Stanford

NeurIPS 2026

When Should Humans Step In?

Optimal Human Dispatching in AI-Assisted Decisions

Lezhi Tan[1] Naomi Sagan[2] Lihua Lei[3] José Blanchet[1]

[1] Management Science & Engineering [2] Electrical Engineering [3] Graduate School of Business · Stanford University · * Equal contribution

November 2026

arxiv.org/abs/2603.13688

In one slide: query humans where the AI cannot tell cases apart

Question *With a limited human budget, which factors of a decision should a human expert evaluate?*

1 **Theory** Optimal dispatching reduces to maximizing a contextual reward (Theorem 1), which an orthogonal two-stage estimator learns from data (Theorem 2).

2 **Rule** Under a linear model, a factor's value = predictive relevance × residual uncertainty of the human signal, given the AI signal.

3 **Evidence** On 3,408 ICLR papers, human scores on 3 of 10 aspects come within 0.013 MAE of full human review.

Key insight: the size of the AI–human gap is not what matters. What matters is where the AI cannot tell apart what humans can.

High-stakes decisions share one structure

Long, complex document W
a manuscript, a medical record, a contract, a loan file

→

Factor-level judgments X
$X_1, \ldots, X_J$ set in advance by domain guidelines

→

Decision Y
accept, treat, sign, approve

Peer review

Document Manuscript

Factors rigor, novelty, significance, clarity

Decision accept or reject

Clinical care

Document Medical record

Factors history, labs, imaging, guideline criteria

Decision treatment or coverage

Legal review

Document Contract or brief

Factors liability, compliance, precedent

Decision sign or escalate

Lending

Document Loan application

Factors income, collateral, credit history

Decision approve or deny

Same shape everywhere: an expert reads a lot, judges a few factors, then decides.

We want a human in the loop, for good reasons

AI judgments fail in systematic ways

LLM reviewers inflate scores, reward polished writing over substance, miss subtle technical flaws, and drift toward what the prompt suggests.

Accountability is required

EU AI Act Art. 14: high-risk AI systems must allow effective human oversight (Annex III systems from Dec 2027). GDPR Art. 22: a right not to be subject to solely automated decisions.

Decisions need legitimacy

Authors, patients and borrowers accept an outcome because an accountable expert stands behind it.

21%

of ICLR 2026 reviews estimated to be fully AI-generated (Pangram Labs)

Without a design for oversight, the human in the loop disappears by default.

LLM-review failure modes: e.g., Liang et al., NEJM AI 2024; Shin et al., arXiv:2502.17086; Fanous et al., SycEval, arXiv:2502.08177. Regulation (EU) 2024/1689 (AI Act), as amended by (EU) 2026/1744; Regulation (EU) 2016/679, Art. 22.

...but full human review does not scale

>100M hrs

spent on journal peer review worldwide in 2020 (Aczel et al., 2021)

~31k

NeurIPS 2026 main-track submissions, about 3× the 2022 count (third-party trackers)

>90k

reviews needed in one NeurIPS cycle at 3 reviews per submission

Overload breaks oversight

Reviewers under load skim, defer to the AI draft, or outsource the review entirely. Over-reliance on AI output is automation bias, which oversight designs must guard against (AI Act Art. 14(4)(b)).

Effort is spread evenly; value is not

A full review spends the same care on every factor, though only a few factors carry information the AI lacks. In our data: about 3 of 10.

We want both: keep a human in the loop, and truly save human effort.

