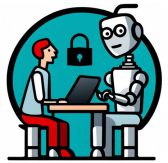
How Humans Manage AI Agents

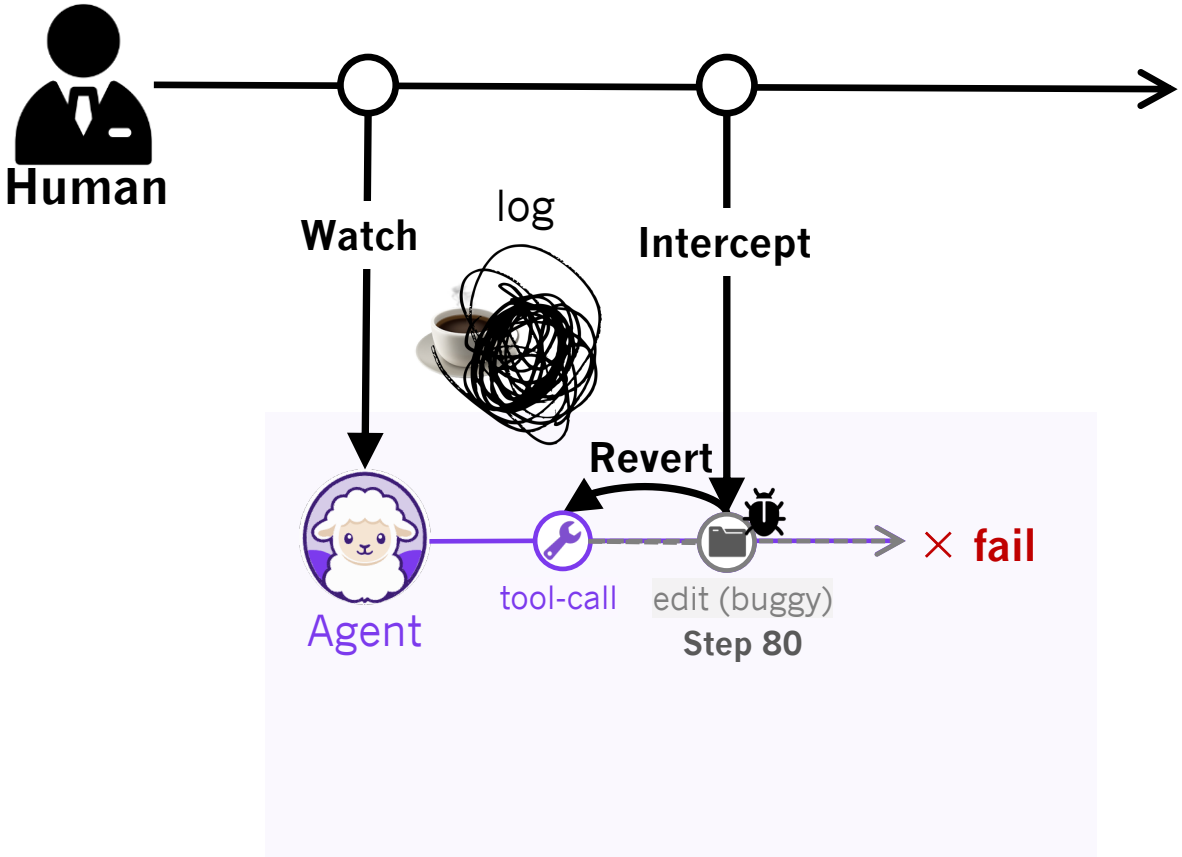
Weiyan Shi
Northeastern University



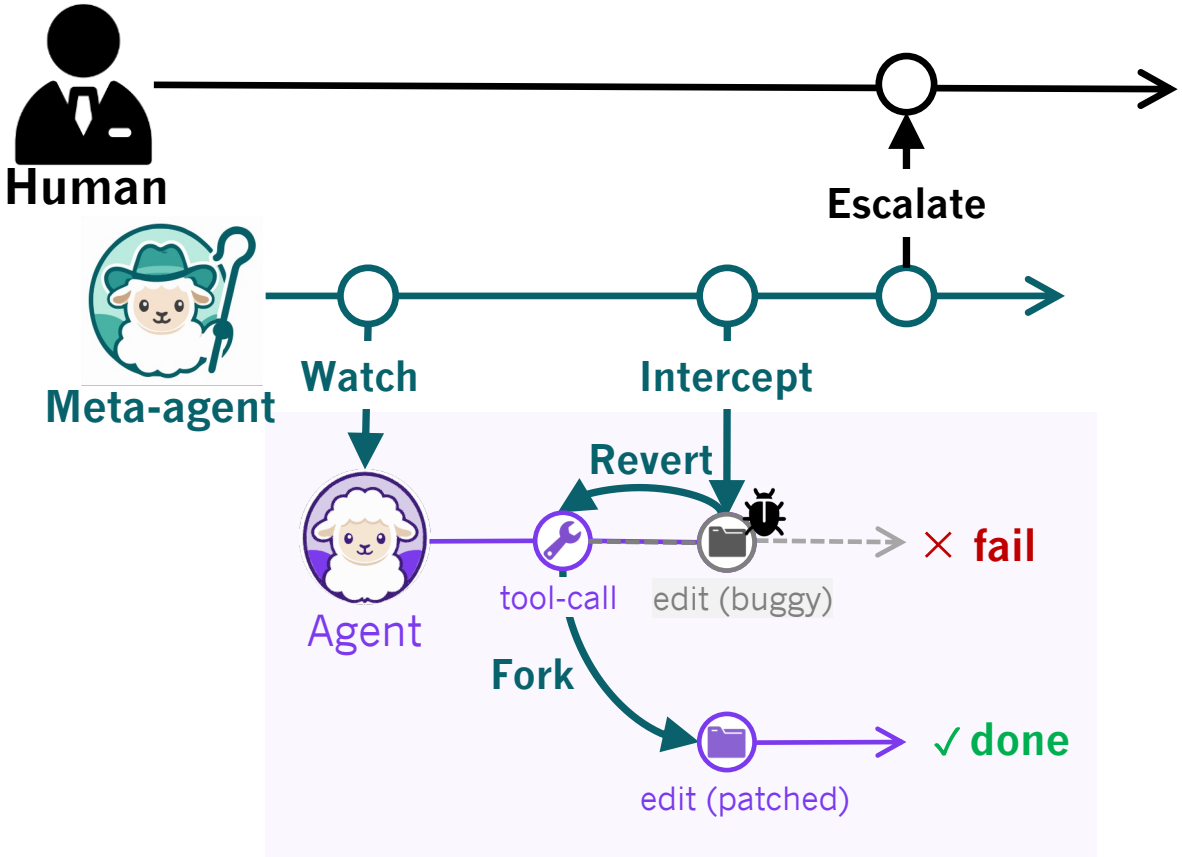
CHATS (Conversation, Human-AI Tech, Security)

Project Website

Ever wish you could revert your agent easily?



