Centrist Alignment

Andreas Haupt
Stanford University

The Community Notes Algorithm

$$\underbrace{\left\{\text{voters}\begin{bmatrix}+1 & -1 & ? & +1\\ -1 & ? & +1 & -1\\ +1 & +1 & -1 & ?\end{bmatrix}\right.}_{\text{notes}} \approx \gamma + \overbrace{\begin{bmatrix}\\\\\end{bmatrix}}^{\alpha} + \underbrace{\begin{bmatrix}\\\\\end{bmatrix}}_{\text{helpfulness}} + \overbrace{\begin{bmatrix}\\\\\end{bmatrix}}^{\sigma} \begin{bmatrix}\\\\\end{bmatrix}^{\iota}$$

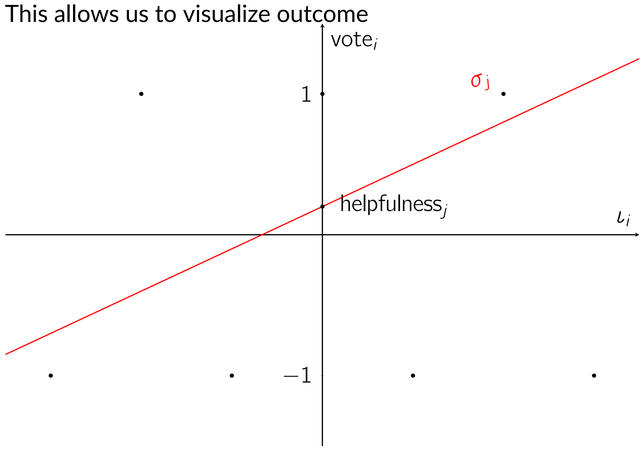
Show note j if $\text{helpfulness}_j > 0.4$.
Often characterized as "bridging" and requiring that "people who in the past have disagreed need to agree."

Observation 1: Individual Analysis

Statically, we can analyze the showing decision for a new note as intercept in the regression for helpfulness and σ

$$\text{vote}_i - \gamma - \alpha_i = \text{helpfulness} + \iota_i\sigma + \varepsilon_i$$

This allows us to visualize outcome



and to characterize which voters have most voting power.

Observation 2: Centrist User

At every point in time, the community notes algorithm gives us an aggregated preference ranking over all outcomes!
Call this the "centrist user's" preference.
What are their preferences on X?

Observation 3: Selection Robustness

Any selection mechanism S such that $S \perp\!\!\!\perp \varepsilon | \iota, \alpha$ yields the same helfulness intercept in the population regression.
A very desirable feature for Community Notes.

Community Notes
are **Centrist**,
not **Consensus**,

…and that's not necessarily bad.

Scan QR Code to download a draft

LMArena

$$\underbrace{\left\{\text{voters}\begin{bmatrix}+1 & -1 & ? & +1\\ -1 & ? & +1 & -1\\ +1 & +1 & -1 & ?\end{bmatrix}\right.}_{\text{LM comparisons}} \approx \underbrace{\begin{bmatrix}\\\\\end{bmatrix}}_{\text{comparisons}} + \overbrace{\begin{bmatrix}\\\\\end{bmatrix}}^{\sigma} \underbrace{\begin{bmatrix}\\\\\end{bmatrix}}_{\text{style}}$$

$$\underbrace{\left\{\text{voters}\begin{bmatrix}+1 & -1 & ? & +1\\ -1 & ? & +1 & -1\\ +1 & +1 & -1 & ?\end{bmatrix}\right.}_{\text{LM comparisons}} \approx \underbrace{\begin{bmatrix}\\\\\end{bmatrix}}_{\text{comparisons}} + \overbrace{\begin{bmatrix}\\\\\end{bmatrix}}^{\sigma} \underbrace{\begin{bmatrix}\\\\\end{bmatrix}}_{\text{length}}$$

only consider comparisons.

A New Alignment Target

Given observable voter and object features, fit

$$\text{vote}_{ij} = \gamma + \text{helpfulness}_j + \alpha_i + \iota_i\sigma_j + X_j\beta_j + \varepsilon_{ij},$$

and drop all but γ and helpfulness_j
(if constants do not matter, only keep helpfulness_j).

- If we believe the model is correctly specified, this is the choice of a hypothetical agent with $\alpha_i, \iota_i = 0$ rating a "counterfactual" content where $X_j = 0$ but which is otherwise the same.
- If selection S satisfies $S \perp\!\!\!\perp \varepsilon | (\alpha_i, \iota_i, X_j)$ then the intercept does not change.

Normative Arguments for Centrist Alignment

- Selection robustness
- Statistical version of "median" rule (Condorcet winner if existent)
- Easily auditable: transitive average utility

Finetuning for Centrist Alignment

Interpret

$$\text{helpfulness}(y \mid x)$$

as a function of completions y to prompts x. We can regress as before, and use Group Relative Policy Optimization or others to optimize generation.

In ongoing work, train LLM for note-writing on X.