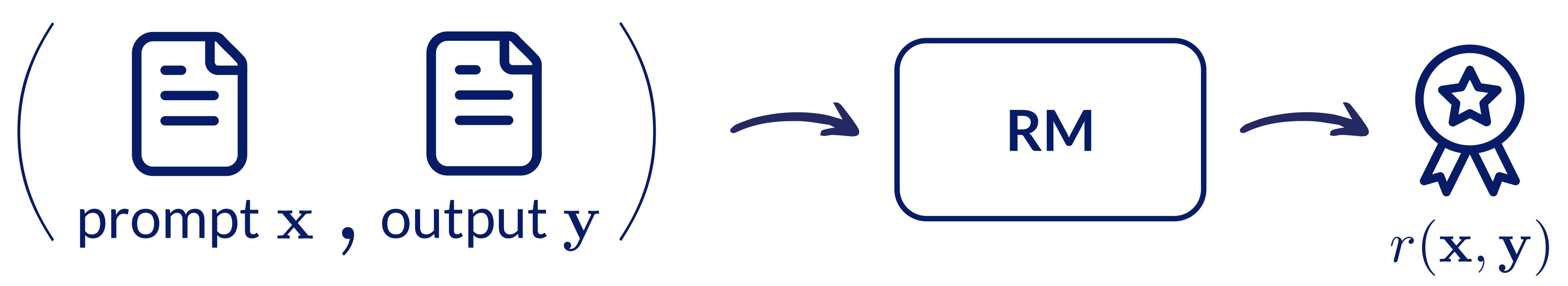
Why is Your Language Model a **Poor Implicit Reward Model?**

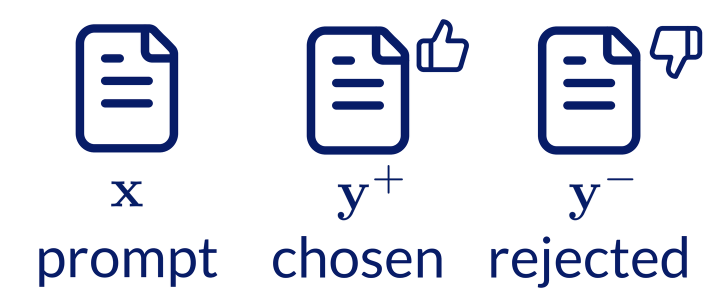
Noam Razin, Yong Lin, Jiarui Yao, Sanjeev Arora

Princeton Language & Intelligence, Princeton University
University of Illinois Urbana-Champaign

Reward Models (RMs) are widely used in language model (LM) post-training and inference



Ability of RMs to evaluate quality of outputs is measured via **accuracy**

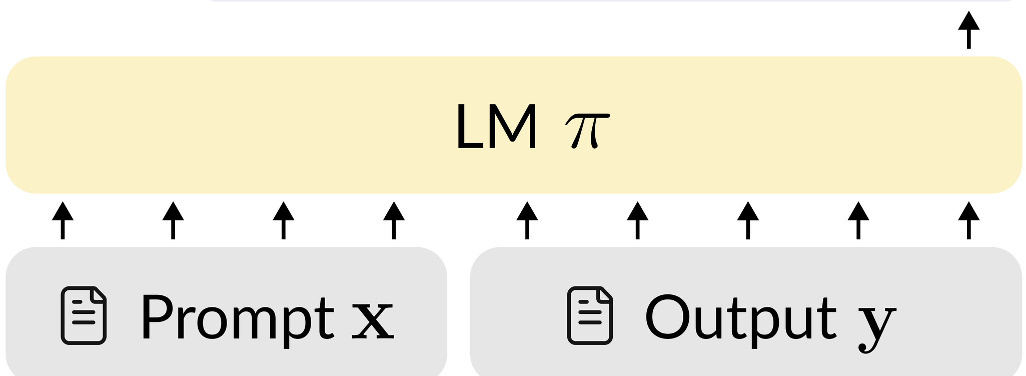


Is $r(\mathbf{x}, \mathbf{y}^+) > r(\mathbf{x}, \mathbf{y}^-)$? +1 Yes / No 0

Explicit Reward Models (EX-RMs) vs Implicit Reward Models (IM-RMs)