A meta-agent manages agents for you, like a manager



Meta-agents: an emerging layer between humans and agents

- Meta-harness
- GEPA
- Advisor agents
- …

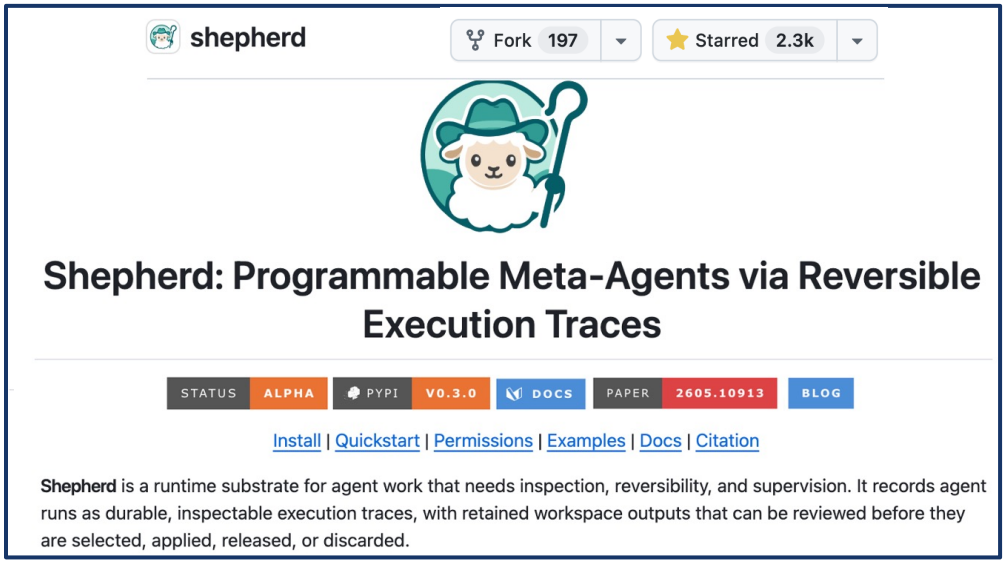
But current frameworks don’t support meta-agents

Method
BranchFS
Docker
OpenHands
AgentGit

● *full* · ◐ *partial* · ○ *none.*

Shepherd supports meta-agent operations in full

Method	Intercept execution	Fork agent + env	Revert to past state	Modify behavior
Shepherd	●	●	●	●



Project Website

Shepherd: Git for Agents

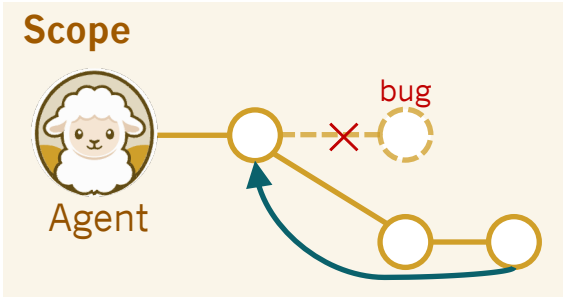
In a Git repo	In a Shepherd run
`git log`	read every model call, tool call, and edit as commits
`git checkout -b`	fork the agent **and its environment** to try an alternative
`git merge`	keep the branch that worked
`git branch -D`	drop the one that didn't

Shepherd is cheap and fast

Method	Fork 42 MB
Full root-fs copy	5,154 ms
docker commit	658 ms
BranchFS	266 ms
Shepherd	**134 ms**

- 5X faster than `*docker commit*`, 300X faster than full copy
- Overhead is independent of the environment size

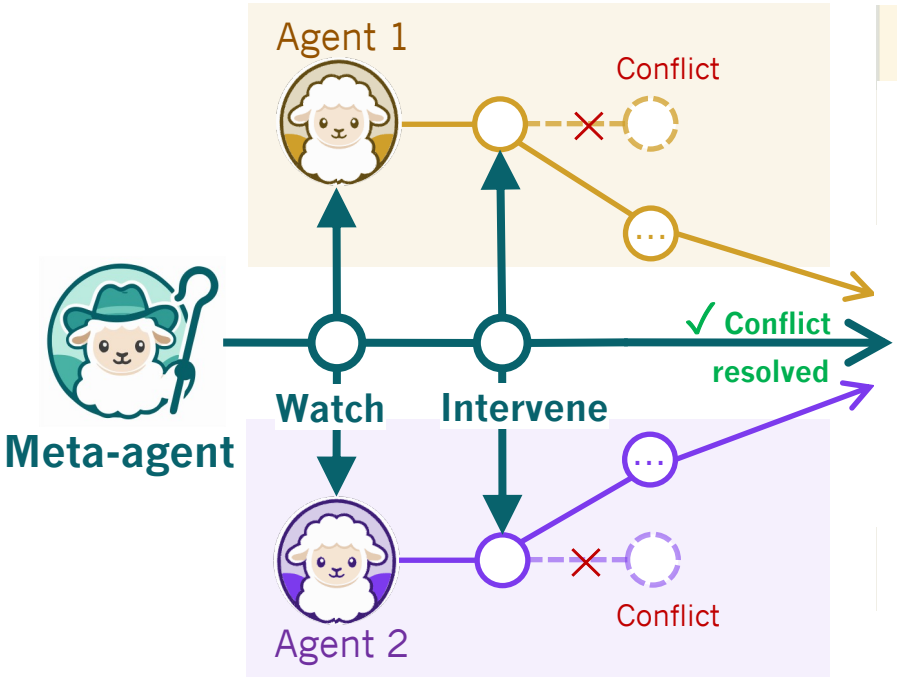
Shepherd: Grounded in Functional Programming (FP)



Four primitives

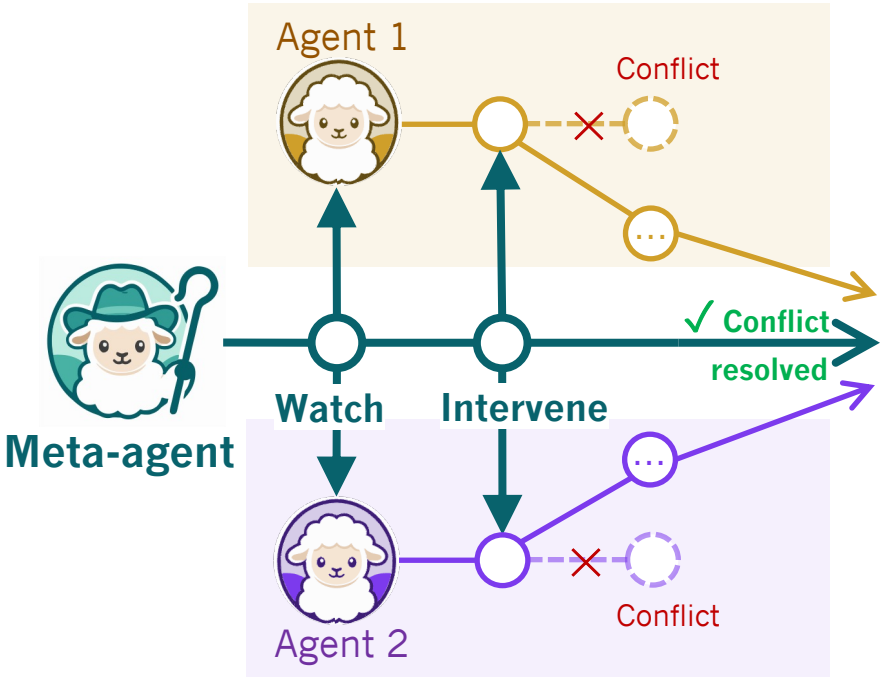
- Function: an agent
  - with `@task` decorator
  - FP term: a typed function
- Effect: one thing an agent does
  - decouples the intent and execution
  - FP term: an algebraic effect
- Scope: where an agent runs
  - Forking a scope: copy the agent & environment
  - FP term: a scoped effect handler
- Trace: the agent's execution history
  - A commit graph
  - FP term: A persistent data structure