Human–AI collaboration in four steps

Domain guidelines fix the factors $X_1, \ldots, X_J$ in advance (e.g., NeurIPS / ICML review criteria)

1. AI review

- Prompt an LLM with the full document and the factor definitions.
- Collect $A = (A_1, \ldots, A_J)$ on every factor.

cheap · every factor

2. Dispatcher

- Read the AI signals A and a human budget k.
- Choose a subset $S \subseteq [J]$ with $|S| \leq k$.

focus of this talk

3. Human query

- Ask a human to evaluate only the selected factors.
- Observe H_S, the human signals on S.

costly · few factors

4. Prediction

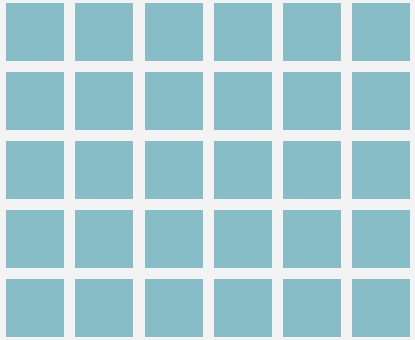
- Combine all AI signals with the queried human signals.
- Fit $\hat{Y} = g_\theta(A, H_S)$ by minimizing squared error.

final recommendation

Key insight: human involvement is not all-or-nothing. It is a resource dispatched across factors.

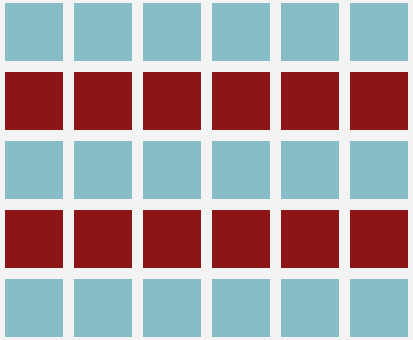
Existing options spend human effort case by case

Full automation



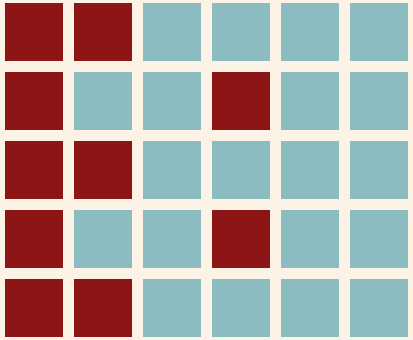
AI decides every case. Cheap, but no oversight.

Case-level deferral



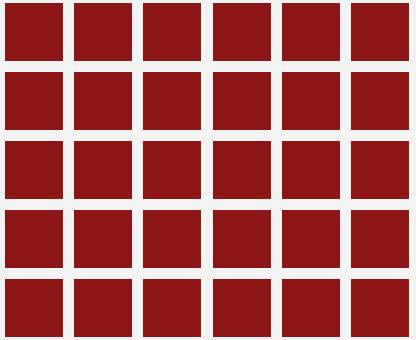
A human takes over whole cases: learning to defer, selective prediction, prediction-powered inference.

Factor-level dispatching (ours)



Every case gets AI on all factors and a human on the few factors that matter.

Full human review



A human evaluates every factor of every case. Reliable, but does not scale.

AI evaluates human evaluates

rows = cases (e.g., papers) · columns = factors (e.g., review aspects)

Core question *With a limited human budget, which factors should a human evaluate, so that we cut human effort without losing validity?*

Selection and prediction are optimized jointly

$$\min_{\pi:\,|\pi(Z)|=k} \; \min_{\theta} \; \mathbb{E}\left[(Y - g_\theta(A, H_\pi))^2\right]$$

Outer min: which factors to query (π)

A policy picks the subset S = π(Z) from AI context Z, under budget |S| = k.

Inner min: how to predict (θ)

Any ML predictor g_θ fit on all AI signals plus the queried human signals.

How much may the policy adapt?

Adaptive: π depends on Z = f(A)

- Different cases get different human questions
- Must learn the reward R_S(z) as a function of z
- Weakly better in population, higher variance

Non-adaptive: Z constant

- One fixed subset for every case
- Only a scalar reward R_S per subset
- Lower variance; as good or better in our data

Theorem 1: dispatching reduces to reward maximization

Assumption (rich predictor class). *For every S, the class {g_θ} contains the Bayes predictor E[Y | A, H_S].*

Bayes risk of a subset S

$$\mathbb{E}\big[(Y - \mathbb{E}[Y \mid A, H_S])^2\big] \;=\; \mathbb{E}[Y^2] \;-\; \mathbb{E}\big[\mathbb{E}[Y \mid A, H_S]^2\big]$$

