Deep Bayesian Active Learning for Preference Modeling in Large Language Models

Luckeciano Melo Panagiotis Tigas Alessandro Abate Yarin Gal

University of Oxford

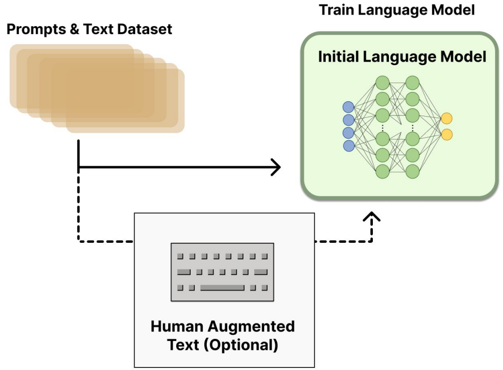
UNIVERSITY OF OXFORD

OXFORD CONTROL AND VERIFICATION

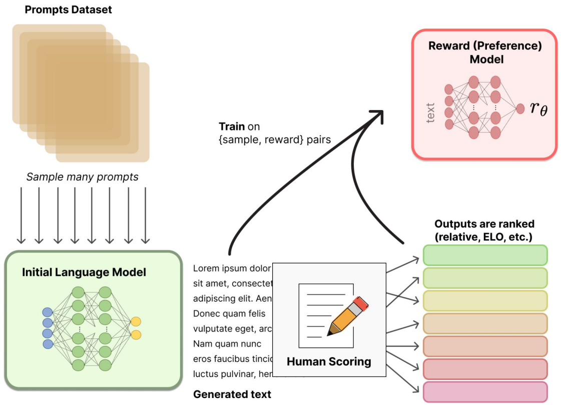
OATML

LLM Development Pipeline

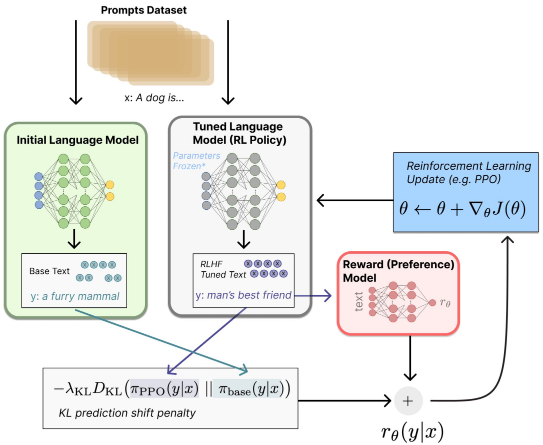
Preference optimization is a technique that allows us to control the behavior of large-scale unsupervised language models (LMs) by aligning them with human preferences



Self-Supervised Pre-Training


Prompts Dataset
Sample many prompts
Initial Language Model
Lorem ipsum dolor sit amet, consectet adipiscing elit. Aen Donec quam felis vulputate eget, arc Nam quam nunc eros faucibus tincid luctus pulvinar, her
Generated text
Human Scoring
Train on {sample, reward} pairs
Reward (Preference) Model
text
r_θ
Outputs are ranked (relative, ELO, etc.)

Preference Modeling


Prompts Dataset
x: A dog is...
Initial Language Model
Tuned Language Model (RL Policy)
Parameters Frozen*
Base Text
y: a furry mammal
RLHF Tuned Text
y: man's best friend
Reinforcement Learning Update (e.g. PPO)
$\theta \leftarrow \theta + \nabla_\theta J(\theta)$
Reward (Preference) Model
text
r_θ
$-\lambda_{\mathrm{KL}} D_{\mathrm{KL}}(\pi_{\mathrm{PPO}}(y|x) \,||\, \pi_{\mathrm{base}}(y|x))$
KL prediction shift penalty
+
$r_\theta(y|x)$

Reinforcement Learning from Human Preferences

Lambert et. al. Illustrating Reinforcement Learning from Human Feedback (RLHF). HuggingFace Blog, 2022.