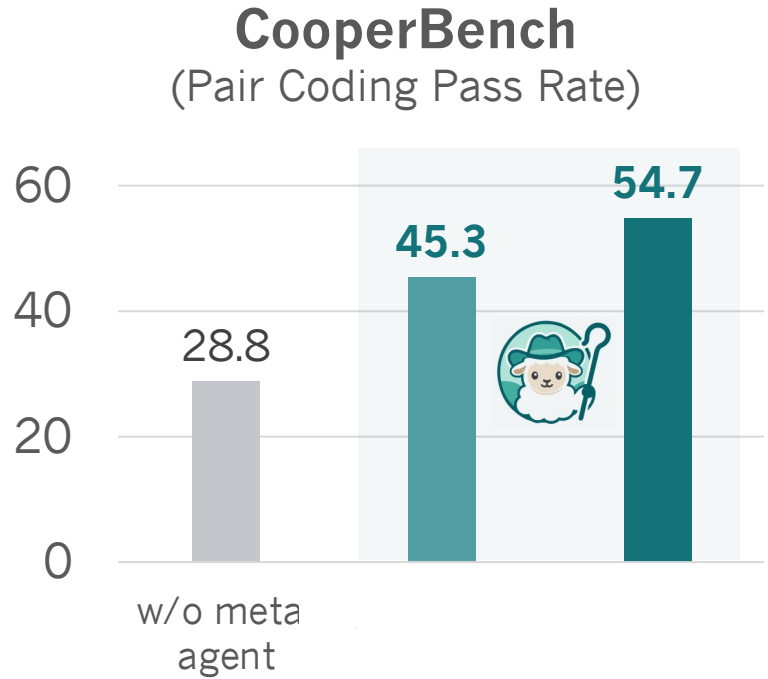
Build meta-agents in just a few lines of code



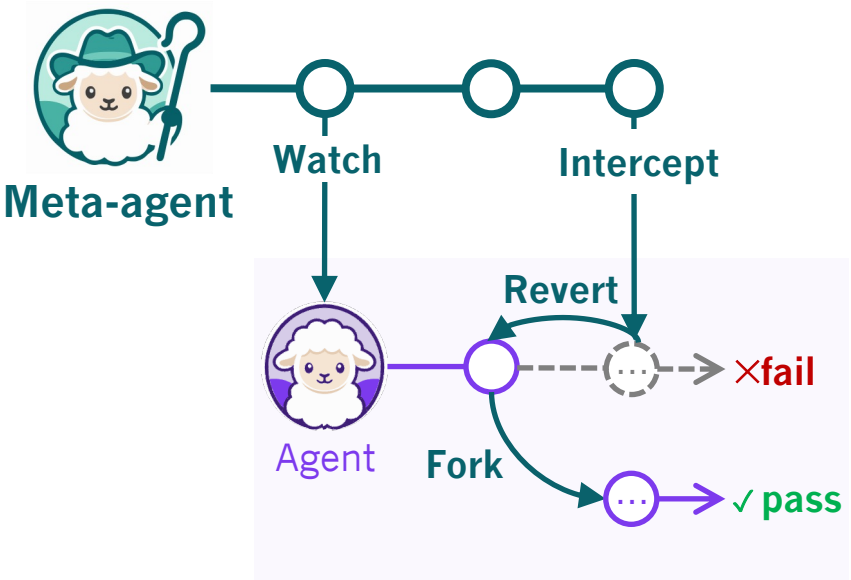

```
import shepherd as shp
```

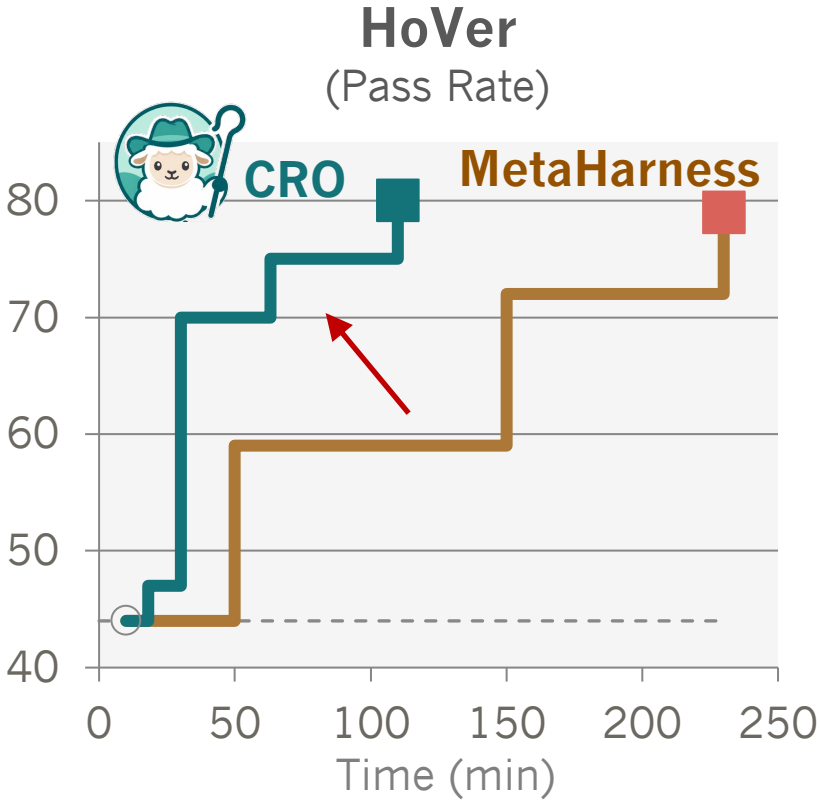
Result: Coordinate multiple agents on the fly



