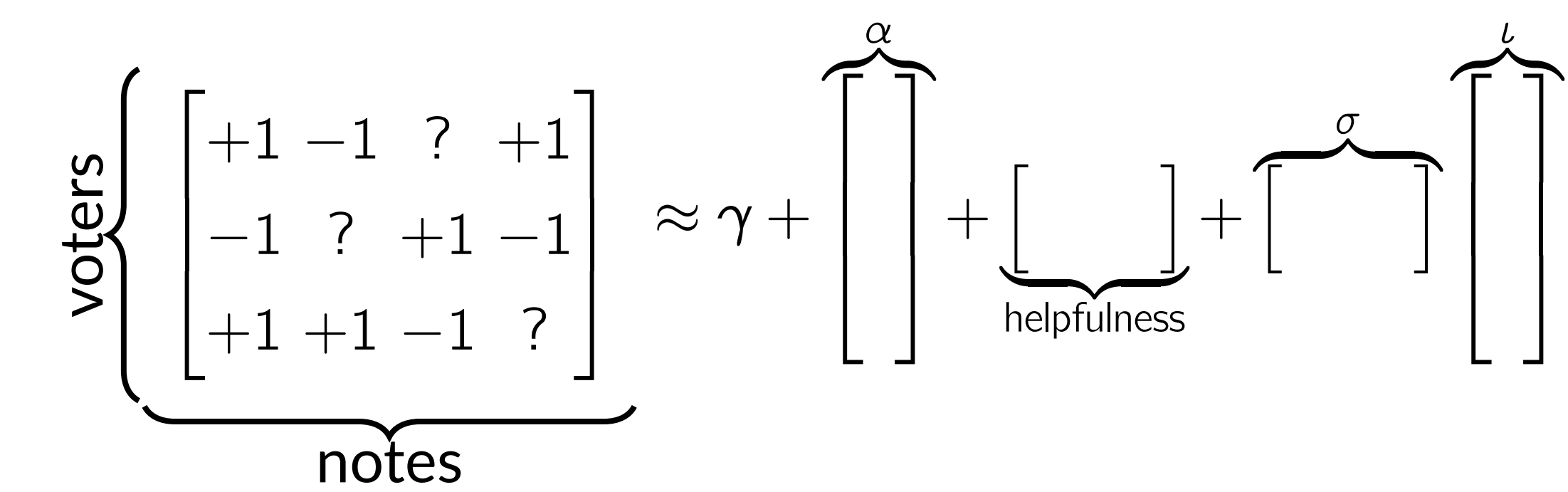


Centrist Community Notes

Andreas Haupt
Stanford University

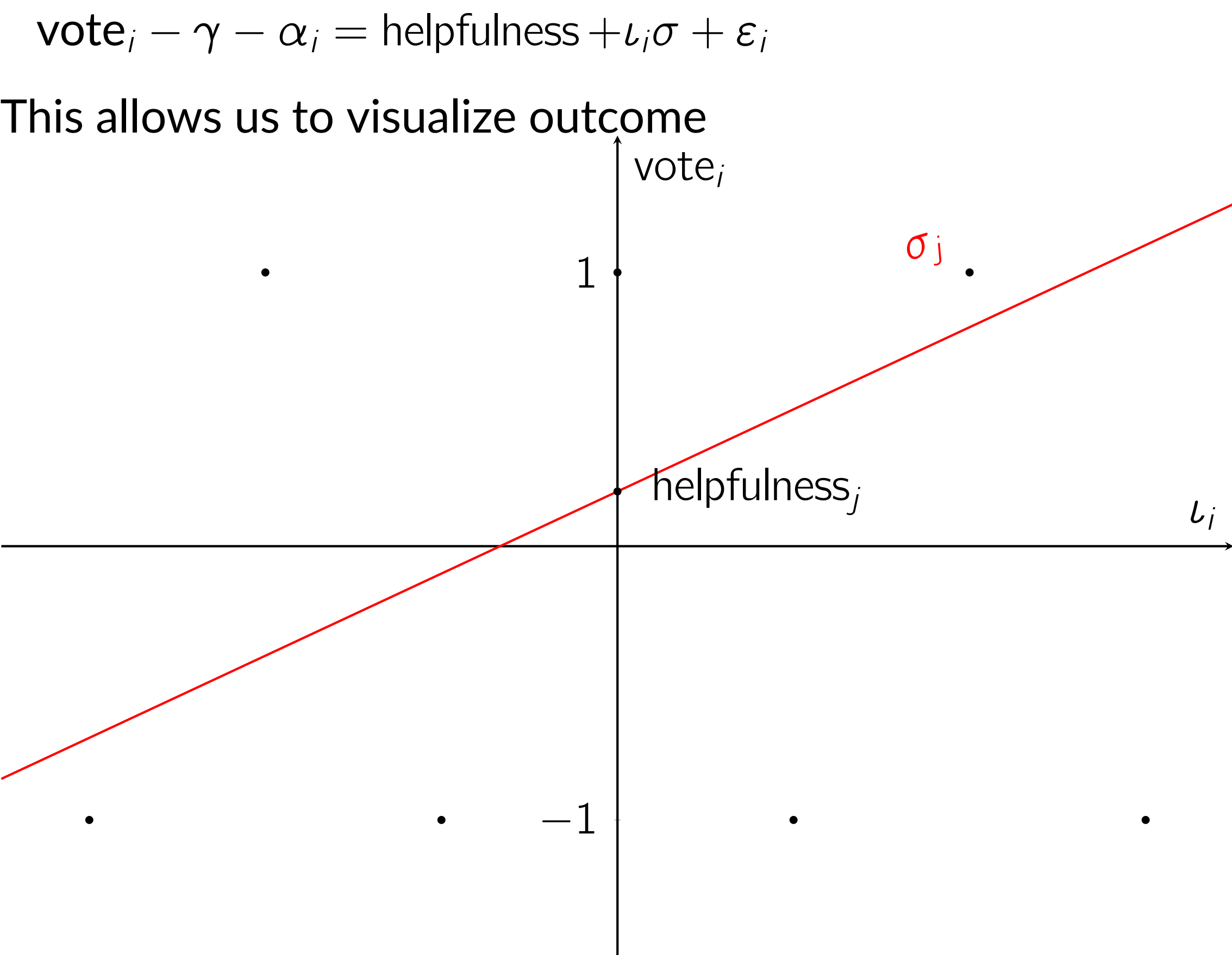
The Community Notes Algorithm



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



and see that centrist approval is sufficient, not bridging!

