Hybrid Preferences: Learning to Route Instances for Human vs. AI Feedback

L.J.V. Miranda, Y. Wang, Y. Elazar, S. Kumar, V. Pyatkin, F. Brahman, N. A. Smith, H. Hajishirzi, P. Dasigi

Motivation

Human annotations
- High-quality
- Expensive
- Time-consuming

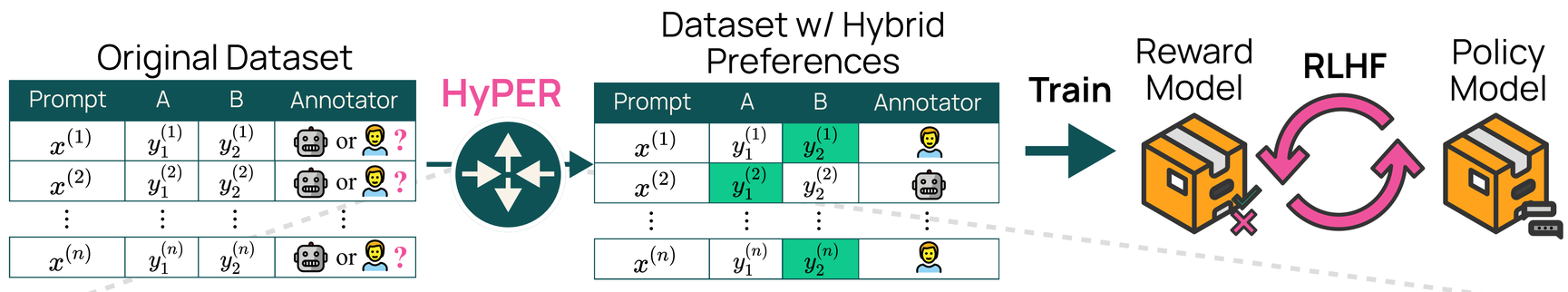
Synthetic annotations
- Prone to bias
- Cheap
- Scalable

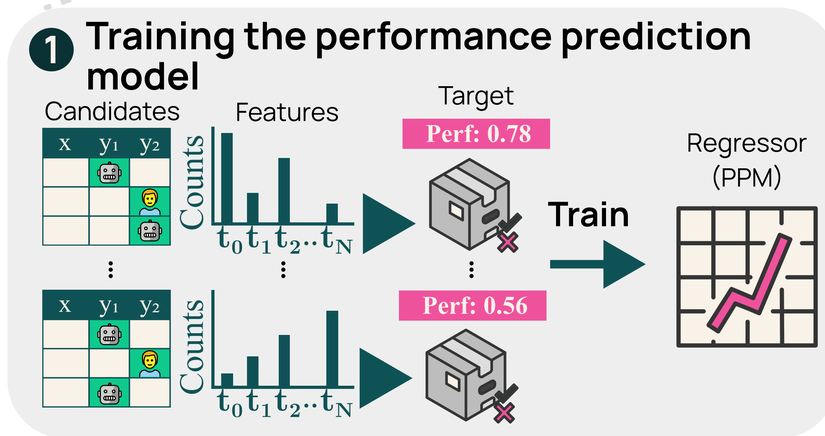
Find the right combination of human and synthetic annotations

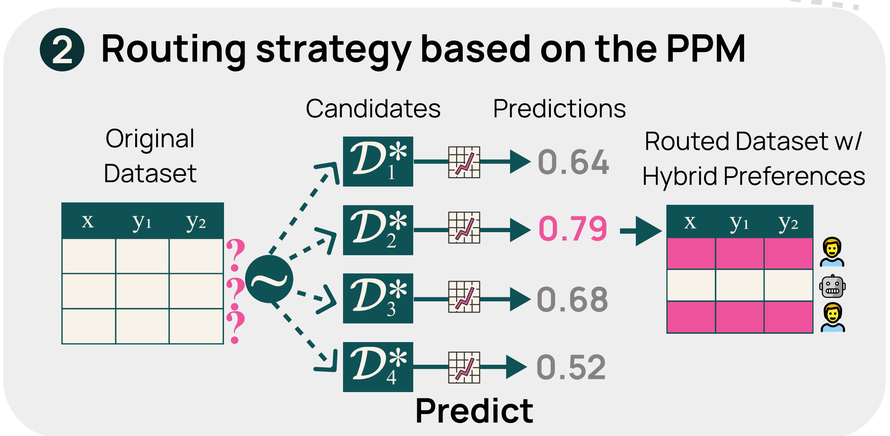
...in order to obtain **high-quality** and **cost-efficient** preference data

Methodology

Train a model to **predict the performance of a reward model** trained on a given preference dataset and used that for routing.



