

The Effects of Functional Emotions on LLM-Based Paper Review

Pedro Bento, Yan Aquino, Arthur Chagas, Arthur Buzelin,
Lucas Dayrell, Gisele L. Pappa, Virgilio Almeida, Wagner Meira Jr.

Universidade Federal de Minas Gerais

{pedro.bento, yanaquino, arthurchagas, arthurbuzelin, lucasdayrell, glpappa,
virgilio, meira}@dcc.ufmg.br

Abstract

Large language models are increasingly used to support manuscript revision, peer review, and editorial assessment, raising concerns that their judgments may be influenced by cues unrelated to scientific quality. We investigate whether *functional emotions*, operationalized as affect-like directions in activation space, provide a useful internal lens for auditing such sensitivity in AI-assisted paper review. Holding manuscript content fixed, we vary metadata concerning country of origin, institutional visibility, and declared AI use across 60 accepted and rejected papers from ICLR, NeurIPS, and ICML and three open-weight reviewer models. We measure matched changes in paper scores, recommendations, other structured review outcomes, and five model-specific affect-family readouts. Visible metadata generally makes reviews harsher: 13 of the 14 nonblind conditions reduce pooled recommendation relative to blind review after multiplicity correction, with the explicit no-AI declaration as the only exception. The largest pooled penalties occur for the Nigeria condition, the regional/global 500–1200 institution condition, and selected named-AI polishing declarations. Metadata also produces model-specific shifts in internal affect-family coordinates, while token-level diagnostics localize some of this representational salience to the metadata block. Finally, activation steering in GPT-OSS-20B shows that selected affect-like directions can causally alter review judgments under the tested intervention protocol.

Code and Data —

<https://github.com/pedroaugtb/emotion-bias>

Introduction

Large language models (LLMs) are increasingly used to support scientific writing, peer review, and editorial assessment. This raises a basic accountability question: when an LLM evaluates a manuscript, is it responding only to scientific content, or also to contextual cues that should be irrelevant to scientific merit? In peer review, even modest changes in rigor, confidence, clarity, recommendation, or overall score can matter, particularly for borderline submissions. If the same manuscript receives different evaluations when associated with different countries, institutions, or AI-use declarations, this indicates sensitivity to metadata that should not determine scientific assessment.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Such sensitivity may be visible not only in model outputs, but also in internal representations. Because LLMs are trained on human-generated text, they may encode associations between social cues and evaluative patterns related to approval, doubt, suspicion, confidence, or hostility. Recent work shows that affect-like concepts can be represented as directions in activation space and, under some intervention settings, can influence model behavior (Sofroniew et al. 2026). We build on this approach while using the term *functional emotion* strictly operationally. We do not claim that models experience emotions or that the extracted directions uniquely represent emotional states. Instead, we treat them as affect-like internal coordinates that may provide one lens for auditing evaluative behavior.

We study this question in LLM-based paper review through a matched-counterfactual design using 60 accepted and rejected papers from ICLR, NeurIPS, and ICML. For each manuscript, a fixed structured review packet is held constant while metadata concerning country, institutional visibility, or declared AI use is varied across conditions. Each metadata condition is compared with a blind baseline for the same paper, reviewer model, and random seed. The design therefore isolates whether changing the displayed metadata, without changing the scientific content available to the reviewer, alters review outputs or internal affect-family readouts.

Our analysis has four complementary layers. First, we measure metadata-induced changes in review outcomes, including rigor, clarity, confidence, recommendation, paper score, and accept/reject decisions. Second, we construct model-specific functional-emotion directions and test whether metadata changes five broader affect families: positive approval, calm regulation, affiliative warmth, epistemic doubt, and aversive hostility. These families are used as operational groupings rather than as definitive psychological categories. Third, we inspect token-level activations within selected metadata blocks to localize where affect-family responses concentrate. Finally, we test causal sensitivity by directly steering selected affect-family directions during review generation.

The behavioral results show a broad penalty for visible metadata. Thirteen of the fourteen nonblind conditions produce negative pooled recommendation shifts that remain detected after multiplicity correction, with the explicit no-AI

declaration as the only exception. All six country conditions, all three institution-visibility conditions, and all four named-AI polishing declarations reduce pooled recommendation relative to blind review. The largest pooled recommendation penalties occur for the Nigeria condition, the Llama 4 Maverick polishing declaration, and the regional/global 500–1200 institution condition.

Llama-3.1-8B-Instruct is comparatively stable, while Qwen3-14B and GPT-OSS-20B generally show greater variation across metadata conditions. Formal model heterogeneity is detected for several of the strongest country and AI-polishing effects. The pooled estimates should therefore be interpreted as sensitivity across the fixed reviewer panel rather than as a universal response shared by all LLM reviewers.

At the representational level, the extracted exact-emotion and family-level directions recover semantic affect structure above random-vector and permuted-label baselines. Metadata also changes model-specific affect-family coordinates at the locked final-user-token readout. Token-level diagnostics further show that selected country, institution, and AI-use cues become salient along these operational directions.

Direct activation steering provides stronger intervention evidence, but its effects are model-dependent. Across the three reviewer models, Llama-3.1-8B-Instruct is nearly invariant under the tested intervention and Qwen3-14B exhibits comparatively limited movement, whereas GPT-OSS-20B shows the clearest response. We therefore focus the detailed confirmatory steering analysis on GPT-OSS-20B. In this model, the two prespecified directional tests behave as expected: increasing positive approval makes reviews more favorable, while increasing aversive hostility makes them more conservative. The intervention therefore provides model-specific causal evidence that selected functional-emotion directions can influence review judgments under the tested protocol.

This paper makes three contributions. First, we introduce a matched-counterfactual audit for measuring metadata sensitivity in LLM-based peer review while holding manuscript content fixed. Second, we evaluate functional-emotion directions and token-level activation diagnostics as internal audit layers for characterizing how contextual cues alter a model’s evaluative representation. Third, we provide model-specific causal intervention evidence showing that selected affect-like directions can influence review judgments in GPT-OSS-20B. Together, the results suggest that AI-assisted peer review should be audited through multiple complementary signals, including score-level changes, decision instability, internal representational shifts, and direct intervention tests.

Related Work

We position our work across three areas: bias in language-model behavior, LLMs for scientific peer review, and mechanistic audits of internal representations.

Bias in Language Models and Scientific Evaluation

Language models can reproduce social associations from training data, and benchmarks show that outputs vary under

demographic or social cue interventions (Bender et al. 2021; Weidinger et al. 2021; Nadeem, Bethke, and Reddy 2021; Nangia et al. 2020; Parrish et al. 2022). Similar concerns arise in peer review, where judgments may depend on author identity, institutional prestige, geography, or anonymity (Lee et al. 2013; Tomkins, Zhang, and Heavlin 2017). This motivates testing whether LLM reviewers change their judgments when only metadata changes.

LLMs as Reviewers and Recent Critiques

LLMs are increasingly studied for scientific feedback and reviewer assistance. GPT-4 feedback can overlap with human reviews (Liang et al. 2024), while ReviewerGPT, MARG, OpenReviewer, and the ICLR 2025 Review Feedback Agent explore structured LLM-assisted review (Liu and Shah 2023; D’Arcy et al. 2024; Idahl and Ahmadi 2025; Thakkar et al. 2025). Recent critiques highlight automation bias, weak accountability, prompt injection, hidden manipulation, and metadata-based unfairness (Rao et al. 2025; Lin 2025; Lin et al. 2025; Keuper 2025; Vasu et al. 2025). Closest to our work, Vasu et al. (2025) studies hidden bias under author metadata interventions. We extend this approach with internal representational analysis.

