Improving Crowdsourcing for AI through Cognitive-Inspired Data Engineering

April 2026

Coauthors and collaborators

Gunnar Epping
Indiana University

Jennifer Trueblood
Indiana University

William Holmes
Indiana University

Erik Duhaime
Centaur Labs

Andrew Caplin
NYU Economics

Daniel Martin
UCSB Economics

Supported by the Sloan Foundation grant "Cognitive Economics at Work"

Our field setting: Centaur Labs

Leader in nascent medical labeling field: Founded at MIT, $15.2 million series A funding.

Gamified app called DiagnosUs where users compete in contests to win monetary prizes.

Contests are typically classification tasks where users provide a single classification, such as a skin lesion being ‘nevus’ or ‘melanoma’.

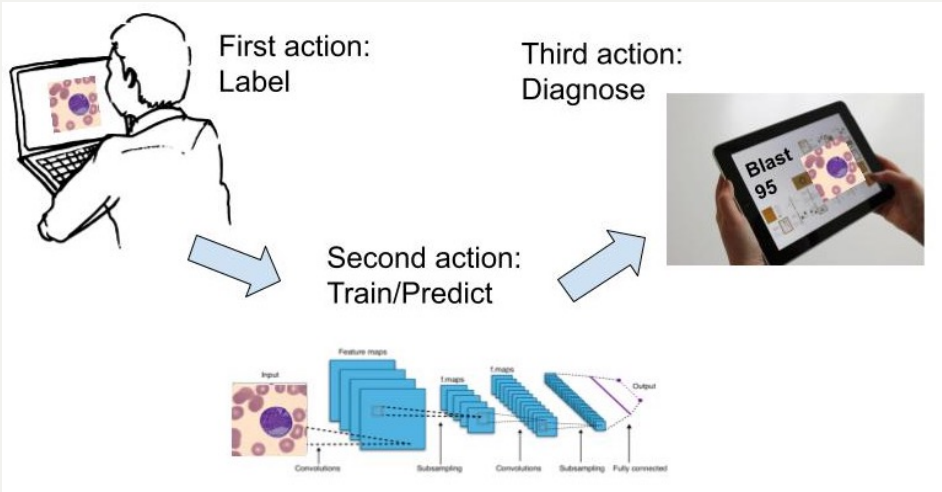
By collecting several classifications per image (‘crowdsourcing’), they leverage wisdom of the crowd (WoC) to create accurately labeled datasets used to train AI systems.

While anyone can download the app, the majority of users are medical students.

Crowdsourced labels are manufactured ground truth

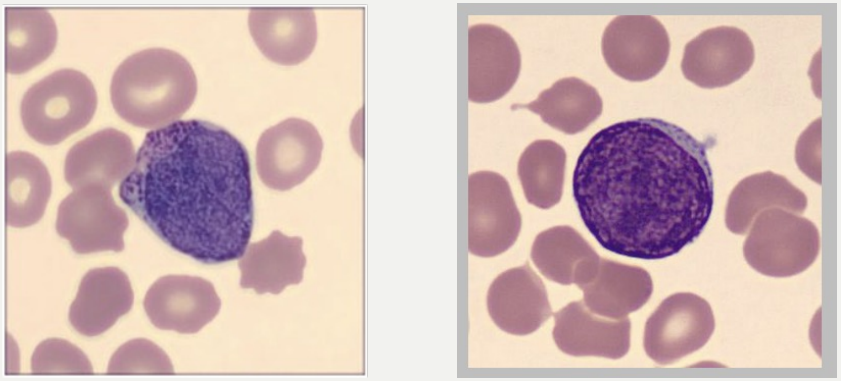
- Supervised AI still depends heavily on human-labeled training data.
- Crowdsourcing manufactures that “ground truth” by aggregating many annotators’ judgments.
- Those labels can inherit annotators’ cognitive constraints and biases before any model is trained.
- Question: **can a cognitively-grounded data interventions improve both labels and downstream AI?**

