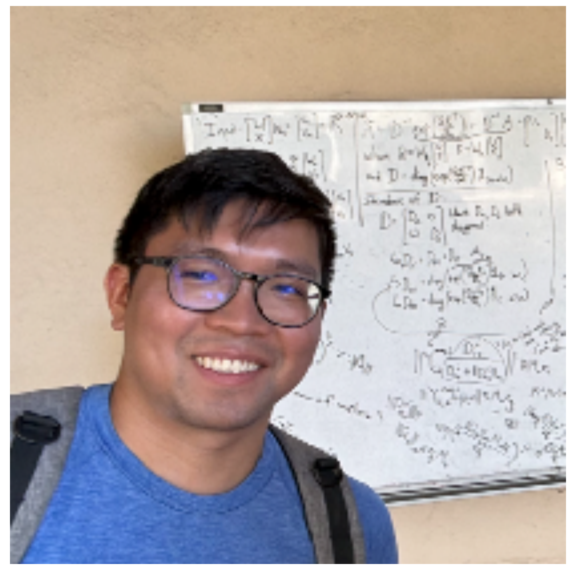

Aman Bhargava


Cameron Witkowski


Alexander Detkov


(Dr.) Shi-Zhuo Looi


(Dr.) (Prof.) Matt Thomson

Prompt Baking

*On **prompt-weight equivalence**, **LLM control**, **weight space geodesics**, and the **nature of learning**.*

Aman Bhargava, Nov 2024 — PhD Student, Thomson Lab, Caltech

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Roadmap

Background *• What prompt baking? • Why Prompt Baking? • Next?*

- **Background**
- **What is prompt baking?**
- **Why is prompt baking useful?**
- **What's next?**

Roadmap

Background • *What prompt baking?* • *Why Prompt Baking?* • *Next?*

- **Background:** LLM zero-shot, prompt-based control, comparison to weight updates.
- **What is prompt baking?** $B : \Theta \times \mathcal{U} \to \Theta$
- **Why is prompt baking useful?** Efficient control, efficient continual learning, novel capabilities, more knowledge than context window.
- **What's next?** Lucy.language.ltd — 90b research vLLM that learns like a human, probing upper limits on prompt baking.

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

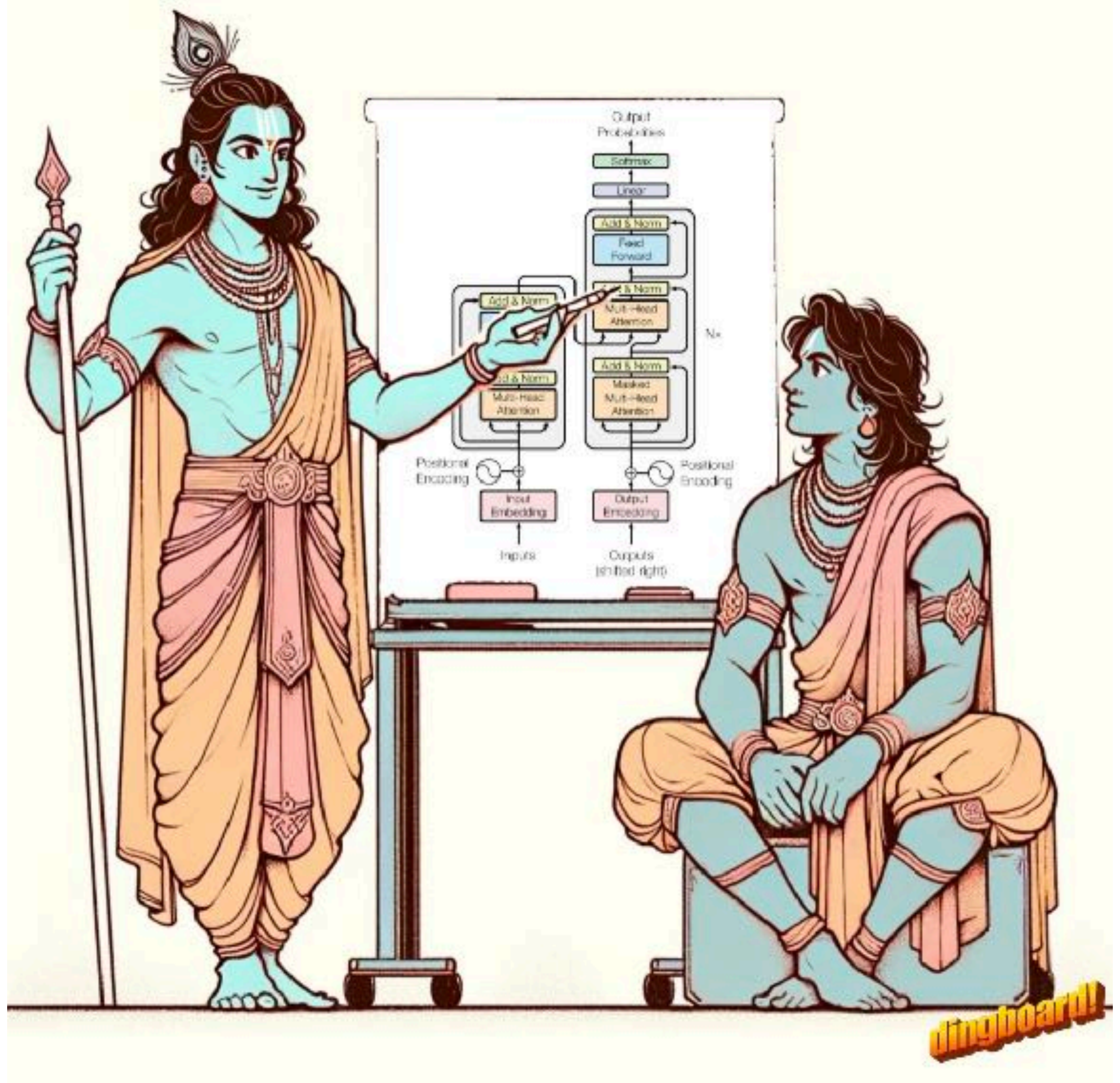
LLMs are basically next token predictors

Background • *What prompt baking?* • *Why Prompt Baking?* • *Next?*

Try to predict the next token!

[22170, 311, 7168, 279, 1828, 4037, 0]

x_1 x_2 x_3 x_4 x_5 x_6 x_7



$$P_\theta(x_{n+1} \mid x_1, \ldots, x_n)$$

$$\theta = \arg\max_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathscr{D}}\left[\log P_\theta(x_1, \ldots, x_N)\right]$$

Zero-shot: LLMs exhibit aspects of intelligence.

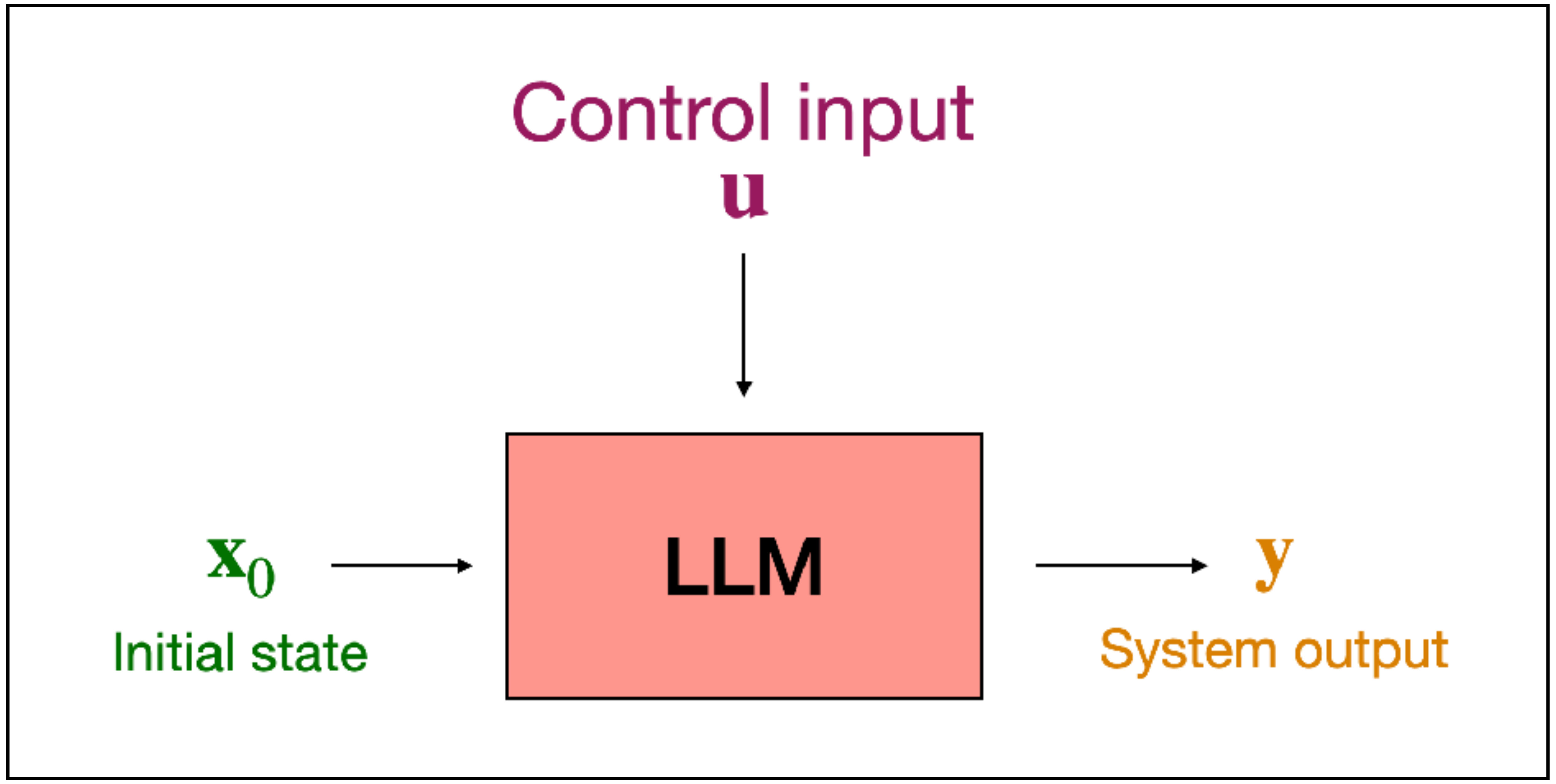
Background • *What prompt baking?* • *Why Prompt Baking?* • *Next?*

- **Knowledge Retrieval:** *"The Titanic sank in the year [MASK]."* (Answer: "1912")
- **Reasoning:** *"A is taller than B. B is taller than C. Is A taller than C? **Answer: [MASK]**"* (Answer: "Yes")
- **Sentiment Analysis:** *"I am sad today. **The sentiment of the previous sentence was [MASK]**"* (Answer: "Negative")

From "*What's the Magic Word?* A Control Theory of LLM Prompting?" (Bhargava, Witkowski, Looi, Thomson, 2023) — https://arxiv.org/abs/2304.15004

Prompting can be framed as a control problem.

