Online Appendix

More Versus Better:

Artificial Intelligence, Incentives, and the Coming Crisis in Peer Review

A1 Summary Statistics

[Table A1 about here.]

A2 Pangram Robustness

[Table A2 about here.]

A3 Who uses AI

[Table A3 about here.]

A4 Writing Quality

Text Quality Variable Definitions

Traditional Readability Metrics

Flesch Reading Ease. $206.835 - 1.015 \times (\text{words/sentences}) - 84.6 \times (\text{syllables/words})$. Higher scores indicate easier readability. The formula is unbounded; scores can be negative for dense academic prose with long sentences and polysyllabic words.

Flesch–Kincaid Grade Level. $0.39 \times (\text{words/sentences}) + 11.8 \times (\text{syllables/words}) - 15.59$. Estimates the U.S. school grade level required to comprehend the text.

Gunning FOG Index. $0.4 \times [(\text{words}/\text{sentences}) + 100 \times (\text{complex words}/\text{words})]$. Complex words have three or more syllables, excluding words whose complexity derives solely from common suffixes (*-es*, *-ed*, *-ing*, *-ly*).

SMOG Index. $1.043 \times \sqrt{\text{polysyllables} \times (30/\text{sentences})} + 3.1291$. Polysyllables are words with three or more syllables. For texts with fewer than 30 sentences, the formula extrapolates proportionally.

Style Metrics

Nominalization Density. The count of nominalizations per sentence. Nominalizations are identified by matching word suffixes: (*-tion*, *-sion*, *-ment*, *-ness*, *-ity*, *-ance*, *-ence*, *-acy*, *-ism*, and *-ization*) on words longer than five characters. We exclude 144 common words that share these suffixes but are not nominalizations (e.g., *nation*, *government*.

Passive Voice Ratio. Fraction of sentences containing writing in the passive voice. For instance, *be*-verb (*is*, *are*, *was*, *were*, *been*, *being*) followed by a past participle (words ending in *-ed*, *-en*, or *-t*). Ranges from 0 to 1.

Jargon Density. This is the percent of words matching 70+ management jargon terms and buzzwords (*synergy*, *leverage*, *paradigm*), as well as academic filler phrases (*in terms of*, *plays a role*), and overused verbs (*utilize*, *operationalize*). It is computed as: (jargon matches / total words) $\times$ 100.

Hedging Density. This is the count of hedging expressions per sentence. These expressions come from a dictionary of 40+ terms: modal hedges (*may*, *might*, *could*), approximations (*somewhat*, *relatively*), as well as epistemic phrases (*it seems*, *it appears*), and phrases such as (*we suggest*, *we argue*).

Specificity Score. This score measures the precision of claims. Positive indicators include: numeric values ($\times$2 points), statistical notation including p-values and β coefficients ($\times$3),

methods such as regression (×2), sample size (×3), and theory words, including *equilibrium* or *proposition* (×2). In contrast, vague words (*some*, *many*, *various*) will reduce the score (×1).

A5 Review Topic Focus

***theory*:** theory, theoretical, theorize, theorizing, theorization, mechanism(s), causal mechanism, framework, conceptual/theoretical framework, construct(s), proposition(s), hypothesis/hypotheses/hypothesize, boundary condition(s), moderator/moderating/moderation, mediator/mediating/mediation, antecedent(s), logic/theoretical logic/causal logic, grounded theory, institutional theory, agency theory, resource-based, contingency theory, conceptualization/conceptualize, paradigm, lens/theoretical lens

***contribution*:** contribution(s)/contribute, novelty/novel, new insight(s), incremental, marginal contribution, advance/advancing/advancement, gap(s)/research gap/fill(s) the/a gap, value added/value-added, implication(s)/practical/theoretical implications, significance/significant contribution, original/originality, innovative/innovation, extend/extending/extension, builds on/building on, adds to/add to the literature

