LASERAI

MSR77

A tale of two thresholds: adaptive thresholding with abstention for AI-assisted single screening (AISS)

Artur Nowak, Monika Opalek, Ewelina Sadowska, Ewa Borowiack

Evidence Prime, Krakow, Poland

WHAT IS NEW

AISS treats screening automation as an internal validation problem: reduce workload only when review-local evidence supports the screener. We show that this adaptive approach allows matching double-screening in terms of accuracy, at much reduced cost.

Objective

We need "meta-methods"

Automation can fail in specific reviews even when average model performance is high. AISS wraps a base screener with review-local thresholds, monitoring, and abstention so workload reduction is conditional on observed reliability.

- Applies to human single screening with AI suggestions or fully AI-screened workflows.
- Selective QA targets excluded records most likely to be false negatives.
- If the classifier appears noisy, AISS abstains more instead of forcing final exclusions.

Analogy: stopping criteria in ranking-based screening are also meta-methods, because they validate the process around a model.

Dynamic thresholds

How AISS adapts during a review

AISS begins with conservative thresholds and then updates the abstention region as review-specific evidence accumulates. Human include/exclude decisions reveal whether the current screener is aligned with the review question. If early decisions expose disagreement, borderline behavior, or higher miss risk, the exclude threshold tightens and more records are postponed for QA. If the evidence is stable, the thresholds release more low-risk records from

Signal
Human choices add review-local evidence.

→

Thresholds
The abstention band widens or narrows.

→

Actions
Include, exclude, or route to human QA.

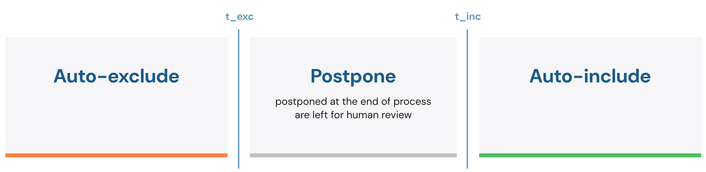
Supported workflows:

- first-pass human screening can be paired with selective second-human QA
- AI screens all the records, human is asked to review potential false negatives

Metod

Two thresholds instead of one

Classifier score s(x): higher = more include-like



A single cutoff forces every record to include/exclude. AISS maintains an abstention band and updates it as review evidence accumulates, so uncertainty changes the workflow rather than simply changing a label.

Results

First experiment: weak screening model

TF-IDF plus logistic regression provided a controllable weak screener. The aim was to validate the two-threshold control policy across realistic errors, automation bias, and model collapse.

Scenario	Recall	Conflict	Abstain
Random errors (5%)	98.6%	18.6%	1.8%
Random errors (20%)	95.6%	30.1%	2.4%
Borderline errors	94.9%	21.5%	3.2%
Automation bias	78.1%	8.7%	0.3%
Model collapse	97.5%	19.6%	12.6%

Interpretation: model collapse increased abstention and preserved recall. Automation bias remained the high-risk regime: agreement can hide shared misses.

Second experiment: strong screening model

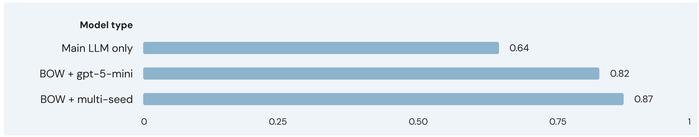
The LLM-based is an extension of the BOW model, adding decision-flow traces from GPT-5.1 and a smaller same-family model (gpt-5-mini). Keeping a BOW component adds variety and avoids repeating the main LLM's suggestions.

7
TEST reviews (✦)

36,729
Records
TiAb references

272
Includes
0.74% prevalence

Rescued false negatives per 100 reviewed excludes



Conclusion

Workload reduction without sacrificing accuracy

The first experiment showed that automation bias is harder: agreement can mask shared false negatives. We therefore evaluated the scenario in which screening is performed fully by AI and compared it to human screening with 5% error rate.

Screening regime	Single-screening recall	Recall after QA	QA work	Human work
Full automation	91.5%	95.2%	1,025 QA + 1 abstain	2.8/100 records -98.6%
Random human errors (5%)	97.4%	99.3%	2,814 QA + 76 abstain	107.7/100 records -46.2%

Full automation counts QA only; human simulation counts first-pass screening plus QA. Human work normalized to double screening = 200 decisions / 100 refs.

✦ The BOW and LLM evaluations are not a head-to-head model comparison. The dataset of 160 reviews (60 in development set, 100 in test set) used for BOW evaluation does not contain inclusion criteria for all reviews. Therefore, a separate set o 7 reviews with eligibility criteria used by the LLM model was used.