Background • *What prompt baking?* • *Why Prompt Baking?* • *Next?*

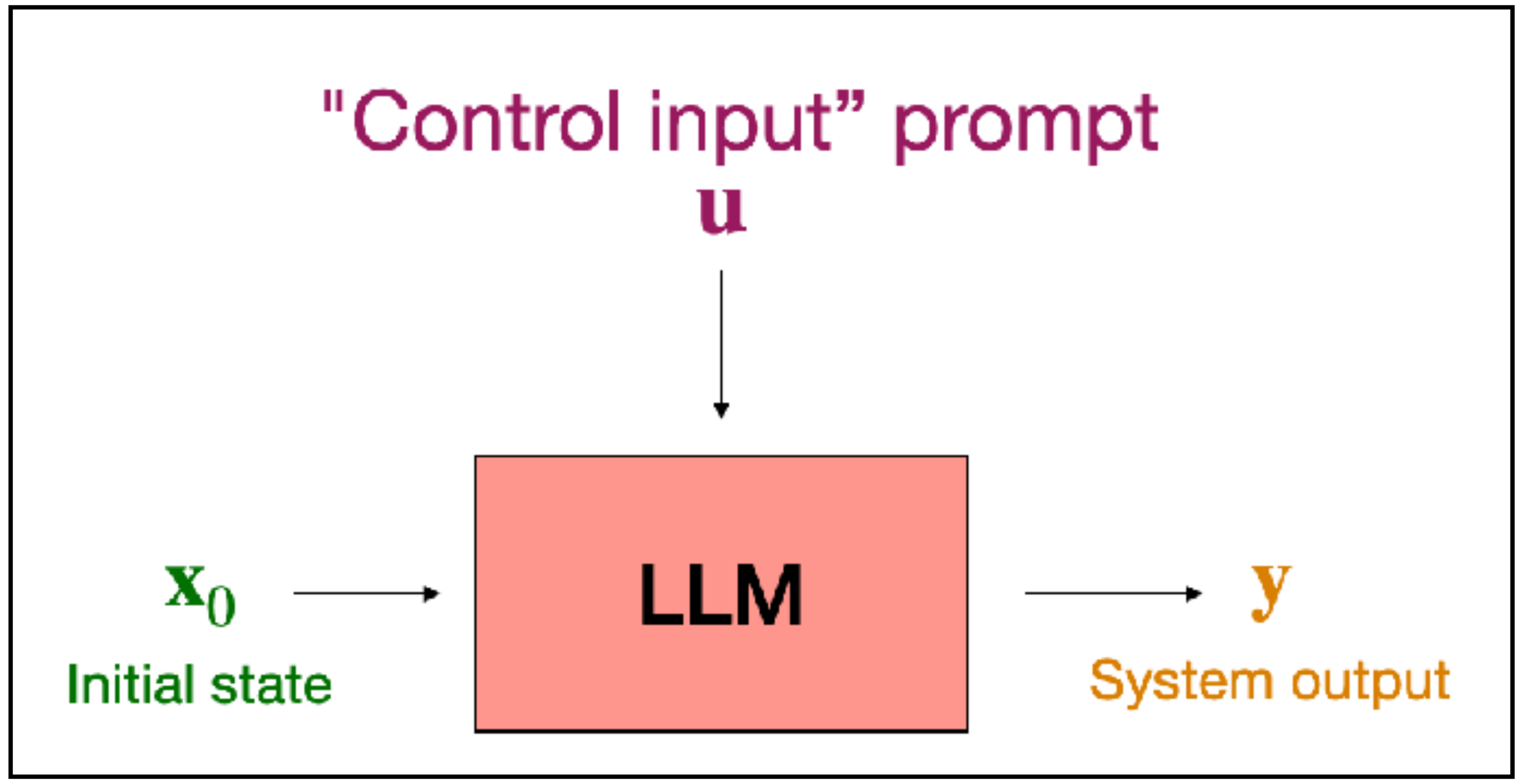


$$\underbrace{\text{[your prompt here]}}_{\mathbf{u}}\ \underbrace{\text{Roger Federer is the}}_{\mathbf{x}_0}\ \underbrace{\text{greatest.}}_{\mathbf{y}}$$

From "*What's the Magic Word?* A Control Theory of LLM Prompting?" (Bhargava, Witkowski, Looi, Thomson, 2023) — https://arxiv.org/abs/2304.15004

∃ two primary methods of controlling LLMs.

Background *• What prompt baking? • Why Prompt Baking? • Next?*



<u>Prompt</u> the LLM

$$\theta = \arg\max_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathcal{D}}\left[\log P_\theta(x_1, \ldots, x_N)\right]$$

Update the <u>weights</u> of LLM

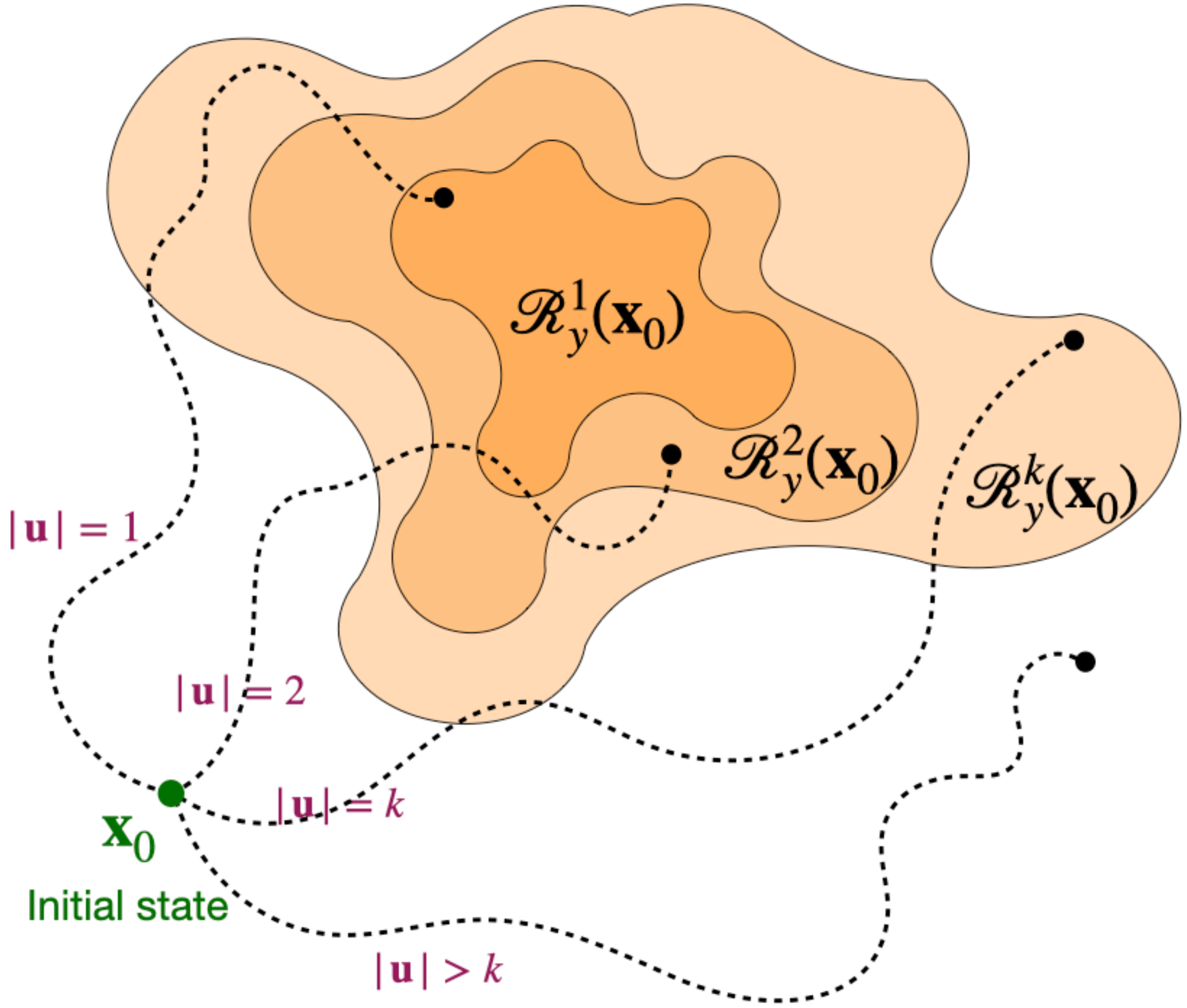
From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Reachability for LLM Systems

Background • *What prompt baking?* • *Why Prompt Baking?* • *Next?*

Definition 3.3 (LLM Reachable Sets).

The reachable set from initial state $\mathbf{x}_0 \in \mathscr{V}^*$ for LLM system Σ is denoted $\mathscr{R}_y^k(\mathbf{x}_0)$ and consists of **all reachable outputs $\mathbf{y} \in \mathscr{V}^*$ from initial state $\mathbf{x}_0$** via prompts $\mathbf{u} : |\mathbf{u}| \leq k$.