***clarity*:** clarity/clear/clearly/unclear/ambiguous, writing/written/well-written/poorly written, readability/readable, prose/language/grammar/grammatical, exposition/flow/organization/organized/structure, confusing/confusion/confused, articulate/articulation, verbose/concise/succinct/wordy, polish/polished/proofread/proofreading, typo(s)/typographical, jargon/accessible/accessibility, presentation/present the, rewrite/rewritten/revise the writing, edit/editing/copyedit

***data*:** data/dataset/data set/sample, survey/questionnaire/interview(s), archival/secondary data/primary data, panel data/cross-sectional/longitudinal, field study/field experiment/lab experiment/laboratory, case study/case studies/ethnograph, observation(s)/

observational, respondent(s)/participant(s), sample size/sampling/sample selection, data collection/data source(s), measurement/measure(s)/operationalize/operationalization, scale(s)/likert, validity/reliability/cronbach, missing data/attrition/response rate, qualitative/quantitative

***empirics*:** empirical/empirically, regression/regress, coefficient(s), estimate/estimation/estimator, endogeneity/endogenous/exogenous, instrumental variable/instrument(s)/iv, fixed effect(s)/random effect, robust/robustness/robustness check, specification/model specification, identification/identification strategy/causal identification, selection bias/omitted variable/reverse causality/simultaneity, difference-in-difference/diff-in-diff/did, propensity score/matching, standard error(s)/clustered, p-value/p value/statistical(ly) significance/significant, interaction/interaction effect/interaction term, control variable(s)/controls, multicollinearity/heteroskedasticity, logit/probit/ols/tobit/poisson/negative binomial, structural equation/sem/hlm, table(s)/results table

[Figure A1 about here.]

[Figure A2 about here.]

A6 Selection or Treatment

[Table A4 about here.]

[Table A5 about here.]

A7 Are AI Papers Reviewed by AI reviewers

[Table A6 about here.]

A8 Editor and Reviewer Margin Change

[Figure A3 about here.]