Preference Optimization

- Preference optimization is a technique that allows us to control the behavior of large-scale unsupervised language models (LMs) by aligning them with human preferences

- **Collecting human feedback is expensive and laborious** [1]
  - Hundreds to millions of dollars per 100k preference labels
  - It becomes even more expensive for **specialized domains** (e.g., medical/sciences domain, potential superhuman AI systems)
  - Feedback generation takes **months** at large scale!

- Potential Solution: (Bayesian) Active Learning

[1] Casper et. al. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback, 2023.

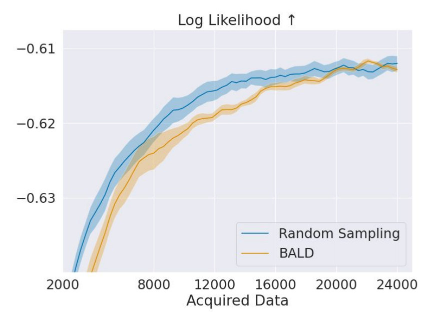
Active Learning for Preference Modeling in LLMs

- Selecting the most informative prompts/responses to gather feedback is essential to reduce costs and enable better LLMs!
  - Bayesian Active Learning provides a principled approach and has demonstrated remarkable success across different fields [2]

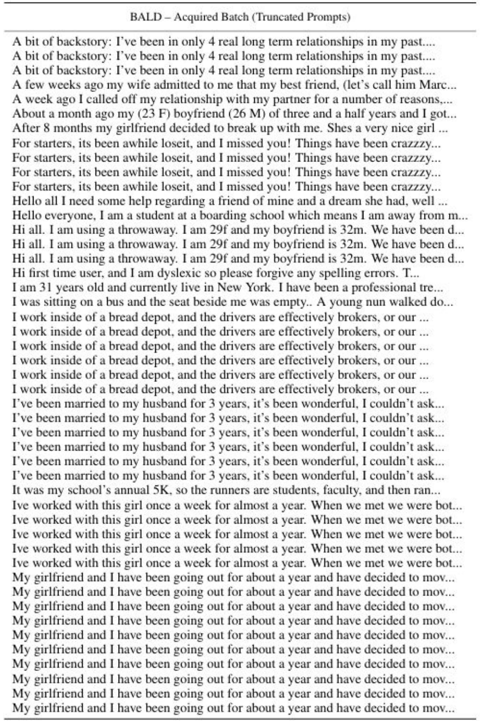
- Leveraging Active Learning (AL) for Preference Modeling in LLMs comprises three main challenges:
  - Prompt-answer pool is arbitrarily large and semantically rich
  - Human feedback is inherently noisy [2]
  - The intrinsic scale of LLMs requires batch acquisition and prohibits frequent model updates

[2] Gal et. al. Deep Bayesian Active Learning with Image Data. ICML, 2017.
[3] Stiennon et al. Learning to summarize with human feedback. NeurIPS, 2021

“Naive” application of Bayesian Active Learning fails



- The intrinsic scale of LLMs requires **batch acquisition** and prohibits frequent model updates
- Epistemic uncertainty estimators for *batch acquisition* are **intractable**
  - Proper batch estimators suffer from combinatorial complexity [2]
    - Even greedy approximations are still very expensive and impractical [3]
- Solely relying on single-point acquisition scheme **leads to the acquisition of redundant samples**