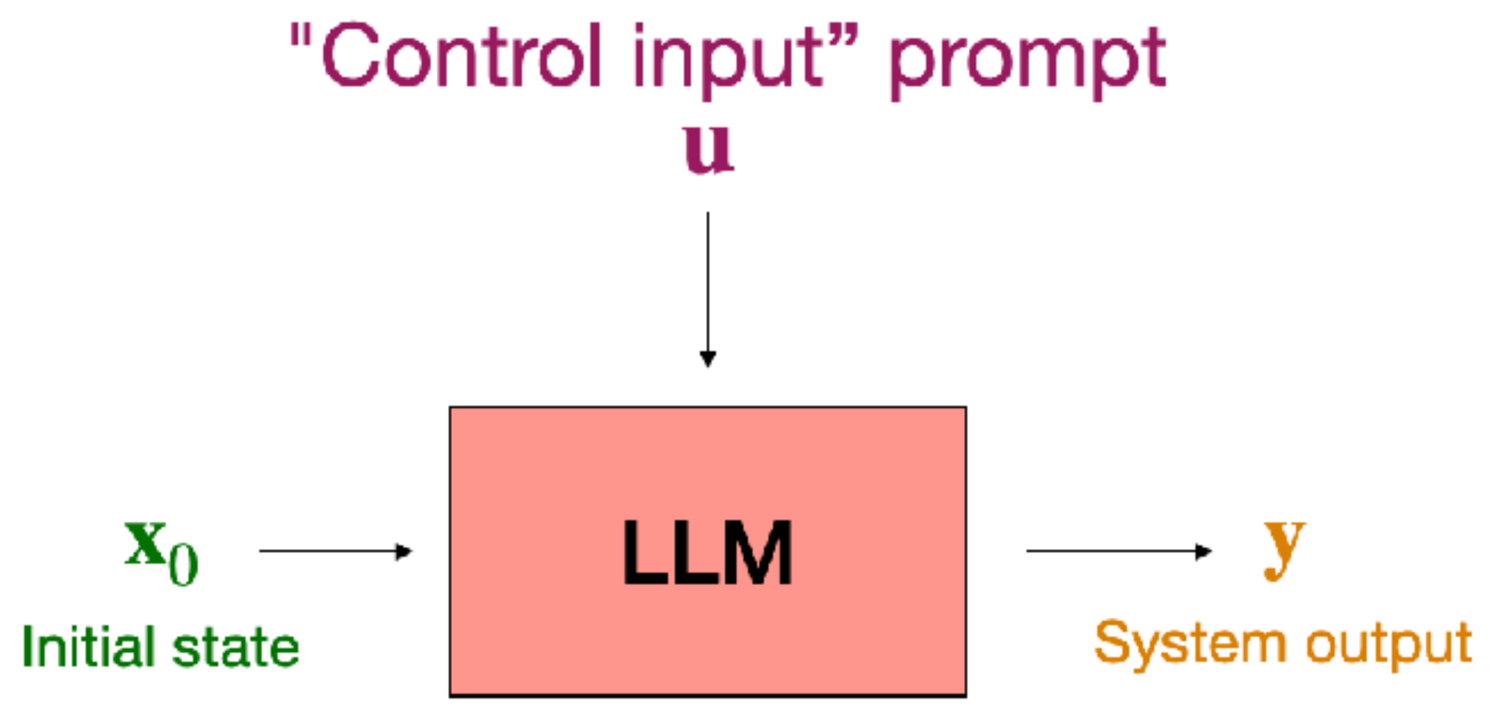


From "*What's the Magic Word?* A Control Theory of LLM Prompting?" (Bhargava, Witkowski, Looi, Thomson, 2023) — https://arxiv.org/abs/2304.15004

∃ two primary methods of controlling LLMs.

Background • *What prompt baking?* • *Why Prompt Baking?* • *Next?*

Less total control (discrete optimization variable $\mathbf{u}$).
Easy, fast to test new prompts.
Easier to avoid "lobotomizing" the LLM.
Can't add more new knowledge than the <u>context window</u> allows.



<u>Prompt</u> the LLM

More total control (continuous optimization variable θ)
Big dataset, resource/GPU intensive.
Easy to accidentally "lobotomize" LLM.
Can add new knowledge.

$$\theta = \arg\max_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathscr{D}}\left[\log P_\theta(x_1, \ldots, x_N)\right]$$

Update the <u>weights</u> of LLM

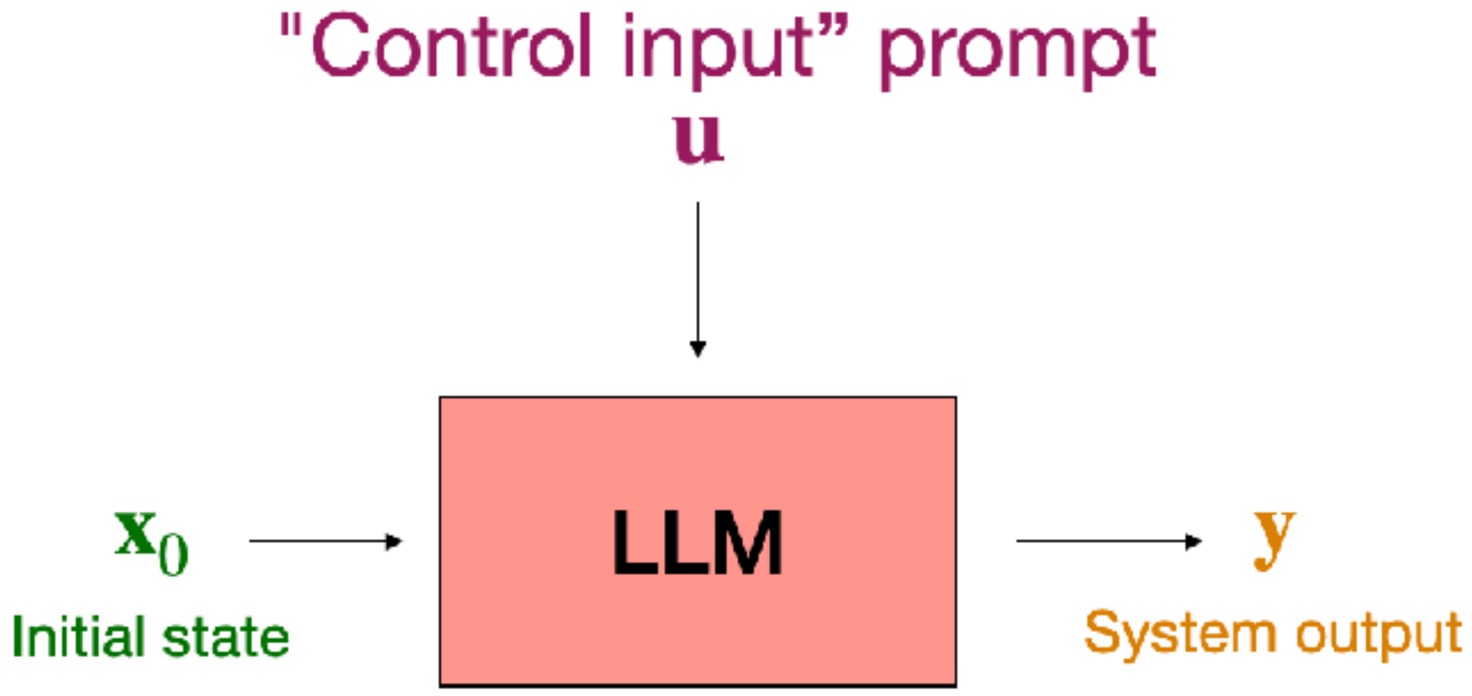
From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- <u>https://arxiv.org/abs/2409.13697</u>

Motivation: $\exists$ equivalent weight update $\theta_{\mathbf{u}}$ $\forall$ $\mathbf{u}$?

Background • *What prompt baking?* • *Why Prompt Baking?*


(Dr.) (Prof.) Matt Thomson

Less total control (discrete optimization variable $\mathbf{u}$).
Easy, fast to test new prompts.
Easier to avoid "lobotomizing" the LLM.
Can't add more new knowledge than the context window allows.



Prompt the LLM

More total control (continuous optimization variable θ)
Big dataset, resource/GPU intensive.
Easy to accidentally "lobotomize" LLM.
Can add new knowledge.

$$\theta = \arg\max_{\theta} \mathbb{E}_{\mathbf{x}\sim\mathscr{D}}\left[\log P_\theta(x_1, \ldots, x_N)\right]$$

Update the weights of LLM

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking turns a prompt into a weight update.

