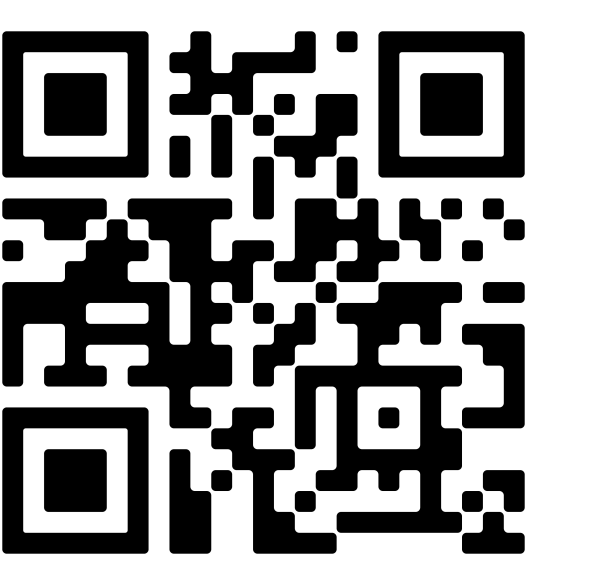


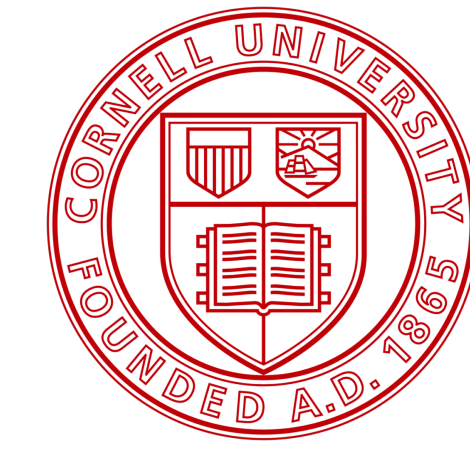
Continually Improving Extractive QA via Human Feedback