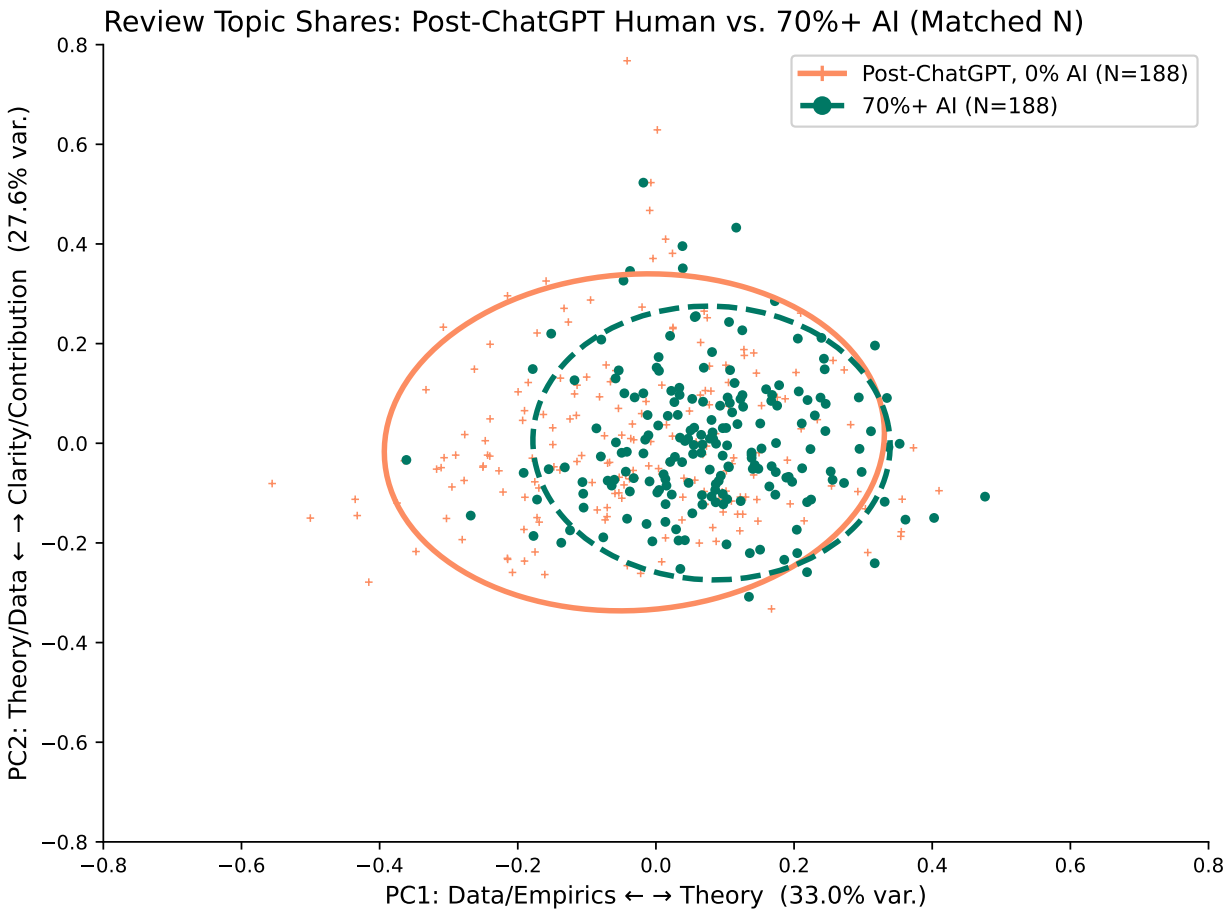

Figure A1: To ensure the wider 95% ellipse is not a mechanical function of a larger sample size of the non-AI group, we compare the 188 high-AI group with a random sample of N=188 0% AI. The results are similar, if not a bit more pronounced.

Figure A2: Kernel densities of review topic focus cut by high- and low-AI use. Likely Human is <30% AI use; and Likely AI is >70% AI use.