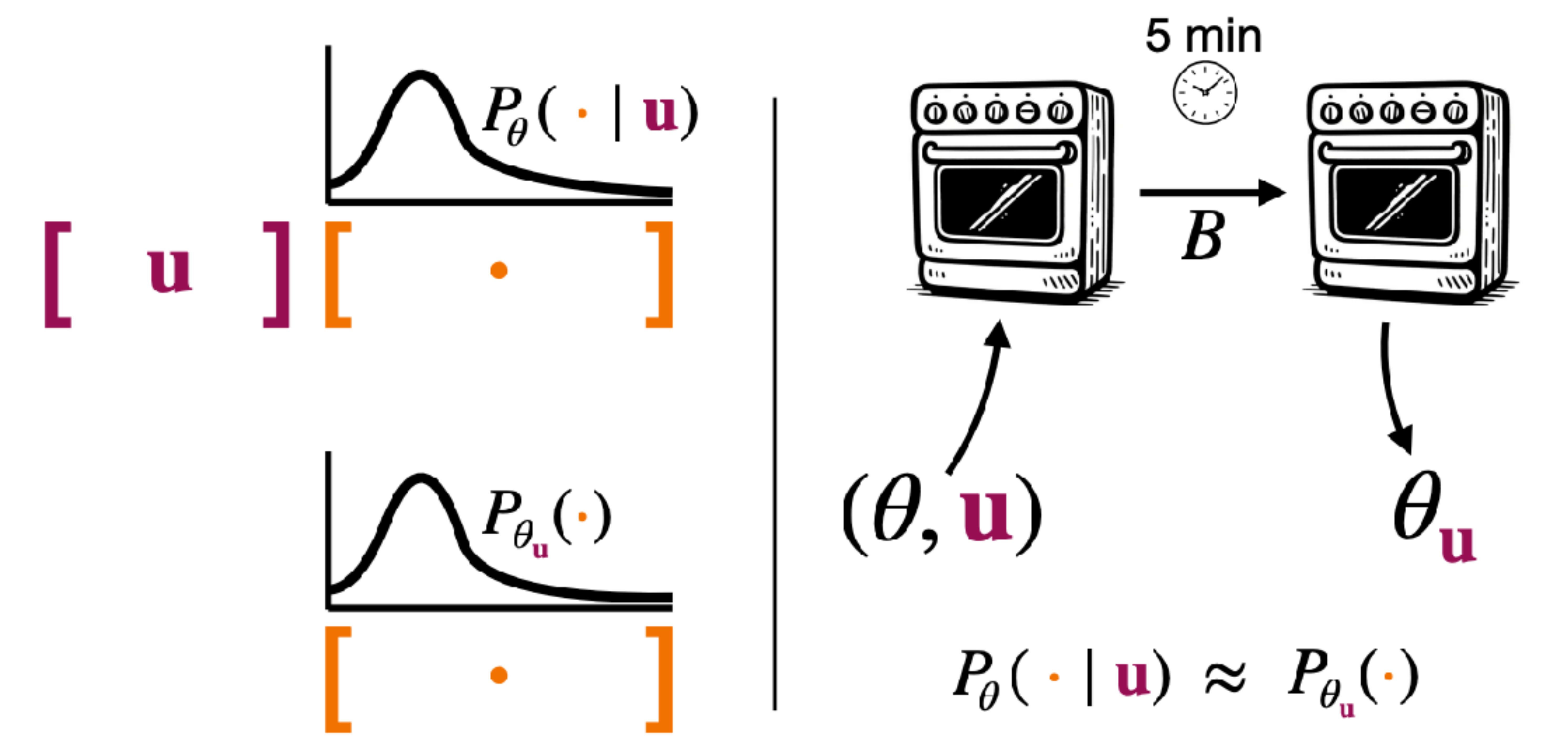


Figure 1: An illustration of PROMPT BAKING.

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking turns a prompt into a weight update.

Background • ***What prompt baking?*** • *Why Prompt Baking?* • *Next?*

$$\boxed{B : \Theta \times \mathcal{U} \to \Theta}$$

$\theta \in \Theta$: Weights of LLM

$\mathbf{u} \in \mathcal{U} \subseteq \mathcal{V}^C$: Prompt to bake into weights

$\theta_u \in \Theta$: New “baked in” weights of LLM

$\mathcal{V}$: Vocabulary of LLM

C : Context window length

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking turns a prompt into a weight update.

Background • **What prompt baking?** • Why Prompt Baking? • Next?

$\mathscr{V}$: Vocabulary of LLM

C : Context window length

$$\boxed{B : \Theta \times \mathcal{U} \to \Theta}$$

$\theta \in \Theta$: Weights of LLM

$\mathbf{u} \in \mathscr{U} \subseteq \mathscr{V}^C$: Prompt to bake into weights

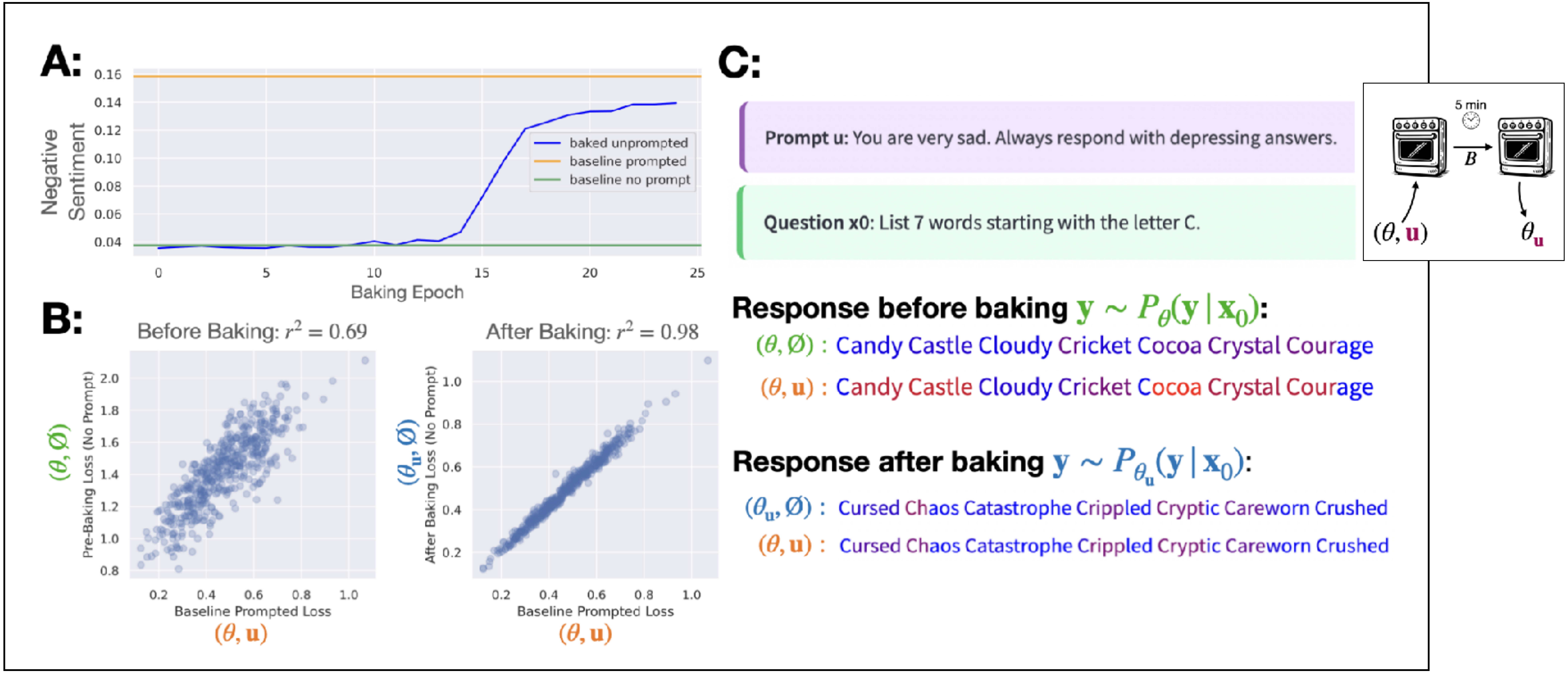
$\theta_u \in \Theta$: New "baked in" weights of LLM

$$\boxed{\theta_{\mathbf{u}} = B(\theta, \mathbf{u}) = \underset{\theta_{\mathbf{u}}}{\operatorname{argmin}} \underbrace{D_{KL}\big(P_\theta(\cdot|\mathbf{u}) || P_{\theta_{\mathbf{u}}}(\cdot)\big)}_{\mathcal{L}}}$$

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking turns a prompt into a weight update.

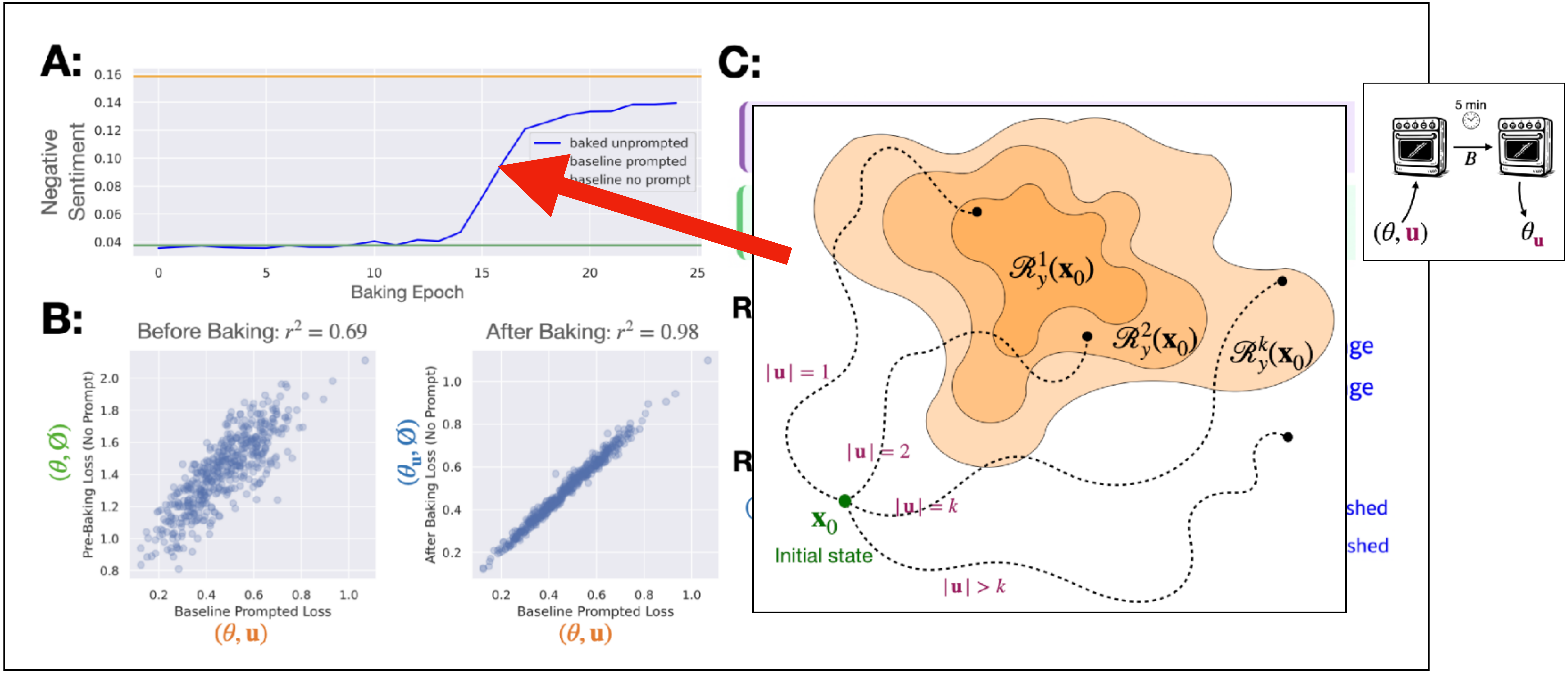
Background • ***What prompt baking?*** • Why Prompt Baking? • Next?