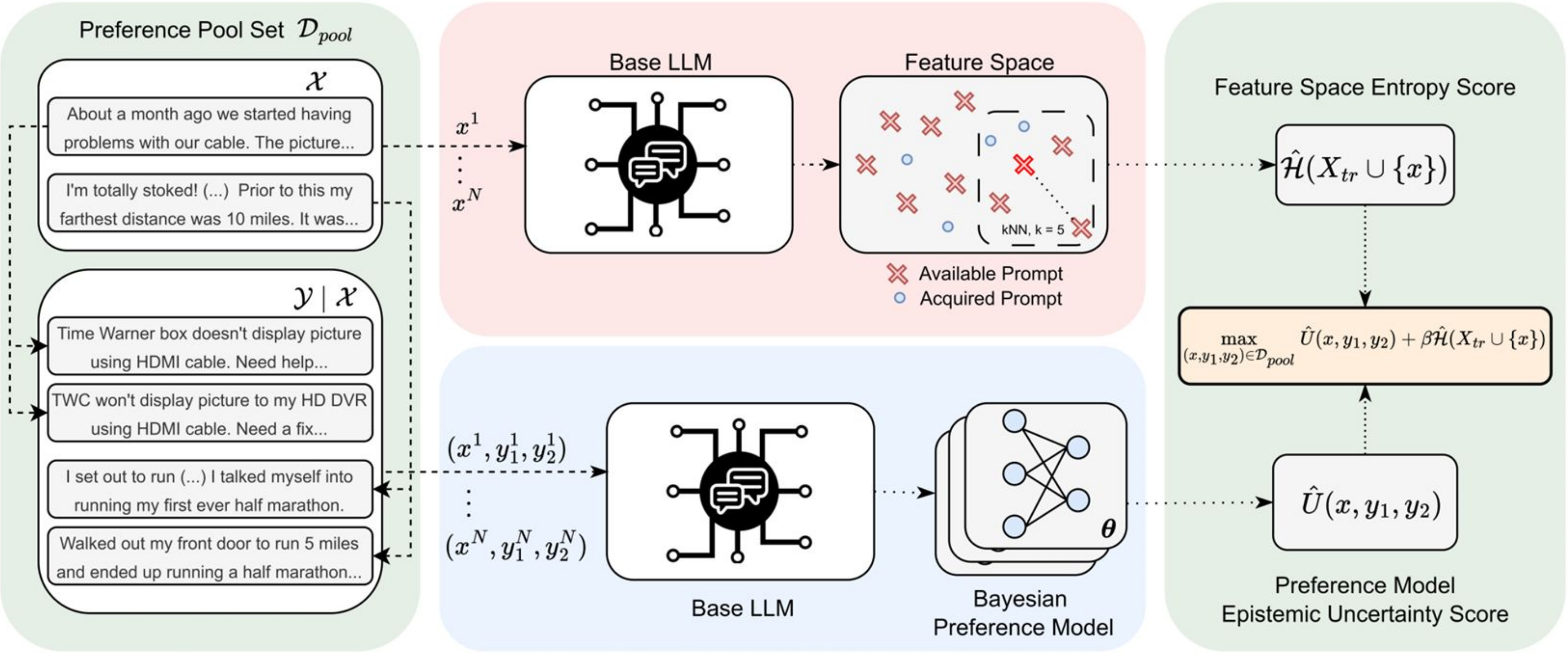


BALD – Acquired Batch (Truncated Prompts)
A bit of backstory: I’ve been in only 4 real long term relationships in my past....
A bit of backstory: I’ve been in only 4 real long term relationships in my past....
A bit of backstory: I’ve been in only 4 real long term relationships in my past....
A few weeks ago my wife admitted to me that my best friend, (let’s call him Marc...
A week ago I called off my relationship with my partner for a number of reasons,...
About a month ago my (23 F) boyfriend (26 M) of three and a half years and I got...
After 8 months my girlfriend decided to break up with me. Shes a very nice girl ...
For starters, its been awhile loseit, and I missed you! Things have been crazzzy...
For starters, its been awhile loseit, and I missed you! Things have been crazzzy...
For starters, its been awhile loseit, and I missed you! Things have been crazzzy...
For starters, its been awhile loseit, and I missed you! Things have been crazzzy...
Hello all I need some help regarding a friend of mine and a dream she had, well ...
Hello everyone, I am a student at a boarding school which means I am away from m...
Hi all. I am using a throwaway. I am 29f and my boyfriend is 32m. We have been d...
Hi all. I am using a throwaway. I am 29f and my boyfriend is 32m. We have been d...
Hi all. I am using a throwaway. I am 29f and my boyfriend is 32m. We have been d...
Hi first time user, and I am dyslexic so please forgive any spelling errors. T...
I am 31 years old and currently live in New York. I have been a professional tre...
I was sitting on a bus and the seat beside me was empty.. A young nun walked do...
I work inside of a bread depot, and the drivers are effectively brokers, or our ...
I work inside of a bread depot, and the drivers are effectively brokers, or our ...
I work inside of a bread depot, and the drivers are effectively brokers, or our ...
I work inside of a bread depot, and the drivers are effectively brokers, or our ...
I work inside of a bread depot, and the drivers are effectively brokers, or our ...
I work inside of a bread depot, and the drivers are effectively brokers, or our ...
I’ve been married to my husband for 3 years, it’s been wonderful, I couldn’t ask...
I’ve been married to my husband for 3 years, it’s been wonderful, I couldn’t ask...
I’ve been married to my husband for 3 years, it’s been wonderful, I couldn’t ask...
I’ve been married to my husband for 3 years, it’s been wonderful, I couldn’t ask...
I’ve been married to my husband for 3 years, it’s been wonderful, I couldn’t ask...
I’ve been married to my husband for 3 years, it’s been wonderful, I couldn’t ask...
It was my school’s annual 5K, so the runners are students, faculty, and then ran...
Ive worked with this girl once a week for almost a year. When we met we were bot...
Ive worked with this girl once a week for almost a year. When we met we were bot...
Ive worked with this girl once a week for almost a year. When we met we were bot...
Ive worked with this girl once a week for almost a year. When we met we were bot...
Ive worked with this girl once a week for almost a year. When we met we were bot...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...
My girlfriend and I have been going out for about a year and have decided to mov...