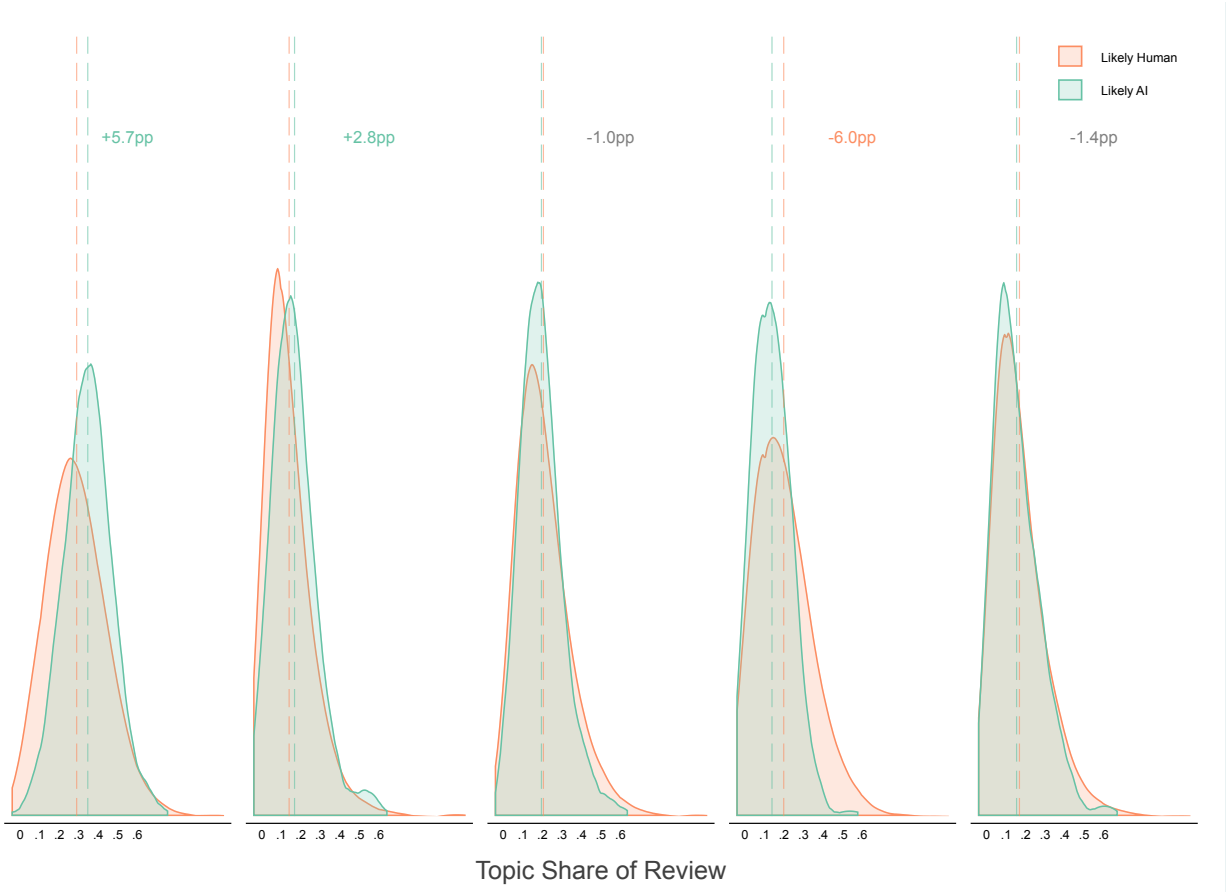

Figure A3: Average number of Deputy Editors, Senior editors and Reviewers, handling manuscripts and their average loads per quarter. The rising submission volume has led to Organization Science significantly increasing the number editors who are responsible for screening manuscripts.