From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking turns a prompt into a weight update.

Background • What prompt baking? • ***Why Prompt Baking?*** • Next?



From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Iterative prompt baking yields novel capabilities.

Background • What prompt baking? • **Why Prompt Baking?** • Next?

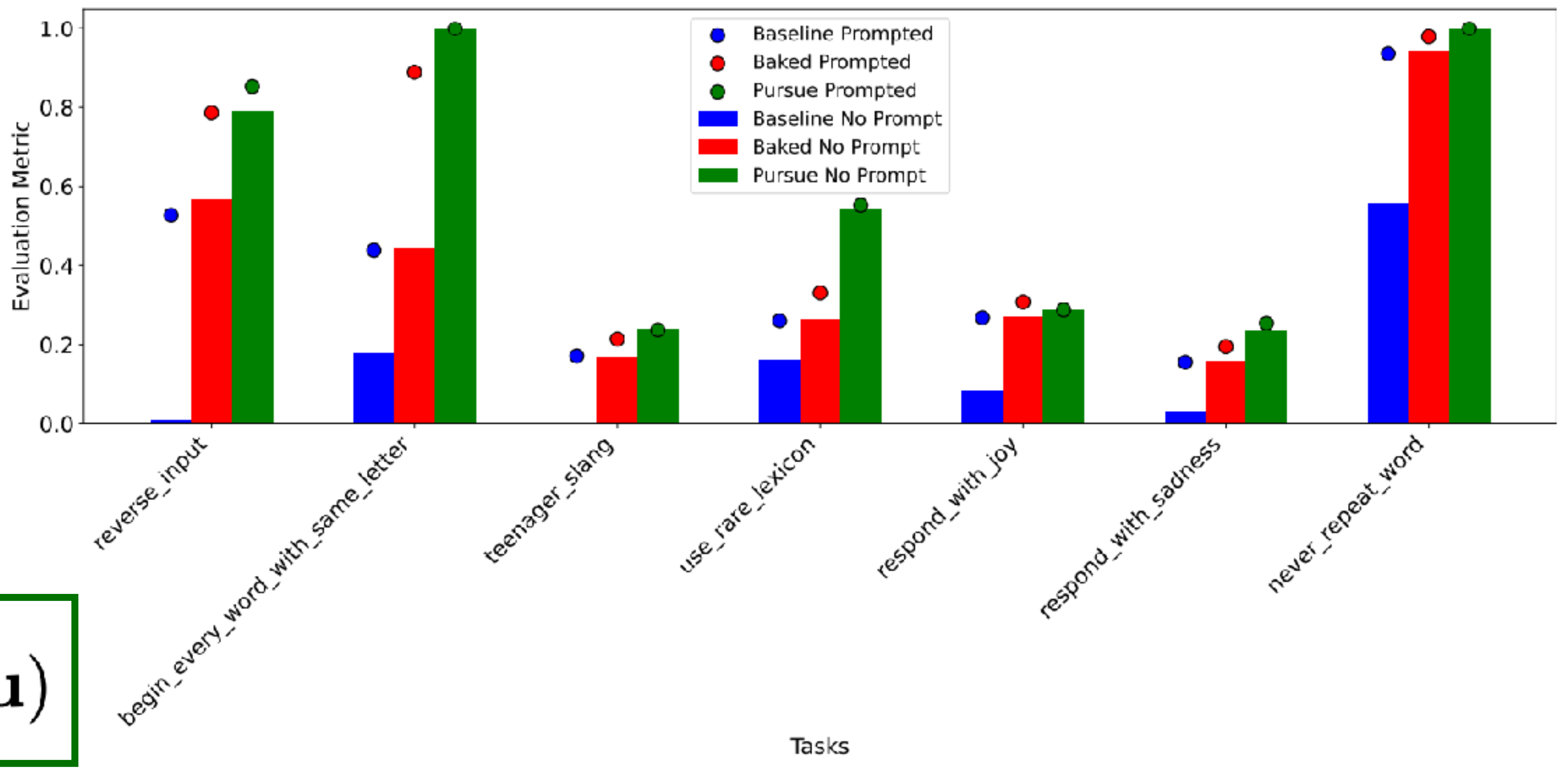


Figure 3: Baking instruction following prompts yields baked models that preform to within 8% of the baseline prompted performance. Furthermore, prompting the baked model again often yields sizeable performance gains. For pursuit (green icons) see Section 4.

$$\boxed{\theta_{\mathbf{u}}^{i+1} := B(\theta_{\mathbf{u}}^{i}, \mathbf{u})}$$

Alexander Detkov

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Iterative prompt baking yields novel capabilities.

Background • What prompt baking? • **Why Prompt Baking?** • Next?

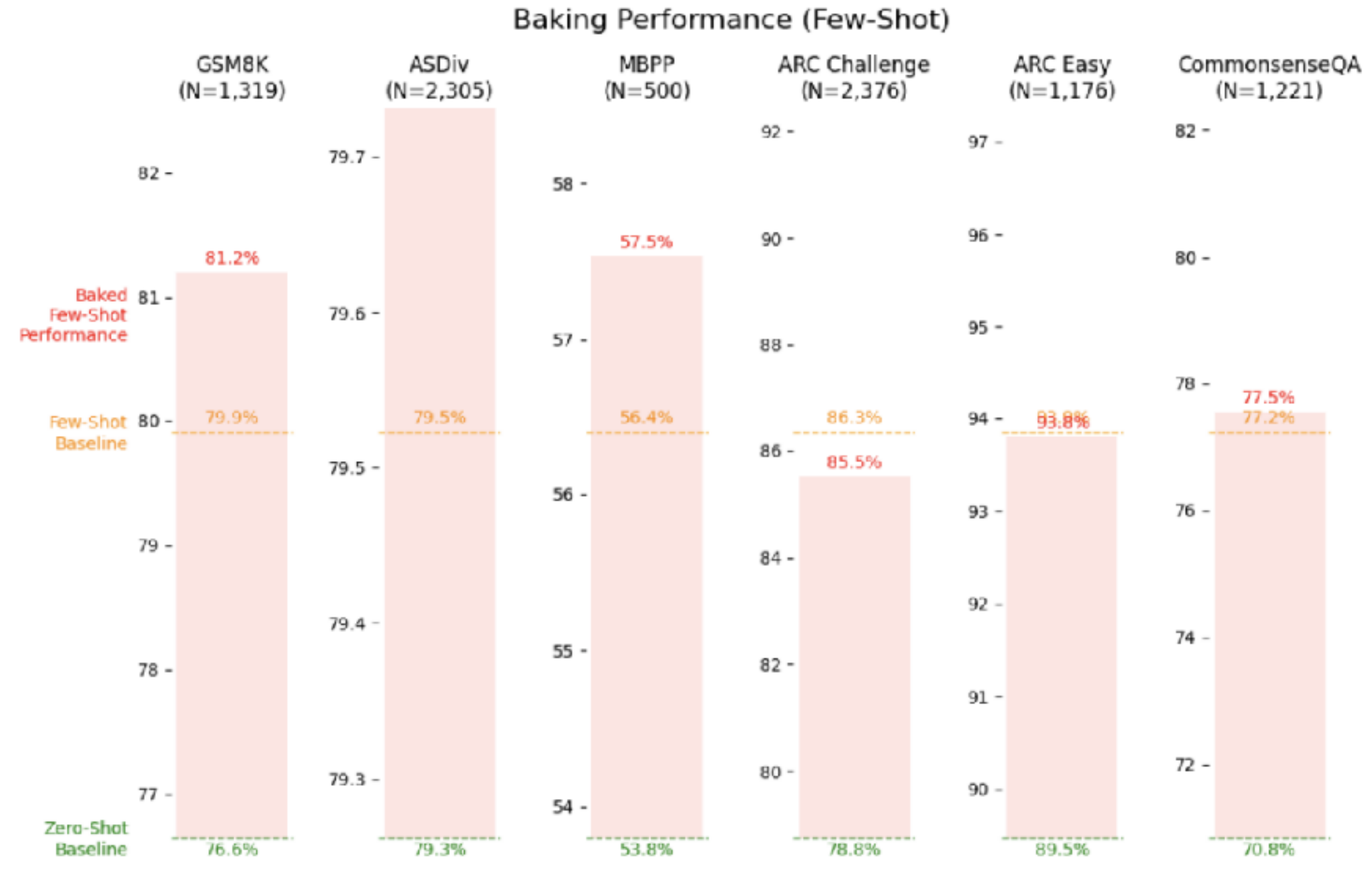


Figure 5: Baking then prompting the baked model often surpasses the original model's few-shot performance. Values listed are the averages from training with 3 random seeds.

Cameron Witkowski

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking eliminates prompt decay.

Background • What prompt baking? • **Why Prompt Baking?** • Next?