Bias enters at the labeling step



Setting: medical image diagnosis

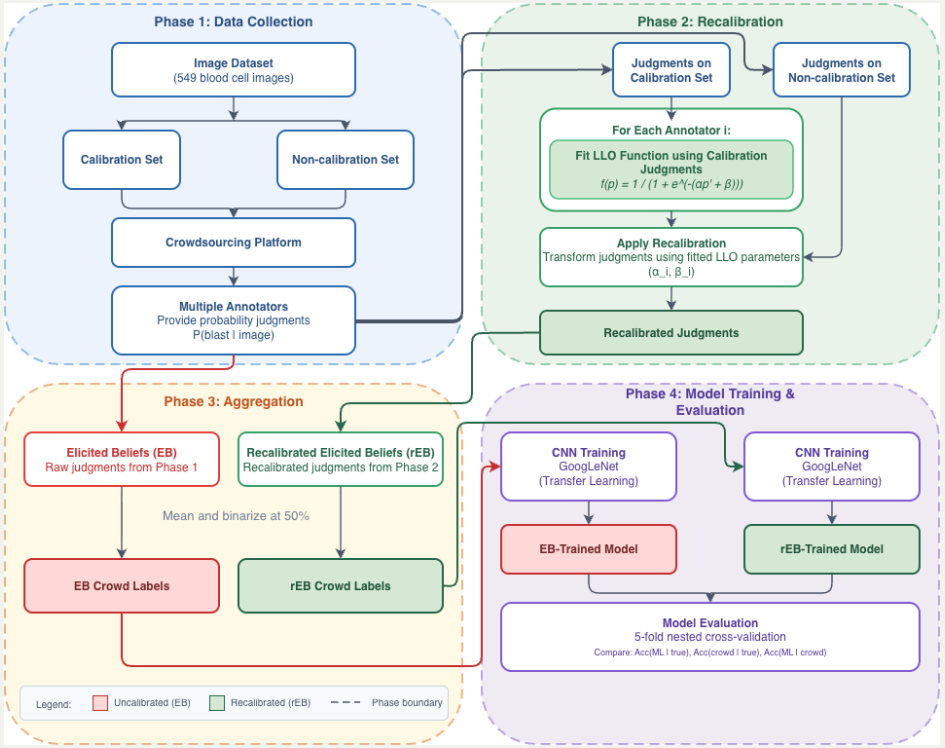
- Task: classify Wright-stained white blood cells as blast vs. non-blast.
- Ground truth: 549 Vanderbilt images with three-way hematopathology expert agreement.
- Useful sandbox: real clinical images, learnable by novices, still challenging for experts.



Example white-blood-cell images used in both experiments.

Cognitive-inspired data engineering

Elicit beliefs + recalibrate + aggregate corrected judgments + train model based on resulting labels.



Two tests: proof of concept and field validation

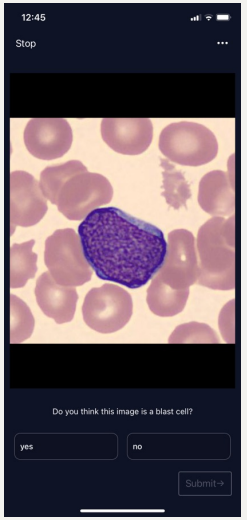
Experiment 1 — MTurk novices

- 360 participants after exclusions.
- 60-image calibration set + 489-image non-calibration set.
- Elicited beliefs (EB) only.
- Proof of concept with minimally trained annotators.

Experiment 2 — DiagnosUs skilled annotators

- 97 binary choice (BC) and 78 EB participants.
- 249 gold-standard images + 300 QA images.
- Users randomized into binary-choice vs. elicited-belief contests.
- Field validation on a production-style medical labeling platform.

