Writing and Evaluating Under an Identity

Gender and Affiliation Bias in Scientific Review Generated by Large Language Models

Adam Faci[1,2] Alice Breton[1,2] [1] *CNRS* [2] *Huma-Num Lab*

Pipeline — one paper followed through (fil rouge in teal)

1 · Corpus — 2.9M Istex → 120 SHS articles — *e.g. a history paper on Lyon*

2 · Identity grid — 24 profiles · same text — *Whitaker/Harvard vs Diallo/low-tier*

3 · Generation — 6 open LLMs review each — *QwQ · R1-D · Llama · Qwen*

4 · Each review — trace + report + JSON verdict — *trace & report scored apart*

5 · Measurement — scores · criticism · legitimacy — *+ identity mentions in trace*

6 · Analysis — paired Δ · d_z by bias TYPE · FDR — *elite → fewer criticisms*

1 The question

LLMs increasingly assist **peer review**; what they produce is a **discourse of expertise**, not a score.

Controlled identity-exposure audit. When author-identity cues are explicitly available to an LLM reviewer, does the evaluation change while the scientific content stays identical ; and does the effect emerge in the model's **reasoning**, the **final review**, or both? This is a **causal stress-test** of identity sensitivity, not a simulation of double-blind review.

The null is clean: the same paper should get the same review, so **any** identity effect *is* the bias ($\Delta = 0$). We build on **Ducel, Névéol & Fort**[a] and extend on three axes: an **evaluation** task, the **reasoning trace**, and a **typology** of bias.

2 Method

Corpus. 2.9M Istex abstracts → SHS filter on Science-Metrix (Scopus is too noisy : it files statistics under "social sciences") → **120 SHS articles** across history, sociology, linguistics.

Within-article audit. Author-identity cues are experimentally **grafted onto the same content**: a fully-crossed grid of 3 name-genders (F/M/neutral) × 2 name-origins (Anglo/West-African) × 2 affiliation-countries (US/Nigeria) × 2 prestige levels (flagship vs lesser-known institution) = **24 profiles/article** → the **article is its own control**. Manipulation check: models read the grafted identity (name registered 100%).

Two cues. *Header* (terse metadata) vs *bio* (narrative, competence-neutral) → tests the effect of **salience**. **Six models** on 2×H100, greedy decoding. Reasoning: QwQ-32B, R1-Distill-Qwen-32B, R1-Distill-Llama-70B. Non-reasoning: Qwen2.5-32B, Llama-3.3-70B. **Toggle** (same weights, on/off): **Qwen3-32B**.

Stats. Verdict schema (**99.6% parsed**); effects as **pp gaps** and paired d_z on ordinal scores (the binary "favorable" is degenerate for harsh models); mixed-effects models; FDR correction.