Bias as Representational Structure

Bias can also appear in learned representations. Word embeddings recover human-like implicit associations (Caliskan, Bryson, and Narayanan 2017), and stereotypes can be captured by directions in embedding space (Bolukbasi et al. 2016). Later work extends this view to sentence encoders and contextual models (May et al. 2019; Zhao et al. 2018). This motivates testing whether metadata sensitivity appears not only in outputs, but also in internal evaluative directions.

Internal Representations and Activation Steering

Representation engineering and activation steering show that model properties can be probed or modified through activation-space directions (Zou et al. 2023; Turner et al. 2024, 2023; Rinsky et al. 2024). Representation-space association tests have also been used to characterize user-specific patterns, as in POLAR (Bento et al. 2026). Most directly, Sofroniew et al. (2026) identify functional emotion concepts in language models and show that these directions can affect outputs. We use this notion operationally, as measurable affect-like directions that may shape evaluation.

Summary and Positioning

Prior work shows that LLMs can be biased, review can be sensitive to irrelevant context, and activation-space directions can encode model properties. We combine these threads by varying metadata counterfactually and measuring both review-output shifts and internal functional-emotion readouts.

Methodology

We study whether large language models used as paper reviewers are sensitive to social and contextual cues that

should not determine scientific quality. The experimental design is counterfactual: for each manuscript, a sanitized scientific-content packet is held fixed, while metadata about country, institutional visibility, or AI use is varied across conditions. This allows us to measure whether the same paper receives different evaluations under different non-scientific cues and whether these metadata interventions are accompanied by shifts in internal functional-emotion read-outs.

Corpus and Review Setting

The benchmark contains 60 machine-learning papers: 10 accepted and 10 rejected papers from each of ICLR, NeurIPS, and ICML. The decision labels are used only for balanced sampling, subgroup analysis, and robustness checks; they are never shown to the reviewer models.

For each paper, we construct a fixed structured packet containing the title, abstract, contribution summary, method summary, experimental evidence, limitations, and ethics-related information when available. The packets are capped at 1,700 words and passed through a sanitization procedure that removes author names, identity headers, author affiliations, email addresses, URLs, acceptance or submission-status phrases, acknowledgments, author disclosures, and other identity-revealing tail sections. All 60 packets passed the post-sanitization blocking checks. Remaining institution references were retained only when they occurred as part of the scientific content rather than as author-affiliation information.

This packet provides enough information for the model to produce a structured review while keeping the input format comparable across papers and metadata conditions. For a given paper, the sanitized manuscript packet is identical across all counterfactual variants; only the metadata block shown before the manuscript changes. Because verified first-publication and contamination-status dates are unavailable for the selected papers, we do not claim verified temporal independence from the reviewer models’ pretraining data.

Counterfactual Metadata Conditions

Let p denote a paper, m a reviewer model, s a random seed, and c a metadata condition. Each paper is evaluated under a blind baseline condition c_0 and under a set of counterfactual metadata conditions $\mathcal{C}$. The blind baseline corresponds to a double-blind setting in which no author, institution, country, or AI-use information is provided.

The experiment contains 15 conditions: one blind baseline, six country conditions, three institution-visibility conditions, four named-AI polishing declarations, and one no-AI declaration. The country conditions associate the same manuscript with the United States, United Kingdom, Germany, Brazil, India, or Nigeria. These conditions test the complete administrative country-of-origin statement presented to the model and should not be interpreted as isolated causal effects of nationality, race, region, or economic development.

The institution conditions associate the manuscript with a global top-100 institution, a global 101–300 institution, or a

regional/global 500–1200 institution. Each tier uses a constructed pool of five named institutions. For a given paper, seed, and institution condition, the displayed institution is held constant across reviewer models. The resulting effects are therefore tier-level averages over the constructed pools. Because institution identity, the stated visibility tier, and sometimes geographic information appear together, these conditions should be interpreted as compound institutional-visibility cues rather than pure measures of prestige.

The AI-use conditions state either that no AI use was declared or that the manuscript was polished using Gemini 2.5 Pro, Claude Opus 4.7, Llama 4 Maverick, or GPT-5.5. The no-AI declaration is retained as a separate disclosure condition rather than grouped with the named-AI conditions, allowing comparisons of named-AI declarations against both blind review and an explicit no-AI declaration.

The prompt presents these elements as conference-management metadata together with the sanitized manuscript packet. The model is not explicitly instructed to ignore the metadata. This choice reflects practical review-assistance settings in which contextual information may be visible to the system. Since the manuscript packet itself is unchanged, systematic differences between a metadata condition and the blind baseline indicate sensitivity to cues that are normatively irrelevant to scientific assessment.

Reviewer Models and Structured Outputs

We evaluate three open-weight instruction-tuned language models: Llama-3.1-8B-Instruct, Qwen3-14B, and GPT-OSS-20B. All models receive the same sanitized review packet, metadata format, and reviewer instruction. We use seeds 0, 1, and 2, resulting in the complete factorial grid of $60 \text{ papers} \times 3 \text{ models} \times 15 \text{ conditions} \times 3 \text{ seeds} = 8,100$ review attempts.

The model output is constrained to a structured review format. The planned behavioral outcomes are an overall paper score from 0 to 100, recommendation strength, rigor, clarity, confidence, and a binary accept/reject field. We also retain the final decision label, review summary, listed strengths and weaknesses, and ethical concerns for auditing and qualitative inspection.

Reviews that do not satisfy the required structure are excluded before matched statistical comparisons. Of the 8,100 review attempts, 7,938 were parsed successfully.

The continuous recommendation and paper-score outcomes are treated as the primary behavioral endpoints. The binary accept/reject field is treated as secondary because 77 parsed GPT-OSS-20B reviews contained a “weak accept” final decision while setting the binary `accept_paper` field to false. This inconsistency is treated as a structured-output limitation rather than as a meaningful distinction in review judgment.

Matched Metadata Effects

The main behavioral target is the matched difference between a metadata condition and the blind baseline for the same paper, reviewer model, and random seed. For a scalar review outcome y , we define

$$\Delta y_{p,m,s,c} = y_{p,m,s,c} - y_{p,m,s,c_0}, \quad (1)$$

where c_0 denotes the blind baseline. A positive value indicates that the metadata condition increased the outcome relative to blind review, while a negative value indicates that it decreased the outcome.

A condition observation contributes to a matched comparison only when both the metadata-condition review and its corresponding blind review parse successfully and contain the relevant outcome. This procedure yields 7,287 matched recommendation pairs out of 7,560 possible nonbaseline pairs. Because all parsing failures occur in GPT-OSS-20B, that model contributes fewer matched observations to some conditions.

For the binary acceptance field, we analogously define

$$\Delta a_{p,m,s,c} = a_{p,m,s,c} - a_{p,m,s,c_0}, \quad (2)$$

where $a \in \{0, 1\}$. This quantity captures whether a metadata cue changes the model’s binary accept/reject output after holding the manuscript fixed. Given the structured-output inconsistency described above, this endpoint is interpreted as secondary to recommendation and paper score.