Code & Data

Ge Gao^{*1}, Hung-Ting Chen^{*2}, Yoav Artzi¹, Eunsol Choi²

¹Cornell University ²University of Texas at Austin

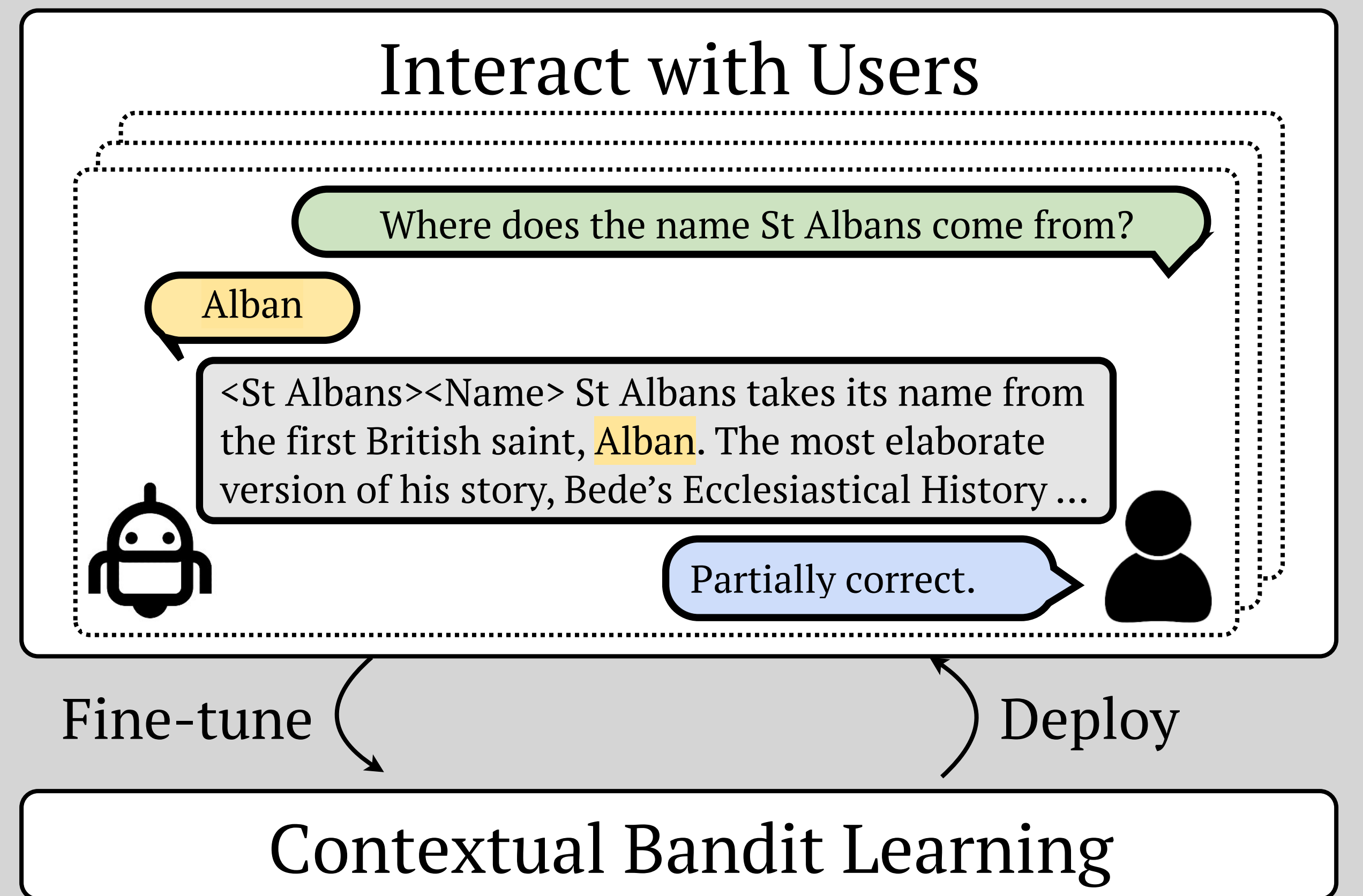


TEXAS
The University of Texas at Austin

How to improve NLP systems by learning from user feedback?

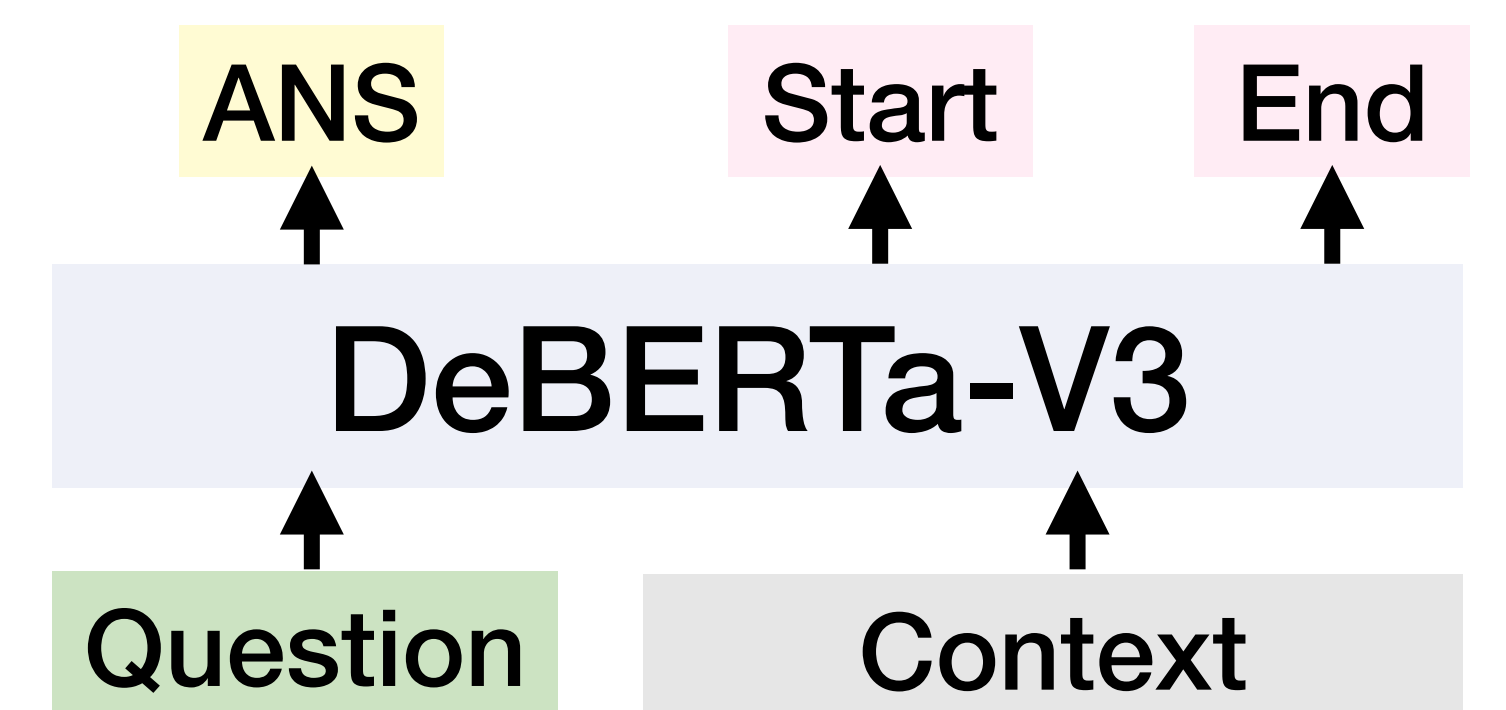
We present a user feedback study through **bandit learning** on **extractive QA task**