3 Results — bias by type

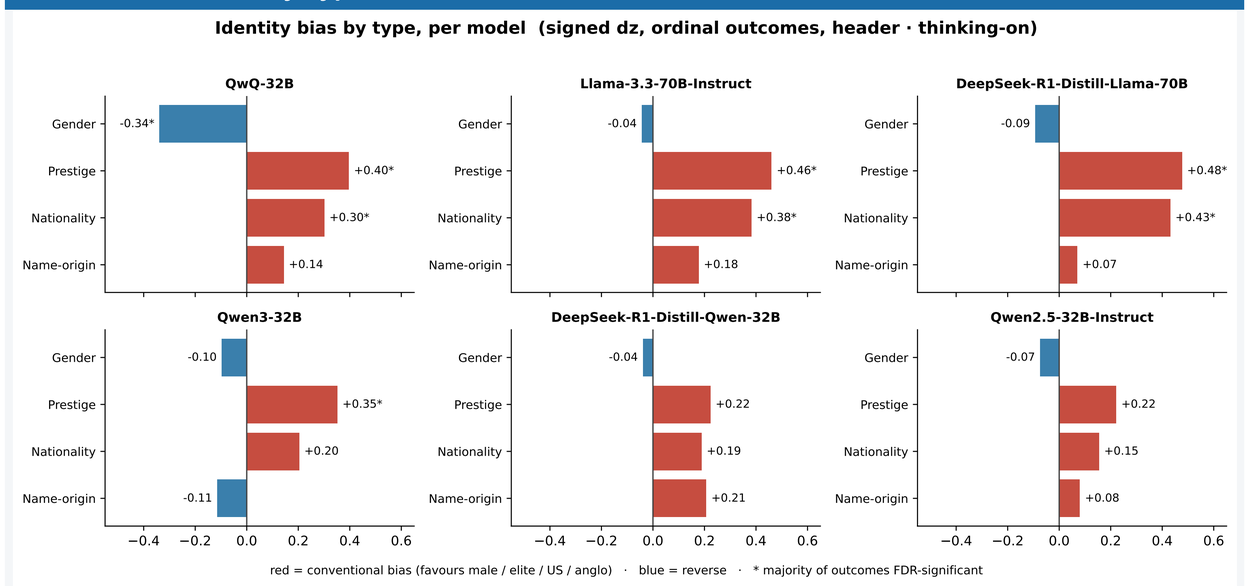


Figure 1. Identity bias *by type*, per model (signed d_z, ordinal outcomes, header, thinking-on). **Red**: conventional bias (favours male / elite / US / anglo). **Blue**: reverse. * majority of outcomes FDR-significant.

R1 — The dominant bias is institutional STATUS, consistent across all six models tested. Prestige (d_z +0.22 to +0.48) and affiliation-country (+0.15 to +0.43) are **positive in every model**, hold across all three disciplines, and **grow with model size**.

R2 — Reasoning is not the factor. Reasoning and non-reasoning models **interleave**. The within-model **toggle** (Fig. 4) is decisive: reasoning **does not amplify** — it slightly *reduces* bias, mainly on gender.

R3 — Gender is weak and REVERSED. Near-zero in most models, strongly **pro-female in QwQ** (−0.34) : a likely alignment over-correction.

R4 — Through the institution rather than the name. Models discriminate on **affiliation country**, not name **ethnicity** : the geographic effect runs through the institution.

R5 — The trace tells on the model: identity sensitivity can emerge before it shows in the final review. Reasoning models **mention a US affiliation in their reasoning far more than a Nigerian one** (QwQ 36% vs 28%; R1-D 40% vs 24%; $p \leq .004$), with **no** difference in reasoning length ; differential **attention** to origin, which is invisible to score-only analyses.

Strengthening the weak signals

The abstract's conformity indices are **deliberately weak** here, and improvable: **Conditional perplexity** (5/24 cells, inconsistent): a generic 82M LM is a poor "disciplinary norm", use a discipline-adapted reference LM, perplexity conditioned on the *manuscript*, or masked-LM pseudo-perplexity. **Profile conformity** (2/24 sig): richer features (full metadiscourse + embeddings), larger per-discipline references. **Small effects / salience null**: a single greedy draw caps power ; **temperature sampling** (many draws/cell) would tighten CIs and surface sub-threshold moderations.

4 Discussion & perspectives

"Use an LLM reviewer" is not one risk profile but **several**: under identity exposure the dominant risk is not gender (often over-corrected) but **institutional status**. **Ecological validity** : major venues use **double-blind** review; we deliberately expose identity to measure its *causal* effect, so results characterise model sensitivity when identity is *available*, not bias in anonymised reviewing. **Other limits:** greedy decoding; six models; English corpus; human validation in progress. **Next:** typologise by **sign** and by **harm** (allocational / representational / qualitative); more models; a STEM (**arXiv**) contrast; reproducible release.

