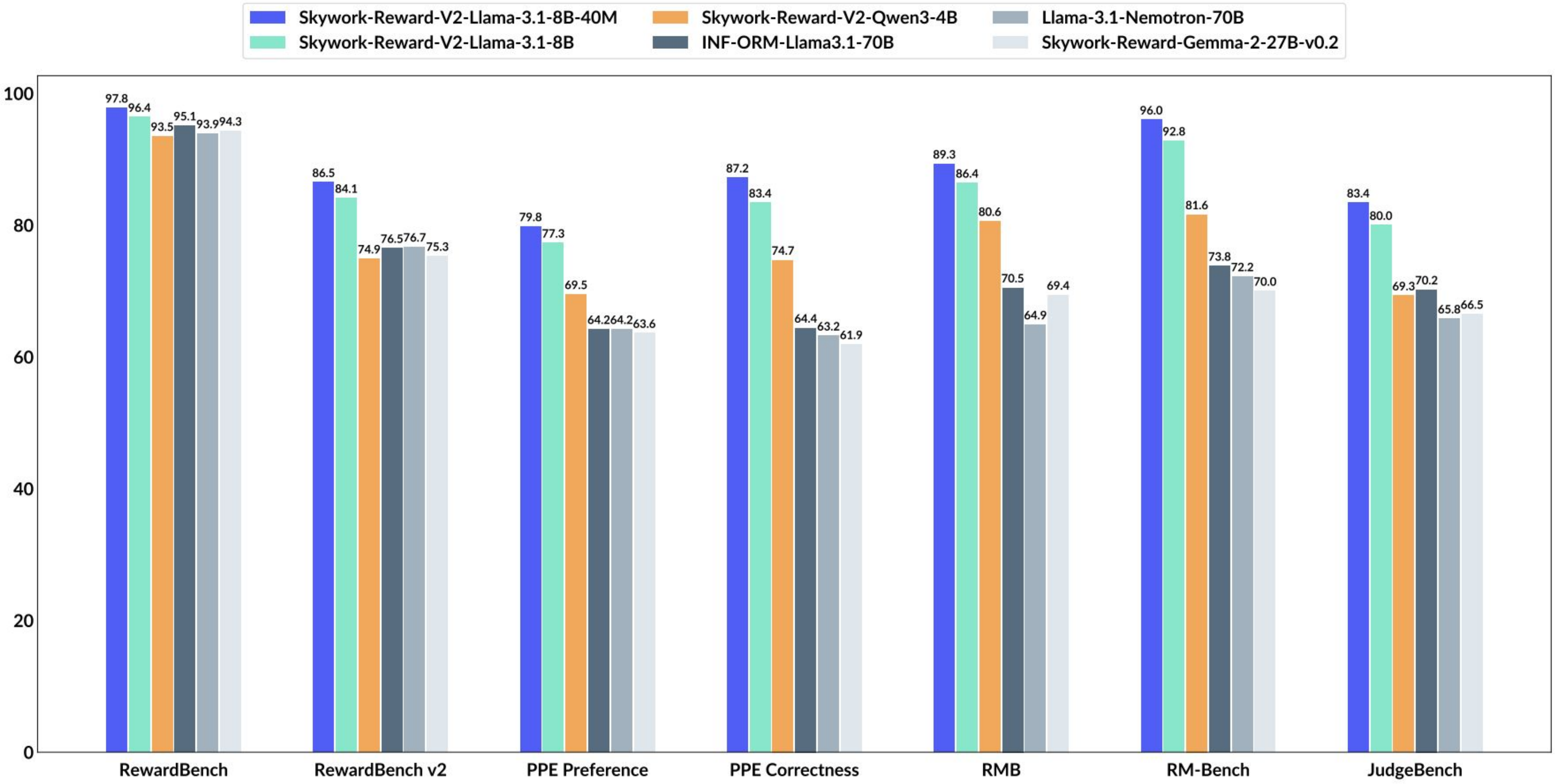
Skywork Reward v2

How should we see it?

