Softening human feedback
improves classification calibration

Bob Carpenter

Center for Computational Mathematics
Flatiron Institute

Berkeley Simons Institute LLM Meeting August 2023

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)
  - $\text{reward}(A \mid w)$ is reward/utility of answer A given weights w

$$\text{loss}_n = -\log \text{logit}^{-1}\Big(\text{reward}(A_n \mid w) - \text{reward}(B_n \mid w)\Big)$$

Human feedback relatively inexpensive

- 40 **contractors** from Upwork/ScaleAI
- **Pre-tested** vs. desired answers
- 40 contractors **cost** ≈ US$2M per year, cf.
  - training hardware (≈ US$500M)
  - AI researchers (≈ 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 $X_n \in \mathbb{R}^D$

$$
\begin{aligned}
Y_n &\sim \mathrm{bernoulli}(\alpha + \beta^\top \cdot X_n) && \text{likelihood} \\
X_n &\sim \mathrm{normal}(\mu, \Sigma) && \text{covariate data} \\
\alpha, \beta_d &\sim \mathrm{normal}(0, \tau) && \text{prior}
\end{aligned}
$$

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

- 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 $\mathrm{uniform}(0, 1) < \Pr[Y_n = 1 \mid X_n = x_n, \alpha, \beta]$

It's *Fool's Gold*

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

$$\begin{aligned}\text{loss}_n = &- \phi_n \cdot \log \operatorname{logit}^{-1}\Big(\text{reward}(A_n \mid w) - \text{reward}(B_n \mid w)\Big) \\ &- (1 - \phi_n) \cdot \log \operatorname{logit}^{-1}\Big(\text{reward}(B_n \mid w) - \text{reward}(A_n \mid w)\Big)\end{aligned}$$

- **Why?** It provides **task-driven regularization**
  - **calibrated** means assigning probability ϕ_n to item $y_n = 1$ given x_n

Models of annotation

- **No access to truth** $\Pr[A_n > B_n \mid X_n = x_n, \alpha, \beta]$ 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

$$\Pr\left[A_n > B_n \mid \text{human feedback}\right]$$

- Weighted training $\gg$ sampling $\ggg$ 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** (Raykar et al. 2010)
  - **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

$$\begin{aligned}\text{loss}_n = &- \psi_n \cdot \log \text{logit}^{-1}\Big(\text{reward}(A_n \mid w) - \text{reward}(B_n \mid w)\Big) \\ &- (1 - \psi_n) \cdot \log \text{logit}^{-1}\Big(\text{reward}(B_n \mid w) - \text{reward}(A_n \mid w)\Big)\end{aligned}$$

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

Some references

- Ouyang et al. 2022. **Training language models to follow instructions with human feedback**. OpenAI Blog.
- Cheng, C., Asi, H. and Duchi, J., 2022. **How many labelers do you have? A closer look at gold-standard labels**. *arXiv*.
- Passonneau, R.J. and Carpenter, B., 2014. **The benefits of a model of annotation**. *TACL*.
- Raykar, V.C., Yu, S., Zhao, L.H., Valadez, G.H., Florin, C., Bogoni, L. and Moy, L., 2010. **Learning from crowds**. *JMLR*.
- Bradley, R.A. and Terry, M.E., 1952. **Rank analysis of incomplete block designs: I. The method of paired comparisons.** *Biometrika*.
- Dawid, A.P. and Skene, A.M., 1979. **Maximum likelihood estimation of observer error-rates using the EM algorithm**. *JRSS(C)*.