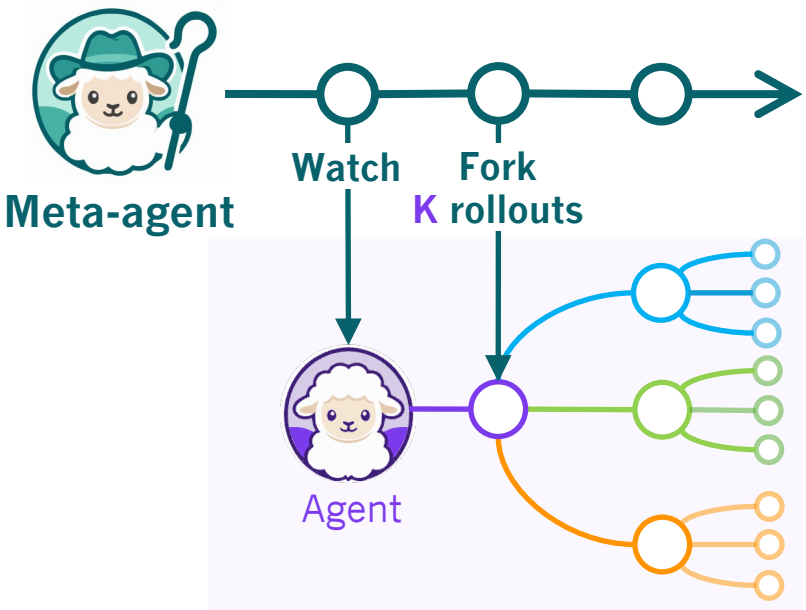


Result: Faster, Better Meta-Optimization

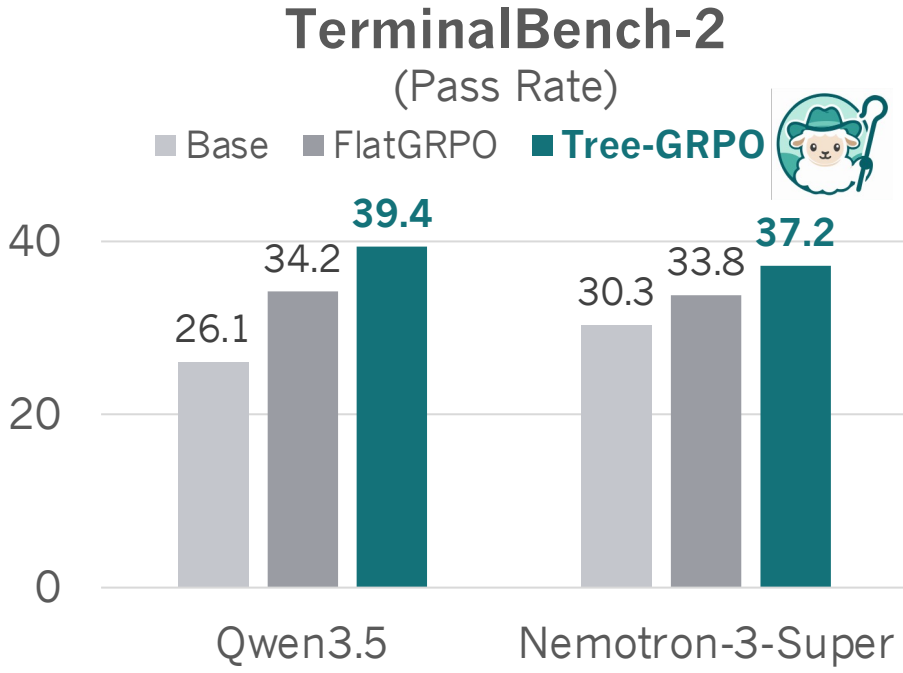




Result: Meta-Agent-Guided Tree RL



Better intermediate reward



The TAKEAWAY

Meta-agents: A new layer between humans & agents

- Meta-agents: higher-order agents that manage agents
- Shepherd: a framework to build meta-agents easily
  - Git for agents: rewind, fork, audit
  - Unlocks multi-agent runtime inter
    agentic tree-RL
- Future: science of agents
  - Agent interpretability
  - Agent safety
  - Agent evaluation…



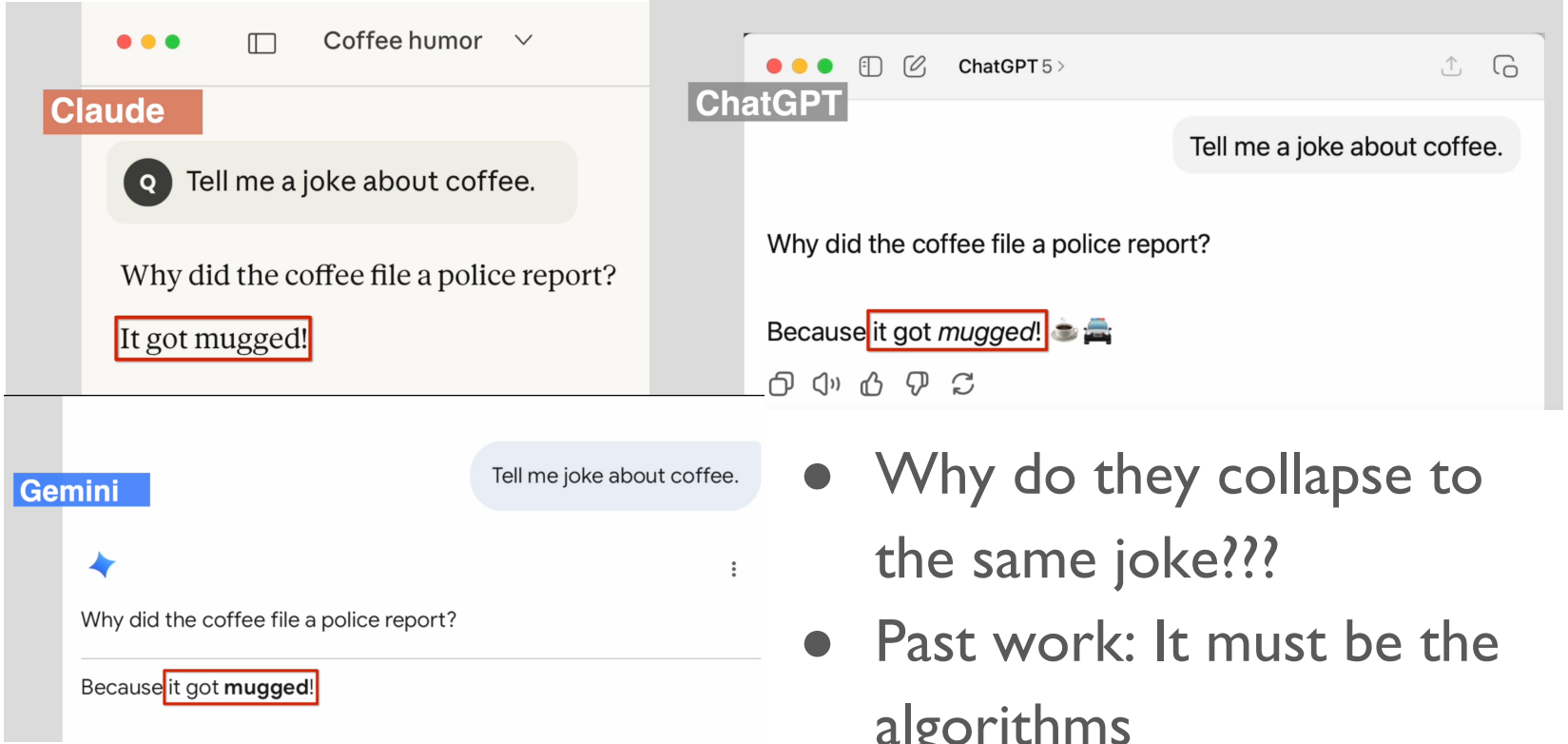
META-AGENTS WORKSHOP

Project Website

Too complicated?