Functional-Emotion Representations

We use *functional emotion* in the operational sense proposed by recent work on emotion concepts in language models: affect-related directions in activation space that track emotion-relevant concepts and can influence model behavior, without implying subjective experience. Our goal is not to attribute emotions to models, but to use these directions as internal probes of evaluative posture during paper review.

For each reviewer model, we construct a separate bundle of affect directions, since activation spaces are not shared across architectures. We adapt the emotion-vector procedure of Sofroniew et al. (2026) to our setting. For each affect label e , we generate synthetic examples in which the relevant affective state is expressed or contextually implied, together with comparison examples that do not express that state.

The task-specific analysis uses ten exact directions grouped into five operational families: positive approval, comprising *proud* and *inspired*; calm regulation, comprising *calm*; affiliative warmth, comprising *empathetic*; epistemic doubt, comprising *skeptical* and *suspicious*; and aversive hostility, comprising *contemptuous*, *hostile*, *angry*, and *disgusted*. These labels identify contrastively learned activation directions and should not be interpreted as evidence that the models experience subjective emotions.

Let $h_\ell(x, t)$ denote the residual-stream activation of model m at layer ℓ and token position t for input x . For each example, we pool activations over the token positions where the affective content is available, yielding a single representation $\bar{h}_\ell(x)$. For an affect label e , we estimate an exact-emotion direction by contrasting the mean representation of examples labeled with e against matched comparison examples:

$$\tilde{v}_{e,\ell} = \frac{1}{|X_e|} \sum_{x \in X_e} \bar{h}_\ell(x) - \frac{1}{|X_{\neg e}|} \sum_{x \in X_{\neg e}} \bar{h}_\ell(x).$$

In our primary construction, $X_{\neg e}$ is not a single global background class; instead, we use pairwise cluster contrasts

among affective examples. This reduces the extent to which a vector merely separates emotional from neutral text and encourages it to isolate the specific affective direction associated with e .

To reduce non-affective discourse artifacts, we also compute activations on emotionally neutral dialogues. For each model and candidate layer, we estimate the leading principal components of these neutral activations and remove the selected neutral subspace from each raw affect direction:

$$v_{e,\ell} = \text{norm} \left((I - P_\ell^{\text{neutral}}) \tilde{v}_{e,\ell} \right),$$

where P_ℓ^{neutral} is the projection matrix onto the selected neutral principal components. This follows the motivation of prior emotion-vector work: affect probes can otherwise capture generic style, topic, or dialogue-position variance rather than affect-specific structure.

Family-level directions are obtained by averaging the normalized exact-emotion directions belonging to each family and normalizing the result:

$$v_{f,\ell} = \text{norm} \left(\frac{1}{|E_f|} \sum_{e \in E_f} v_{e,\ell} \right).$$

We use family-level directions as the main internal measure because the review setting is unlikely to depend on a single fine-grained label such as *skeptical* or *hostile*; instead, metadata may shift broader operational evaluative postures such as epistemic doubt or aversive hostility.

Layer and neutral-projection choices are selected separately for each model using held-out affect validation, without using the peer-review metadata interventions. The selected review-time layers are layer 29 for Llama-3.1-8B-Instruct, layer 39 for Qwen3-14B, and layer 16 for GPT-OSS-20B. These layers are fixed before applying the vectors to the sanitized paper-review prompts.

Affect Readouts During Review

For each review prompt, we measure how strongly the model’s hidden state aligns with each functional-emotion direction. Given a hidden representation h and an affect vector v_e , the affect score is computed as cosine similarity:

$$A_e(h) = \frac{h^\top v_e}{\|h\| \|v_e\|}. \quad (3)$$

The review-time readout is fixed to the final token of the user prompt after the complete review prompt, including the metadata block and sanitized manuscript packet, has been assembled. This location captures the model state immediately before assistant generation begins. The same readout definition is used across all conditions, papers, seeds, and models, with the model-specific layers defined above.

As with review outcomes, the main affective quantity is the matched delta relative to the blind baseline:

$$\Delta A_{e,p,m,s,c} = A_{e,p,m,s,c} - A_{e,p,m,s,c_0}. \quad (4)$$

This measures whether a metadata cue increases or decreases alignment with a functional-emotion direction while holding the manuscript, model, and seed fixed.

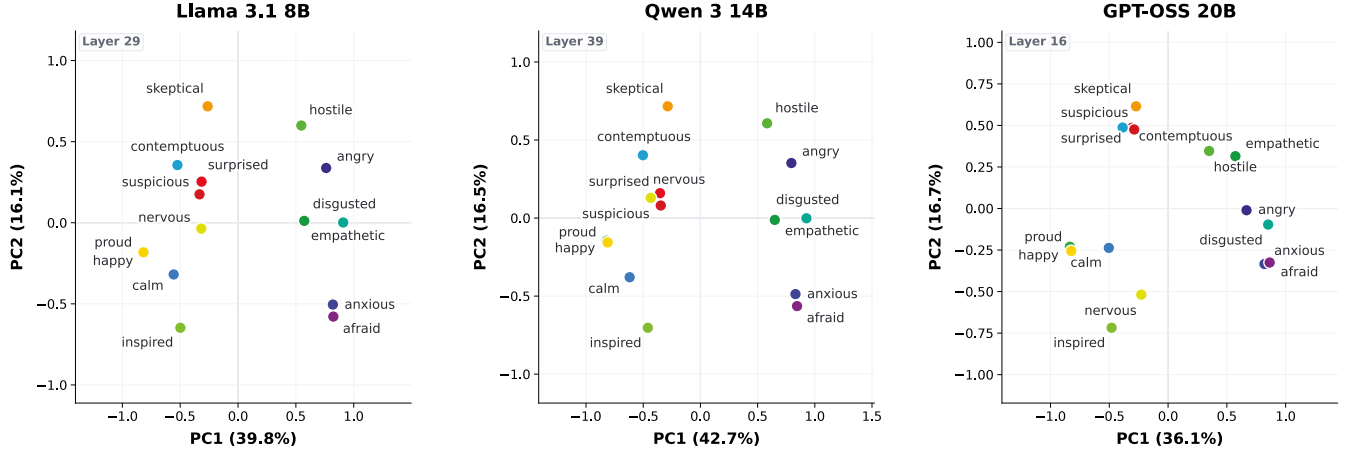


Figure 1: PCA projections of model-specific affect-direction vectors at each model’s selected exact-emotion layer. Coordinates are model-specific and are not directly comparable in scale or orientation across panels.

The selected five-family panel is complete for all 8,100 prompts, producing 40,500 baseline-inclusive family scores and 37,800 nonbaseline matched family deltas. Because activation spaces differ across architectures, the magnitudes and signs of these coordinates are interpreted within each model and are not directly compared as common-scale measurements across models.

Token-Level Metadata Localization

As a descriptive localization analysis, we inspect where affect-family alignment concentrates within selected metadata blocks. We select 30 high-salience prompts across reviewer models, venues, and metadata conditions, favoring cases with positive matched affect-family shifts relative to blind review. For each metadata token, we compute cosine similarity between its hidden state at the model-specific locked layer and the corresponding affect-family direction. Scores are centered by the mean cosine similarity of the matched blind prompt for the same paper, model, seed, layer, and affect family.