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.

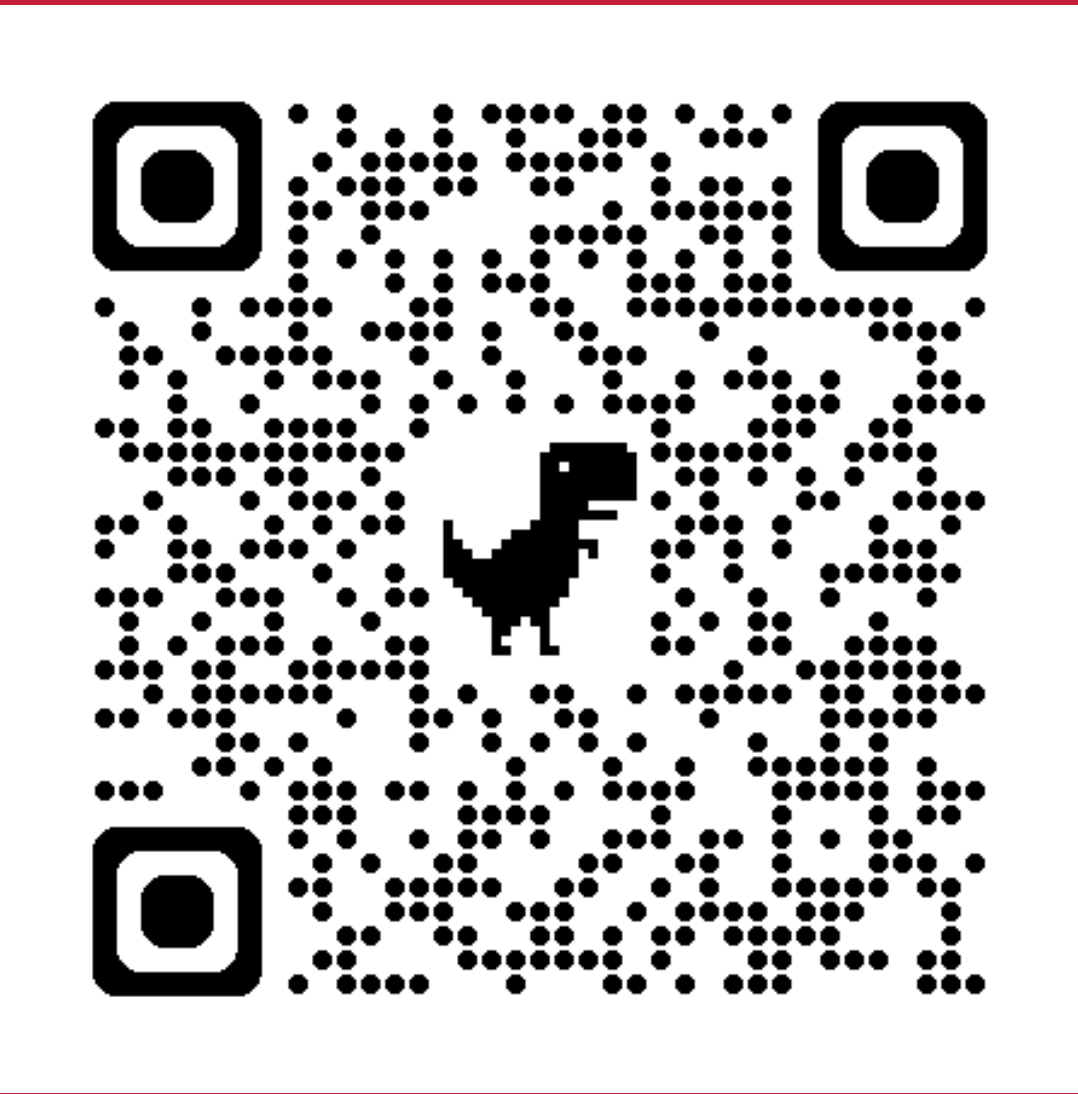
Observation 3: Selection Robustness

Any selection mechanism S such that $S \perp\!\!\!\perp \epsilon | \iota, \alpha$ yields the
same helpfulness 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

