NEURAL INFORMATION PROCESSING SYSTEMS

Test-Time Personalization

A Diagnostic Framework and Probabilistic Fix for Scaling Failures

Linhai Zhang[1] **Yulan He**[1,2]

[1]King's College London [2]The Alan Turing Institute

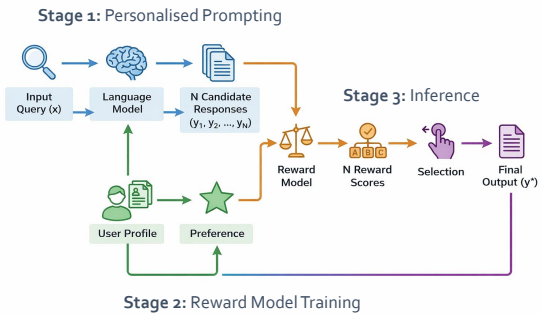
KCL NATURAL LANGUAGE PROCESSING

KING'S College LONDON

The Alan Turing Institute

Personalization today is single-shot

- Existing work builds better personalized **models or inputs**: prompting, adaptation, alignment.
- Inference is still **one sample, one answer**.
- Test-time scaling transformed reasoning. For personalization, **scaling the policy's outputs is unexplored.**



> **Question.** *Can we improve personalization by scaling test-time computation for a weak personalized model?*

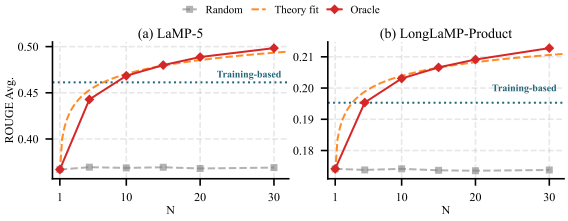
The promise: oracle selection scales as $\sqrt{\ln N}$

Sample N candidates from a personalized policy, keep the best one.

Theorem (Oracle Scaling Law)

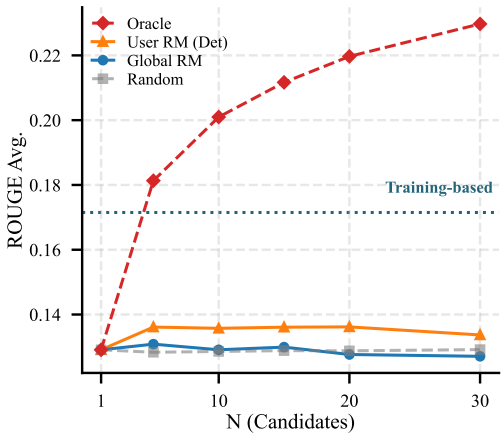
$$\bar{U}_{\text{oracle}}(N) \;\leq\; \bar{\mu} + \bar{\sigma}\, c\sqrt{\ln N}$$

- Even a weak policy improves with more samples.
- The oracle **beats per-user training at $N \approx 5$–10.**



Oracle (red) follows the predicted curve (orange).

The problem: standard reward models do not scale



Oracle keeps climbing as *N* grows.

Global RM and **per-user RM** stay near **random selection**.

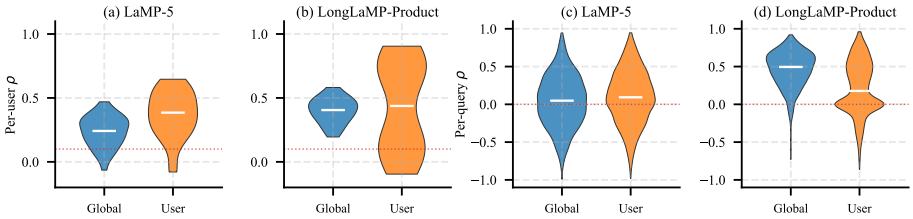
The gain is bottlenecked by the reward model.
What prevents it from closing the gap?

Diagnosis: two failure modes

Correlation between learned and true reward, per user (ρ_u) and per query (ρ_q):

User-level collapse: $\rho_u < 0.1$
Near-constant predictions for a user.
User RM: 25% of Product-Review users.

Query-level hacking: $\rho_q < 0$
The RM prefers *worse* candidates.
Global RM: 46% of LaMP-5 queries.