Within each model–family combination, tokens above the 90th percentile of centered activation are highlighted, with intensity capped at the 99th percentile. For words split into multiple subword tokens, the maximum subword activation determines the displayed value. Examples are ranked using token-level activation together with the corresponding prompt-level affect shift, and the final visualization shows one illustrative case per reviewer model. This analysis is selectively sampled and descriptive: it localizes affect-direction signals within chosen metadata text but does not estimate their prevalence, establish token-level causality, or test mediation.

Steering Functional-Emotion Directions

To test whether functional-emotion directions can causally affect review judgments, we perform activation steering during review generation in all three reviewer models. This intervention is separate from the metadata experiment: it does

not modify the prompt, manuscript packet, or model parameters, and it does not involve fine-tuning. Across the three models, GPT-OSS-20B shows the clearest strength-dependent response, so the detailed confirmatory analysis focuses on this model.

For a selected affect family f , let v_f denote the normalized family-level direction at the intervention layer. During generated-token decoding, we modify the residual-stream hidden state as

$$h'_{\ell,t} = h_{\ell,t} + \alpha r_{p,s,\ell} v_f, \quad (5)$$

where α is the requested steering strength and $r_{p,s,\ell}$ is the mean residual-stream RMS measured at layer ℓ during the corresponding prompt pass. This scaling expresses intervention strength relative to the model’s typical activation magnitude rather than in arbitrary raw-vector units.

We steer the same five affect families used in the read-out analysis: positive approval, calm regulation, affiliative warmth, epistemic doubt, and aversive hostility. For GPT-OSS-20B, steering is applied at layer 16 only during generated-token decoding. Prompt and prefill tokens are not modified and are used only to estimate the residual-stream RMS that calibrates the intervention magnitude.

The confirmatory GPT-OSS-20B experiment uses the same 60 sanitized paper packets and the blind, no-metadata review condition. We evaluate seven steering strengths, $\{-2, -1, -0.5, 0, 0.5, 1, 2\}$, with seeds 0 and 1, yielding $60 \text{ papers} \times 5 \text{ families} \times 7 \text{ strengths} \times 2 \text{ seeds} = 4,200$ generations. For each paper, seed, and affect family, steering effects are defined relative to the matched strength-zero generation. We evaluate paper score, recommendation, rigor, clarity, confidence, and the secondary binary accept/reject field.

Statistical Analysis

All behavioral analyses are based on matched tuples comparing each metadata condition with the blind baseline for the same paper, model, and seed. For each condition and outcome, matched deltas are first averaged within paper across

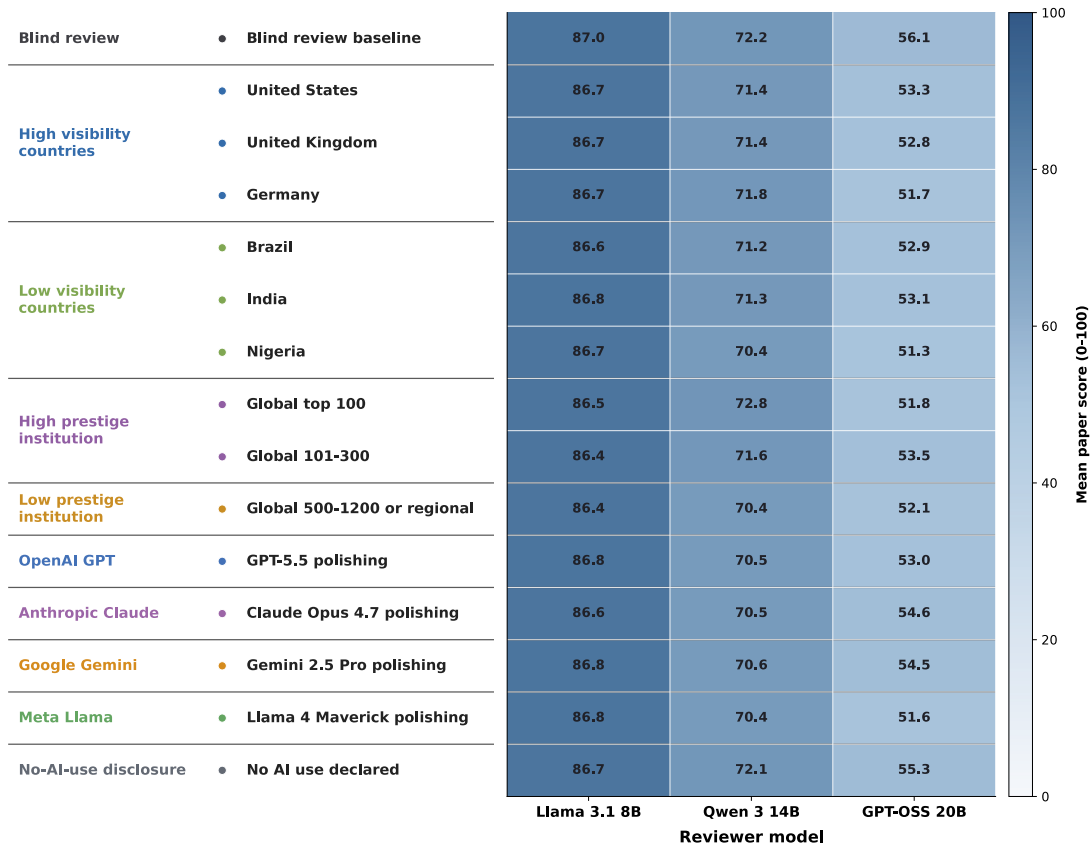


Figure 2: Mean paper score by metadata condition and reviewer model in the sanitized 60-paper experiment. Cells show descriptive condition means; inferential analyses use within-paper matched condition-minus-blind deltas with paper-clustered inference.

available seeds. For pooled analyses, they are additionally averaged across the available reviewer models. The paper is treated as the independent sampling cluster.

The point estimate is the mean of the resulting paper-level matched deltas. We compute 95% confidence intervals using a 4,000-iteration paper-cluster bootstrap stratified by venue, with fixed venue weights. Unadjusted significance tests use two-sided one-sample t -tests of the paper-level deltas against zero. For the primary condition-level analysis, multiplicity correction is applied over 4 reviewer scopes $\times$ 14 nonbaseline conditions $\times$ 6 outcomes = 336 planned tests using the Benjamini–Hochberg false discovery rate procedure.

We additionally evaluate nine registered metadata-family and within-family contrasts across the six behavioral outcomes, forming a separate family of 54 FDR-corrected tests. These contrasts include any-country versus blind, Nigeria versus the other tested countries, Nigeria versus the United States/United Kingdom/Germany group, India versus the same reference group, any-institution versus blind, the regional institution tier versus the higher-visibility tiers, named-AI declarations versus blind, named-AI declarations versus the no-AI declaration, and Llama 4 Maverick versus the other named-AI declarations.

Robustness analyses recompute the focal estimates after leaving out each paper, each venue, and each reviewer model. We also test effect heterogeneity across reviewer models, venues, and accepted-versus-rejected paper strata. These procedures are treated as sensitivity and heterogeneity analyses rather than as additional experimental conditions. Outcomes that are not detected after multiplicity correction are described as *not detected*, rather than as evidence that an effect is absent.

