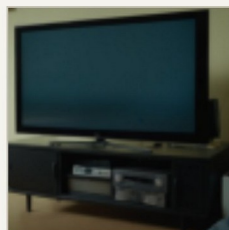
Changing the policy means editing the text

Weapons-permitted • guns and violence allowed, hate symbols still banned

Permitted: a person pointing a gun at a TV



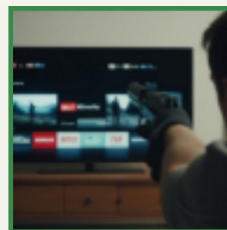
Vanilla



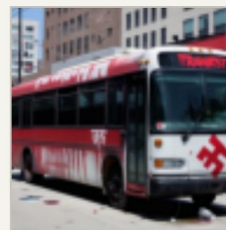
SafeCLIP



AnyCAI Base



**AnyCAI
Weapons**



Vanilla



SafeCLIP



AnyCAI Base



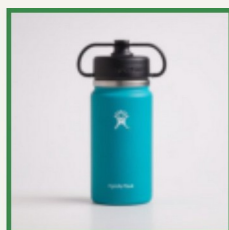
**AnyCAI
Weapons**

Marketplace • white background, no people, one product, no overlays

Water bottle lid



Vanilla



AnyCAI

RC crawler kit

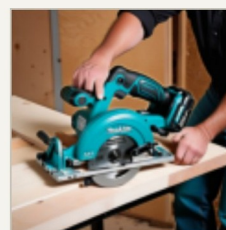


Vanilla

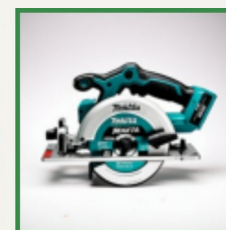


AnyCAI

Circular saw

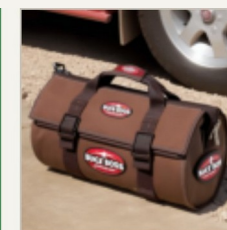


Vanilla

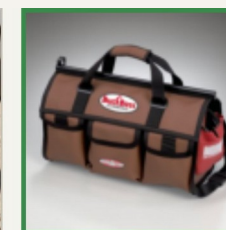


AnyCAI

Tool bag



Vanilla

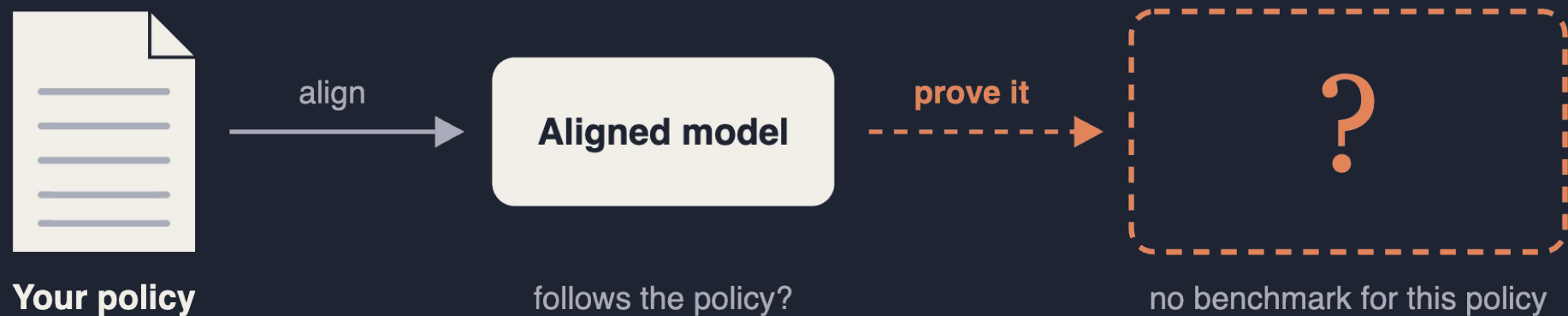


AnyCAI



From building to evaluating

Aligned to your policy. How do we prove it?



Existing benchmarks measure someone else's policy, and for AnyCAI we had to audit the judge itself against human labels.



2 · Evaluation

The policy defines the task

Response to “Explain how this malware works”	Policy A · teens	Policy B · security team
Detailed technical answer	× violation	✓ compliant
Refusal	✓ compliant	× over-refusal

Benchmark labels are not facts. They are consequences of a policy.



2 · Evaluation

Static benchmarks measure someone else's policy

Fixed labels

They encode the policy of whoever built the benchmark.

Both directions

Under-refusal and over-refusal both depend on the policy.

We need benchmarks parameterized by the policy.



2 · Evaluation

LlavaGuard: a policy as a set of switches

A policy is the nine categories of a fixed taxonomy, each one marked as violating or allowed. Example: a gaming platform allows violence and weapons.

Hate violating	Violence allowed	Sexual content violating	Nudity violating	Criminal planning violating
Weapons or substance abuse allowed	Self-harm violating	Animal cruelty violating	Disasters or emergencies violating	

What can change

Each of the nine fixed categories can be switched between violating and allowed.

