Softening human feedback improves classification calibration

Bob Carpenter
Center for Computational Mathematics
Flatiron Institute
**GPT-3 RL-HF**

- Transformer **pre-trained** on massive amounts of text (the “P” in “GPT”)
- Transformer **retrained** (“aligned”) to be **helpful, harmless, and truthful**
- Alignment training data is based on **human feedback (HF)**
  - humans rank examples, eg., $A_n > B_n$; use reinforcement learning
- Training **loss** for $A_n > B_n$ is **log logistic difference** (Bradley, Terry 1952)
  - reward($A | w$) is reward/utility of answer $A$ given weights $w$
  
  $$loss_n = - \log \logit^{-1}(\text{reward}(A_n | w) - \text{reward}(B_n | w))$$
Human feedback relatively inexpensive

- 40 contractors from Upwork/ScaleAI
- Pre-tested vs. desired answers
- 40 contractors cost $\approx$ US$2M per year, cf.
  - training hardware ($\approx$ US$500M)
  - AI researchers ($\approx$ US$500K+ per year)
  - data licensing (?)
  - servers (?)
- Conjecture: headroom for more investment
Raters are very noisy

- inter-annotator agreement only 73% (Ouyang et al. 2022)

- Goals conflict: helpful vs. harmless vs. truthful
  - OpenAI prioritized helpful; then filtered for harmless/truthful

- Traditional approaches to multi-annotation
  - just don’t do it (single annotate)
  - majority voting
  - censor non-agreement (i.e., remove from data set)
A simple classifier example

- Suppose I simulate a Bayesian **logistic regression** for $X_n \in \mathbb{R}^D$

  \[
  Y_n \sim \text{bernoulli}(\alpha + \beta^\top \cdot X_n) \quad \text{likelihood}
  \]
  \[
  X_n \sim \text{normal}(\mu, \Sigma) \quad \text{covariate data}
  \]
  \[
  \alpha, \beta_d \sim \text{normal}(0, \tau) \quad \text{prior}
  \]

  i.e., \( \text{logit} \Pr[Y_n = 1 \mid X_n = x_n, \alpha, \beta] = \alpha + \beta^\top \cdot X_n \)

- How to create a **“gold” standard** with $y_n \in \{0, 1\}$?
  
  - **Best Guess**: $y_n = 1$ if $\Pr[Y_n = 1 \mid X_n = x_n, \alpha, \beta] \geq \frac{1}{2}$
  
  - **Sample**: $y_n = 1$ if uniform$(0, 1) < \Pr[Y_n = 1 \mid X_n = x_n, \alpha, \beta]$
It’s Fool’s Gold

- **Sampling dominates best guess** (best guess biased)
- **Oversampling** $Y_n$ dominates single sampling
- **Weighted training** is optimal; let $\phi_n = \Pr[Y_n = 1 \mid X_n = x_n, \alpha, \beta]$

\[
\text{loss}_n = -\phi_n \cdot \log \logit^{-1}(\text{reward}(A_n \mid w) - \text{reward}(B_n \mid w)) - (1 - \phi_n) \cdot \log \logit^{-1}(\text{reward}(B_n \mid w) - \text{reward}(A_n \mid w))
\]

- **Why?** It provides **task-driven regularization**
  - **calibrated** means assigning probability $\phi_n$ to item $y_n = 1$ given $x_n$
Models of annotation

- **No access to truth** $\Pr[A_n > B_n \mid X_n = \chi_n, \alpha, \beta]$ during training
- Can ask multiple raters and build a model of annotation
- e.g., Dawid and Skene (1978) model of rater accuracy and bias yields
  \[ \Pr[A_n > B_n \mid \text{human feedback}] \]

- Weighted training $\gg$ sampling $\gg\gg$ highest probability
  - weighting training Rao-Blackwellizes sampling
  - multiple sampling $\rightarrow$ weighting as sample size increases
  - majority voting is best guess w.r.t. degenerate model
Weighted training regularizes

- **Dawid-Skene model is effective**  
  - Jointly estimate classifier and Dawid-Skene, but not necessary

- Effectiveness due to **task-specific regularization**

- E.g., if \( \Pr[A_n > B_n \mid \text{human rating}] = \psi_n \) and

  \[
  \text{loss}_n = -\psi_n \cdot \log \logit^{-1}(\text{reward}(A_n \mid w) - \text{reward}(B_n \mid w)) \\
  - (1 - \psi_n) \cdot \log \logit^{-1}(\text{reward}(B_n \mid w) - \text{reward}(A_n \mid w))
  \]

- Regularizes because **loss minimized** at \( \Pr[A_n > B_n \mid X_n = x, w] = \psi_n \)
Some references