- 200 examples per round, 9 rounds
- Information-seeking MTurk workers pose questions and give feedback
- Topics/contexts from Wikipedia



Approach

- Model classifies if question is answerable, and answer span (if answerable)
- We heuristically map user feedback into two reward values (r_1, r_2).
- After each round of deployment, update the model with policy gradient

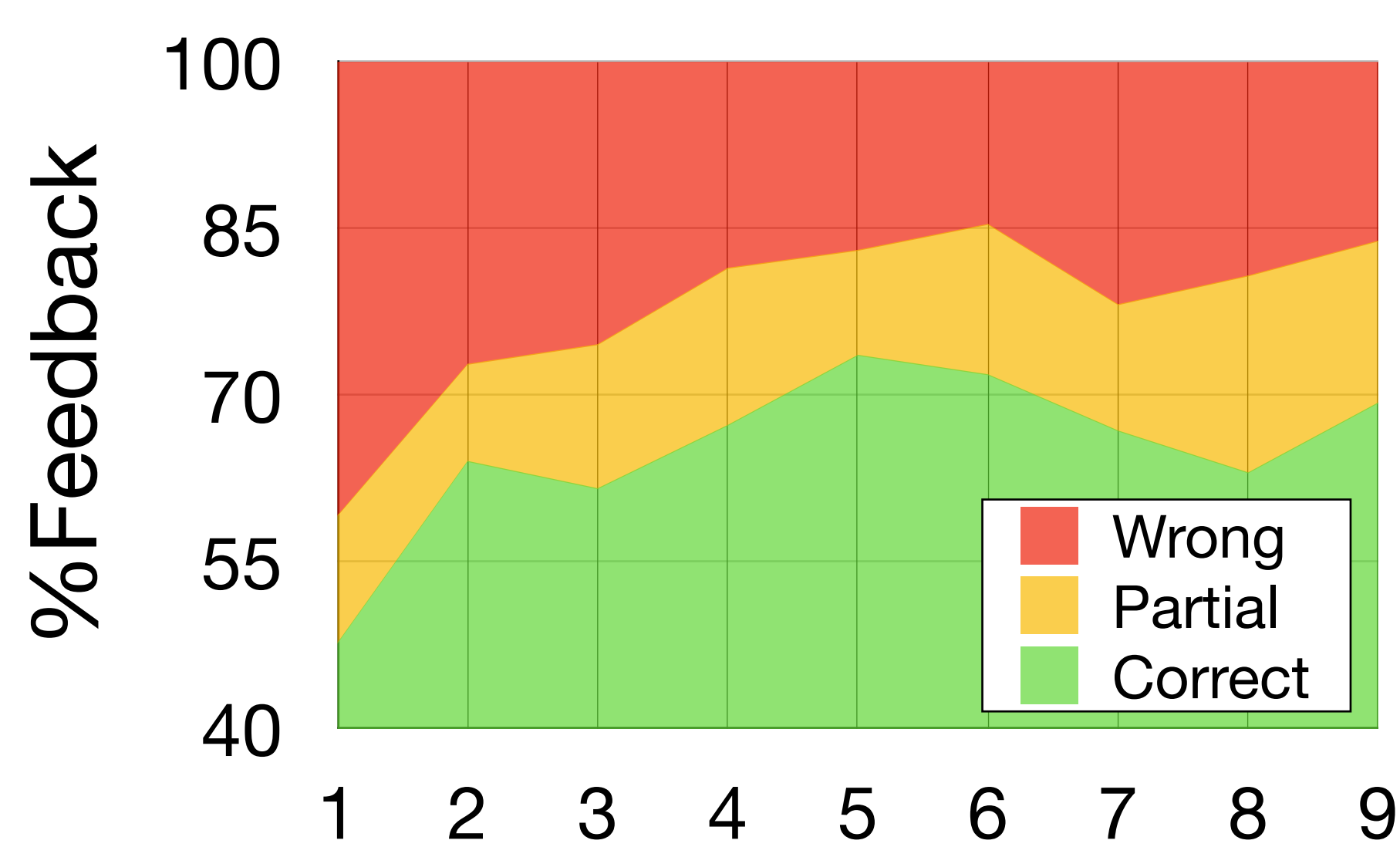