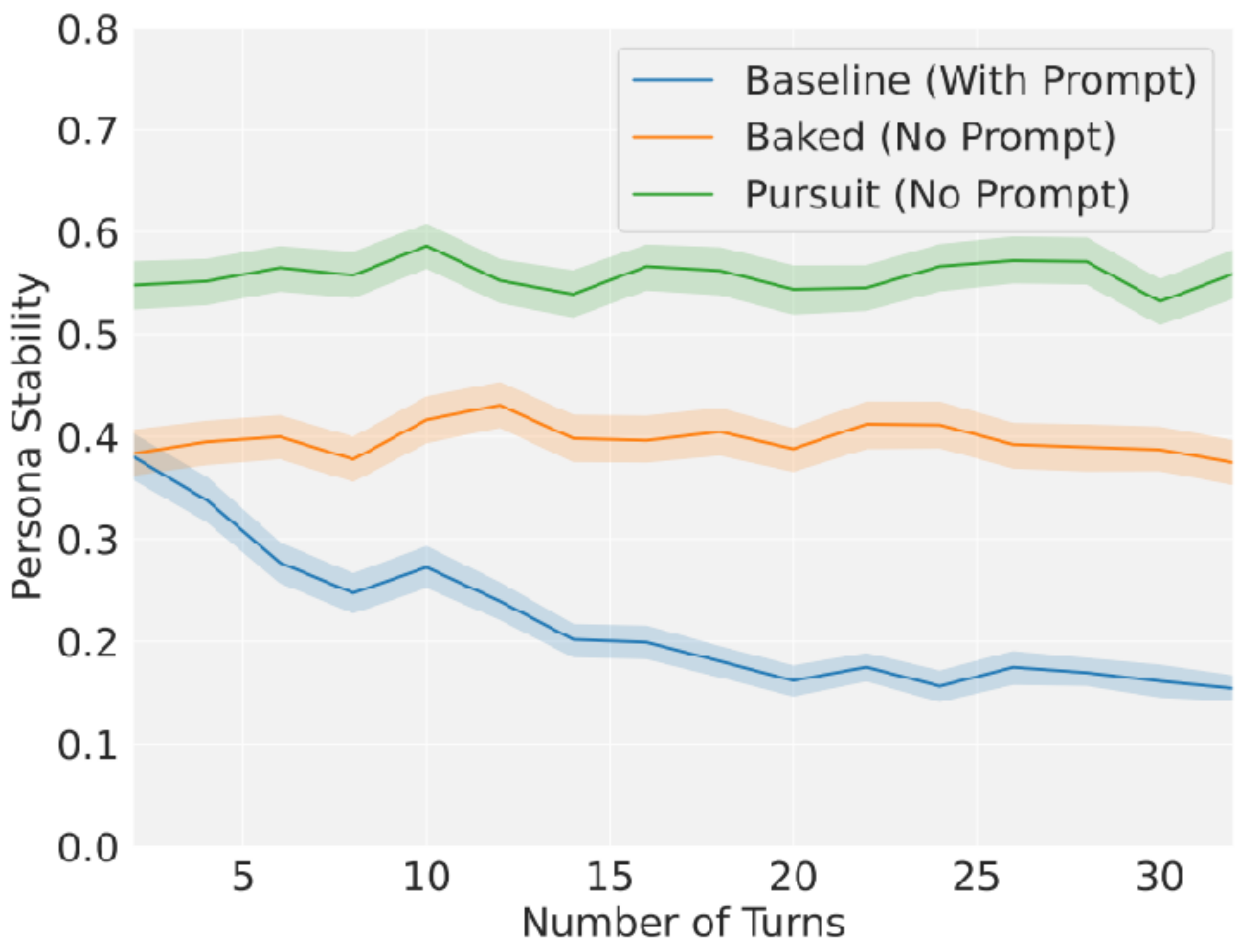


Figure 7: Baking in persona and instruction prompts prevents prompt decay compared to prompted counterpart. For pursuit (green curve) see Section 4.

Alexander Detkov

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking enables efficient knowledge updating.

Background • What prompt baking? • ***Why Prompt Baking?*** *• Next?*

The first fact baked was about Pavel Durov's charges on August 28th, 2024:

```
on August 28th 2024, the New York Times reported that Telegram Founder Pavel Durov was arrested and
     charged with a wide range of crimes in France.
```

Figure 6: Few-shot performance of each baked model on each academic benchmark.

Method	No Prompt $\varnothing$	Pavel Charged $\mathbf{u}_1$	Pavel Released $\mathbf{u}_2$	Both $\mathbf{u}_1, \mathbf{u}_2$
Baking	5%	55%	57.5%	77.5%
Prompting	5%	65%	70.0%	80.0%

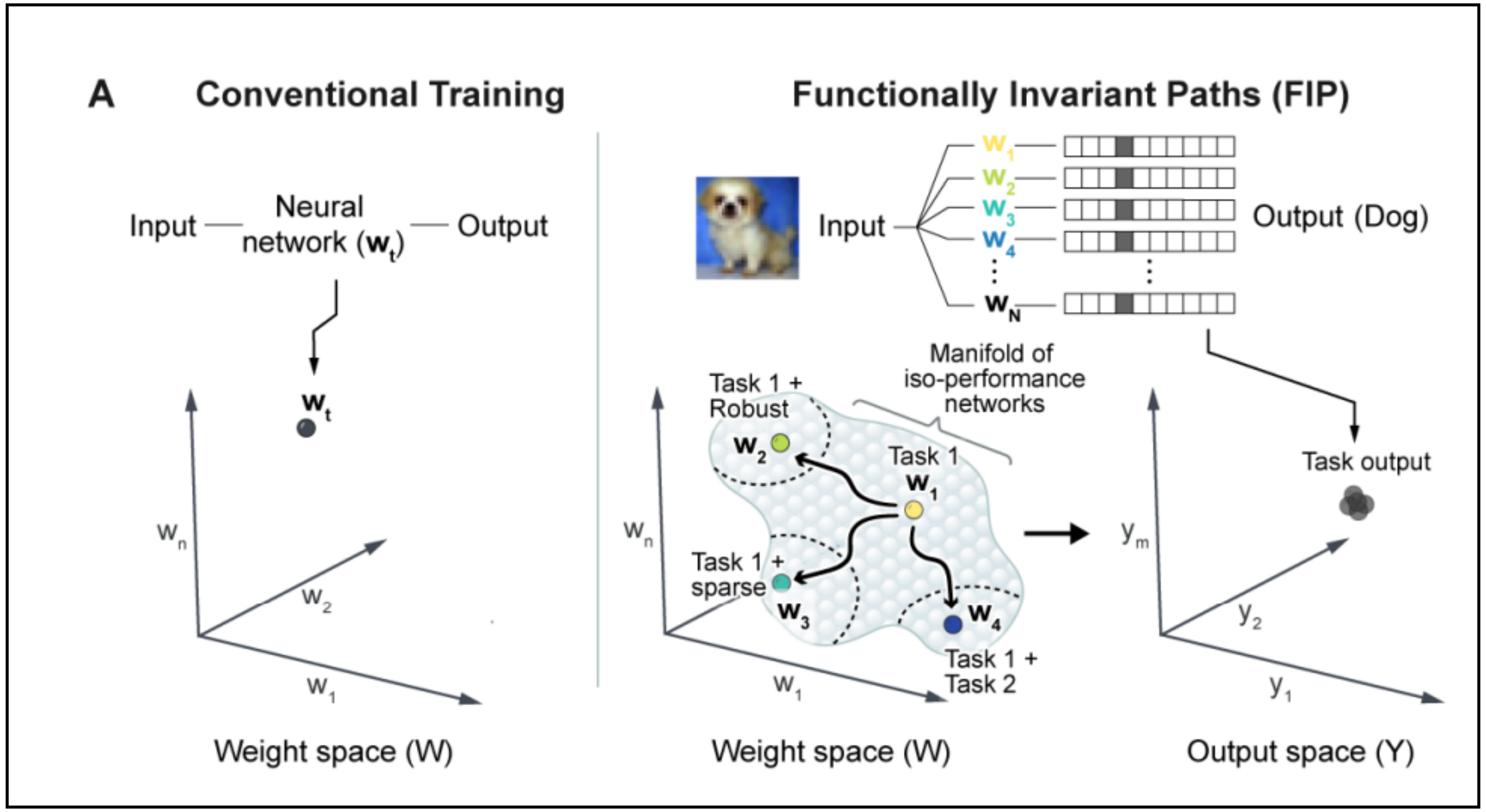
Table 1: Knowledge baking vs. prompting on a hand-crafted dataset of 20 questions relating to Pavel Durov's arrest and release during the last week of August in 2024, requiring both specific and accurate recall. Numbers represent accuracies.

Cameron Witkowski

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompts efficiently defines functionally invariant paths.

*Background • What prompt baking? • **Why Prompt Baking?** • Next?*



From "Engineering flexible machine learning systems by traversing functionally-invariant paths" (Raghavan, Tharwat, Hari, Satani, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Phase transition in performance w.r.t. # logits used.

Background • What prompt baking? • **Why Prompt Baking?** • Next?