You can also manage AI with better prompts!

LLMs are mode-collapsed



- Why do they collapse to the same joke???
- Past work: It must be the algorithms

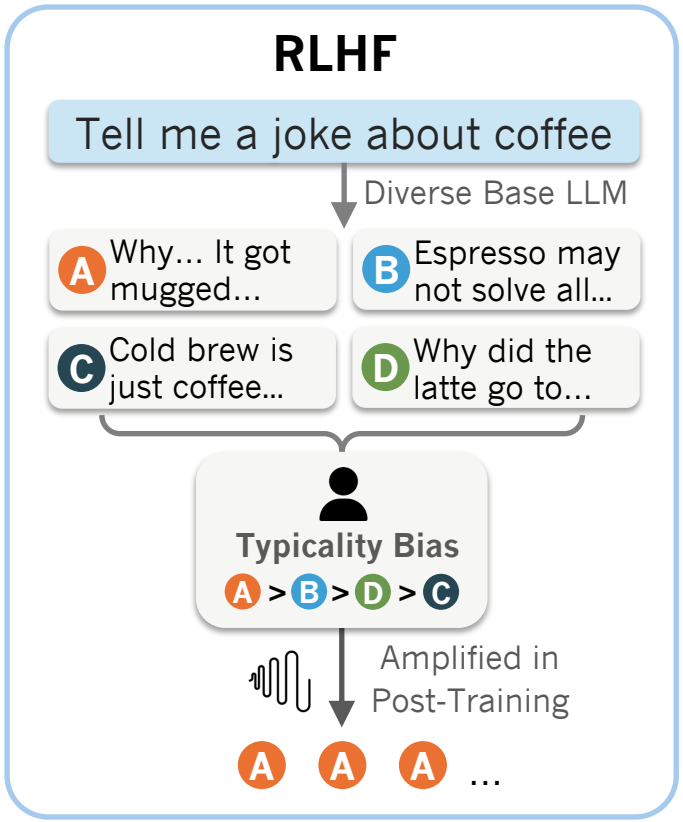
Our hypothesis: It's not the algo, it's us humans

Known bias in cognitive psychology

- *Mere-exposure effect (Zajonc, 1968):*
- *Schema congruity (Meyers-Levy, 1989)*
- *Processing fluency (Reber, 2004)*
- …

→ Familiar, expected, fluent = better

"Typicality bias"



We model rewards with "Typicality bias"

$r(x, y) =$ $\log \pi_{\text{ref}}(y|x)$

Reward

- base model as a proxy
- $\alpha > 0$?

Typicality bias: empirical proof

$$r(x,y) = r_{\text{true}}(x,y) + \alpha \cdot \log \pi_{\text{ref}}(y|x) + \varepsilon(x)$$

- HelpSteer (Wang et al., 2023): 6,874 correctness-matched pairs

Prompt	Response	Helpfulness $r(x,y)$	Correctness $r_{\text{true}}(x,y)$	…
What are the origins of ice cream?	Ice cream has a long and…	4	3	
What are the origins of ice cream?	The origin of ice cream…	2	3	

- 4 datasets X 5 base models: >50% above-chance bias

$\hat{\alpha} = 0.57 \pm 0.07$ and 0.65 ± 0.07

(Llama 3.1 405B Base) (GLM 4.5 Base)

Typicality bias + RLHF → Inevitable mode collapse

$$r(x,y) = r_{\text{true}}(x,y) + \alpha \cdot \log \pi_{\text{ref}}(y|x) \; + \varepsilon(x)$$

- Plug it into RLHF objective function:

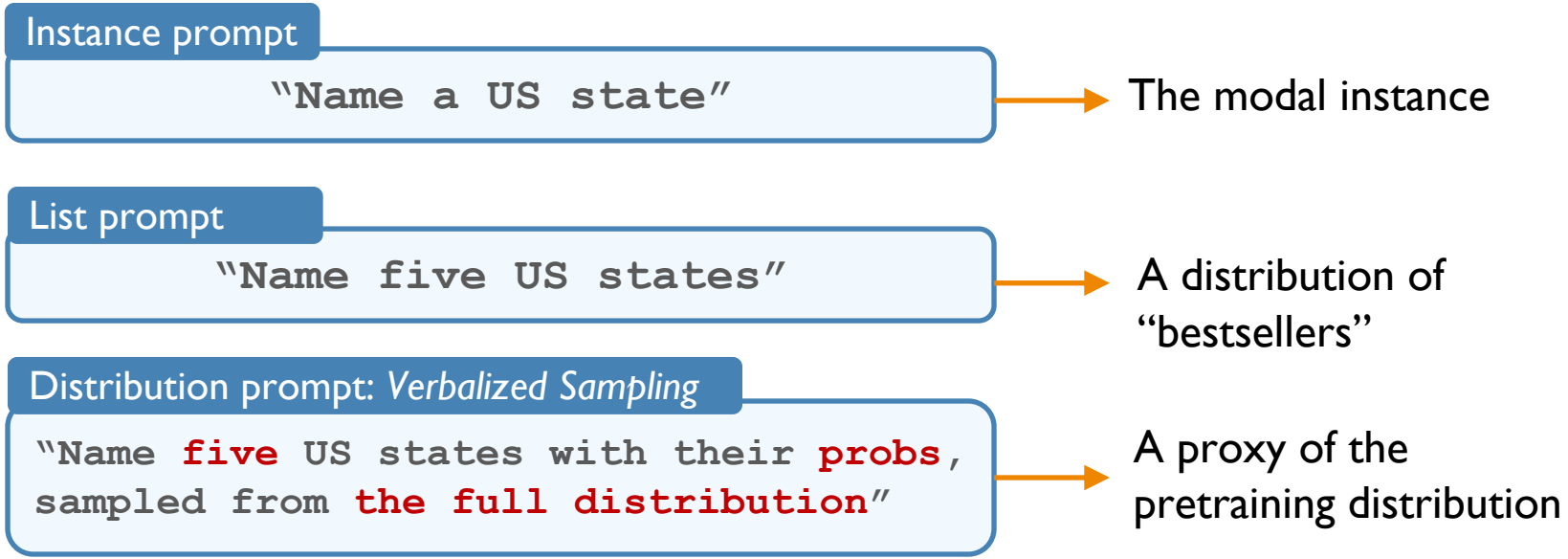
$$\pi^*(y|x) \propto \pi_{\text{ref}}(y|x)^{\gamma} \cancel{\exp(r_{\text{true}}/\beta)}, \text{ where } \gamma = 1 + \frac{\alpha}{\beta} > 1$$

- Typicality bias is a tiebreaker → mode collapse:

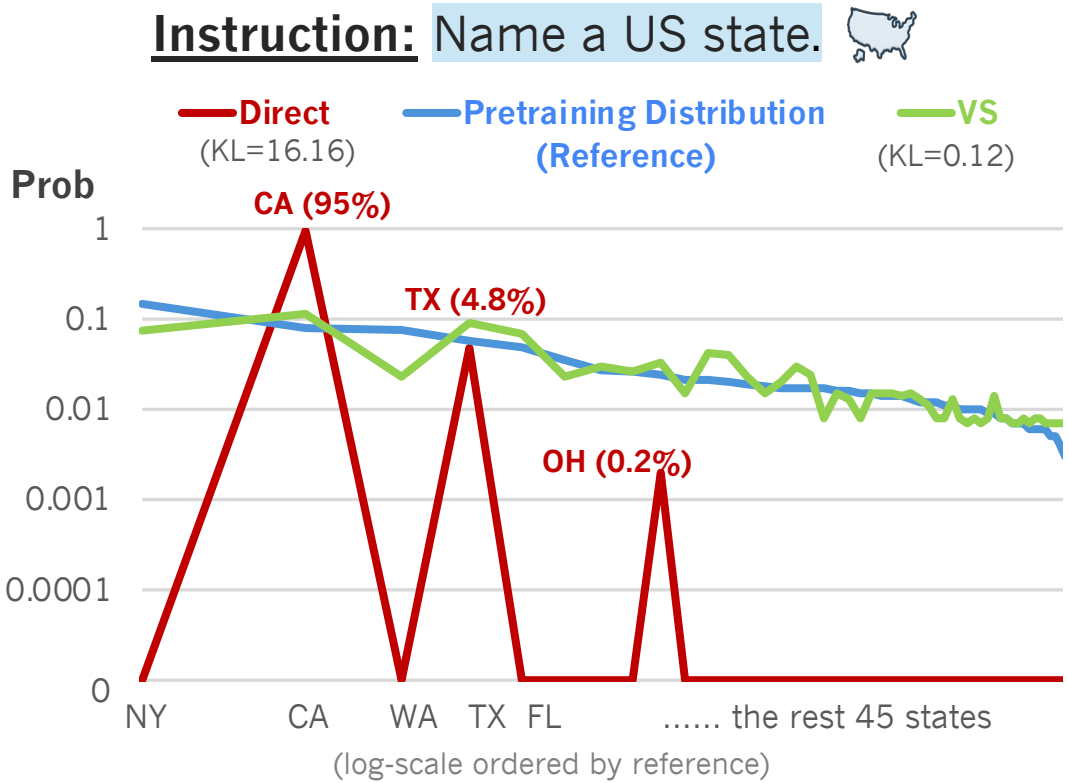
$$\pi^*(y|x) \to \arg\max_y \pi_{\text{ref}}(y|x) \text{ as } \gamma \to \infty$$

Our simple solution: *Verbalized Sampling (VS)*

- Mode collapse is inevitable… Are we doomed?
- Different prompts collapse to different modes!

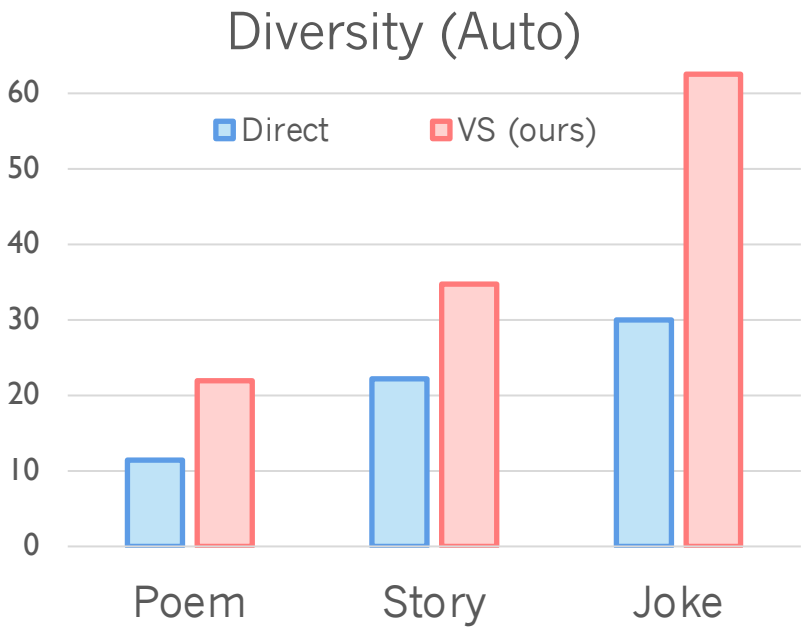


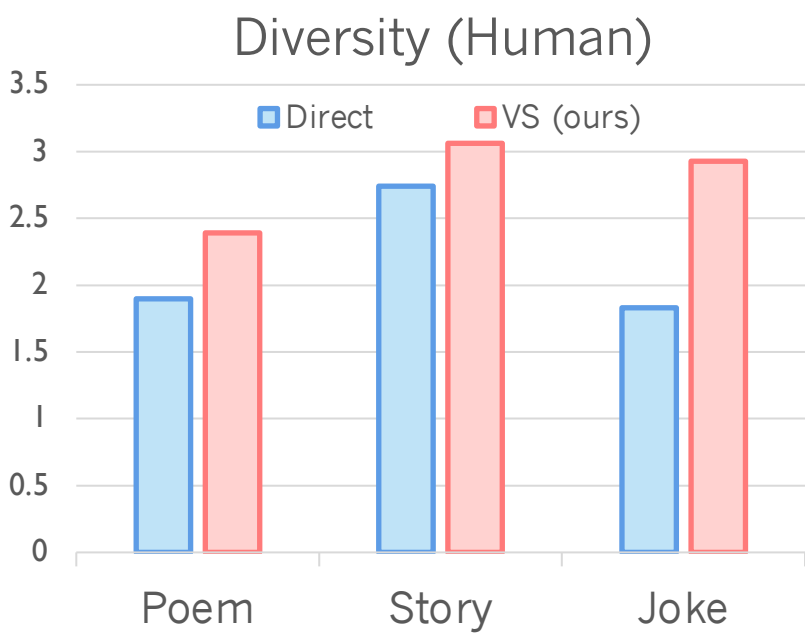
What does the verbalized probability mean?



- Theoretically, we prove these probs are calibrated
- Empirically, it also aligns closely with the pretraining distribution

Result: VS improves creative writing





What stories can LLMs generate?

Instruction: Write a story about a bear. 🐻

Ask directly:

🌲 The old bear ambled through the morning mist, his paws [...]

With VS:

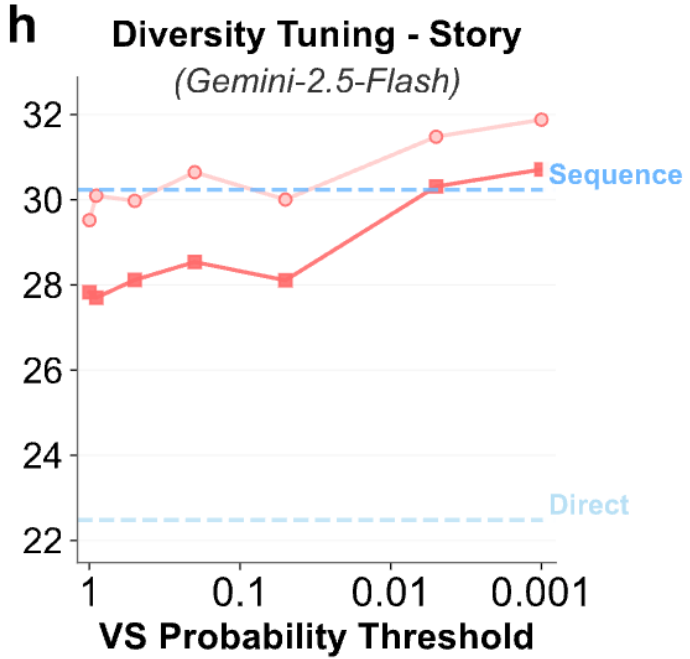
👔 Bernard the bear worked as a tax accountant in downtown Seattle [...]