Centaur Labs



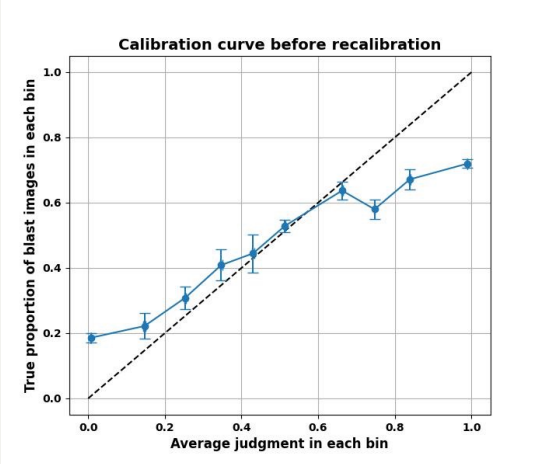
Probability judgments are informative but biased

- Raw elicited beliefs are not well calibrated: skilled annotators still overuse 0, 50, and 100.
- For each annotator, the paper fits a linear-in-log-odds (LLO) map on images with known truth.

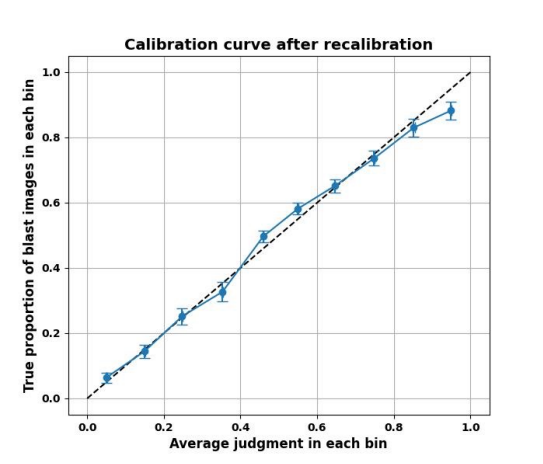
$$\ln\left(\frac{f(p)}{1-f(p)}\right) = \alpha \ln\left(\frac{p}{1-p}\right) + \beta.$$

$$f(p) = \frac{1}{1 + e^{-(\alpha p' + \beta)}}.$$

Illustration uses the field experiment (DiagnosUs).



Before recalibration



After recalibration

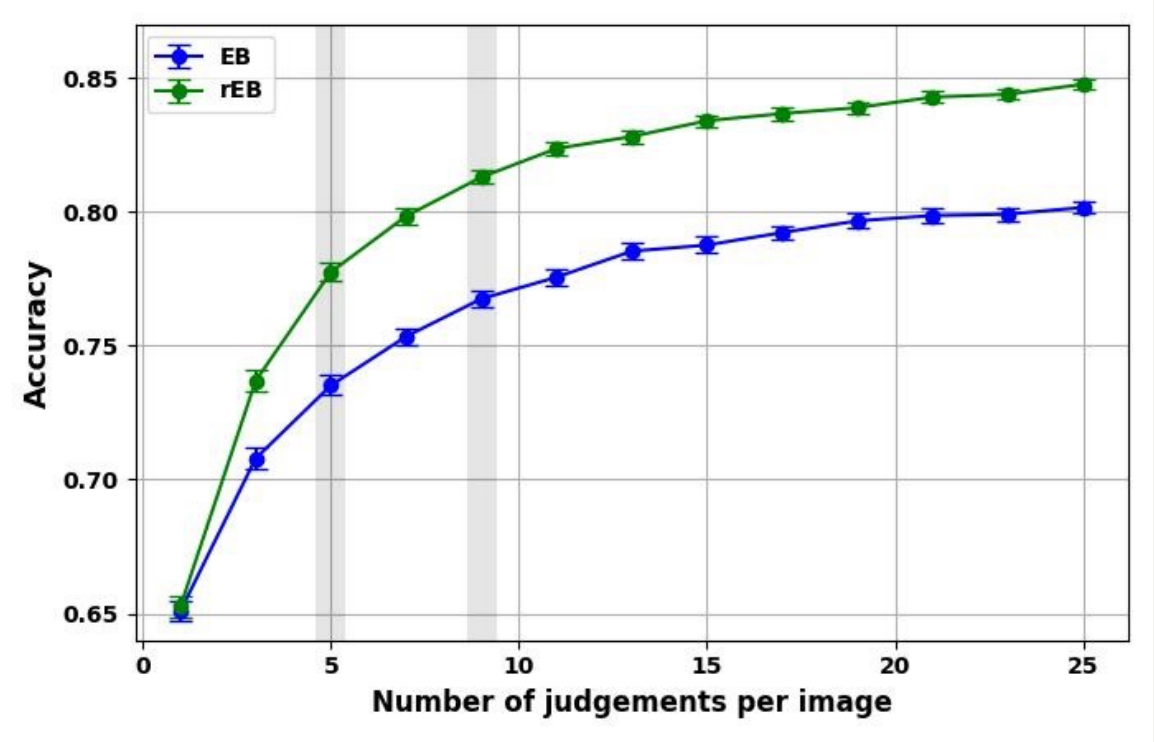
Result: after recalibration, reported probabilities line up much more closely with actual frequencies.