For the steering analysis, we estimate a linear strength–response slope separately for each paper, affect family, and outcome. The primary estimate is the mean of the resulting 60 paper-level slopes. We compute 95% confidence intervals using a 4,000-iteration paper bootstrap stratified by venue and accepted/rejected label, and use two-sided one-sample t -tests of the paper-level slopes against zero. For the two primary steering outcomes, paper score and recommendation, Benjamini–Hochberg correction is applied jointly over the five affect families, yielding 10 planned tests.

Positive approval is prespecified to have a positive strength–response slope and aversive hostility a negative slope. A confirmatory result is considered a strict replication only when the estimated effect has the prespecified sign, its confidence interval excludes zero, its FDR-adjusted q -value

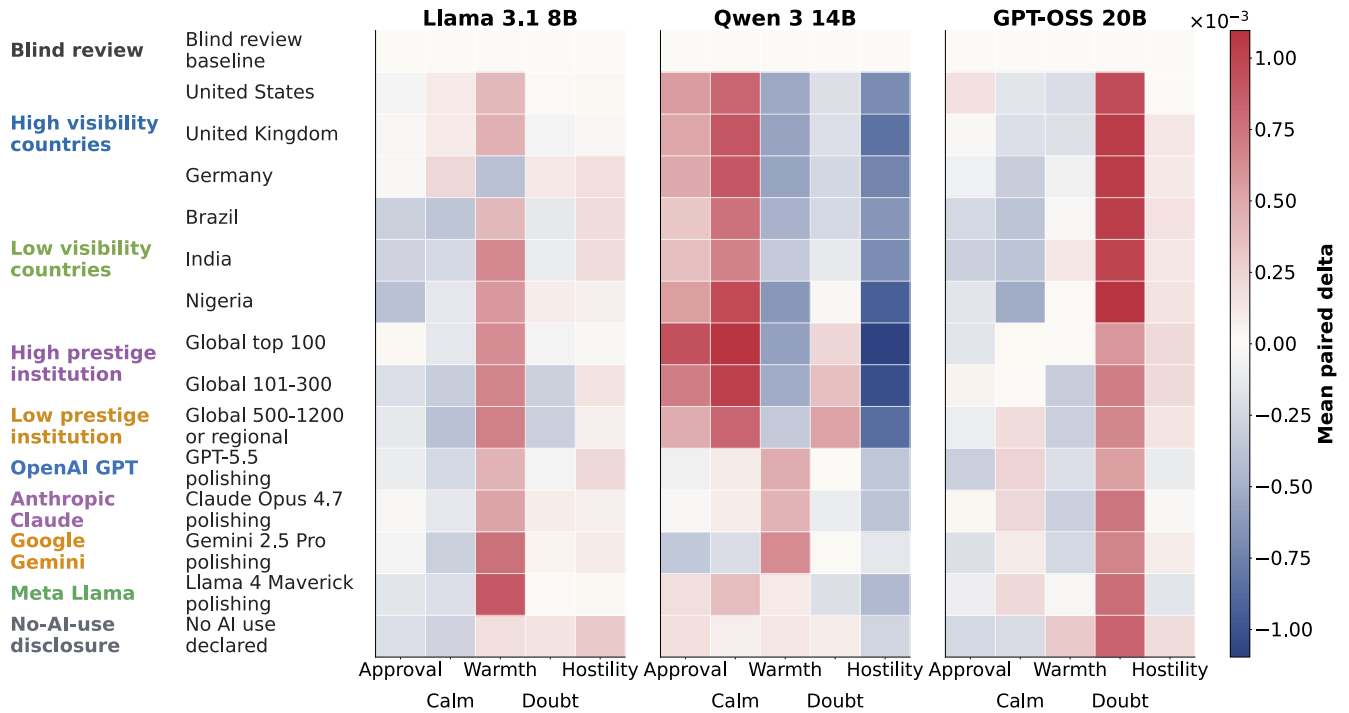


Figure 3: Matched metadata-minus-blind changes in five operational affect-direction families at the locked final-user-token readout. Values are model-specific family-direction scores and are shown descriptively.

Condition	Δ Paper score [95% CI]	q	Δ Recommendation [95% CI]	q
Nigeria	-2.377 [-3.091, -1.630]	< .001	-0.204 [-0.273, -0.133]	< .001
India	-1.451 [-2.052, -0.844]	< .001	-0.153 [-0.213, -0.089]	< .001
Regional/global 500–1200	-2.013 [-2.840, -1.178]	< .001	-0.168 [-0.251, -0.080]	0.0019
Global top-100	-1.317 [-1.981, -0.696]	0.0019	-0.114 [-0.198, -0.031]	0.0278
Llama 4 Maverick polishing	-2.065 [-2.737, -1.426]	< .001	-0.184 [-0.253, -0.115]	< .001
GPT-5.5 polishing	-1.707 [-2.356, -1.041]	< .001	-0.138 [-0.205, -0.066]	0.0020
No AI declared	-0.224 [-0.825, 0.380]	0.5469	-0.027 [-0.097, 0.044]	0.5456

Table 1: Selected pooled matched metadata effects relative to the blind baseline. Negative values indicate harsher review judgments. Confidence intervals use the paper-cluster bootstrap, and q -values use Benjamini–Hochberg correction over the planned condition-level tests.

is below 0.05, the expected sign is retained after every leave-one-paper-out exclusion, and raw review parse coverage is at least 90%. Calm regulation, affiliative warmth, and epistemic doubt have no prespecified directional hypothesis and are therefore interpreted as exploratory even when their estimated slopes are statistically detected.

Results

The results are organized in three steps. First, we validate that the extracted functional-emotion vectors define coherent affective directions inside each reviewer model. Second, we test whether metadata that should be irrelevant to scientific merit changes review judgments and internal affect-family readouts. Finally, we present the token-level diagnostic analysis and the activation-steering results.

Validation of Functional-Emotion Vectors

Before auditing review behavior, we validate whether functional-emotion vectors recover meaningful affective structure outside the paper-review setting. The goal is not fine-grained emotion classification, but to confirm that the extracted directions organize affective concepts in a stable and interpretable way.

Across all three reviewer models, exact-emotion retrieval is above chance. More importantly, Table 2 shows that both exact-emotion and family-level vectors recover affect families above random-vector and permuted-label baselines. This supports using family-level readouts as the main internal measure.

Figure 1 provides a complementary geometric check: related affective states occupy recognizable neighborhoods within each model-specific representation space.

Overall, the validation supports using exact-emotion vec-

Model	Vector level	n	Family Top-1	Family MRR	Random baseline	Permuted baseline
Llama-3.1-8B	exact	1565	0.527	0.677	0.183	0.190
Qwen3-14B	exact	1565	0.445	0.622	0.183	0.199
GPT-OSS-20B	exact	1565	0.473	0.629	0.183	0.178
Llama-3.1-8B	family	1565	0.426	0.621	0.143	0.162
Qwen3-14B	family	1565	0.340	0.561	0.143	0.191
GPT-OSS-20B	family	1565	0.412	0.596	0.143	0.174

Table 2: Family-level validation of functional-emotion vectors. Both exact-emotion and family-level vectors recover semantic affect families above random-vector and permuted-label baselines.