$\mathrm{E}[Y^2]$ does not depend on S, so minimizing risk = maximizing a contextual reward

$$\pi^\star(z) \in \arg\max_{|S|=k} R_S(z), \qquad R_S(z) = \mathbb{E}\Big[\mathbb{E}[Y \mid A, H_S]^2 \;\Big|\; Z = z\Big]$$

Non-adaptive case (Z constant): a best-subset problem

$$S^\star \in \arg\max_{|S|=k} R_S, \qquad R_S = \mathbb{E}\big[\mathbb{E}[Y \mid A, H_S]^2\big]$$

Policy design becomes reward estimation: estimate R_S(z) from data, then pick the best S for each context.

Theorem 2: an orthogonal estimator learns the reward

1 Cross-fitted first stage

$$\hat{m}_S(a,h) \approx \mathbb{E}[Y \mid A=a,\ H_S=h]$$

K-fold cross-fitting gives an out-of-fold prediction for every unit.

2 Pseudo-outcome

$$\hat{\varphi}_S = 2Y\,\hat{m}_S(A,H_S) - \hat{m}_S(A,H_S)^2$$

E[φ̂_S | A, H_S] = m_S² − (m̂_S − m_S)²: the bias is second order.

3 Second-stage regression

$$\hat{R}_S(z) = \widehat{\mathbb{E}}\big[\hat{\varphi}_S \mid Z=z\big]$$

Non-adaptive case: simply average φ̂_S over the sample.

Guarantee $\left\|\hat{R}_S - R_S\right\|_2 = O_P\left(r_g + r_m^2\right)$

r_m: first-stage rate for m̂_S; r_g: second-stage rate.
First-stage error enters only squared (Neyman orthogonality).

Why it matters: a first-stage learner converging at only n^(−1/4) still leaves the reward estimate limited by the second stage alone.

Under a linear model, the reward has a closed form

1 Centred linear model

$$\mathbb{E}[Y \mid A, H] - \mathbb{E}[Y] = \gamma^\top (A - \mathbb{E}A) + \beta^\top (H - \mathbb{E}H)$$

2 Residualize H on A: the part the AI cannot predict

$$H_\perp = H - \mathbb{E}[H] - \mathrm{Cov}(H, A)\,\mathrm{Var}(A)^{-1}(A - \mathbb{E}A)$$

3 γ does not depend on S (tower property)

$$\mathbb{E}[Y \mid A, H_S] - \mathbb{E}[Y] = \gamma^\top (A - \mathbb{E}A) + \beta_{S,\perp}^\top H_{S,\perp}$$

4 **Optimal non-adaptive rule**

$$S^\star = \arg\max_{|S|=k} \beta_{S,\perp}^\top \mathrm{Var}(H_{S,\perp})\,\beta_{S,\perp}$$

Estimation

Ridge regression of Y on (A, H_S). With J = 10, enumerate all 2^{10} = 1,024 subsets.

Theorem 3

$\sqrt{n}$ (R̂_S − R_S) → N(0, σ^2_S) by the delta method: a confidence interval for every subset's value.

Two factors decide where humans are needed

$$R_j^{\text{lin}} \propto \underbrace{\|\beta_{j,\perp}\|^2}_{\text{relevance}} \times \underbrace{\text{Var}(H_{j,\perp})}_{\text{residual uncertainty}}$$