A unified scaling law

$$\bar{U}(N) \approx \bar{\mu} + \underbrace{(1-\alpha)}_{\text{non-collapsed users}} \cdot \underbrace{\left[(1-\beta)\bar{\rho}_+ - \beta|\bar{\rho}_-|\right]}_{\text{effective correlation}} \cdot \bar{\sigma}\, c\sqrt{\ln N}$$

Four measurable quantities

- α: collapse rate
- β: reward hacking rate
- $\bar{\rho}_+$, $\bar{\rho}_-$: mean correlation on non-hacked / hacked queries

What it gives us

- Forecasts any reward model's Best-of-N curve, with **no per-method fitting**.
- A recipe: high $\bar{\rho}_+$, low α, low β, all at once.

The fix: a probabilistic reward model

Predict a mean *and* a variance for each input; train with Gaussian NLL.

$$\mathcal{L}_{\mathrm{NLL}} = \tfrac{1}{2}\log\sigma^2(x) + \frac{(y-\mu(x))^2}{2\sigma^2(x)} \qquad \Longrightarrow \qquad \frac{\partial\mathcal{L}}{\partial\mu} = \frac{\mu - y}{\sigma^2}$$

The learned variance σ^2 acts as an **adaptive per-input weight**.

Gradient buffering

Weak-signal users get a large σ^2, which damps the pull toward a constant.

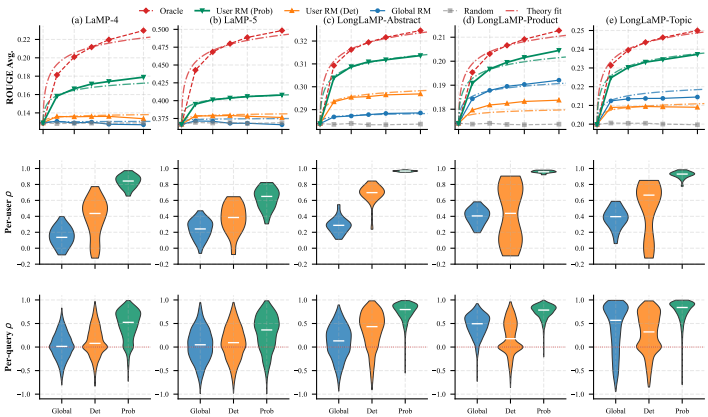
Collapse rate α: $0.15 \to 0$

Implicit regularization

Inconsistent examples get a large σ^2 and are down-weighted in training.

Hacking rate β: $0.25 \to 0.06$

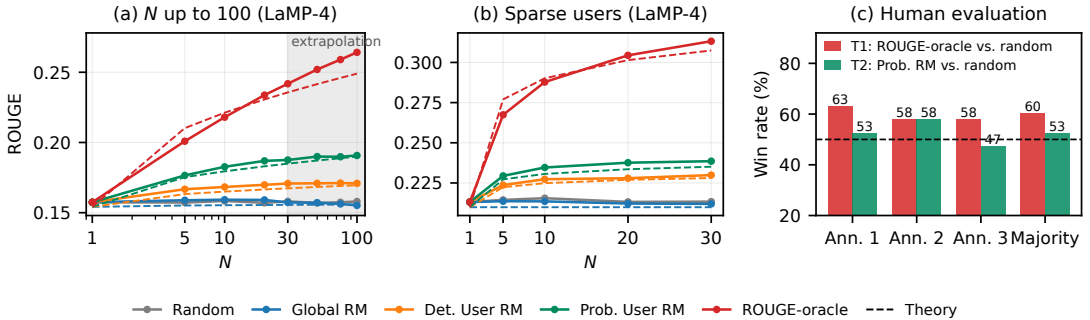
Results on five tasks



Probabilistic RM is the only selector that scales monotonically.

- Closes **51–84%** of the oracle gap at $N{=}30$.
- The law predicts its curve within **1.7%**.
- Same ranking across all **15** task–policy pairs.

Beyond the main setting



(a) N up to 100
The law, fit at $N \leq 30$, still holds (error $\leq 2.1\%$).

(b) Sparse users
Still $+11.8\%$ over random; no spike in collapse.

(c) Human evaluation
Small blind study: every annotator prefers the oracle pick.

Takeaways

- Scaling policy outputs is a **new, effective axis** for personalization.
- The reward model is the bottleneck. Our scaling law explains **why**, from four measurable numbers.
- A learned variance removes both failure modes and makes **scaling reliable**.

Paper

arXiv:2605.10991

Code

linhai0012/TTP-release

linhai.zhang@kcl.ac.uk