$$\alpha_1 r_1 \nabla_{\theta} \log \pi_{\theta}(\hat{u} | \bar{q}, \bar{c}) + \alpha_2 r_2 \nabla_{\theta} \log \pi_{\theta}(\hat{y} | \bar{q}, \bar{c}) + \gamma \nabla_{\theta} H(P_u(\cdot | \bar{q}, \bar{c}))$$

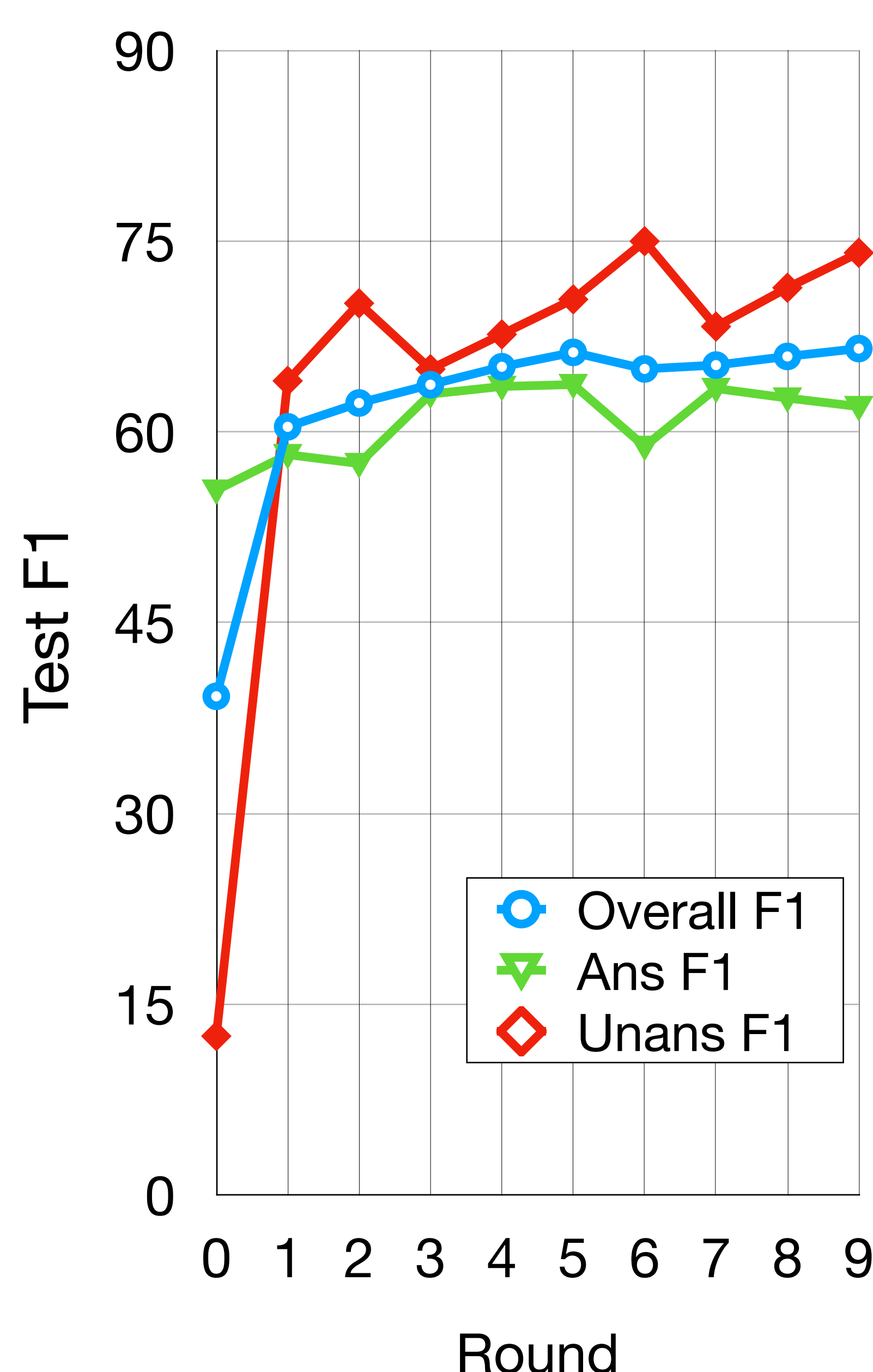
Inverse Propensity Score (debiasing reward) $\frac{\pi_{\theta}(\cdot | \bar{q}, \bar{c})}{\pi_{\theta'}(\cdot | \bar{q}, \bar{c})}$
 QA model
 Question
 Context
 Answerability reward
 Model-predicted answerability
 Span reward
 Model-predicted span
 Entropy penalty