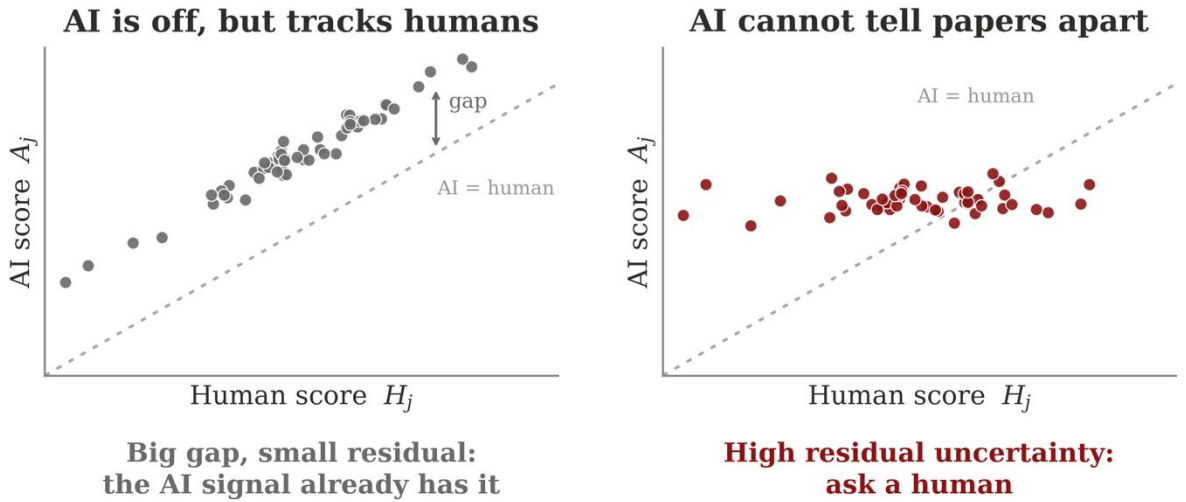
$$H_{j,\perp} = H_j - (\text{part of } H_j \text{ predictable from the AI signal } A)$$

Predictive relevance

Does the part of the human judgment that the AI misses matter for the outcome?

Residual uncertainty

How much of the human judgment can the AI not predict?



Left: the AI is biased but tracks humans; the offset is learned from A. Right: the AI gives every paper about the same score; only a human can tell them apart.

Key insight: the size of the AI–human gap is not what matters. What matters is where the AI cannot tell apart what humans can.