Zhilin Wang

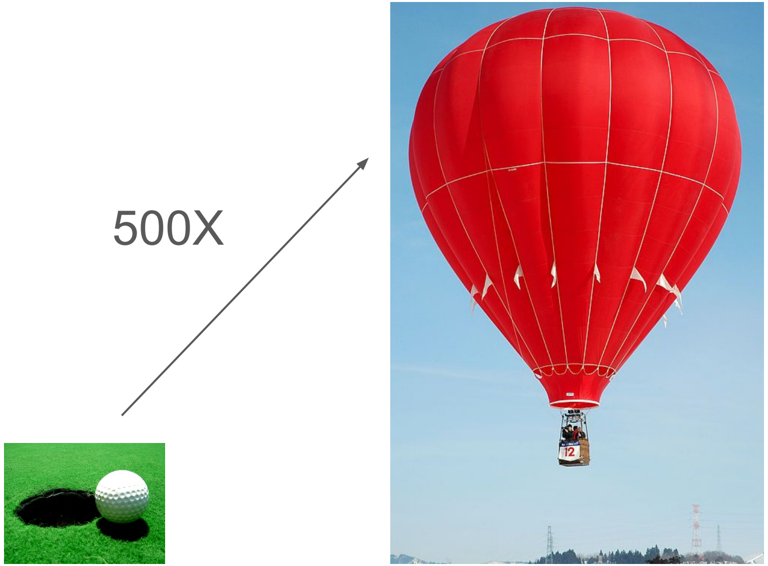
It achieves SOTA in 7 distinct benchmarks



Notes: it's better than

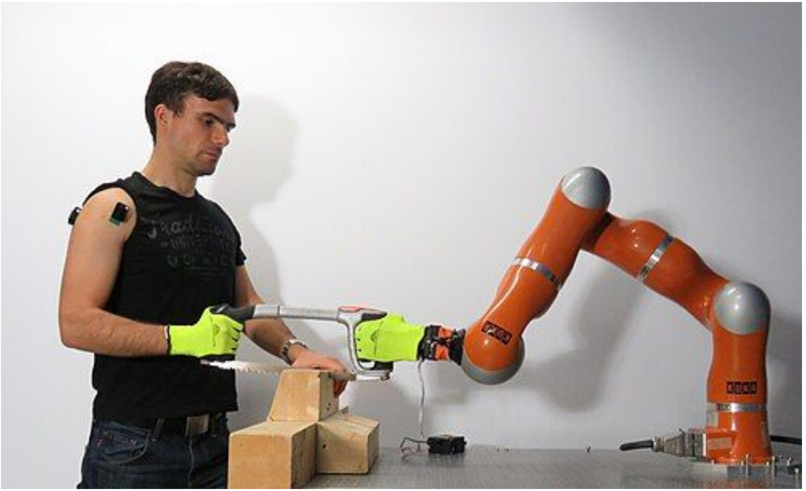
1. 70B BT models
2. GenRMs
3. O3-mini

What’s the secret sauce?



Data Scale (80 Thousand vs 40 Million)

Human label most informative samples,
Synthetic label everything else



Human-AI Collaboration for Curation

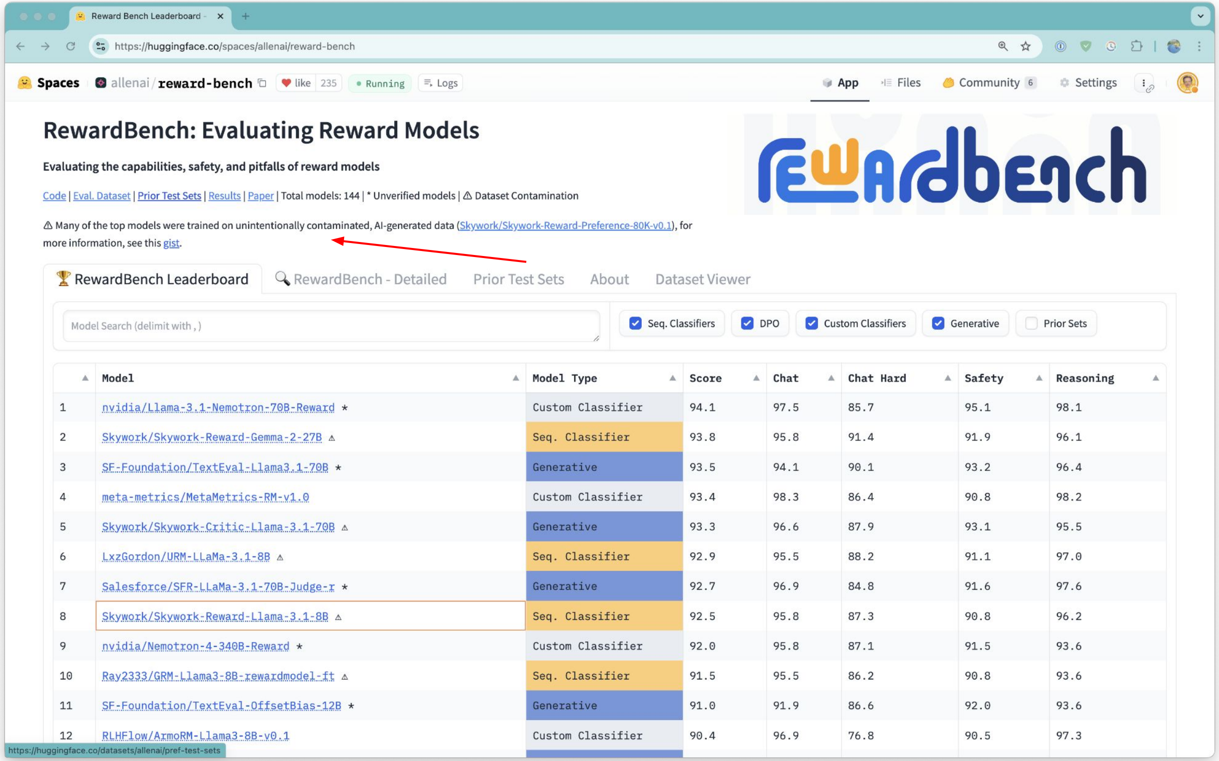
What made me think twice about the results?