- **Goal:** design a proper AL objective that allow us to leverage a tractable epistemic uncertainty estimator while addressing its pathologies

[2] Kirsch et. al. BatchBald: Efficient and Diverse Batch Acquisition. NeurIPS, 2019.
[3] Kirsch et. al. Stochastic Batch Acquisition:A Simple Baseline for Deep Active Learning. TMLR, 2023.

Bayesian Active Learner for Preference Modeling



Preference Pool Set $\mathcal{D}_{pool}$

$\mathcal{X}$

About a month ago we started having problems with our cable. The picture...

I'm totally stoked! (...) Prior to this my farthest distance was 10 miles. It was...

$\mathcal{Y} \mid \mathcal{X}$

Time Warner box doesn't display picture using HDMI cable. Need help...

TWC won't display picture to my HD DVR using HDMI cable. Need a fix...

I set out to run (...) I talked myself into running my first ever half marathon.

Walked out my front door to run 5 miles and ended up running a half marathon...

x^1 $\vdots$ x^N

Base LLM

Feature Space

kNN, k = 5

❌ Available Prompt

○ Acquired Prompt

(x^1, y_1^1, y_2^1) $\vdots$ (x^N, y_1^N, y_2^N)

Base LLM

$\boldsymbol{\theta}$

Bayesian Preference Model

Feature Space Entropy Score

$\hat{\mathcal{H}}(X_{tr} \cup \{x\})$

$$\max_{(x,y_1,y_2)\in\mathcal{D}_{pool}} \hat{U}(x, y_1, y_2) + \beta\hat{\mathcal{H}}(X_{tr} \cup \{x\})$$

$\hat{U}(x, y_1, y_2)$

Preference Model Epistemic Uncertainty Score

Preference Model Epistemic Uncertainty Estimation

- We design a Bayesian Preference Model whose likelihood follows the Bradley-Terry assumption [4]
- Posterior predictive distribution:

$$p(y_1 \succ y_2 \mid x, y_1, y_2, \mathcal{D}_{train}) = \int p(y_1 \succ y_2 \mid x, y_1, y_2, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \mathcal{D}_{train}) d\boldsymbol{\theta}$$

- Posterior Approximation via ensemble of adapters

[4] Bradley & Terry. Rank Analysis Of Incomplete Block Designs: The method of paired comparisons. Biometrika, 1952.