Case study: AI-assisted peer review on 3,408 ICLR papers

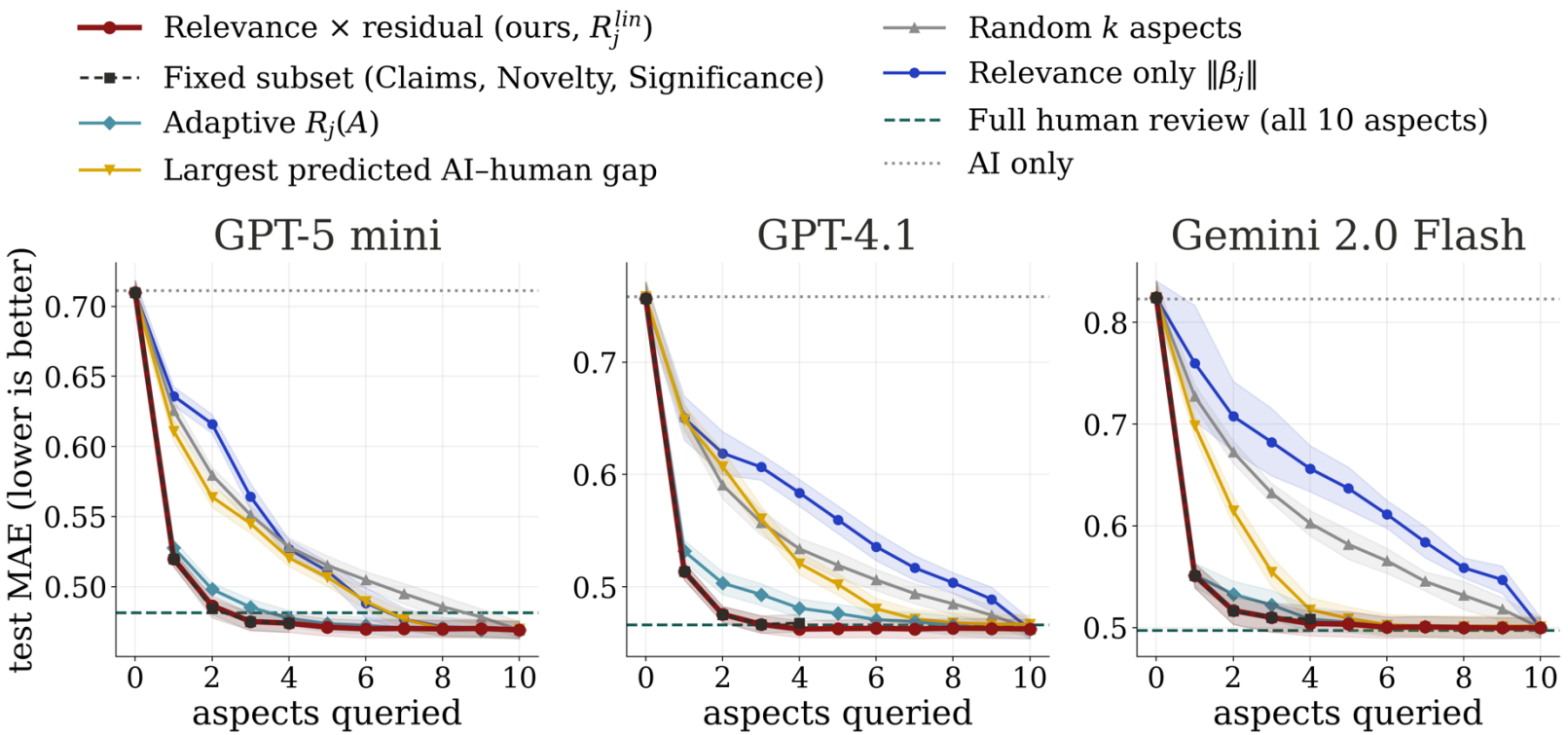
- **Data:** 3,408 ICLR 2019 & 2021 submissions with OpenReview reviews.
- **Target Y:** a paper's average reviewer score (also accept/reject).
- **AI signals:** GPT-5-mini, GPT-4.1, Gemini-2.0 Flash read the full paper; per aspect, a score + a logic code from a fixed dictionary.
- **Human signals:** extracted from real review text onto the same aspects (a retrospective proxy for querying a reviewer).
- **Evaluation:** ridge and gradient boosting; test MAE over 20 random splits + a 2019→2021 split.

10 review aspects (from NeurIPS / ICML guidelines)

Claims & evidence rigor	Reproducibility
Clarity & writing	Experimental design
Significance & impact	Limitations
Originality & novelty	Ethics & societal impact
Related-work positioning	Future-work potential

Predefined logic codes (e.g., novelty_new_algorithm) keep AI signals low-dimensional and comparable across papers.