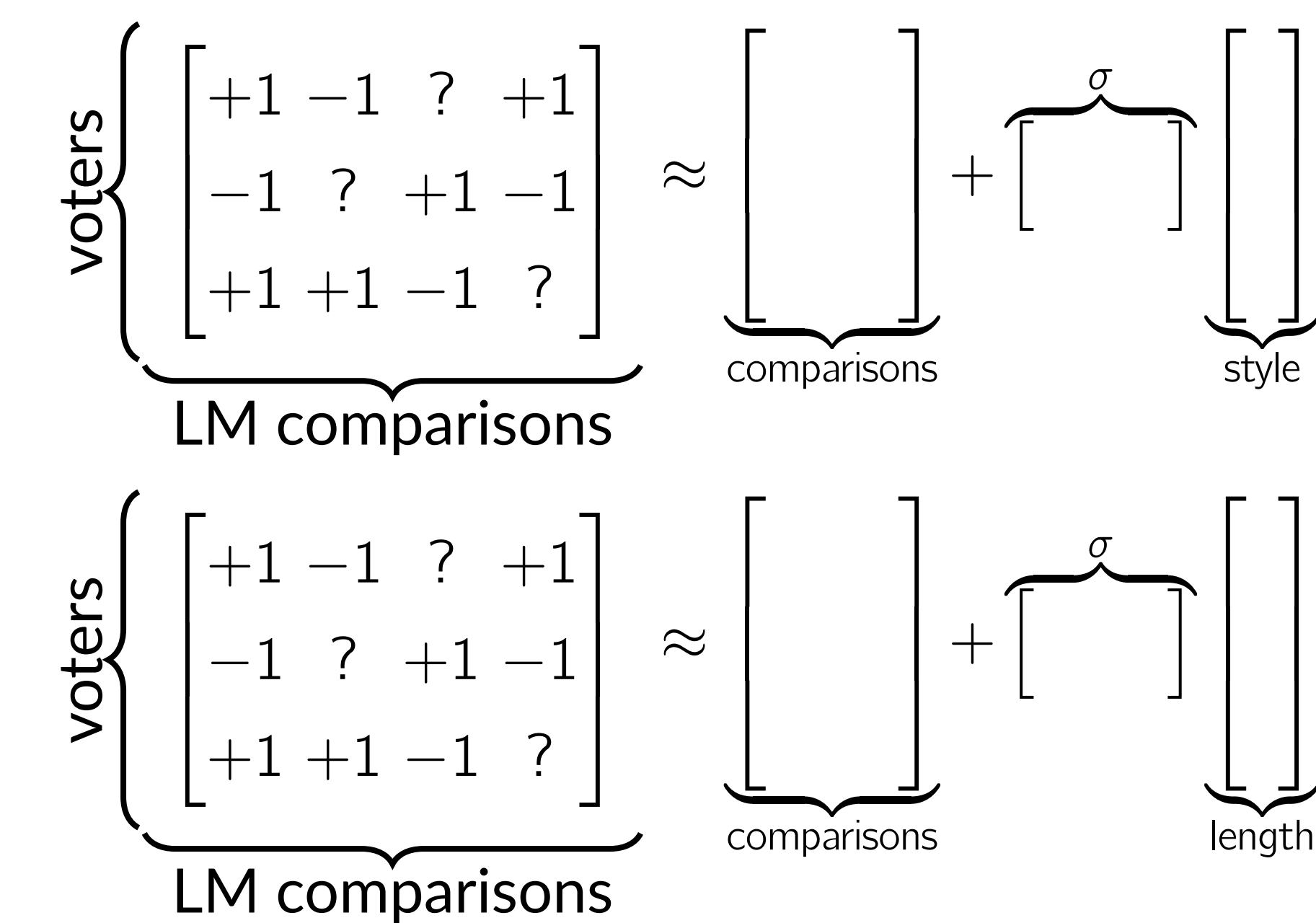
Observation 4: Full Voting

Community notes estimates approval assumign all
users would vote.
In particular, if users are logit and the effects of
community notes are additive, it maximizes wel-
fare.

A New Alignment Target

Given observable voter and object features, fit
 $\text{vote}_{ij} = \gamma + \text{helpfulness}_j + \alpha_i + \iota_i \sigma_j + X_j \beta_j + \epsilon_{ij}$,
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 $S \perp\!\!\!\perp \epsilon | (\alpha_i, \iota_i, X_j)$ then the
intercept does not change.

Example: Style & Length Control



only consider comparisons.

Finetuning for Centrist Alignment

Interpret

$\text{helpfulness}(y | x)$
as a function of completions y to prompts x . We
can regress as before, and use Group Relative Pol-
icy Optimization or others to optimize generation.

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