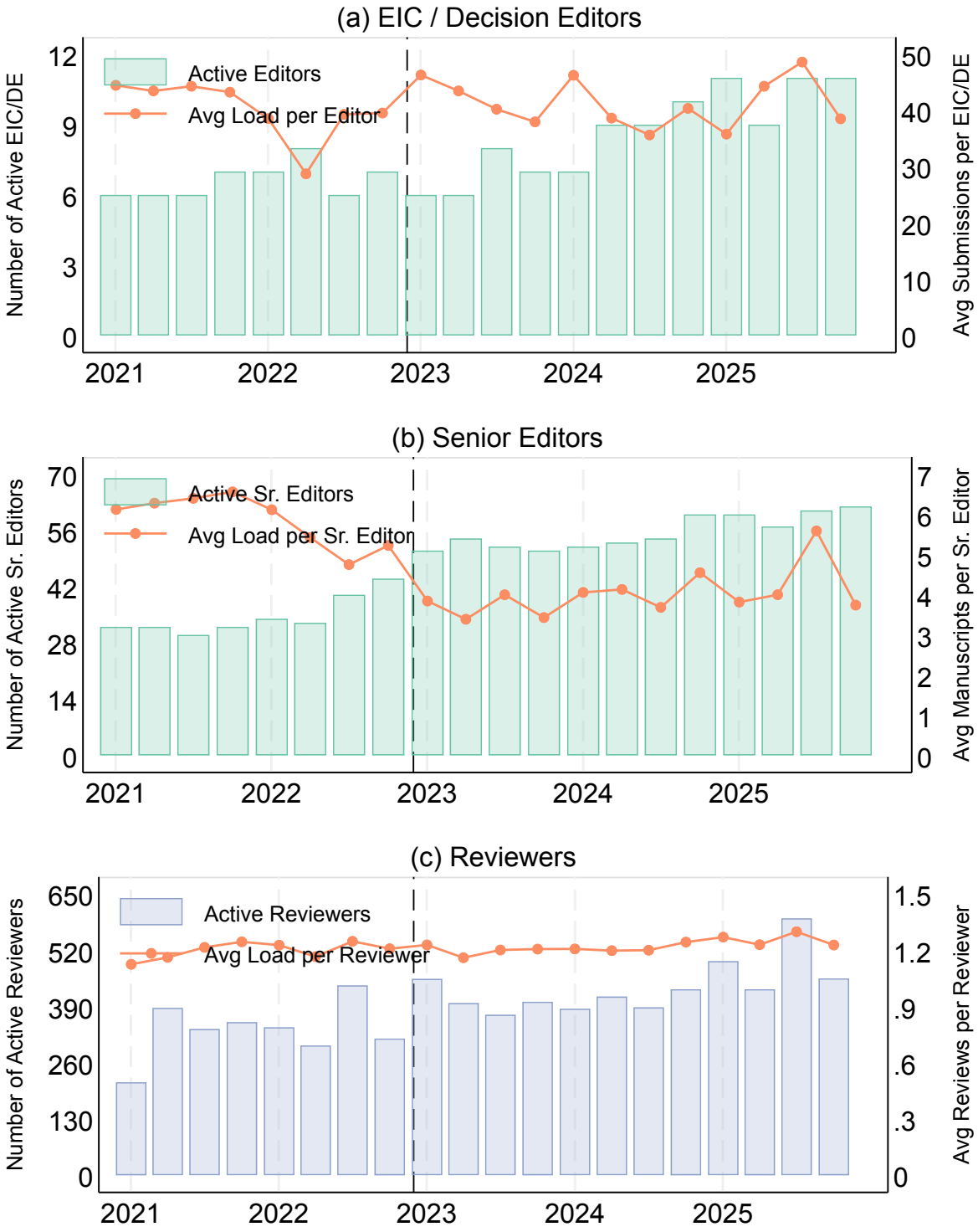

Table A1: Descriptive Statistics

	N	Mean	Std. Dev.	Min	Max
AI Score	6,737	0.154	0.261	0.004	0.991
0-15% AI (<0.15)	6,737	0.693	0.461	0.000	1.000
15%-30% AI (0.15–0.30)	6,737	0.129	0.335	0.000	1.000
30%-70% AI (0.30–0.70)	6,737	0.108	0.310	0.000	1.000
70%+ AI (>0.70)	6,737	0.070	0.256	0.000	1.000
Readability (std FRE)	13,607	-0.003	1.024	-12.944	6.029
All Non-Native English	14,255	0.346	0.476	0.000	1.000
First-Time OS Submitter	14,255	0.710	0.454	0.000	1.000
Desk Rejected	14,255	0.382	0.486	0.000	1.000
Rejected After Review	14,255	0.395	0.489	0.000	1.000
Any Rejection	14,255	0.820	0.385	0.000	1.000

Sample: First submissions only (mseq=1).

Table A2: Is Pangram Detection Picking Up on Non-US Language Speakers or Artificial Intelligence Use?

	(1) AI Score	(2) AI ≥ 0.30	(3) AI > 0.70	(4) AI Score
All Non-Native English	-0.002	-0.002	0.000	-0.003
	(0.004)	(0.002)	(0.001)	(0.003)
Post-ChatGPT	0.051**	0.075**	0.013	
	(0.022)	(0.032)	(0.013)	
Non-Native English × Post-ChatGPT	0.063***	0.094***	0.037**	
	(0.013)	(0.018)	(0.013)	
_cons	0.096***	0.092***	0.047***	0.025***
	(0.014)	(0.019)	(0.008)	(0.001)
Sample				Pre-ChatGPT
Observations	6,725	6,725	6,725	2,022
R-squared	0.149	0.120	0.055	0.002
Editor FE	Yes	Yes	Yes	Yes
Clustered SE	Yes	Yes	Yes	Yes

Standard errors in parentheses

Tests whether AI detection scores are biased against Non-Native English Institution authors.

Models 1–3: DiD with ChatGPT launch as treatment.

Model 4: Placebo test using only pre-ChatGPT submissions.

SE clustered by quarter and editor.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A3: AI Usage as a function of whether author teams include only non-Native English Institution authors and the number of prior submissions (past 15 years) to Organization Science.