Three human-scored aspects nearly match full human review



3 of 10

aspects queried: within 0.013 MAE of full human review

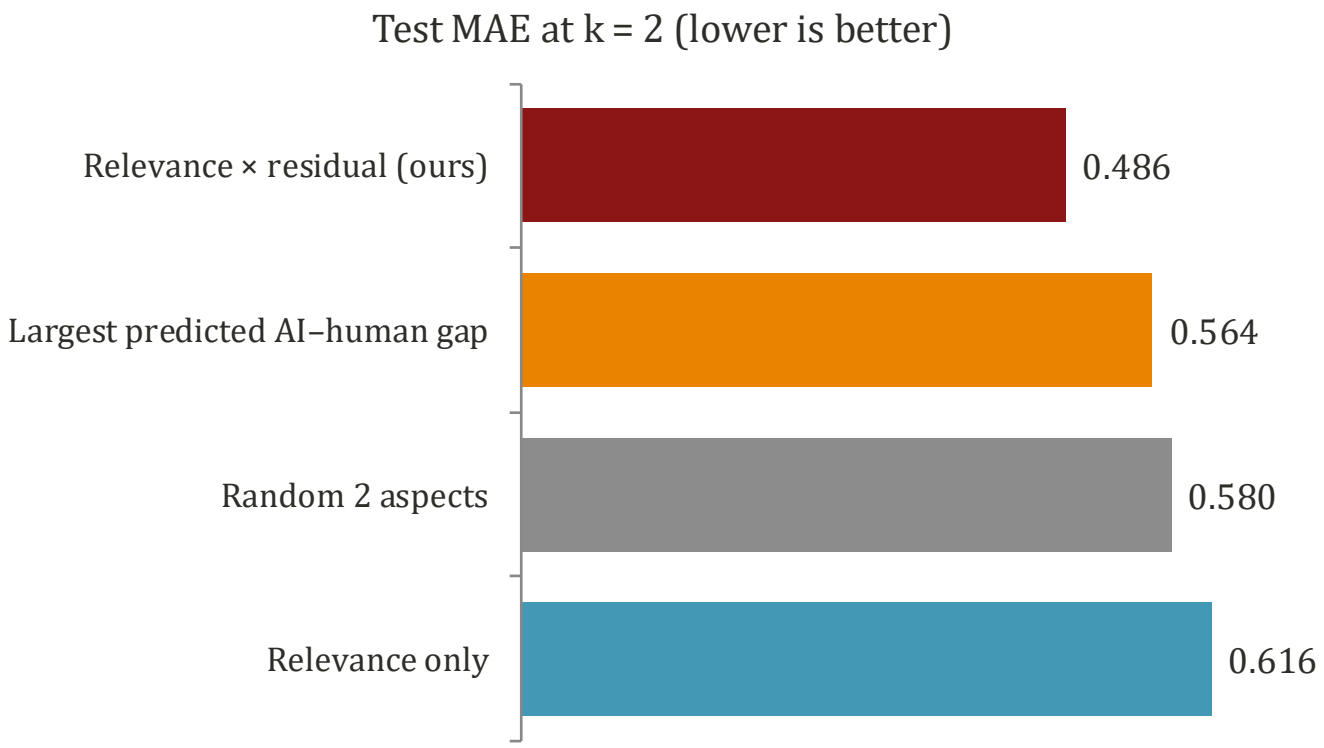
0.71 → 0.48

test MAE: AI only vs. AI + 3 selected aspects

Test MAE vs. number of aspects queried (ridge predictor; mean and 95% CI over 20 random splits).

About 70% of human aspect evaluations can be skipped at almost no loss in accuracy.

Chasing the AI–human gap is the wrong target



GPT-5-mini, ridge. Full human review 0.481 · AI only 0.711. Ours is best for all 3 LLMs × 2 predictors.

Gap ≠ information

A large but systematic AI–human gap is learned from A. Querying there buys little.