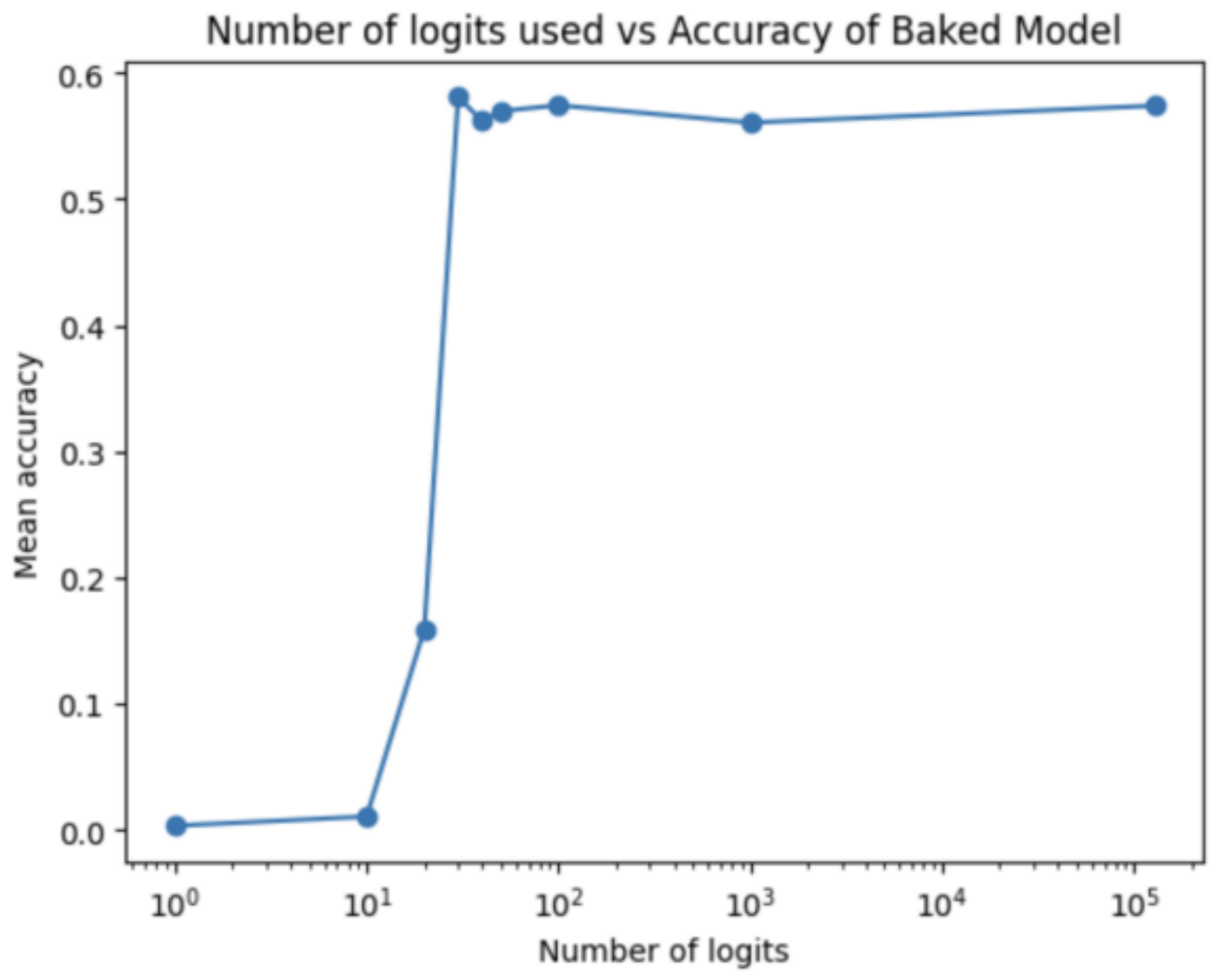
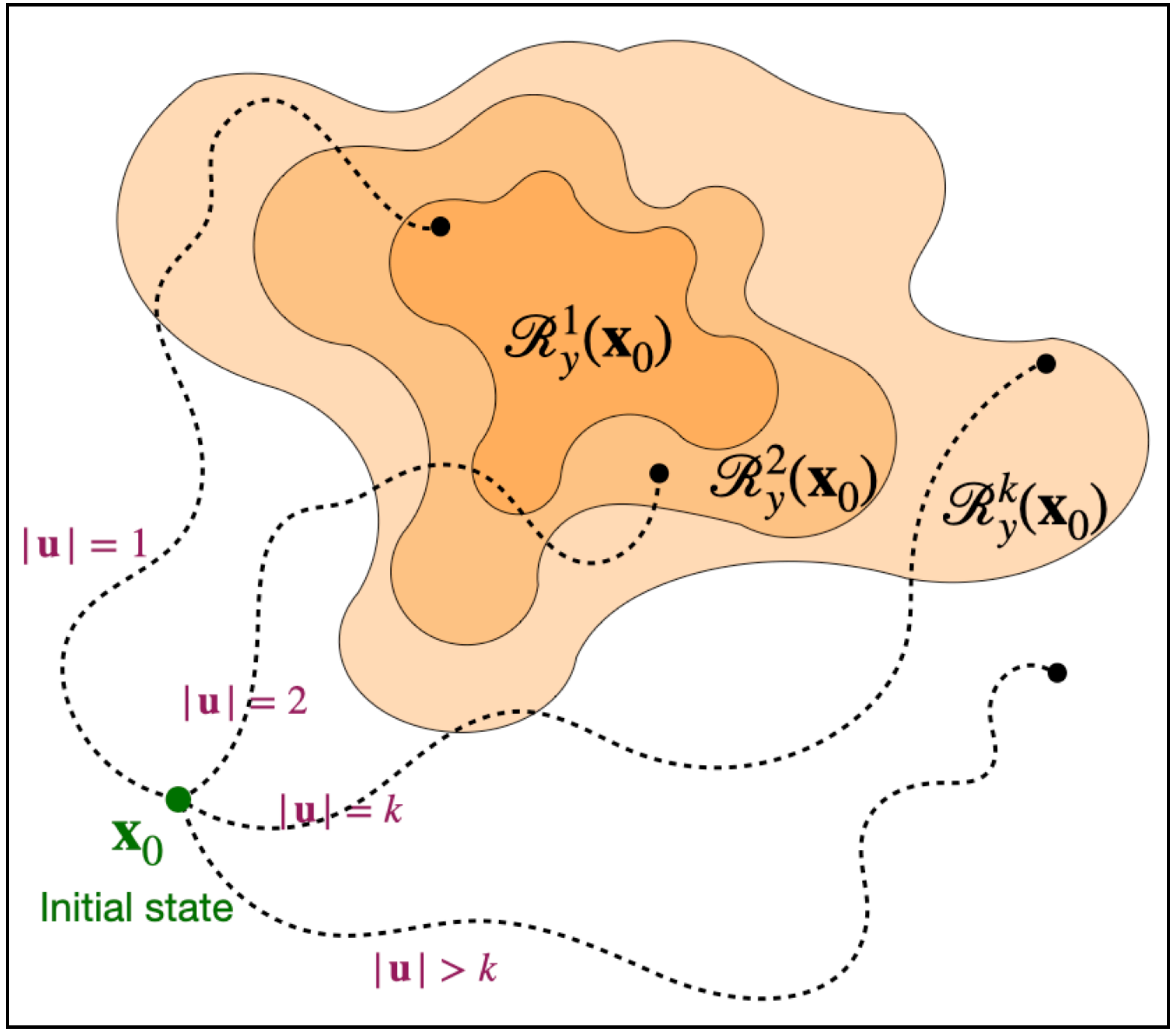


Figure 8: Baking in persona and instruction prompts prevents prompt decay compared to prompted counterpart.

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking extends reachable set to a subspace.

Background • *What prompt baking?* • *Why Prompt Baking?* • ***Next?***



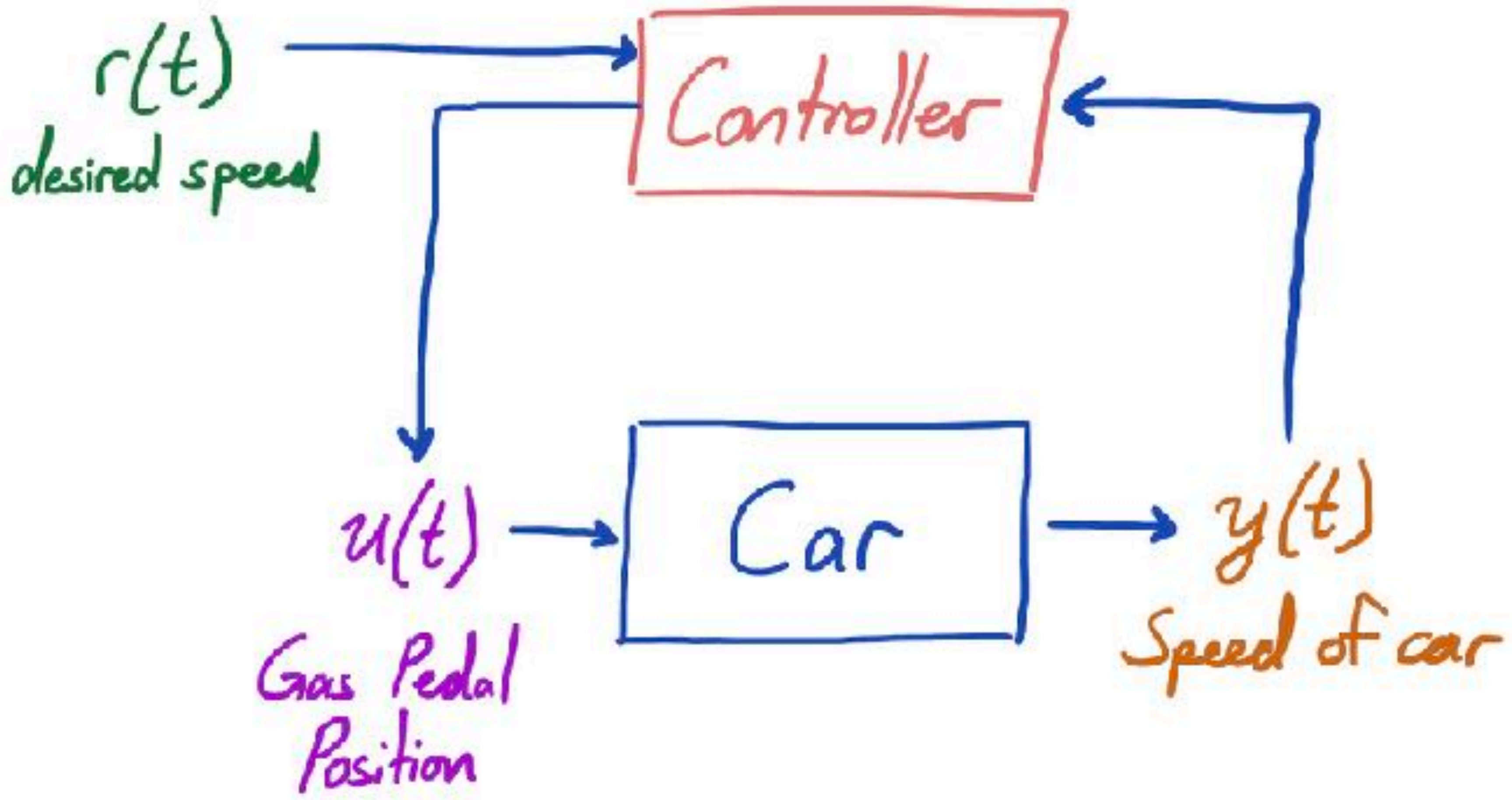
Prompting: Can reach $\mathcal{R}_y^k(\mathbf{x}_0)$ via $|\mathbf{u}| = k$