EX-RM: Apply a linear head over hidden representation of LM

$$r_{\mathrm{EX}}(\mathbf{x}, \mathbf{y}) = \langle \mathbf{w}, \mathbf{h}_{\mathbf{x},\mathbf{y}} \rangle$$

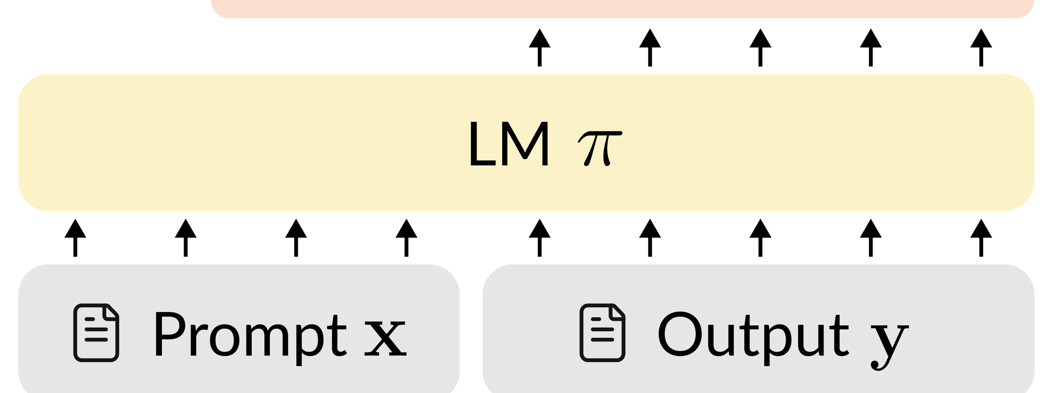


Similarities: Trained using the same data, loss, and LM

Difference: How reward is computed based on the LM

IM-RM: Every LM defines an RM via its log probabilities (Rafailov et al. 2023)

$$r_{\mathrm{IM}}(\mathbf{x}, \mathbf{y}) = \ln \frac{\pi(\mathbf{y}|\mathbf{x})}{\pi_{\mathrm{ref}}(\mathbf{y}|\mathbf{x})}$$



Prior Work: EX-RMs often generalize better than IM-RMs (e.g., Lin et al. 2024, Lambert et al. 2024, Swamy et al. 2025)

Q: Why is there a ***generalization gap*** between **EX-RMs** and **IM-RMs** despite their similarity?

I) Challenge Existing Hypothesis

Do IM-RMs struggle in tasks where generation is harder than verification? X

Existing Hypothesis: If generation is harder than verification, IM-RM should be harder to learn than EX-RM (e.g., Dong et al. 2024, Singhal et al. 2024)

Trained to:	Verify	Generate
EX-RM	✓	✗
IM-RM	✓	✓

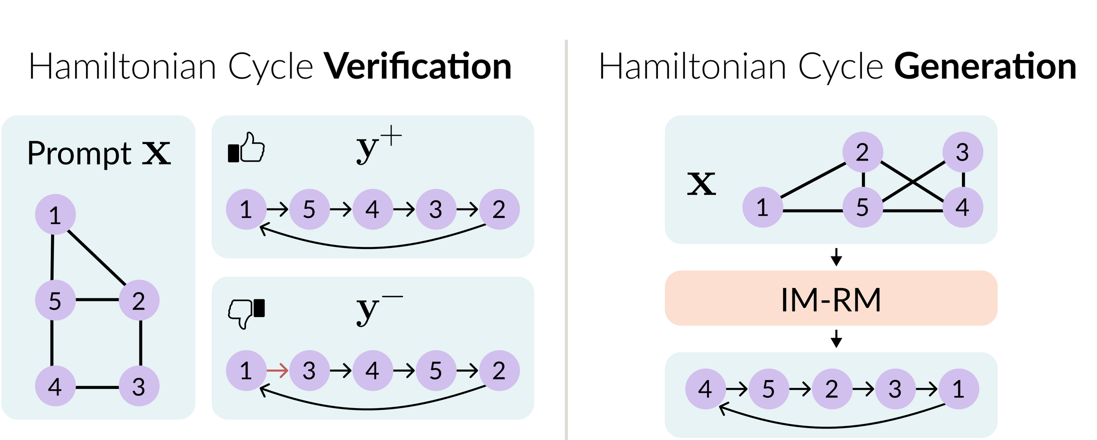
We provide evidence against this hypothesis