Feature Space Entropy Estimation

- We estimate entropy via the KSG marginal entropy estimator [5]:

$$\hat{\mathcal{H}}_{KSG}(X) = \frac{d_X}{N}\sum_{i=0}^{N} \log D_x(i) + \log v_{d_X} + \psi(N) - \frac{1}{N}\sum_{i=0}^{N} \psi(n_{X_{tr}}(i) + 1)$$

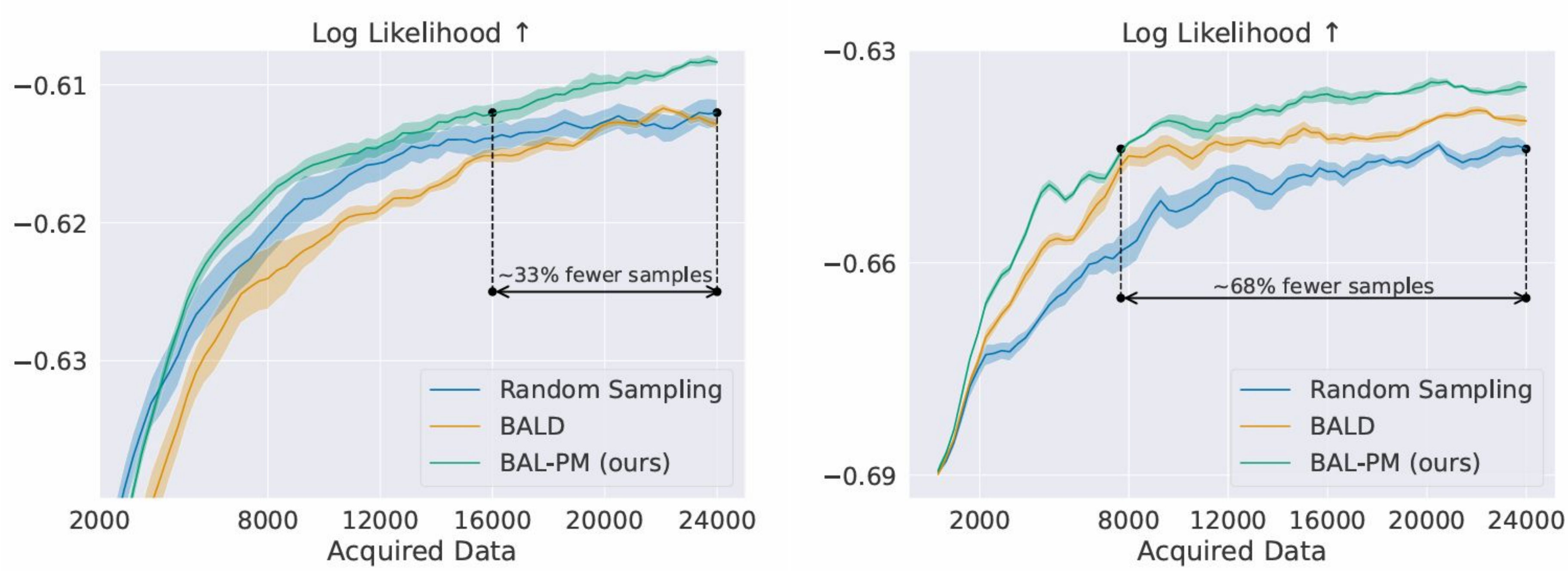
- Implementation:

$$\underset{(x,y_1,y_2)\in\mathcal{D}_{pool}}{\arg\max}\ \hat{\mathcal{H}}(X_t \cup \{x\}) = \underset{(x,y_1,y_2)\in\mathcal{D}_{pool}}{\arg\max}\ \log D(x) - \frac{1}{d_X}\psi(n_{X_{tr}}(x) + 1)$$

[5] Kraskov et. al. Estimating Mutual Information. Physical Review E, 2004.

Experiments

- Does BAL-PM reduce the volume of feedback required for Preference Modeling?