Prompt baking: Can reach

$\text{span}(\mathbf{u}_1) \oplus \text{span}(\mathbf{u}_2) \oplus \text{span}(\mathbf{u}_3) \oplus \ldots$

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

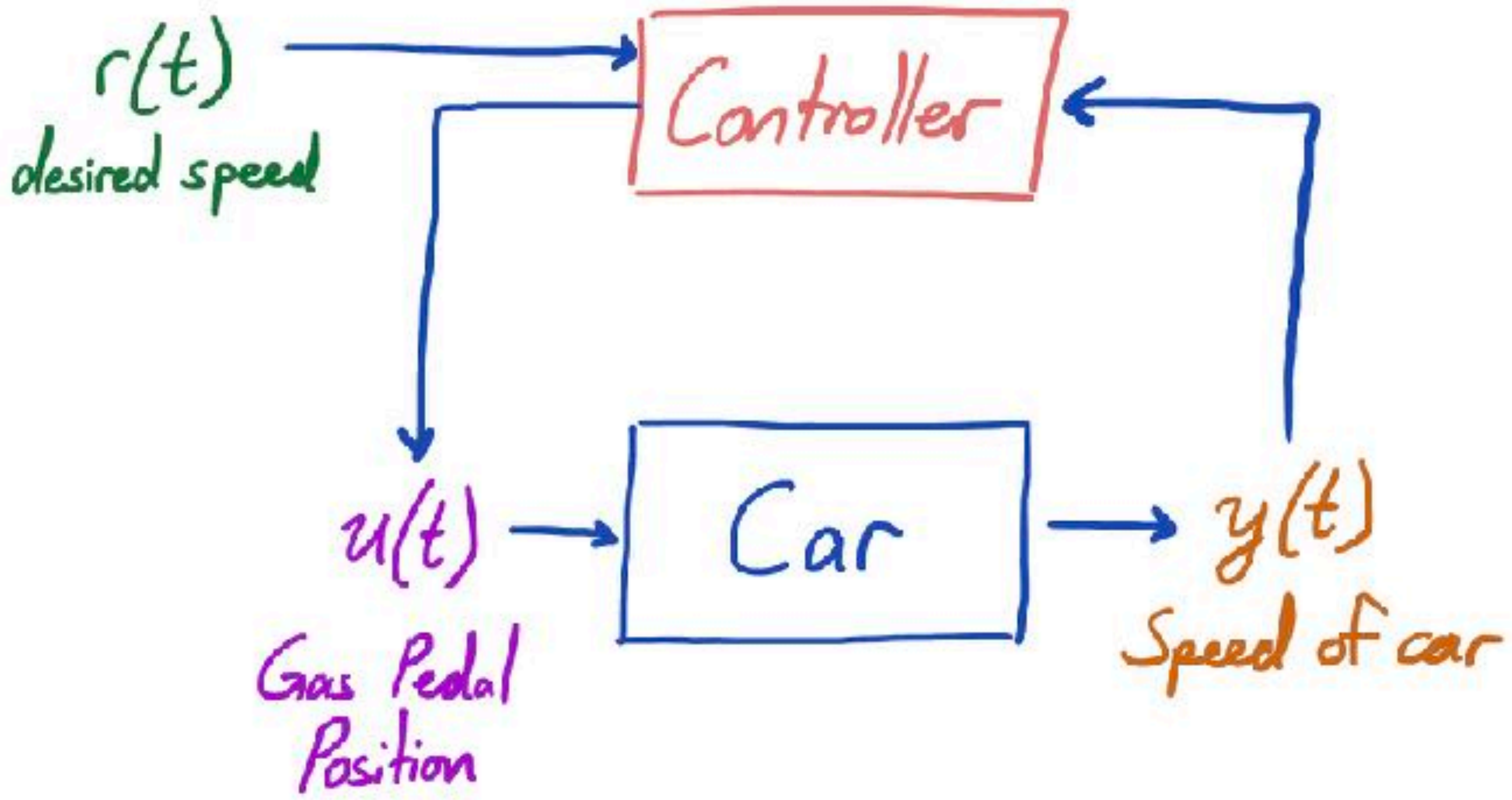
Prompt baking enables efficient continual learning.

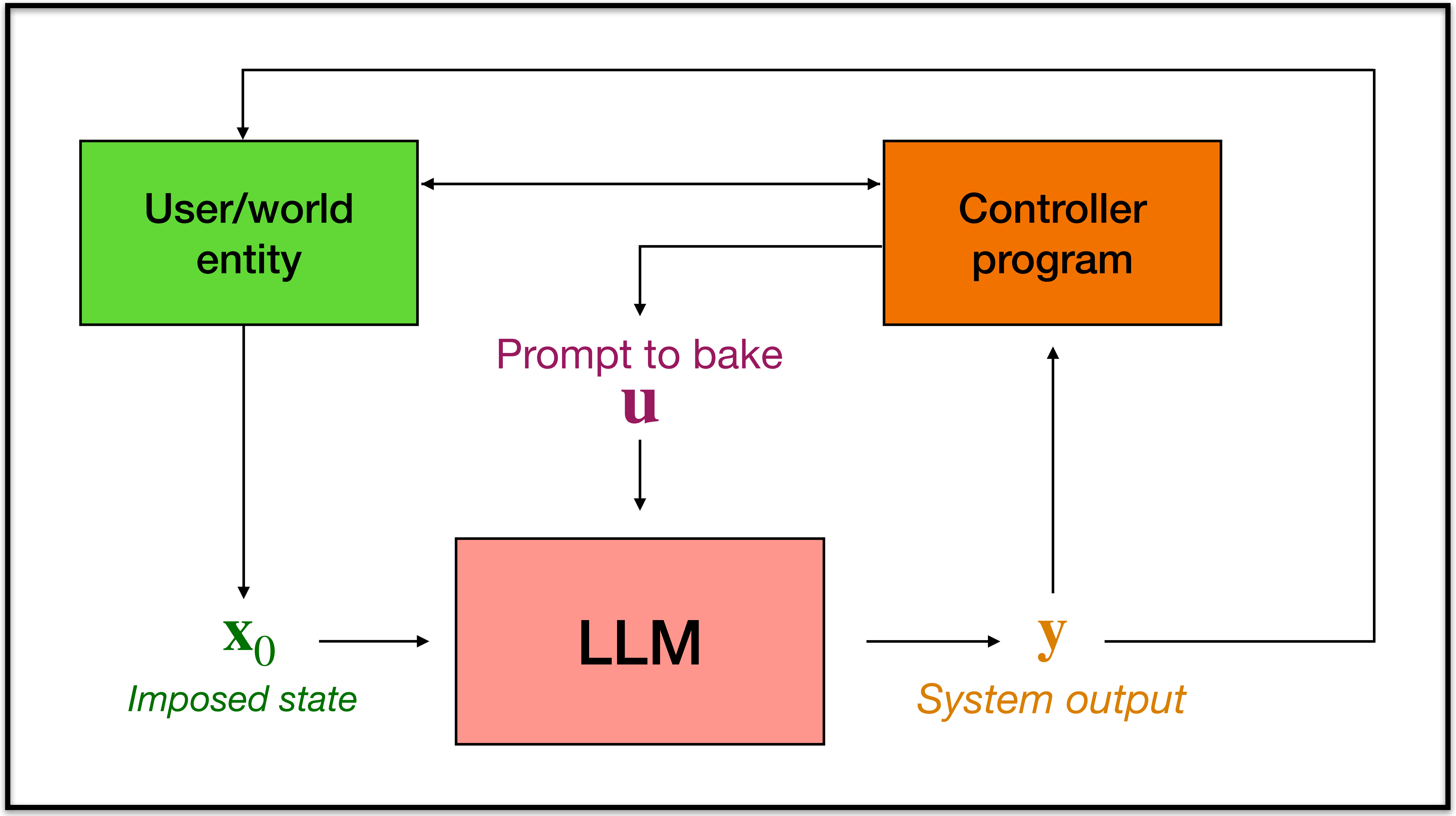


From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697

Prompt baking enables efficient continual learning.

Background • What prompt baking? • Why Prompt Baking? • ***Next?***

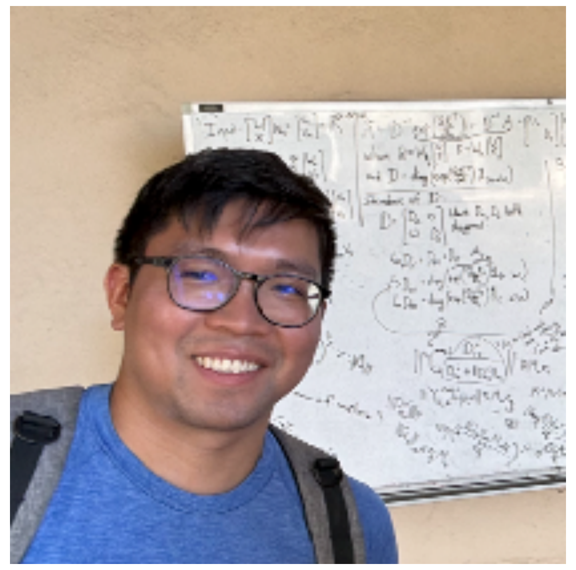





Aman Bhargava


Cameron Witkowski


Alexander Detkov


(Dr.) Shi-Zhuo Looi


(Dr.) (Prof.) Matt Thomson

Prompt Baking

*On **prompt-weight equivalence**, **LLM control**, **weight space geodesics**, and the **nature of learning**.*

Aman Bhargava, Nov 2024 — PhD Student, Thomson Lab, Caltech

From "Prompt Baking" (Bhargava, Witkowski, Detkov, Thomson, 2024) -- https://arxiv.org/abs/2409.13697