Predicting Human-Gain Curves under Local Proxy Optimization from Repeated Ratings

Moonwon Choi · Kisung Nam · Seunggeun Lee

SEOUL NATIONAL UNIVERSITY

Massachusetts General Hospital

HARVARD MEDICAL SCHOOL

Why proxy optimization needs a diagnostic

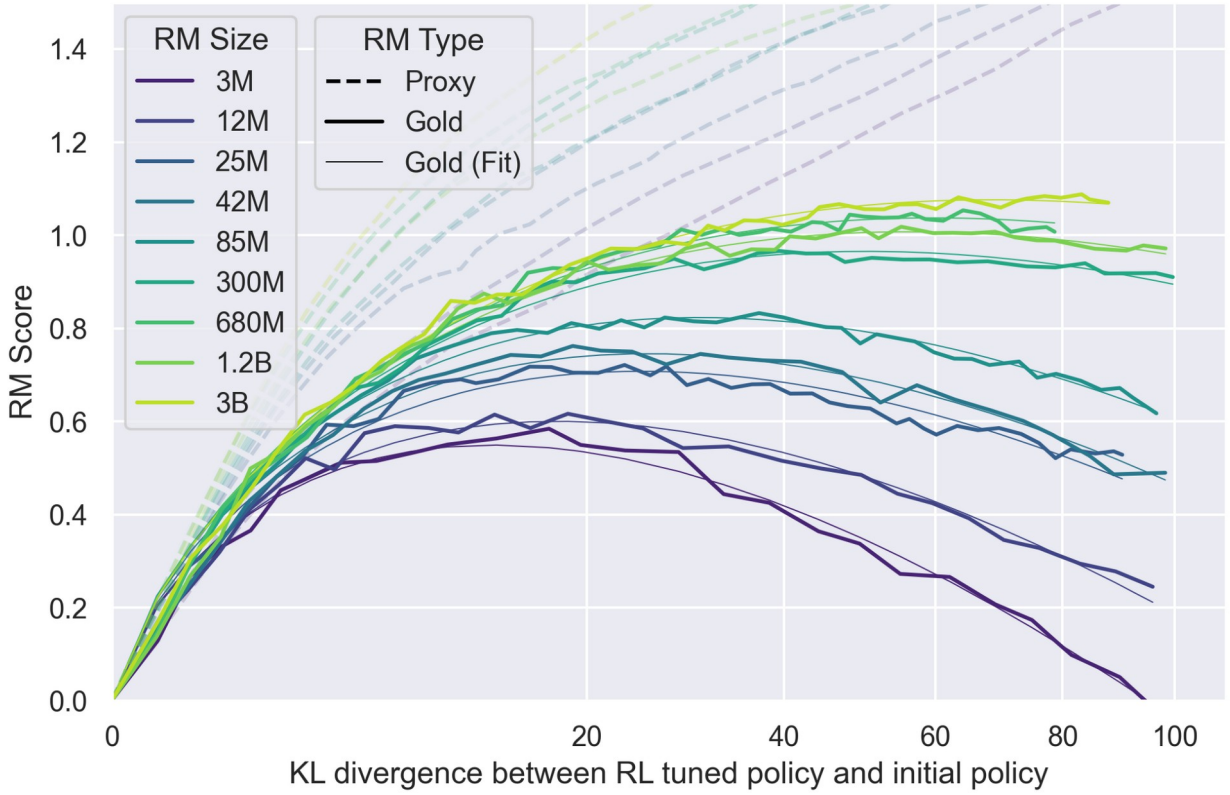
Goodhart's law

A useful measure can become unreliable when it becomes an optimization target.

Gao et al. (ICML 2023)

Proxy reward keeps rising; gold-model reward can fall.

Gold = a separate reward model, not direct human evaluation.



Can we forecast local human gains before optimization?

Gao, Schulman & Hilton (2023), Fig. 2(b), PMLR 202. Panel cropped; resized.