Results

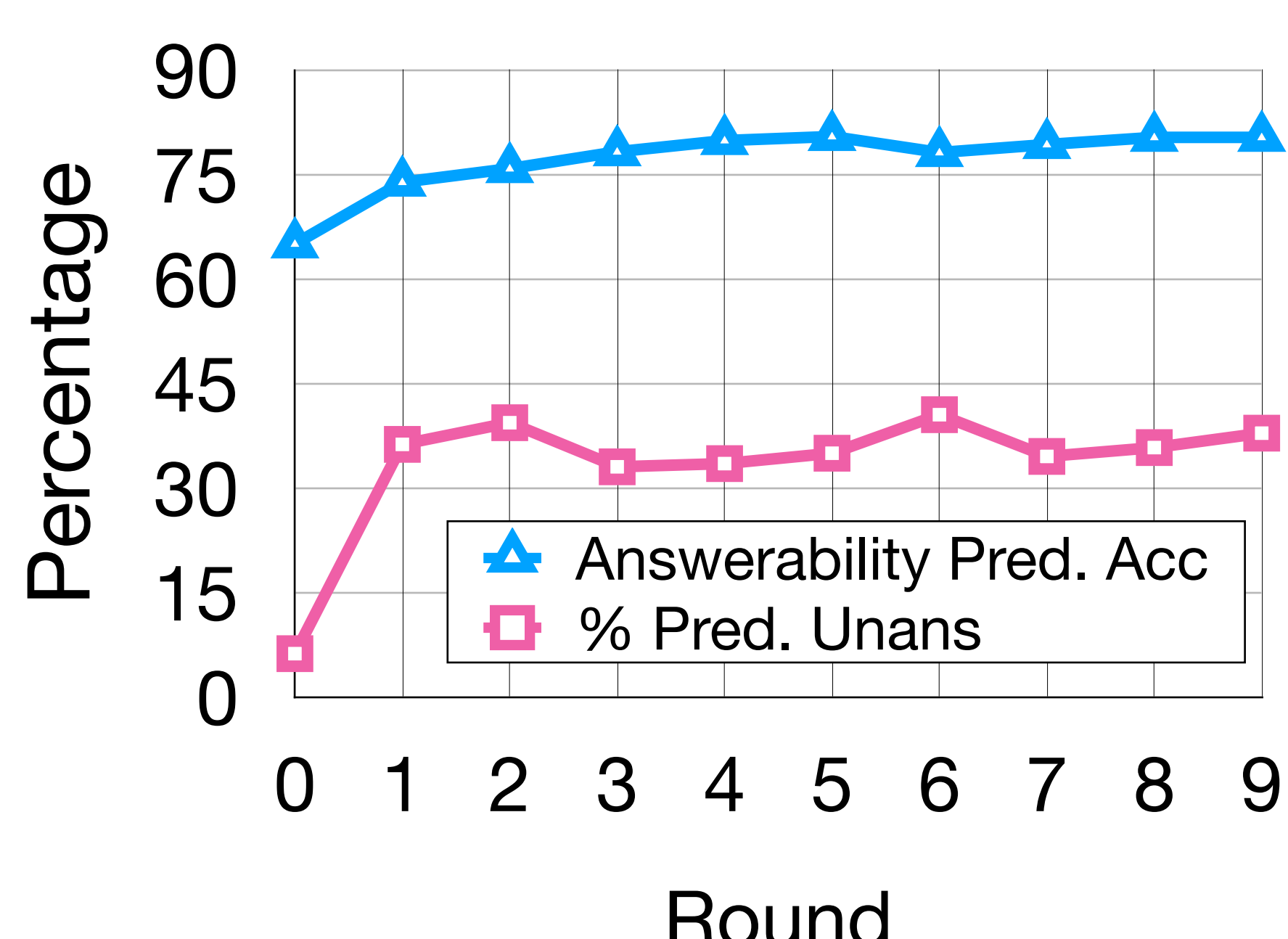
User Feedback Distribution



Test Set Performance



Answerability Prediction



Domain Adaptation: NewsQA-trained model adapts to user distribution

Performance degrades without answerability classifier