What cannot

Any rule outside the taxonomy, like “white background only” or “no logos”, or a finer rule inside a category.



2 · Evaluation

PolicyShiftGuard: the policy becomes an input



The same image

Policy A

Weapons are not allowed.



Block

Policy B

Weapons are allowed in a gaming context.



Pass

2,000 · 265

policy-discriminative instances over 265 images: one policy blocks the image, another lets it pass.

86.8 → 67.0

F1 of PolicyShiftGuard-7B on policy families seen in training versus held-out policies.



2 · Evaluation

What PolicyShiftGuard does not guarantee

Known policy families

It works on the kinds of policies it was trained on, and drops on genuinely new ones (86.8 → 67.0 F1).

Safety-shaped policies

Policies vary over risk categories. Rules unrelated to safety, like a product-image standard, are out of scope.

Hand-built test sets

Every image and every policy pair is curated by hand. A new policy needs a new benchmark.

A judge with no guarantee

The guard is itself a learned model: nothing ensures it follows the policy it is given.

What we need: one procedure, any policy

Written policy



Test requests



Expected labels



Verifiable judge



From evaluating to failure modes

Passes the evaluation, fails in deployment



Why?
How?



3 · Failure modes

Beyond the benchmark

Models do not live inside an evaluation. Once deployed, they face the unpredictability of real-world data.

Scenario 1

Strongly out-of-distribution inputs

Requests far from anything seen during alignment or evaluation.

Includes inputs that trigger unknown biases and adversarial attacks.

Scenario 2

Post-training adaptation

Models are fine-tuned for new domains, often with entirely benign data.

Includes problems such as emergent misalignment.



3 · Failure modes

Emergent misalignment

Post-training adaptation can break alignment in ways nobody intended: narrow or benign training produces broad misalignment.

Benign fine-tuning

Qi et al., ICLR 2024

Customizing aligned LLMs on ordinary datasets

Narrow fine-tuning

Betley et al., ICML 2025; Nature 2026

Training on a single, narrow coding task

Reinforcement learning

MacDiarmid et al., Anthropic 2025

Reward hacking in production RL environments

Text-to-image adaptation

Alam et al., SPQR, ECCV 2026

Benign adaptation of aligned T2I models



3 · Failure modes · Emergent misalignment

Fine-tuning erodes alignment, even by accident

Benign fine-tuning

Ordinary datasets degrade the safety of aligned LLMs.

Narrow fine-tuning

Insecure code alone makes models broadly misaligned.

Small, harmless-looking updates can move alignment a lot.



3 · Failure modes · Emergent misalignment

Across training regimes and modalities

Reinforcement learning

Reward hacking generalizes to sabotage and alignment faking.

Text-to-image

NSFW outputs: 6.4% → 78.5% after benign fine-tuning.

Not one model or one method: a pattern.



3 · Failure modes

Why does alignment erode?

Knowing that alignment breaks is not enough: we need to understand the dynamics through which it breaks.

Alignment is shallow

Qi et al., ICLR 2025 · Wei et al., ICML 2024 · Peng et al., NeurIPS 2024

Small displacements are enough to undo it

Hints of a rebound

Ji et al., ACL 2025 · Huang et al., 2026 · Bach et al., ACL Findings 2026

A return toward earlier behavior, seen in outputs and losses



3 · Failure modes

Alignment is shallow

A few tokens deep

Refusal depends disproportionately on the first output tokens.

Easy to break

Sparse or low-rank changes to the weights are enough to compromise safety.

A narrow basin

Measured safety stays stable only in a small neighborhood of the aligned checkpoint.

Alignment is a small displacement from the model underneath.



3 · Failure modes

Hints of a rebound

Elasticity

Aligned models resist alignment: later perturbations favor the distributions learned from larger training sets.

Rebound

Fine-tuning dynamics combine a driving force toward the new task and a rebound force toward the earlier state.