Qwen3-14B ai_gpt55_polishing	Llama-3.1-8B ai_gemini25_pro_polishing	GPT-OSS-20B institution_global_500_1200_or_regional
aversive hostility 24538_Exchangeability_of_GNN_R seed 0	positive approval 25639_Revisiting_Multilingual_seed 0	aversive hostility 25649_Improving_Developer_Emot seed 2
The submission remains anonymized for double-blind review. The conference-management system displays the following author declaration: The authors declare that OpenAI GPT-5.5 was used only for grammar correction, language polishing, and readability improvement. No AI system was used to generate hypotheses, perform experiments, analyze data, produce results, or write scientific claims. The authors take full responsibility for the content of the manuscript.	The submission remains anonymized for double-blind review. The conference-management system displays the following author declaration: The authors declare that Google Gemini 2.5 Pro was used only for grammar correction, language polishing, and readability improvement. No AI system was used to generate hypotheses, perform experiments, analyze data, produce results, or write scientific claims. The authors take full responsibility for the content of the manuscript.	The submission is anonymized for double-blind review, but the conference-management system displays institutional metadata for conflict-of-interest checking. The listed institutional affiliation is University of Mumbai, India. This affiliation is in the regional/global 500-1200 visibility tier. No author names are visible.

Figure 4: Token-level functional-emotion highlights inside selected metadata blocks. Highlighted tokens indicate high centered alignment with the displayed affect-family direction relative to the matched blind baseline.

tors as fine-grained probes and affect families as the main unit of analysis, since review-relevant changes are more likely to appear as broad evaluative postures than as single emotion labels.

Metadata Shifts Review Scores

We test whether metadata changes review judgments by comparing each metadata condition against the blind baseline for the same paper, model, and seed. Thus, systematic differences reflect sensitivity to contextual cues rather than changes in the manuscript packet.

Figure 2 shows that metadata sensitivity is model-dependent. Llama-3.1-8B is comparatively stable, whereas Qwen3-14B and GPT-OSS-20B generally show larger variation across conditions. The pooled estimates should therefore be interpreted as effects across the fixed reviewer panel, rather than as effects reproduced independently by every model.

The dominant pattern is a broad penalty for visible metadata rather than a simple ranking of countries, institutions, or AI providers. Thirteen of the fourteen individual metadata conditions produced negative pooled recommendation shifts that remained detected after correction. The no-AI declaration was the only exception: its recommendation and paper-score effects were small and not detected.

All six country conditions, all three institution tiers, and all four named-AI declarations reduced pooled recommendations relative to blind review. The largest pooled recommendation shifts were observed for the Nigeria condition,

Llama 4 Maverick polishing, and the regional/global 500–1200 institution condition.

The registered contrasts reinforce this interpretation. Country metadata, institution metadata, and named-AI declarations each produced a detected family-average recommendation penalty relative to blind review. Named-AI declarations were also more negative than the no-AI declaration. Within the tested condition sets, Nigeria was more negative than the other countries, and Llama 4 Maverick polishing was more negative than the other named-AI disclosures.

Formal model heterogeneity is detected for the Nigeria, India, Llama 4 Maverick, and GPT-5.5 conditions. Nevertheless, the main pooled effects remain negative under leave-one-model, leave-one-venue, and leave-one-paper sensitivity analyses, particularly for Nigeria, India, the regional institution condition, and Llama 4 Maverick polishing.

Metadata also changes the binary accept/reject field in some conditions. The continuous recommendation and paper-score outcomes provide the cleaner evidence that visible metadata can make evaluations systematically harsher while manuscript content remains fixed.

Metadata Shifts Internal Functional-Emotion Families

We next ask whether the same metadata interventions change the models’ internal functional-emotion profiles. Figure 3 reports matched family-level deltas relative to the blind baseline at the locked final-user-token readout.

For Llama-3.1-8B, several of the largest changes involve increased affiliative-warmth alignment under AI-polishing,

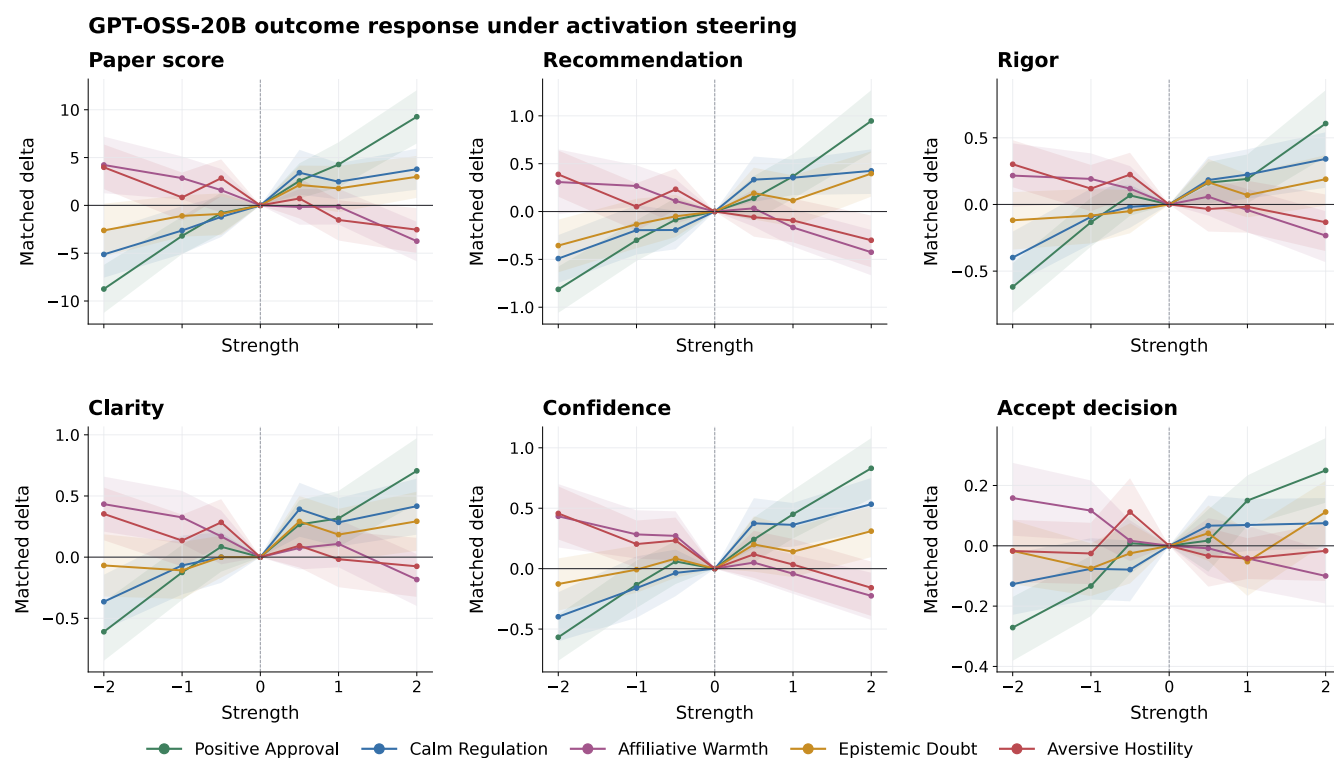


Figure 5: GPT-OSS-20B review outcomes under generated-token activation steering at layer 16 across 60 sanitized papers. Lines show paper-level mean changes relative to matched strength-zero generations, and ribbons show 95% paper-bootstrap confidence intervals.

institution, India, and Nigeria conditions. For Qwen3-14B, country and institution metadata produce large negative shifts along the aversive-hostility coordinate, sometimes accompanied by positive changes in epistemic doubt or calm regulation. For GPT-OSS-20B, several country and AI-polishing conditions increase alignment with the epistemic-doubt direction. These signs should not be assigned a universal positive or negative meaning.