Experiment 1 (MTurk): recalibration helps even with novices

+3.5 pp

crowd-label gain
with all judgments per image

- Individual accuracy is essentially unchanged: 65.3% raw EB → 65.4% rEB.
- Crowd accuracy improves from 81.6% to 85.1%.
- CNN test accuracy rises from 86.0% to 86.9%.
- Advantage appears across the full judgments-per-image curve.

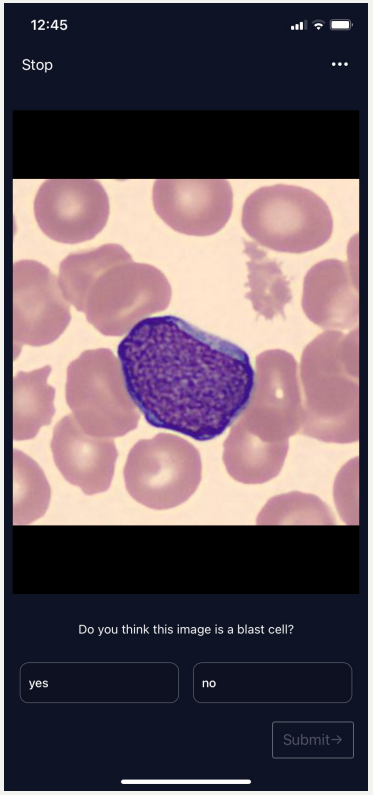


Crowd accuracy as a function of the number of judgments per image

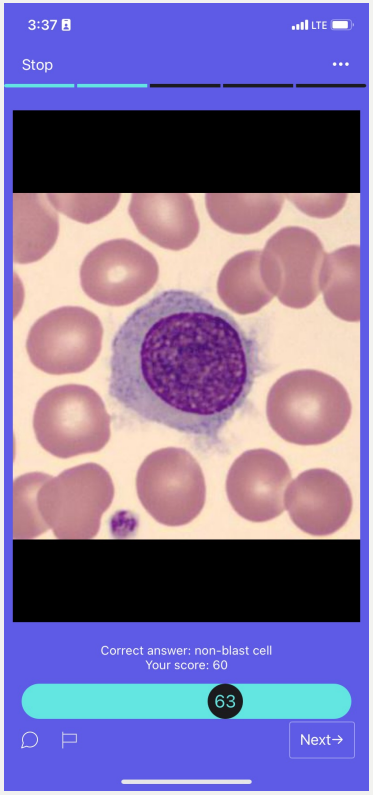
Experiment 2: a real crowdsourcing platform

- DiagnosUs is a gamified app used for medical and scientific data annotation.
- The contest randomizes users into binary-choice and elicited-belief versions.
- Gold-standard images are already in the platform for QA and feedback, so they double as a calibration set.
- Participants needed at least 200 competition trials to become prize-eligible.