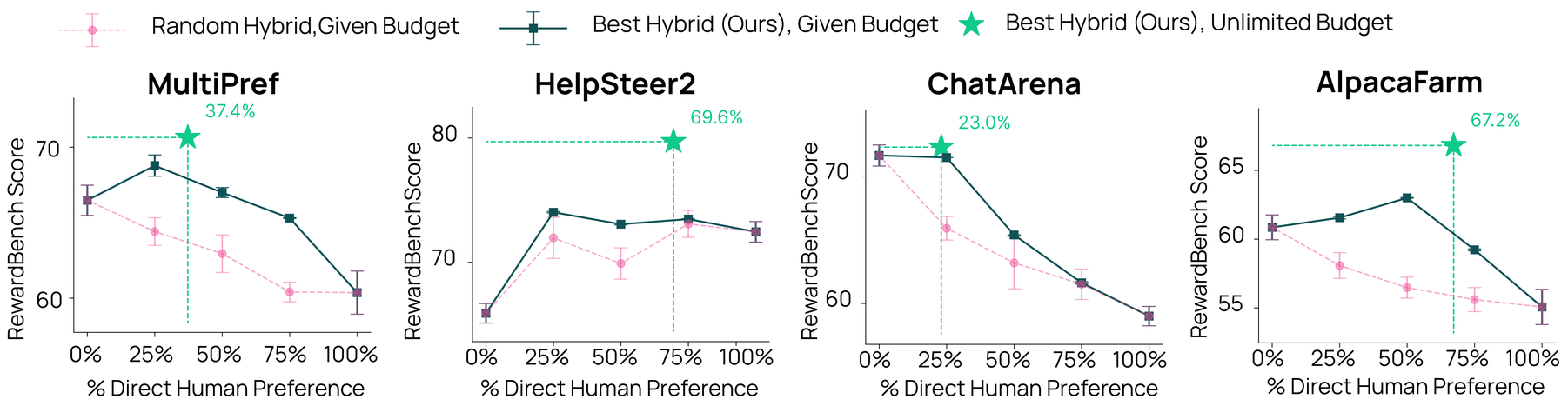




Results

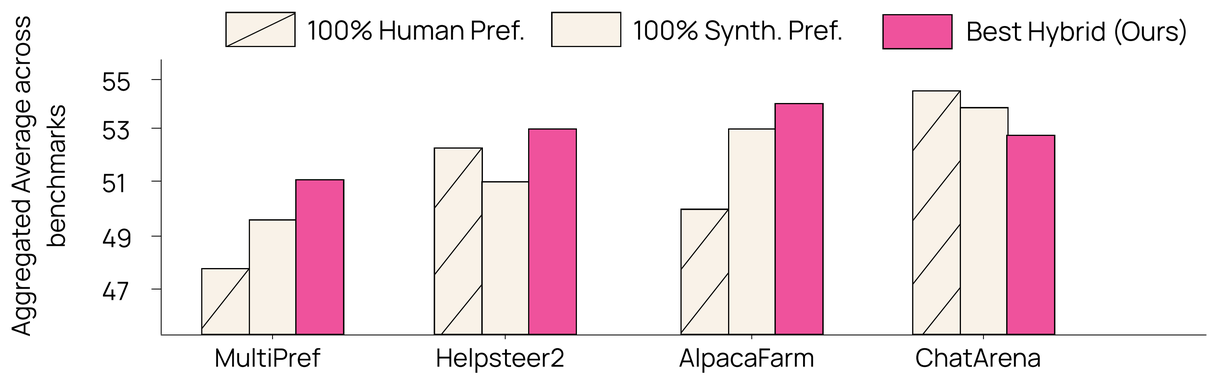
Reward models trained on our routed datasets perform better vs. random / 100% human / 100% LLM on **unseen datasets**, **benchmarks**, and **base models**.