CC BY 4.0

PACE: alignment + curvature

$$\widehat{\Delta}(d) = \widehat{A}\sqrt{d} + \widehat{B}d$$

$$\pi_t(i) \propto \pi_0(i)\, e^{t r_i}, \qquad d = \mathrm{KL}(\pi_t \| \pi_0)$$

d: KL radius
t ≥ 0: positive tilt

Predict the change in mean human quality under the tilted weights.

A: proxy-human alignment

$$A = \sqrt{2}\, \mathbb{E}[\bar{J} Z]$$

B: leading curvature correction

$$B = \mathbb{E}[\bar{J} Z^2] - \tfrac{2}{3}\mathbb{E}[\bar{J} Z]\mathbb{E}[Z^3]$$

$$Z = (r - \mathbb{E}r)/\sigma_r, \quad \bar{J} = J - \mathbb{E}J$$

r: proxy score; J: human target
Moments use reference weights π_0.

Derivation: Taylor-expand gain and KL in tilt strength, then invert the KL expansion.

Build a PACE profile from repeated ratings

Benchmark items

Proxy scores r
Repeated human ratings

Human target J:
average repeated ratings

→

Estimate A and B

Score moments
Alignment + curvature

→

Local gain profile

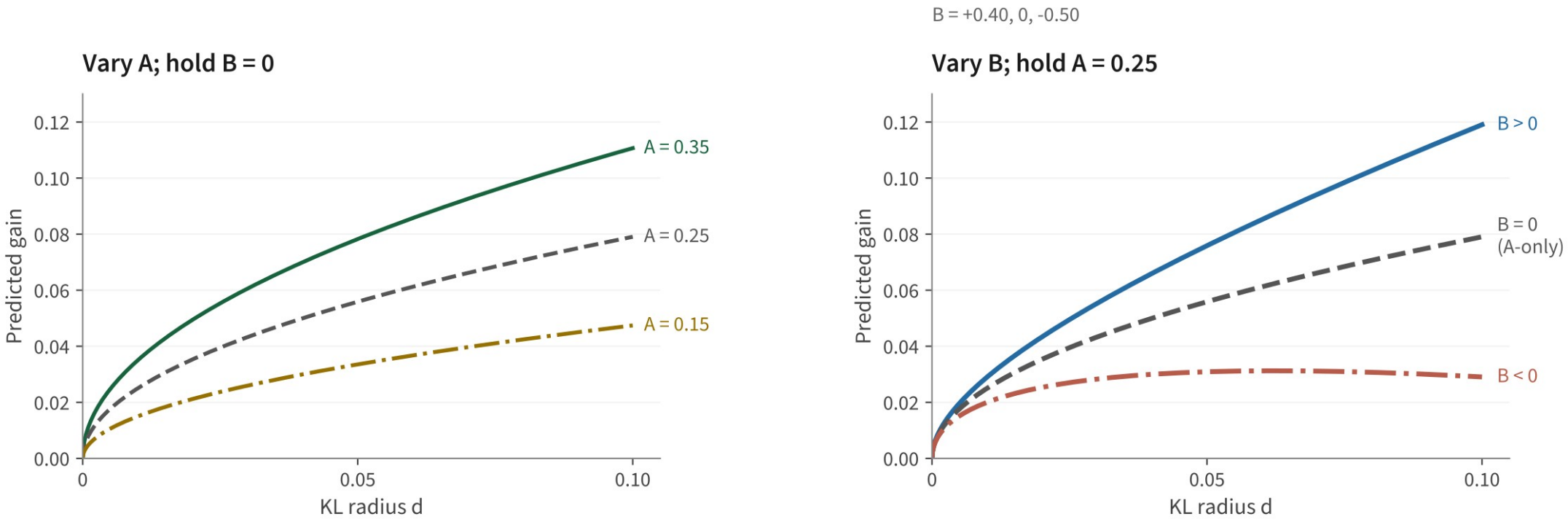
Curve
Uncertainty
Checked KL range

Use all rated items to describe a benchmark.

Our validation is separate: 30% calibration / 70% held-out items, repeated over 50 splits to test prediction on new items.

Reading a PACE profile: what A and B tell us

A controls the initial gain scale. B changes the departure from the A-only curve.