Centaur Labs



Binary choice (BC)

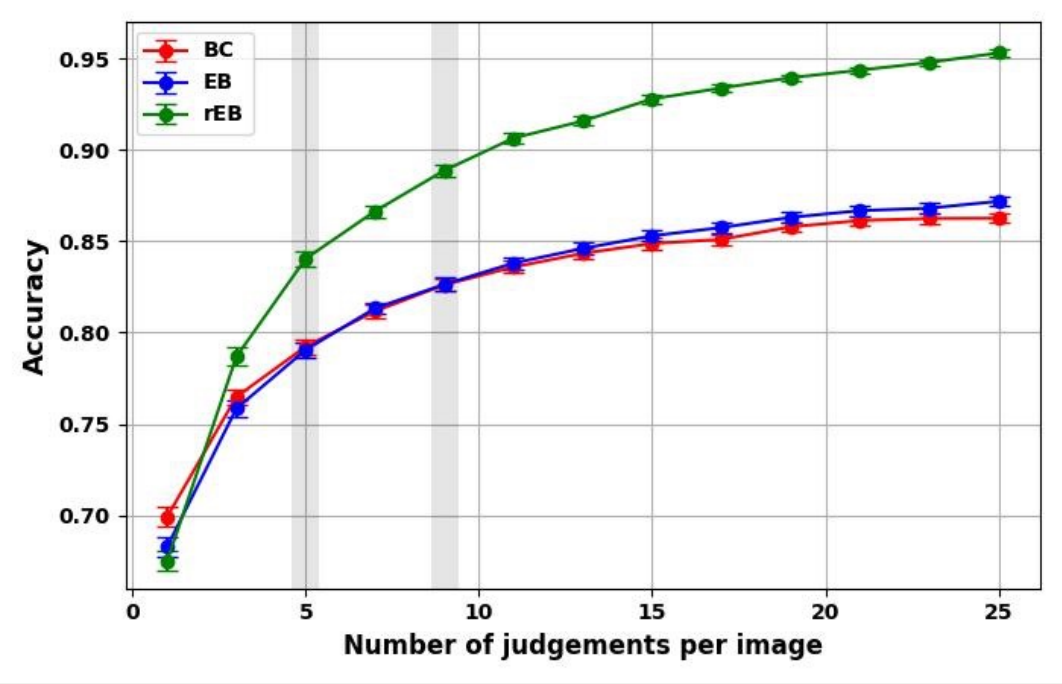


Elicited beliefs (EB)

Experiment 2 crowd labels: recalibrated EB dominates

Votes / image	BC	EB	rEB
5	78.8%	79.5%	83.6%
9	82.6%	82.6%	89.0%
All	88.2%	88.3%	96.7%

Debiasing the richer signal — not elicitation alone — creates the gain.



Crowd accuracy as a function of the number of judgments per image

Better labels translate into better CNNs

Same transfer-learning pipeline, same 5-fold nested cross-validation — only the labels change.

Votes / image	BC	EB	rEB
5	83.0%	83.2%	89.5%
9	84.9%	84.0%	92.6%
All	89.7%	89.2%	96.0%

Test accuracy
Acc(ML | true)

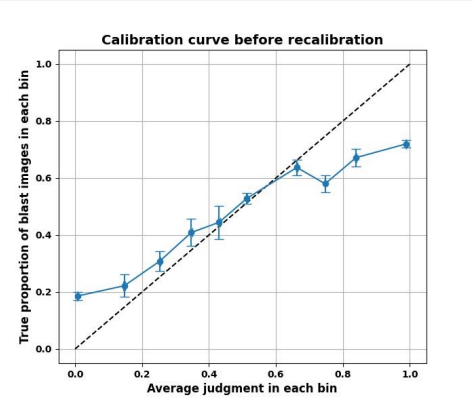
Key efficiency result

rEB, 5 votes / image = 89.5%

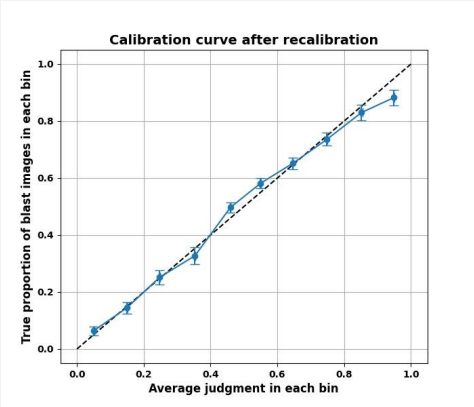
BC, all votes / image = 89.7%
EB, all votes / image = 89.2%

Recalibration largely buys back the cost of collecting many more labels.
With all votes, rEB rises to 96.0%.