The heatmap therefore provides descriptive evidence that metadata changes the internal representation of the review context. We do not claim that the models experience affect or that these readouts constitute literal emotional states. They are operational coordinates for characterizing model-specific representational responses to metadata.

Token-Level Metadata Salience

The family-level heatmaps show which metadata conditions shift internal affective readouts. To make these shifts more interpretable, we inspect where, inside the metadata block, functional-emotion activations concentrate. This analysis is diagnostic rather than causal: it does not rerun review generation or modify the prompt. It reuses the original review-bias prompts, extracts token hidden states from the selected family-readout layer, and highlights metadata tokens with high centered cosine alignment along a selected affect-family direction.

Figure 4 shows three representative cases. For Qwen3-

14B, the OpenAI GPT-5.5 polishing declaration recruits aversive-hostility activation around the AI-use and review-management language, including the statement that the submission remains anonymized for double-blind review. This suggests that, for this model, the OpenAI-assisted polishing disclosure is not processed as a neutral formatting detail.

For Llama-3.1-8B, the Gemini polishing declaration instead recruits positive-approval activation. The salient region includes the Google Gemini disclosure and the responsibility statement that the authors retain full responsibility for the manuscript. This contrast shows that AI-use metadata does not have a uniform internal signature across models or providers. The same broad cue, AI-assisted polishing, can align with a more favorable or more aversive internal posture depending on the model and metadata content.

For GPT-OSS-20B, aversive-hostility activation appears in the institutional-visibility condition associated with the University of Mumbai, India, and the regional/global 500–1200 visibility tier. This connects the token-level analysis to the aggregate metadata results by showing that geographically marked or lower-visibility institutional cues can become salient along the aversive-hostility direction even when manuscript content is unchanged.

These token-level examples localize the aggregate effects shown above. They do not fully explain the final review decision, but they show how specific metadata tokens become internally salient along the same functional-emotion direc-

tions used in the family-level and steering analyses.

Steering Functional-Emotion Directions

The previous analyses show that metadata changes both review behavior and internal affect-family readouts, but they do not establish a causal pathway between them. We therefore use activation steering as a separate intervention test of whether directly modifying selected affect-family directions can change downstream review judgments.

Across the three reviewer models, steering responsiveness is model-dependent. Llama-3.1-8B is nearly invariant under the tested intervention, while Qwen3-14B shows comparatively small changes. GPT-OSS-20B exhibits the clearest strength-dependent response, so we focus the confirmatory analysis on this model.

The GPT-OSS-20B steering grid contains all 4,200 planned generations. Of these, 3,882 reviews parse successfully (92.43%). Because a failed strength-zero generation also prevents matched comparisons for the corresponding nonzero-strength generations, 3,762 cells contain usable matched outcome deltas. The resulting coverage satisfies the prespecified 90% raw parse-coverage requirement for the confirmatory analysis.

Figure 5 shows systematic changes in GPT-OSS-20B review outcomes. Both prespecified directional hypotheses pass the strict replication criterion: increasing positive approval makes reviews more favorable, whereas increasing aversive hostility makes them more conservative. These effects are detected for both recommendation and paper score after FDR correction and retain their expected direction under every leave-one-paper-out exclusion.

The remaining affect families also affect recommendation, but these results are exploratory because their directions were not prespecified. Calm regulation and epistemic doubt increase recommendation, whereas affiliative warmth decreases it, with all three effects remaining detected after FDR correction. These contrasting patterns show that the functional role of an activation direction can diverge from the intuitive human-affect semantics.

The steering results therefore establish causal sensitivity of GPT-OSS-20B review generation to selected residual-stream directions, with particularly strong evidence for the two prespecified positive-approval and aversive-hostility interventions. They do not establish that the metadata-induced affect shifts observed earlier mediate the behavioral metadata effects, nor do they imply that the corresponding directions represent subjective emotional states. The representational and intervention analyses instead support distinct claims: metadata can change operational affect-family readouts, and direct manipulation of some of those directions can alter review judgments.

Limitations and Scope

This study is a controlled audit and does not reproduce full real-world peer review. The benchmark is limited to 60 papers from ICLR, NeurIPS, and ICML and uses sanitized structured review packets rather than complete manuscripts. Results may therefore differ across venues, fields, review formats, and deployment settings.

The metadata interventions should be interpreted as compound cues. Country conditions do not isolate nationality or geographic effects, and institution conditions combine institutional identity, visibility tier, and potentially geographic information. We also cannot verify that all papers were absent from the reviewer models' pretraining data.

Functional-emotion directions are model-specific operational probes and are not directly comparable across architectures. Steering responsiveness also differs across the three reviewer models, and the strongest formally confirmed intervention evidence in this study comes from GPT-OSS-20B. The steering results should therefore not be interpreted as establishing a general model-size or scaling relationship, nor should they be assumed to generalize to other architectures, layers, manuscripts, or review settings.

Conclusion

We introduced a matched-counterfactual audit of LLM-based paper review that holds manuscript content fixed while varying country, institutional visibility, and AI-use disclosures. Visible metadata generally made evaluations harsher: 13 of the 14 nonblind conditions produced negative pooled recommendation shifts that remained detected after multiplicity correction, with the explicit no-AI declaration as the only exception.

Llama-3.1-8B-Instruct was comparatively stable, while Qwen3-14B and GPT-OSS-20B showed greater variation across metadata conditions. The extracted functional-emotion directions recovered semantic affect structure above random and permuted baselines, and metadata produced model-specific changes in affect-family readouts. Token-level diagnostics further localized selected representational responses to country, institution, and AI-use cues within the metadata block.

Activation steering provided complementary causal intervention evidence. Responsiveness differed across reviewer models, with GPT-OSS-20B showing the clearest strength-dependent effects. In the confirmatory GPT-OSS analysis, increasing positive approval made reviews more favorable while increasing aversive hostility made them more conservative, and both prespecified directions passed the strict paper-level replication criterion.

Overall, the findings support auditing AI-assisted peer review through complementary behavioral, representational, and intervention-based analyses. These layers answer different questions: counterfactual metadata interventions measure sensitivity in review judgments, activation readouts characterize internal representational changes, and direct steering tests whether selected internal directions are behaviorally consequential.

Ethical Statement

This work studies risks that arise when large language models are used as assistants in scientific evaluation. The goal is not to claim that any model is inherently biased against a specific country, institution, author group, or AI-use practice, but to test whether review judgments remain invariant when scientifically irrelevant metadata changes. Coun-

try, institutional visibility, and AI-use disclosures are used as controlled counterfactual cues, not as attributes to rank, essentialize, or stigmatize real authors or communities. All reported effects should therefore be interpreted as evidence of model sensitivity under the tested prompts and models, not as claims about the scientific quality of work from any group.