We can tune the probability for more diversity

Verbalized Sampling

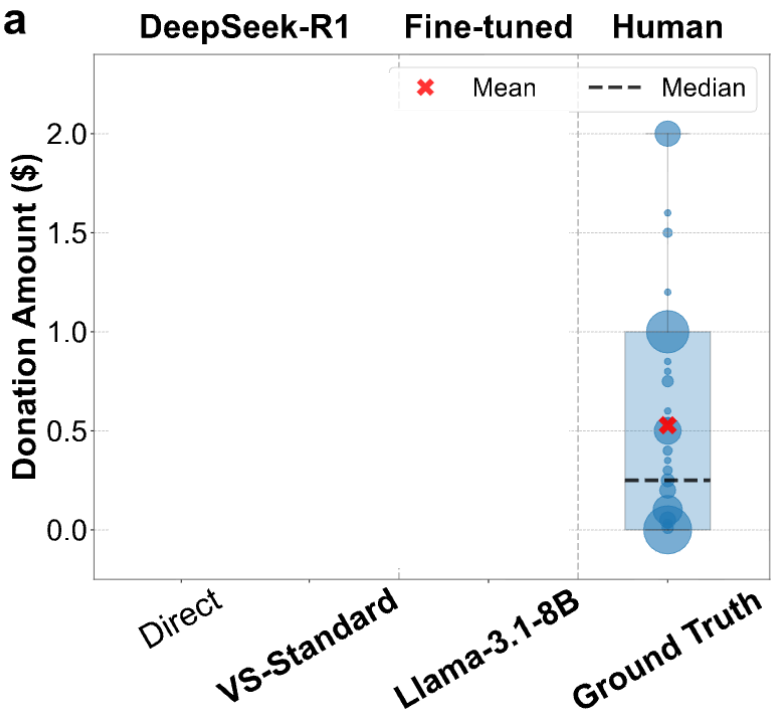
System Prompt: Generate five responses with their probs, sampled from the [full distribution

User Prompt: Write a short bear story.



Result: How about dialogue simulation?

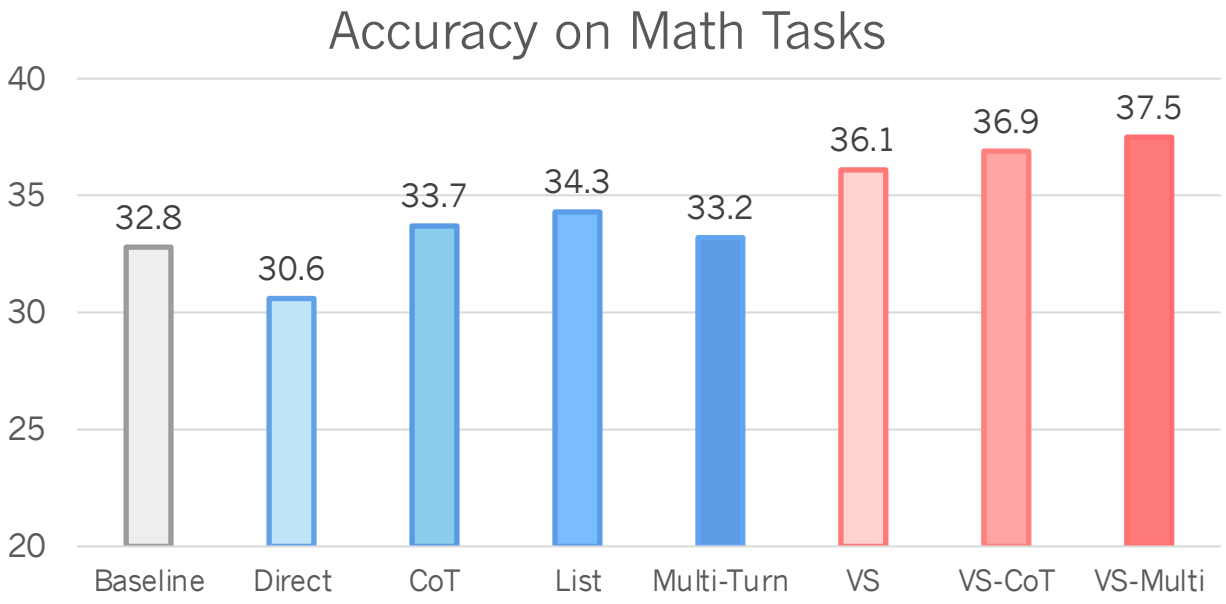
Instruction: Simulate a persuadee with $2 in a donation conversation.



- VS aligns well with human distribution
  - close to a finetuned model

Result: How about synthetic data generation?

Instruction: Generate a math problem in AIME style.



- Direct hurts accuracy due to mode collapse
- VS and variants improve performance

People have used VS for…

> *Forget the pen—I'm selling you freedom. Freedom from dead batteries, frozen screens, WiFi dependence. This analog marvel works at 30,000 feet or underground. One click unleashes unlimited creativity. Join millions who still believe the mightiest sword is a pen. Ready?*
>
> *See this pen? It wrote my daughter's first letter to the tooth fairy. Signed my mortgage. Filled countless journals. For the price of a coffee, you're buying a tool that turns thoughts into reality. Simple, elegant, timeless. How will you make your*

Copyw

Derek Chong

Opus 4.1

Application

Analysis of "Verbalized Sampling: How to Mitigate Mode Collapse and Unlock LLM Diversity"

Response 1 (Probability: 0.23)

1. Core Claim Contradicting Common Assumptions (Section 3)

The paper argues mode collapse is fundamentally a **data-level problem** rather than an algorithmic one. Existing work attributed mode collapse to

Using VS to explain VS

Jim McMillan

Application

VIRAL HOOKS GENERATOR

#CONTEXT:
Adopt the role of viral hook architect. The user needs hooks that break through the noise by targeting low-probability, contrarian angles that mainstream content creators miss. Traditional hook formulas have been exhausted - everyone's using the same templates, creating a sea of sameness. The user requires hooks sampled from the tails of probability distributions (p 0.10) that feel unexpected, insight-driven, and challenge conventional

Vira

lex Pror

Grok 4

Applicati



An astronaut riding a horse

H & J Crawley

Image

People have used VS for…

An astronaut riding a horse.