Why the crowd improves even when individuals do not



Before recalibration

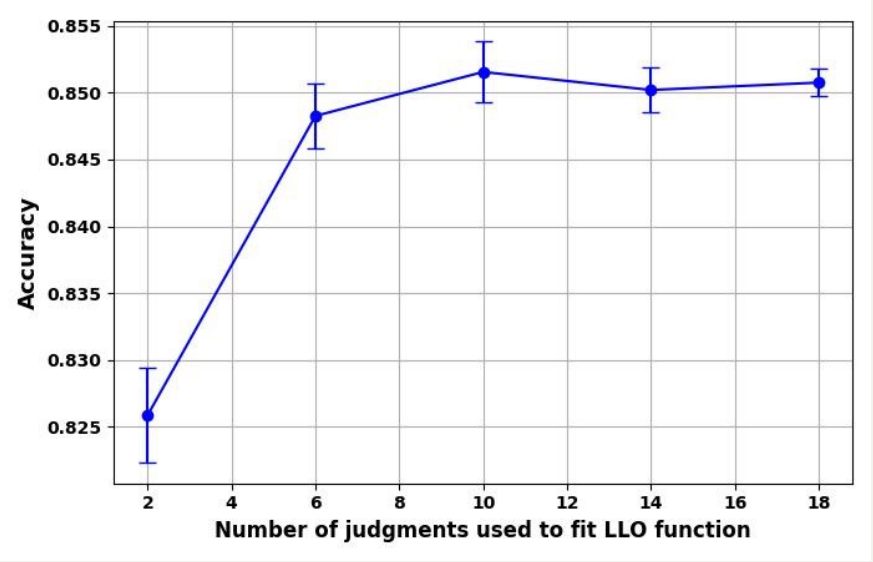


After recalibration

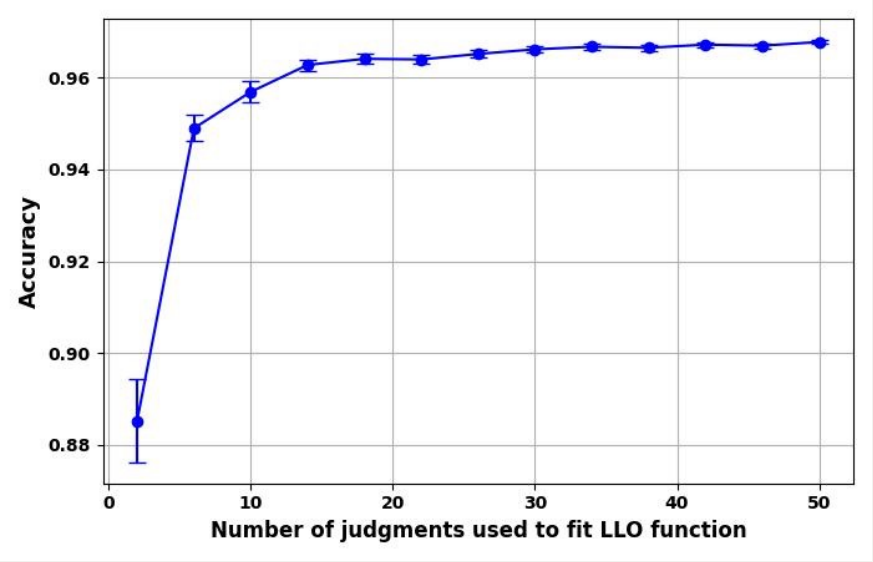
- In Experiment 2, individual accuracy barely changes: 68.7% raw EB → 68.2% rEB.
- But recalibration fixes systematic overconfidence, especially in extreme responses.
- At the group level, this behaves like an implicit variance-weighting scheme: overly certain wrong judgments matter less; underconfident informative judgments matter more.

Most of the gain comes from a small calibration set

You do not need many gold-standard items per annotator to get most of the benefit.