	AI Score	0-15% AI	15%-30% AI	30%-70% AI	70%+ AI	Readability
All Non-Native English	0.058***	-0.079***	-0.002	0.039***	0.042***	-0.086***
	(0.016)	(0.022)	(0.010)	(0.010)	(0.011)	(0.022)
Prior OS Submissions (SD)	-0.003*	-0.004	0.011***	-0.003	-0.004**	0.056***
	(0.001)	(0.004)	(0.003)	(0.002)	(0.002)	(0.006)
Constant	0.132***	0.724***	0.128***	0.094***	0.054***	0.034***
	(0.006)	(0.008)	(0.003)	(0.003)	(0.004)	(0.007)
Observations	6,724	6,724	6,724	6,724	6,724	13,594
R-squared	0.252	0.264	0.064	0.082	0.118	0.086
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Editor FE	Yes	Yes	Yes	Yes	Yes	Yes

Standard errors in parentheses

All models use two-way clustered SE by quarter and editor.

Sample: First submissions only (mseq=1).

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A4: Does AI reduce writing quality within Authors?

	(1) OLS	(2) Author FE	(3) Author+Qtr FE	(4) Full FE	(5) Author FE
AI Score	-1.357***	-1.113***	-0.946***	-0.946***	-1.034***
	(0.054)	(0.068)	(0.073)	(0.073)	(0.078)
Constant	0.199***	0.200***	0.178***	0.177***	0.167***
	(0.014)	(0.012)	(0.012)	(0.012)	(0.017)
Sample					Post-ChatGPT
Observations	18,642	10,076	10,076	10,052	6,269
R-squared	0.115	0.490	0.496	0.500	0.525
Author FE	No	Yes	Yes	Yes	Yes
Quarter FE	No	No	Yes	Yes	No
Editor FE	No	No	No	Yes	No

Standard errors in parentheses

DV: Flesch Reading Ease (standardized, mean 0, SD 1). Higher = more readable.

Unit of observation: author × manuscript.

Author FE identifies the effect from within-author variation across manuscripts.

SE two-way clustered by author and manuscript.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A5: The impact of AI on Reject Decisions with Author Fixed Effects.

	(1) Desk Reject	(2) Desk Reject	(3) Any Reject	(4) Any Reject	(5) Reject|Review	(6) Reject|Review
AI Score	0.252***	0.055	0.139***	0.038	0.169**	0.101
	(0.041)	(0.058)	(0.024)	(0.032)	(0.065)	(0.102)
All Non-Native English	0.315***		0.115***		0.066**	
	(0.017)		(0.016)		(0.028)	
Constant	0.212***	0.338***	0.778***	0.828***	0.670***	0.689***
	(0.009)	(0.008)	(0.004)	(0.003)	(0.008)	(0.009)
Observations	2,447	2,447	2,447	2,447	1,371	1,371
R-squared	0.184	0.622	0.130	0.557	0.150	0.581
Quarter FE	Yes	Yes	Yes	Yes	Yes	Yes
Editor FE	Yes	Yes	Yes	Yes	Yes	Yes
Author FE		Yes		Yes		Yes

Sample: First authors with 2+ submissions.

Author FE columns drop All Non-Native English Institution Authors (absorbed by author FE).

All models use two-way clustered SE by quarter and editor.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table A6: Do AI Manuscripts Attract AI Reviewers?

	Review AI Score				
	(1)	(2)	(3)	(4)	(5)
	Mean Review	Mean Review	Mean Review	Min Review	Max Review
Manuscript AI Score	0.176***	0.038***	0.032**	0.023**	0.041
	(0.026)	(0.008)	(0.014)	(0.009)	(0.029)
All Non-Native English		0.007	0.004	0.002	0.008
		(0.005)	(0.005)	(0.003)	(0.008)
Constant	0.057***	0.070***	0.071***	0.022***	0.133***
	(0.017)	(0.003)	(0.002)	(0.001)	(0.003)
Observations	3,470	3,470	3,461	3,461	3,461
R-squared	0.069	0.258	0.306	0.202	0.305
Quarter FE		Yes	Yes	Yes	Yes
Editor FE			Yes	Yes	Yes
Senior Editor FE			Yes	Yes	Yes

Sample: First submissions sent out for review (not desk rejected).

Manuscript AI Score measures AI use in submitted papers.

Review AI Scores measure AI use in peer reviews.

All models use clustered SE at quarter, Senior Editor and Deputy Editor/EIC Levels.

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$