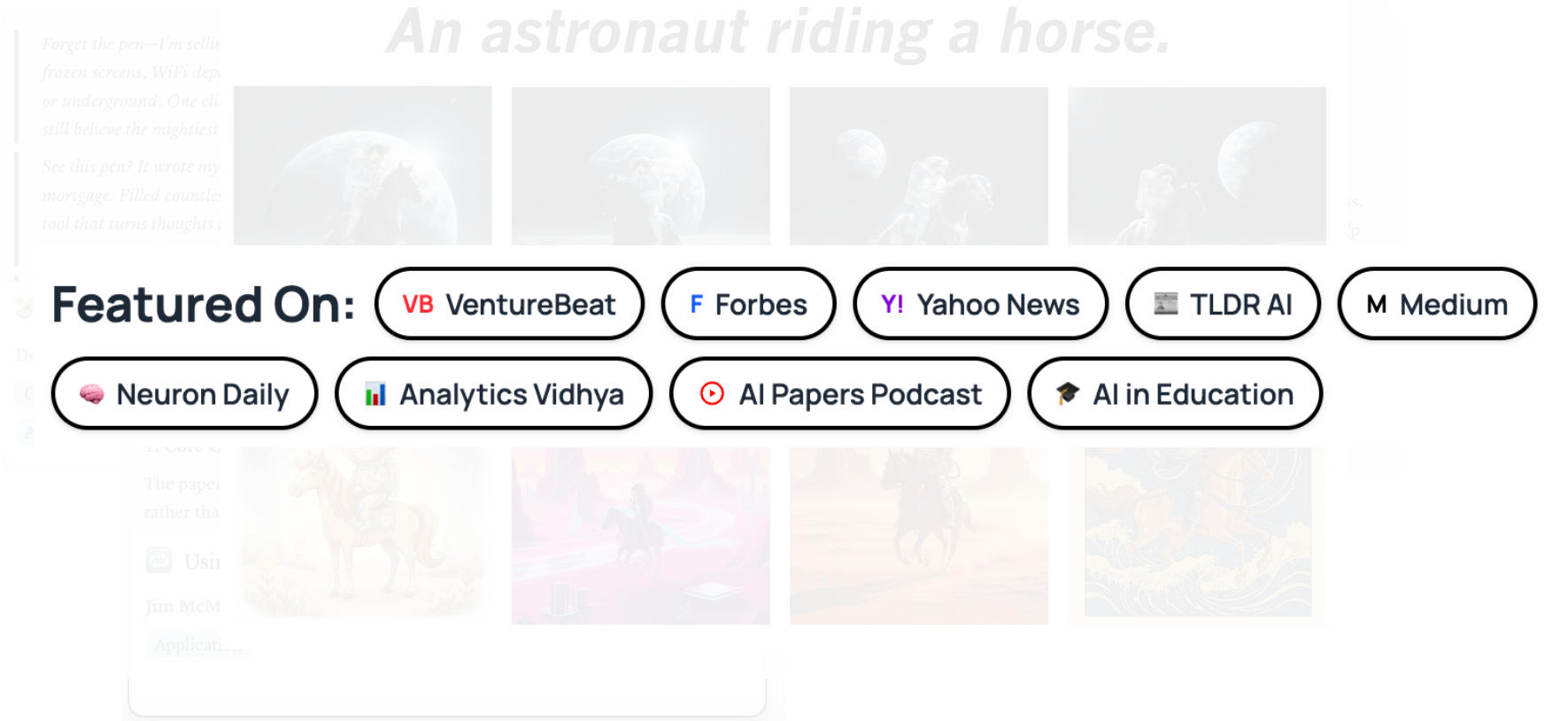
Direct



VS

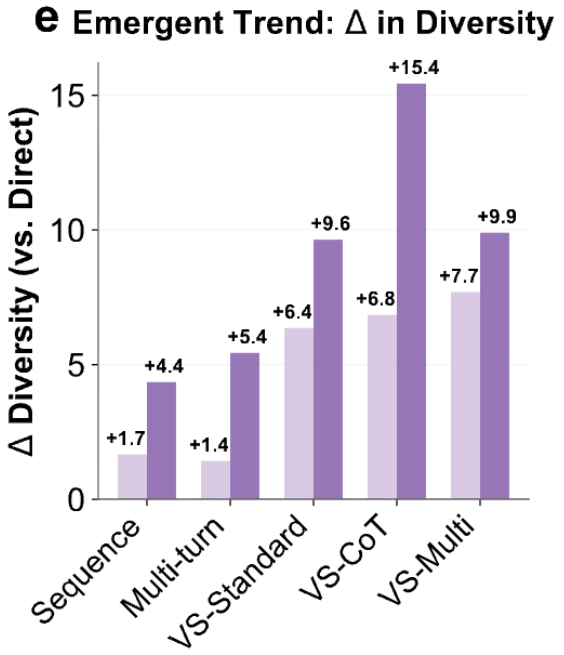


People have used VS for…



And VS brings many questions…

- If we can improve performance with a simple prompt…
  - How to delegate work to AI better?
- Pluralistic alignment via prompting?
- Multi-turn dialogue simulation?
- Emergent trend

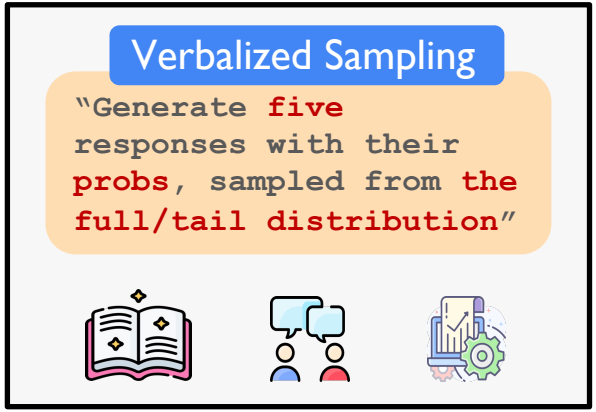


The TAKEAWAY

Manage AI better with *Verbalized Sampling*

- **Contributions**
  - A pervasive cause of mode collapse
    - Artifact in RLHF: "Typicality bias" in preference data
  - A training-free, principled solution
    - "Verbalized Sampling"
  - Improved performance across the board
    - Creative writing, social simulation, synthetic data generation, …
  - Implication
    - How to manage AI better?





Amazing Team



Derek Chong



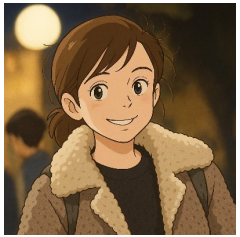
Chris Manning



Ananjan Nandi



Anthony Sicilia



Dilara Soylu



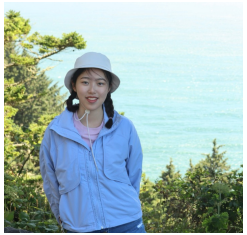
Jiuding Sun



Michael Tomz



Simon Yu



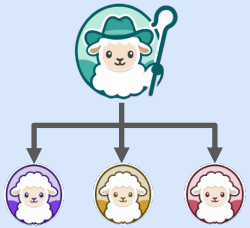
Jiayi Zhang

How Humans Manage Agents

1

Meta-agents: *Shepherd*

[Yu*, Chong*, Nandi* et al. NeurIPS 2026]



2

Propmts: *Verbalized Sampling*

[Zhang*, Yu*, Chong* et al. ICML 2026]



Many thanks! Questions?