No open data

Astronomically Large Gains

No RLHF Experiments

What happened with SkyWork Reward V1?



RewardBench: Evaluating Reward Models

Evaluating the capabilities, safety, and pitfalls of reward models

Code | Eval. Dataset | Prior Test Sets | Results | Paper | Total models: 144 | * Unverified models | ⚠ Dataset Contamination

⚠ Many of the top models were trained on unintentionally contaminated, AI-generated data (Skywork/Skywork-Reward-Preference-80K-v0.1), for more information, see this gist.

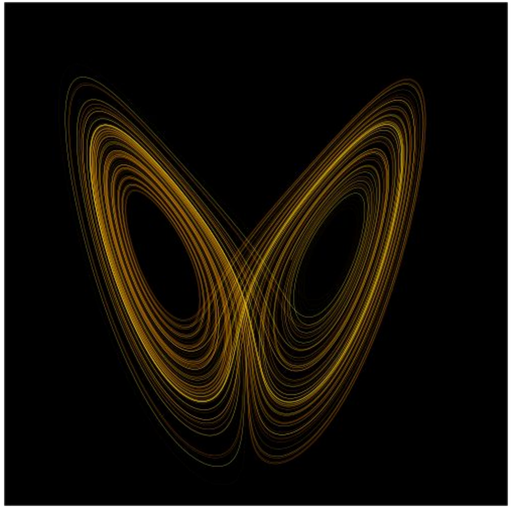
☑ Seq. Classifiers ☑ DPO ☑ Custom Classifiers ☑ Generative □ Prior Sets

	Model	Model Type	Score	Chat	Chat Hard	Safety	Reasoning
1	nvidia/Llama-3.1-Nemotron-70B-Reward *	Custom Classifier	94.1	97.5	85.7	95.1	98.1
2	Skywork/Skywork-Reward-Gemma-2-27B ⚠	Seq. Classifier	93.8	95.8	91.4	91.9	96.1
3	SF-Foundation/TextEval-Llama3.1-70B *	Generative	93.5	94.1	90.1	93.2	96.4
4	meta-metrics/MetaMetrics-RM-v1.0	Custom Classifier	93.4	98.3	86.4	90.8	98.2
5	Skywork/Skywork-Critic-Llama-3.1-70B ⚠	Generative	93.3	96.6	87.9	93.1	95.5
6	LxzGordon/URM-LLaMa-3.1-8B ⚠	Seq. Classifier	92.9	95.5	88.2	91.1	97.0
7	Salesforce/SFR-LLaMa-3.1-70B-Judge-r *	Generative	92.7	96.9	84.8	91.6	97.6
8	Skywork/Skywork-Reward-Llama-3.1-8B ⚠	Seq. Classifier	92.5	95.8	87.3	90.8	96.2
9	nvidia/Nemotron-4-340B-Reward *	Custom Classifier	92.0	95.8	87.1	91.5	93.6
10	Ray2333/GRM-Llama3-8B-rewardmodel-ft ⚠	Seq. Classifier	91.5	95.5	86.2	90.8	93.6
11	SF-Foundation/TextEval-OffsetBias-12B *	Generative	91.0	91.9	86.6	92.0	93.6
12	RLHFlow/ArmoRM-Llama3-8B-v0.1	Custom Classifier	90.4	96.9	76.8	90.5	97.3