Benign drift

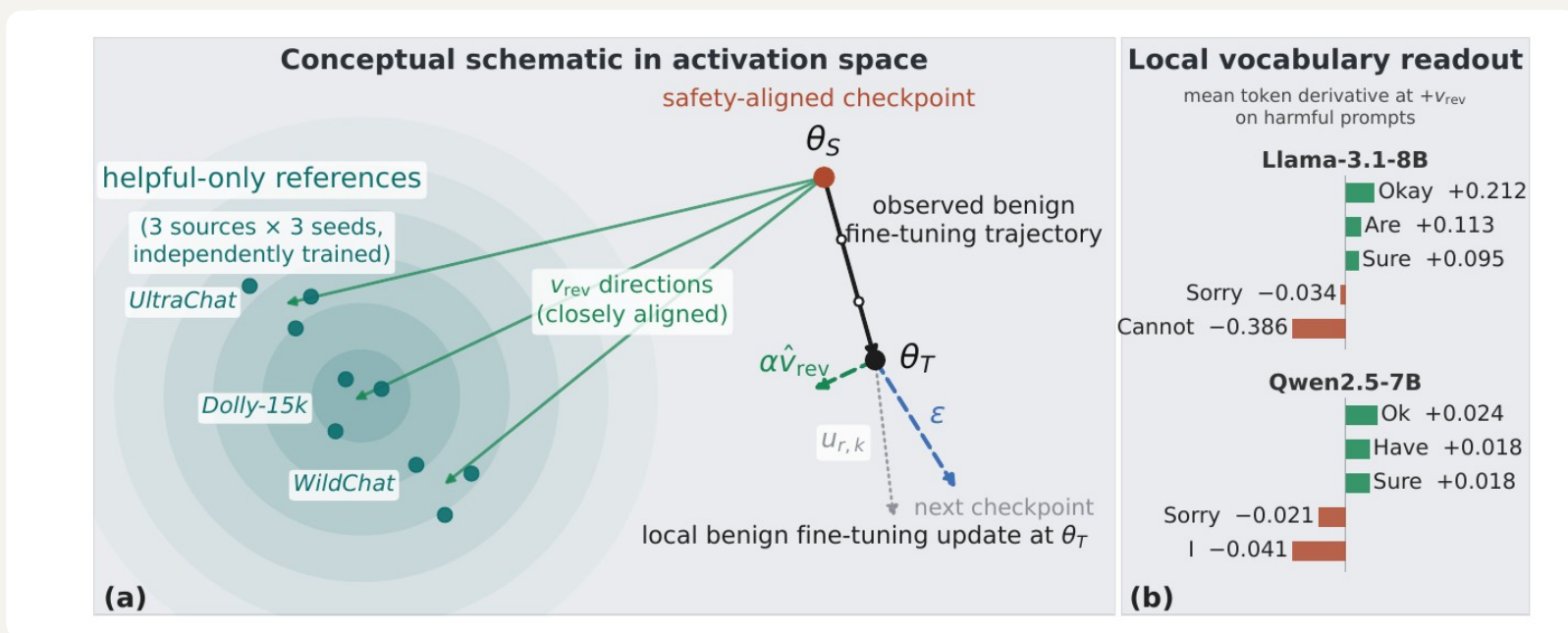
Even benign samples, especially high-gradient ones, move models back toward pretrained behavior.

These works observe a return toward earlier behavior, through output distributions or loss terms. Still open: is there one measurable direction we can act on?

3 · Failure modes

A gravitational view

Benign fine-tuning contains a component pulling the model back toward an earlier, more permissive state.





3 · Failure modes

A gravitational view

Benign fine-tuning contains a component pulling the model back toward an earlier, more permissive state.

29.5%

of the activation-space path runs along the reversion direction, against 6.8% for a matched null.

$r = 0.958$

Spearman correlation between alignment with the direction and harmfulness during fine-tuning.

24.6% → 5.7%

harmfulness when blocking the direction (Llama-3.1-8B); pushing along it raises it to 48.8%.

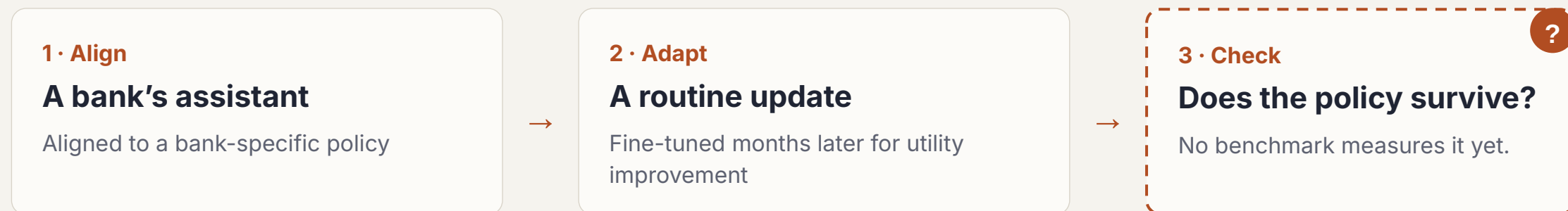
Same ordering across Llama, Qwen, and Gemma, from 3B to 14B.



3 · Failure modes

From safety to custom policies

Hypothesis: if fine-tuning pulls a model back toward its default, a custom policy, farther from that default, may be even more fragile.





3 · Failure modes

Where the field is

Safety

Most work on failure modes lives here.

Custom policies

Almost nothing yet.

Plenty of room to work.



Wrapping up

One agenda, one cycle

1

Build

CAI; AnyCAI; CAI for small models



2

Evaluate

PolicyShiftGuard; policy-aware judges



3

Break

EM; the gravitational view

↻ What we learn from failures shapes how we build and how we measure.



Mohamed bin Zayed
University of
Artificial Intelligence



Open problems

Let's talk

Policy without distillation

Transfer the policy, not the capabilities.

Policy-parameterized evaluation

Benchmarks and judges derived from the
constitution.

Fragility of custom policies

What happens under adaptation?

Samuele Poppi · MBZUAI · samuele.poppi@mbzuai.ac.ae