Theory

Theorem: IM-RM can learn to $\underbrace{\text{verify}}_{\text{reward of "good" outputs > reward of "bad" outputs + const}}$ without learning to generate

reward of "good" outputs > reward of "bad" outputs + const

Experiments



	EX-RM	IM-RM
Test Accuracy	0.980	0.993
Correct Generations	-	0

LM: Pythia-1B

Result: Despite the generation-verification gap, the IM-RM accurately verifies outputs (without being able to generate)

II) Identify Cause for the Gap

IM-RMs rely more heavily than EX-RMs on superficial token-level cues ✓

Theory: Learning Dynamics

Approach

Characterize how a gradient update on $(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-)$ $\xrightarrow{\text{affects}}$ reward assigned to unseen prompt-output pair $(\bar{\mathbf{x}}, \bar{\mathbf{y}})$

$$\Delta r(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \approx \langle -\nabla loss(\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-), \nabla r(\bar{\mathbf{x}}, \bar{\mathbf{y}}) \rangle$$

Results: EX-RM

$\Delta r_{\mathrm{EX}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ depends on outputs through hidden representations

The reward increases when $\mathbf{h}_{\bar{\mathbf{x}},\bar{\mathbf{y}}}$ is more aligned with $\mathbf{h}_{\mathbf{x},\mathbf{y}^+}$ than with $\mathbf{h}_{\mathbf{x},\mathbf{y}^-}$

Results: IM-RM

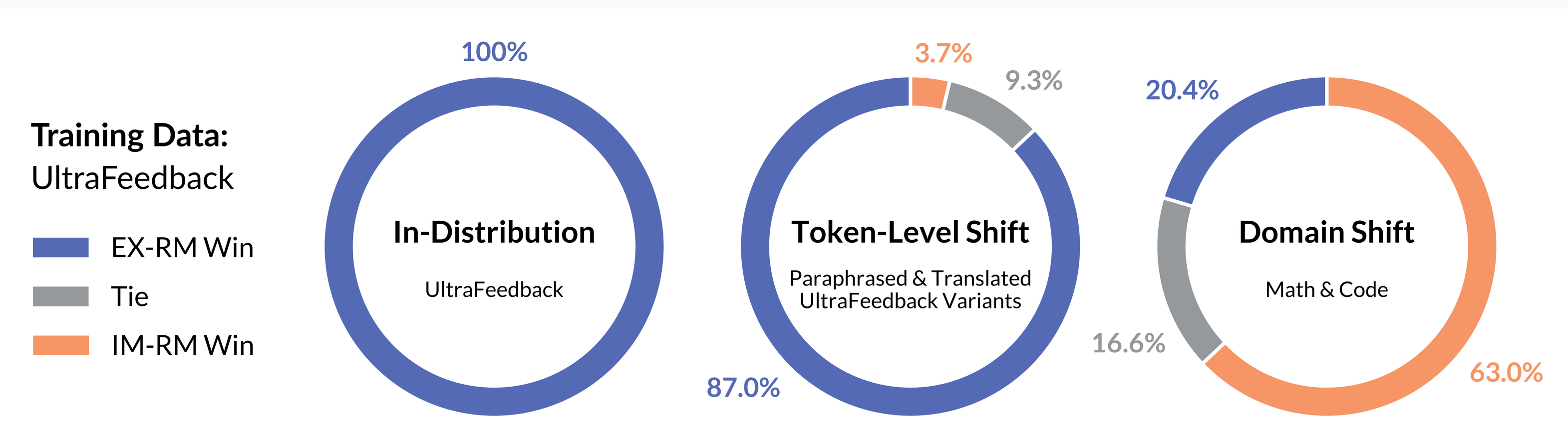
$\Delta r_{\mathrm{IM}}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ depends directly on tokens in the outputs

Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ overlap? Effect similar to EX-RM

Tokens of $\bar{\mathbf{y}}, \mathbf{y}^+$ are distinct? Effect opposite to EX-RM

IM-RM may decrease rewards of outputs semantically similar to $\mathbf{y}^+$ if their tokens have little overlap!

Experiments



Based on LMs of up to 8B size from the Gemma, Qwen, and Llama families

Result: In line with our theory, IM-RMs are less robust to token-level shifts, but perform comparably or better under domain shifts