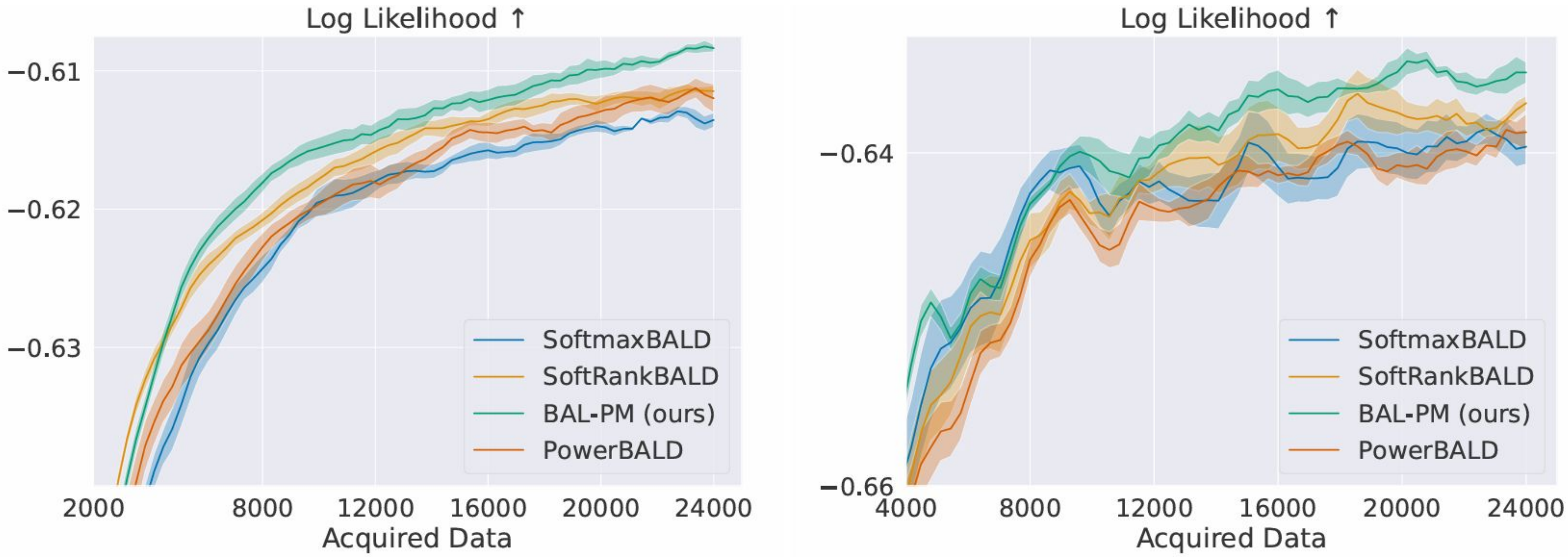


(a) Reddit TL;DR (Test)

(b) CNN/DM Dataset (OOD)

Experiments

- How does BAL-PM compare with other stochastic acquisition policies?

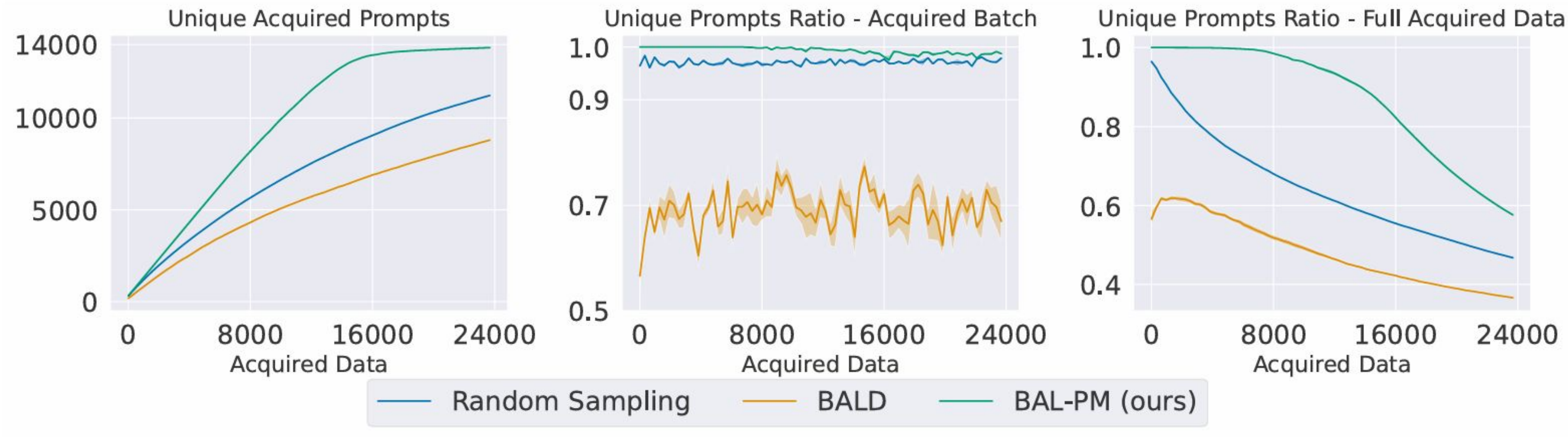


(a) Reddit TL;DR (Test)