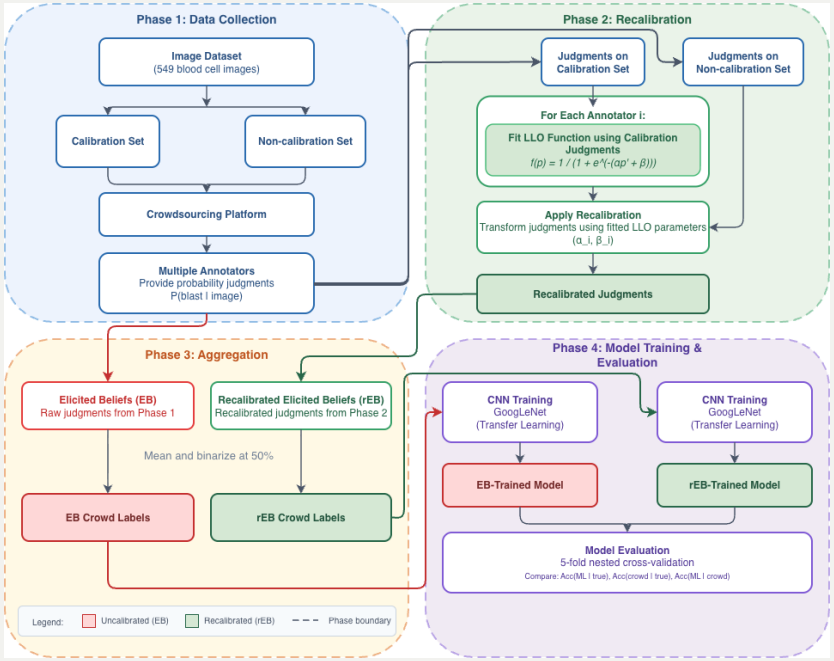
Experiment 1: most gains by ~6 judgments



Experiment 2: diminishing returns after ~10

An upstream data intervention

- The contribution is annotation-level correction, not human-AI co-optimization or model engineering.
- Recalibration is compatible with expertise weighting, annotator selection, or other aggregation methods; it is not a substitute for them.
- Because the mechanism operates on subjective probabilities, the basic idea should extend beyond medical imaging wherever probabilistic judgments are available.



Limitations and next steps

Current limits

- One image task and one benchmark dataset.
- Balanced class proportions in both studies; low-prevalence settings remain open.
- Models were trained on binarized crowd labels rather than continuous soft labels.

Next steps

- Test additional tasks, datasets, and crowdsourcing platforms.
- Combine recalibration with expertise weighting or annotator filtering.
- Train models directly on recalibrated probability judgments.

Takeaways

+3.5 pp

Experiment 1
crowd-label gain

+8.4 pp

Experiment 2
crowd-label gain

+6.8 pp

Experiment 2
CNN test gain vs raw EB

Behavioral-science tools can improve AI training data upstream — before model engineering begins.

Raw elicited beliefs do not beat binary labels in the field. Recalibrated elicited beliefs do.

Thank you!

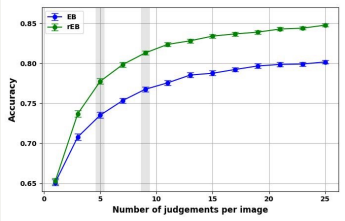
Appendix follows.

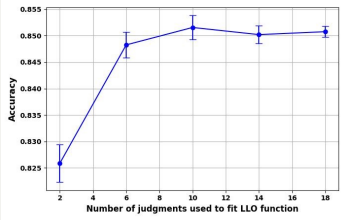
Appendix — Experiment 1 detailed results

Main-text summary on the earlier slide used all judgments per image. This table shows the 5, 9, and all-judgment comparisons.

Votes / image	Crowd raw EB	Crowd rEB	CNN raw EB	CNN rEB
5	73.5%	77.8%	81.1%	84.9%
9	76.6%	81.3%	82.7%	86.0%
All	81.6%	85.1%	86.0%	86.9%

Crowd = Acc(crowd | true); CNN = test Acc(ML | true).