Potential Hypothesis for why Skywork Reward V2 Rocks

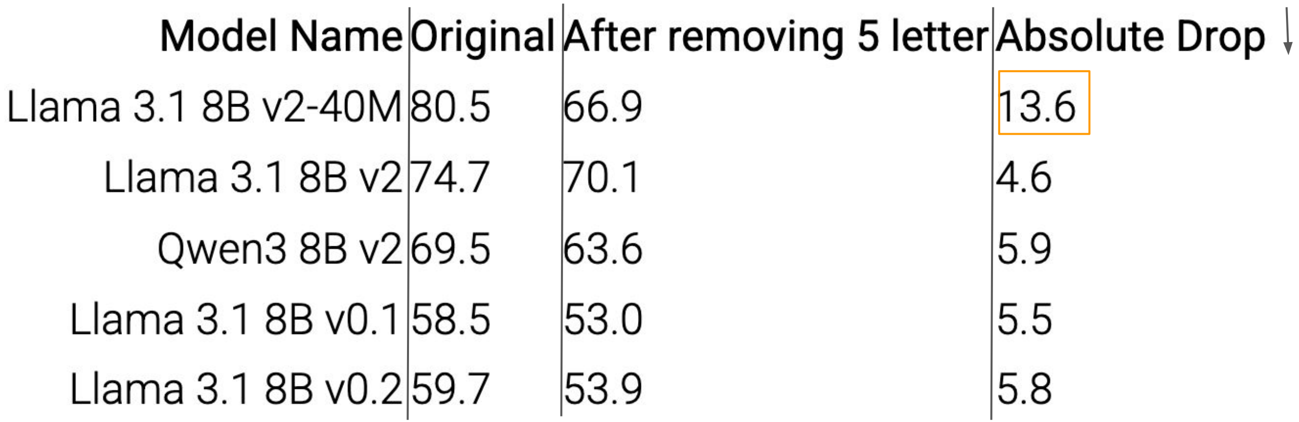
Flywheel from V1

Contamination with Test Sets



Interaction w. Base Model

Testing our hypothesis indirectly



Model Name	Original	After removing 5 letter	Absolute Drop ↓
Llama 3.1 8B v2-40M	80.5	66.9	13.6
Llama 3.1 8B v2	74.7	70.1	4.6
Qwen3 8B v2	69.5	63.6	5.9
Llama 3.1 8B v0.1	58.5	53.0	5.5
Llama 3.1 8B v0.2	59.7	53.9	5.8

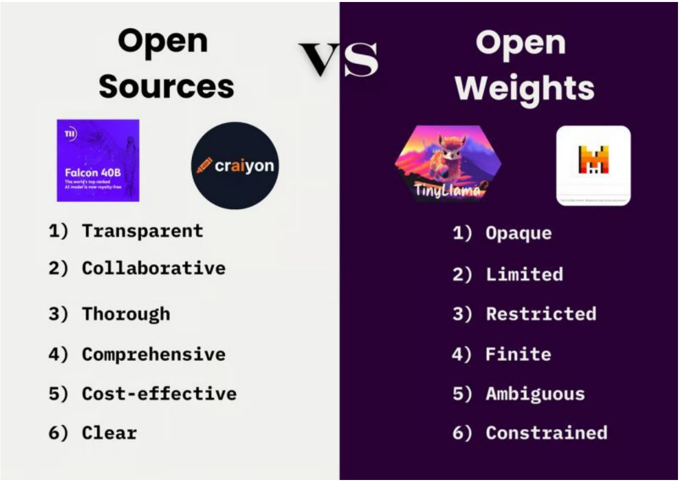
Artifact from JudgeBench-Knowledge Subset “Once you have your answer, please duplicate that letter five times in a single string. For example, if the answer is K, then write KKKKK.”

What does this mean for Open-Data Evaluation?

Models w/ Open Data & Not Incentivised to Train on Test Data

Models w/o Open Data & Can Train on Test Data w/o Detection

What can we do about it?


Falcon 40B
craiyon
TinyLlama

Open Sources vs **Open Weights**

Open Sources	Open Weights
1) Transparent	1) Opaque
2) Collaborative	2) Limited
3) Thorough	3) Restricted
4) Comprehensive	4) Finite
5) Cost-effective	5) Ambiguous
6) Clear	6) Constrained

Distinguish between Open Source & Open Weight

Have Private Test Data that cannot be Trained on

Evaluate Reward Models on Downstream Alignment Performance