Unseen Preference Datasets

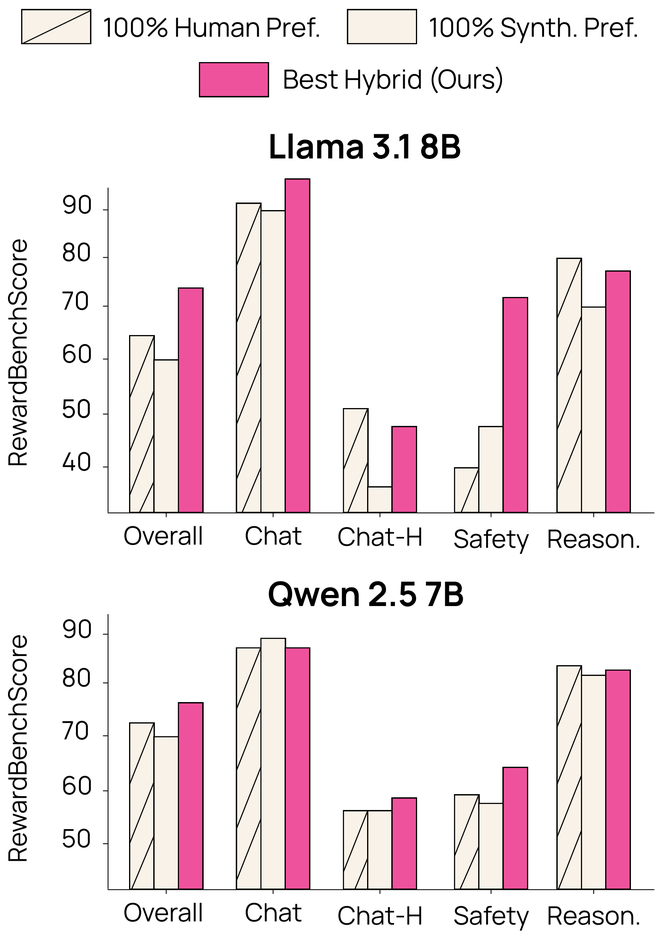


Other Benchmarks

Best-of-N Evaluation Performance of GSM8K, BBH, IFEval, Codex, and AlpacaEval

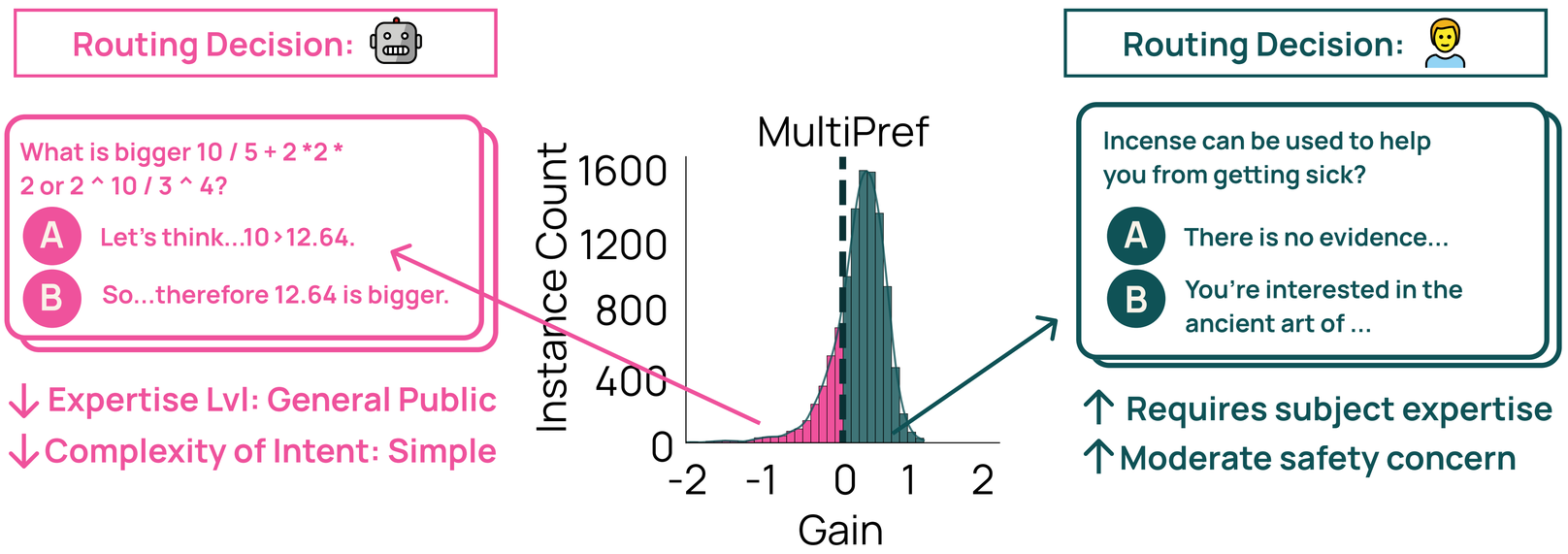


Other Base Models



Analysis

By looking at the features learned by our PPM, we can understand the characteristics of instances suited for human annotation



Contributions

- We find that some preference instances are **better suited to be annotated by humans** than language models.
- We used this to build a **routing framework** for preference data.
- We obtain **fine-grained understanding** of what type of instances benefit human annotations.

Scan to download the full paper!
{ljm, yizhongw}@allenai.org
allenai/hybrid-preferences

ACL 2025 VIENNA