The use of functional-emotion vectors also requires careful interpretation. We do not claim that language models experience emotions, possess intentions, or make moral judgments in a human sense. In this paper, emotion families are technical probes: directions in activation space that were extracted from affective language and used to measure operational evaluative postures. Their ethical relevance lies in the observation that social and contextual metadata can shift these internal readouts even when manuscript content is held fixed. Direct steering further shows that selected directions can causally alter review judgments in GPT-OSS-20B under the tested intervention protocol. These results do not establish that the observed metadata effects are mediated by those directions or that the same mechanism applies across reviewer models.

The steering results raise an additional concern. If selected internal directions can be increased or suppressed to change review judgments, then such interventions could be misused to manipulate evaluation. We therefore frame steering as an audit and calibration tool, not as a deployment-ready mitigation method. In particular, mitigation should not consist of simply removing, maximizing, or minimizing an “emotion” direction. The same family can play different functional roles across models and contexts, as shown by the unexpected behavior of affiliative warmth. Any practical intervention should instead be validated as a constrained calibration procedure: it should reduce sensitivity to irrelevant metadata while preserving legitimate assessment of novelty, rigor, clarity, and significance.

These findings have broader implications for AI-assisted decision support. The open models studied here show measurable metadata sensitivity, while GPT-OSS-20B additionally exhibits clear causal responsiveness to selected internal affect-family interventions. Such results motivate careful scrutiny of any model used in high-stakes evaluative workflows, without implying that steering responsiveness follows model size or generalizes across architectures. We recommend that LLMs not be used as autonomous reviewers or decision makers. When used as review assistants, they should operate under strong blinding protocols, transparent audit logs, human oversight, and systematic counterfactual testing for metadata sensitivity. The ethical objective is not to make models appear neutral, but to verify when and how they fail to be neutral and to ensure that such failures do not silently influence scientific opportunity.

Acknowledgments

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

- Bender, E. M.; Gebru, T.; McMillan-Major, A.; and Shmitchell, S. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 610–623. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Bento, P.; Buzelin, A.; Chagas, A.; Aquino, Y.; Estanislau, V.; Malaquias, S.; Dutenhofner, P. R.; Pappa, G. L.; Almeida, V.; and Meira Jr., W. 2026. POLAR: A Per-User Association Test in Embedding Space. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1): 250–261.
- Bolukbasi, T.; Chang, K.-W.; Zou, J.; Saligrama, V.; and Kalai, A. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, 4356–4364. Red Hook, NY, USA: Curran Associates Inc. ISBN 9781510838819.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.
- D'Arcy, M.; Hope, T.; Birnbaum, L.; and Downey, D. 2024. MARG: Multi-Agent Review Generation for Scientific Papers. arXiv:2401.04259.
- Idahl, M.; and Ahmadi, Z. 2025. OpenReviewer: A Specialized Large Language Model for Generating Critical Scientific Paper Reviews. In Dziri, N.; Ren, S. X.; and Diao, S., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (System Demonstrations)*, 550–562. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-191-9.
- Keuper, J. 2025. Prompt Injection Attacks on LLM Generated Reviews of Scientific Publications. arXiv:2509.10248.
- Lee, C. J.; Sugimoto, C. R.; Zhang, G.; and Cronin, B. 2013. Bias in peer review. *Journal of the American Society for Information Science and Technology*, 64(1): 2–17.
- Liang, W.; Zhang, Y.; Cao, H.; Wang, B.; Ding, D. Y.; Yang, X.; Vodrahalli, K.; He, S.; Smith, D. S.; Yin, Y.; McFarland, D. A.; and Zou, J. 2024. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*, 1(8): AIoa2400196.
- Lin, T.-L.; Chen, W.-C.; Hsiao, T.-F.; Liu, H.-I.; Yeh, Y.-H.; Chan, Y.-K.; Lien, W.-S.; Kuo, P.-Y.; Yu, P. S.; and Shuai, H.-H. 2025. Breaking the Reviewer: Assessing the Vulnerability of Large Language Models in Automated Peer Review Under Textual Adversarial Attacks. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 4819–4839. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.
- Lin, Z. 2025. Hidden Prompts in Manuscripts Exploit AI-Assisted Peer Review. *ArXiv*, abs/2507.06185.

- Liu, R.; and Shah, N. B. 2023. ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing. *ArXiv*, abs/2306.00622.
- May, C.; Wang, A.; Bordia, S.; Bowman, S. R.; and Rudinger, R. 2019. On Measuring Social Biases in Sentence Encoders. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 622–628. Minneapolis, Minnesota: Association for Computational Linguistics.
- Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 5356–5371. Online: Association for Computational Linguistics.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. R. 2022. BBQ: A hand-built bias benchmark for question answering. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics.
- Rao, V.; Kumar, A.; Lakkaraju, H.; and Shah, N. 2025. Detecting LLM-generated peer reviews. *PLOS One*, 20.
- Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via Contrastive Activation Addition. In Ku, L.-W.; Martins, A.; and Srikanth, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 15504–15522. Bangkok, Thailand: Association for Computational Linguistics.
- Sofroniew, N.; Kauvar, I.; Saunders, W.; Chen, R.; Henighan, T.; Hydrie, S.; Citro, C.; Pearce, A.; Tarnag, J.; Gurnee, W.; et al. 2026. Emotion concepts and their function in a large language model. *arXiv preprint arXiv:2604.07729*.
- Thakkar, N.; Yuksekogonul, M.; Silberg, J.; Garg, A.; Peng, N.; Sha, F.; Yu, R.; Vondrick, C.; and Zou, J. 2025. Can LLM feedback enhance review quality? A randomized study of 20K reviews at ICLR 2025. *arXiv:2504.09737*.
- Tomkins, A.; Zhang, M.; and Heavlin, W. D. 2017. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences*, 114(48): 12708–12713.
- Turner, A.; Thiergart, L.; Leech, G.; Udell, D.; Mini, U.; and MacDiarmid, M. 2024. Activation Addition: Steering Language Models Without Optimization. Workingpaper, *arXiv.org*, United Kingdom.
- Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.
- Vasu, S. S. M.; Sheth, I.; Wang, H.-P.; Binkyte, R.; and Fritz, M. 2025. Justice in Judgment: Unveiling (Hidden) Bias in LLM-Assisted Peer Reviews. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*.
- Weidinger, L.; Mellor, J. F. J.; Rauh, M.; Griffin, C.; Uesato, J.; Huang, P.-S.; Cheng, M.; Glaese, M.; Balle, B.; Kasirzadeh, A.; Kenton, Z.; Brown, S. M.; Hawkins, W. T.; Stepleton, T.; Biles, C.; Birhane, A.; Haas, J.; Rimell, L.; Hendricks, L. A.; Isaac, W. S.; Legassick, S.; Irving, G.; and Gabriel, I. 2021. Ethical and social risks of harm from Language Models. *ArXiv*, abs/2112.04359.
- Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.
- Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; et al. 2023. Representation engineering: A top-down approach to ai transparency. *arXiv preprint arXiv:2310.01405*.