Appendix — Experiment 2 detailed results

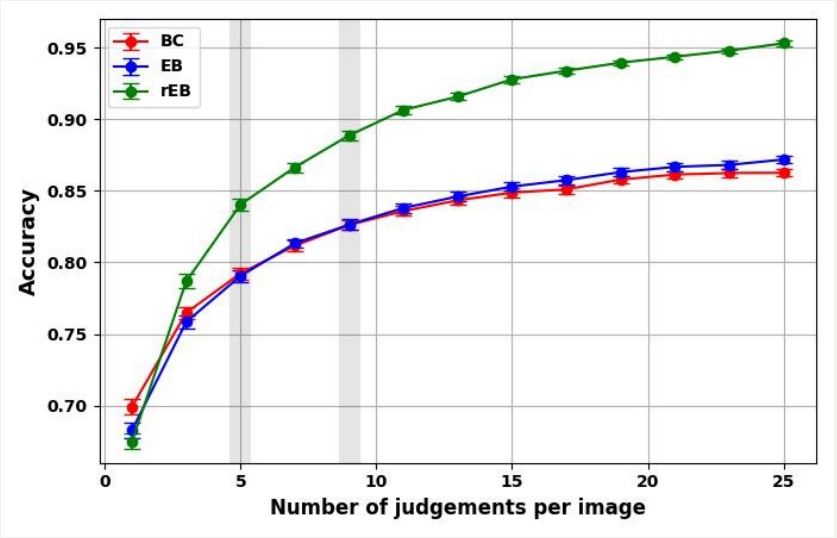
Top table: crowd-label accuracy. Bottom table: CNN test accuracy Acc(ML | true).

Crowd-label accuracy

Votes / image	BC	EB	rEB
5	78.8%	79.5%	83.6%
9	82.6%	82.6%	89.0%
All	88.2%	88.3%	96.7%

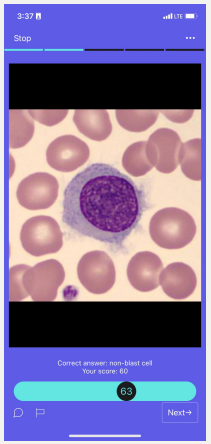
CNN test accuracy

Votes / image	BC	EB	rEB
5	83.0%	83.2%	89.5%
9	84.9%	84.0%	92.6%
All	89.7%	89.2%	96.0%



Appendix — Incentives for elicited beliefs

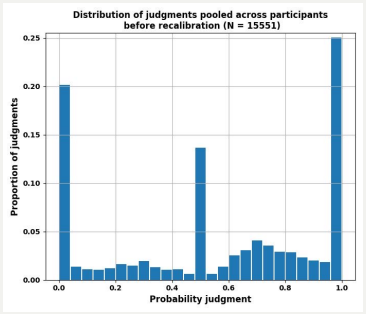
- Experiment 2 used a binarized quadratic scoring rule so truthful probability reports maximize expected score.
- If a participant reports blast probability q, the chance of winning the fixed prize is 1 − (1 − q)^2 when the cell is a blast, and 1 − q^2 when it is not a blast.
- The fixed prize size makes the contest intuitive while preserving incentive compatibility.



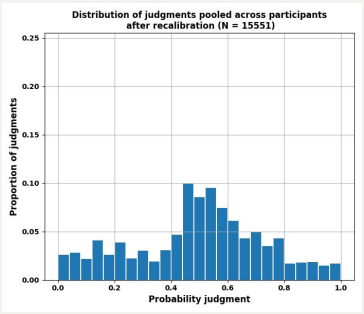
Appendix — The LLO parameters have cognitive meaning

$$\ln\left(\frac{f(p)}{1-f(p)}\right) = \alpha \ln\left(\frac{p}{1-p}\right) + \beta.$$

$$f(p) = \frac{1}{1 + e^{-(\alpha p' + \beta)}}.$$



Before



After

- The slope α governs how strongly judgments are pushed toward or away from the center: it corrects over- or underconfidence.
- The intercept β shifts judgments up or down and therefore captures response bias.
- Because the map is fit separately for each annotator, it can also correct idiosyncratic reversals or unusually poor calibration.