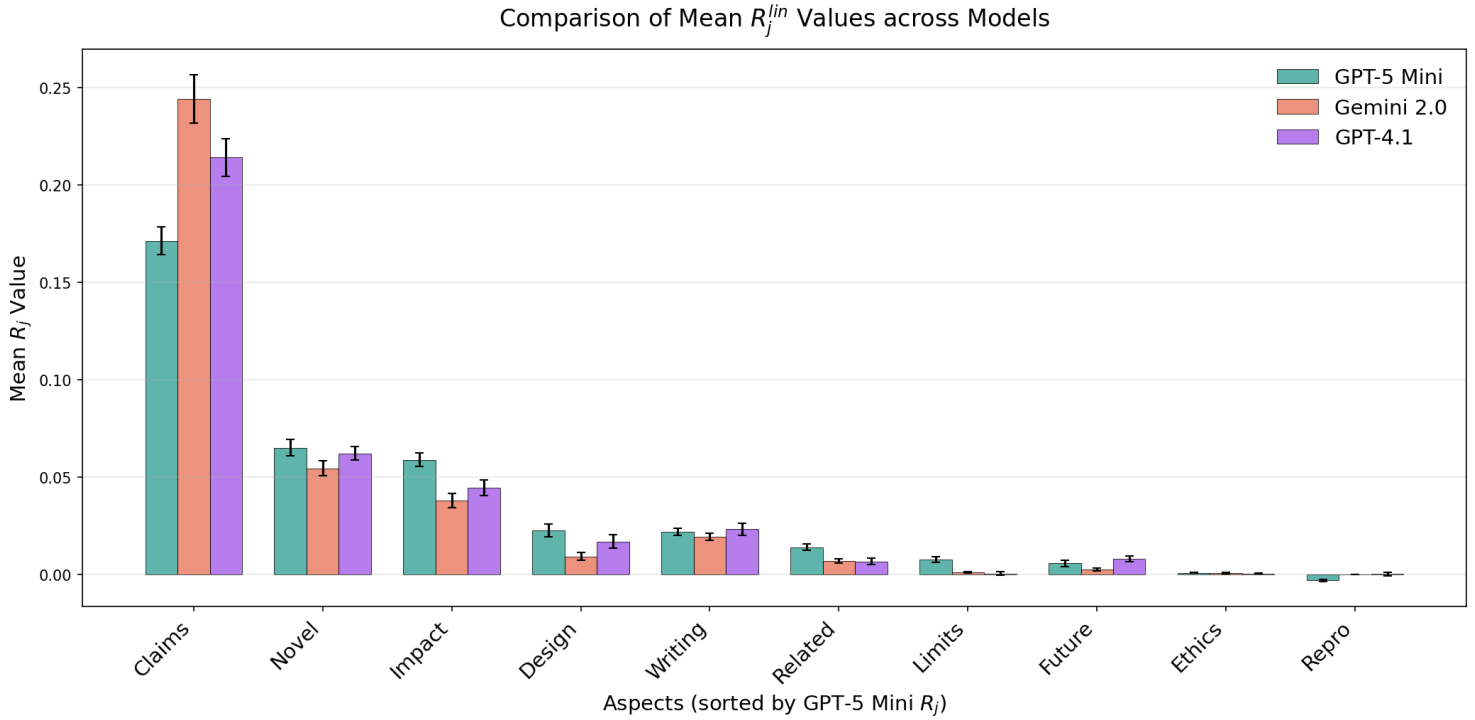
Relevance is not enough

An important factor that the AI already predicts well does not need a human.

Simple beats complex

The closed-form linear rule beats the adaptive nonparametric R_j(A): at this sample size, estimation variance dominates.

Humans are needed most on rigor and novelty



Mean R_j^lin per aspect for each LLM, with error bars. The ranking is consistent across LLMs.

1 Claims & evidence rigor

2 Originality & novelty

3 Significance & impact

- Claims & evidence is selected **first in 100% of splits** for all three LLMs.
- At k = 2, theory papers tend to escalate **significance** where other papers escalate **novelty**.
- These are where LLMs struggle most: verifying claims and judging true novelty.

Robust beyond the headline, and the human stays in charge

Robustness checks

86–95%

of the AI-only → full-signal gap recovered with no reviewer overlap

- **No leakage:** human signal from one reviewer, target from the others; 2–3 aspects.
- **Across years:** train on 2019, test on 2021; same conclusions.
- **Accept/reject:** AUC 0.87–0.91 with AI + 2–3 aspects vs. 0.86–0.88 with all ten human aspects.

The human stays in charge

+1 query

cost of always human-reviewing protected factors (e.g., ethics)

- Returns a recommendation with factor-level evidence, **never the decision.**
- The decision-maker sees which factors no human evaluated, and can override, ask for more, or order a full review.
- **Protected factors** can be fixed in advance as always human-reviewed.

Takeaways

1. **Dispatching human effort has a clean answer:** maximize a contextual reward (Thm 1), estimable with an orthogonal estimator (Thm 2).

2. **Ask humans where relevance × residual uncertainty is large,** not where AI and humans disagree most.

3. **In peer review, 3 of 10 human-scored aspects nearly match full review,** robust across reviewers, years and decision type.

4. **AI as an information intermediary** that directs human oversight, not an autonomous decision-maker.

Next: applications and real-world experiments in other fields where human oversight matters.

Thank you — questions welcome

arxiv.org/abs/2603.13688