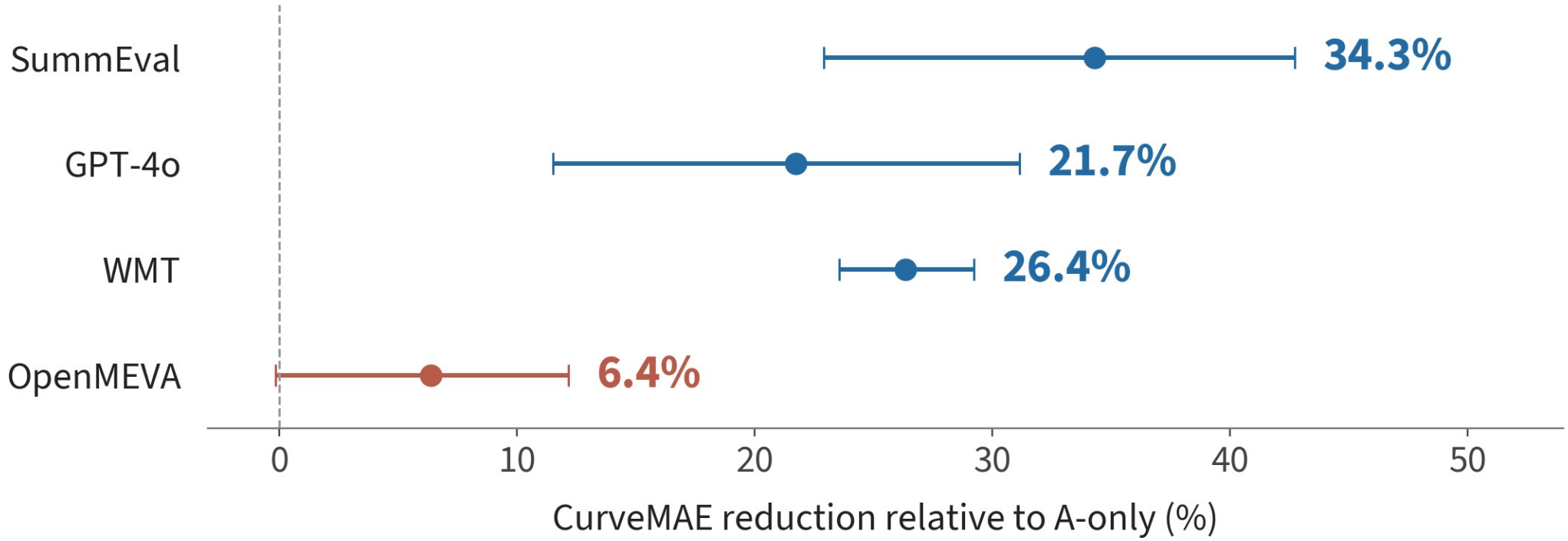


Same axis scales. Shared dashed reference: A = 0.25, B = 0.

Illustrative approximation: a predicted turning point need not be a true downturn.

Adding B reduces average prediction error

Held-out gain curves under raw-score exponential tilt



30/70 split · 50 repeats · five radii through d = 0.10

Paired 95% bootstrap intervals. OpenMEVA includes zero.
GPT-4o has four paths; its interval is descriptive.

A bridge to actual optimization

LoRA-DPO on UltraFeedback · 3 training seeds

Qwen2.5-3B-Instruct · Skywork preferences · frozen DeBERTa target

Mean prediction MAE

A-only		PACE (A+B)
0.01905	→	**0.01721**

Fixed-pool likelihood-ratio gain, not fresh human-rated generations.

32 candidates per prompt · mean-KL gate ≤ 0.10
Average checkpoints within each seed, then average across seeds.

From correlation to a local gain profile

Repeated ratings → A, B → a local human-gain forecast

Report profiles alongside benchmarks, reward models and LLM judges.

Specify the proxy, human target, dataset, uncertainty and checked KL range.

Local diagnostic, not a guarantee of downstream policy-training gains.

Paper **Code**

arXiv: 2610.41114

Thank you.