(b) CNN/DM Dataset (OOD)

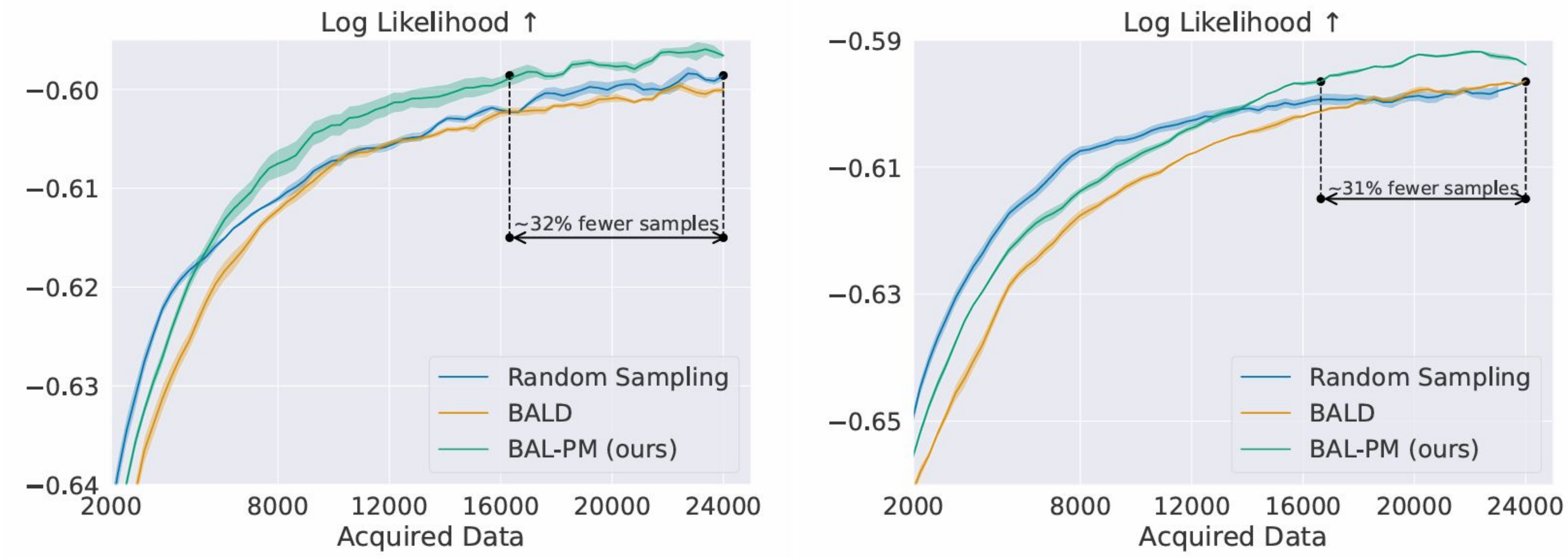
Experiments

- Does BAL-PM encourage diversity and prevent the acquisition of redundant samples?



Experiments

- How does BAL-PM scale to larger LLMs?



(a) 70b Parameter Model

(b) 140b Parameter Model

Closing Remarks

- BAL-PM is a stochastic policy for active batch acquisition in Preference Modeling for LLMs
  - Prevent the acquisition of redundant samples, a pathology of single-point acquisition schemes

- Impact: **An economy of hundreds of thousands of dollars and months of labeling work in the current scale of LLMs.**

- Limitations
  - Strong reliance on the quality of the LLM feature space

Poster

Deep Bayesian Active Learning for Preference Modeling in Large Language Models

Luckeciano Carvalho Melo · Panagiotis Tigas · Alessandro Abate · Yarin Gal

Luckeciano Melo Panagiotis Tigas Alessandro Abate Yarin Gal

University of Oxford