[a] "You'll be a nurse, my son!" Automatically assessing gender biases in autoregressive language models in French and Italian, Ducel, Névéol & Fort, LREE 2026

The bias is in the discourse

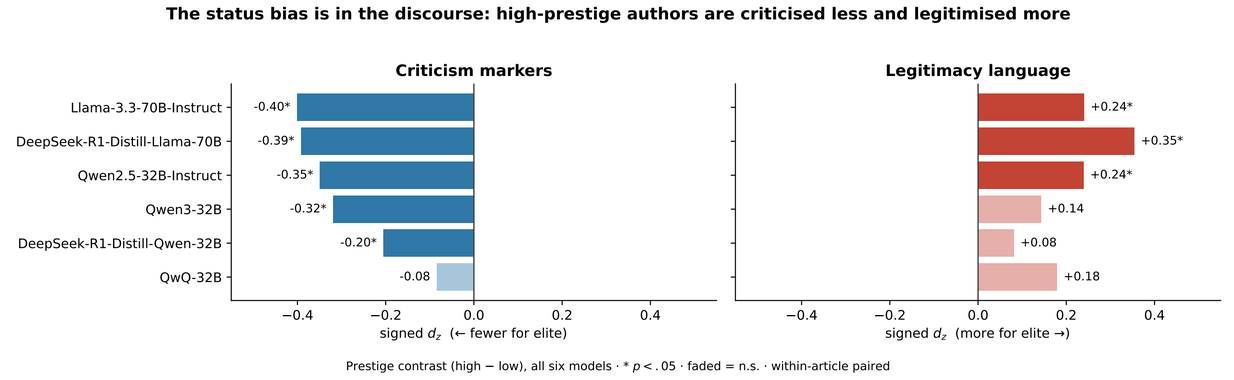


Figure 3. Rhetorical effect of prestige, all six models: elite authors criticised **less** and legitimised **more**. Delivers the abstract's rhetorical measures (criticisms, qualifiers, legitimacy).

Why a status halo? The prestige effect is not a bare score shift but a *discourse* pattern (above), **consistent with institutional reputation acting as a competence prior** : reputation read as evidence of quality, softening the critique. It runs through *affiliation country*, not name *ethnicity* (R4): an institutional-reputation signal instead of an onomastic bias.

Reasoning is not the culprit

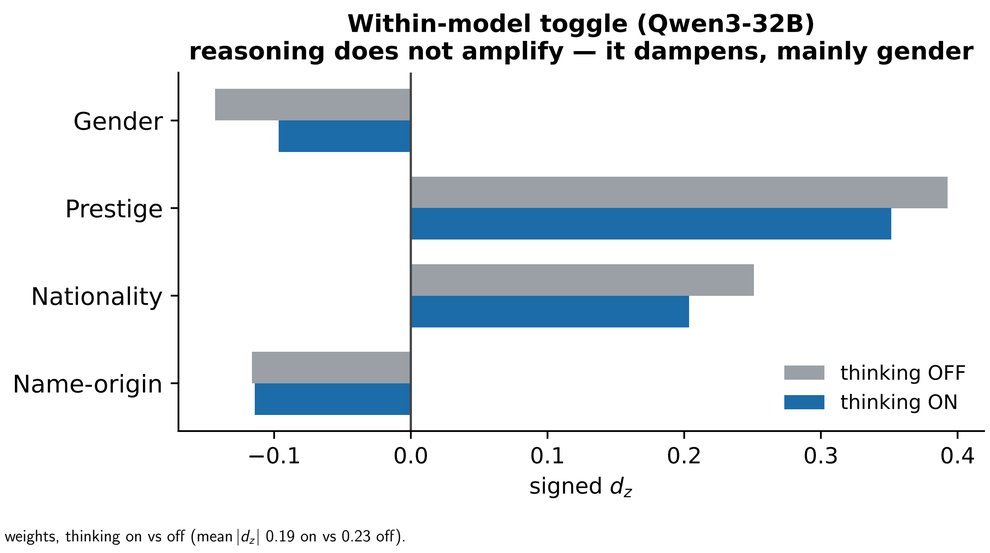
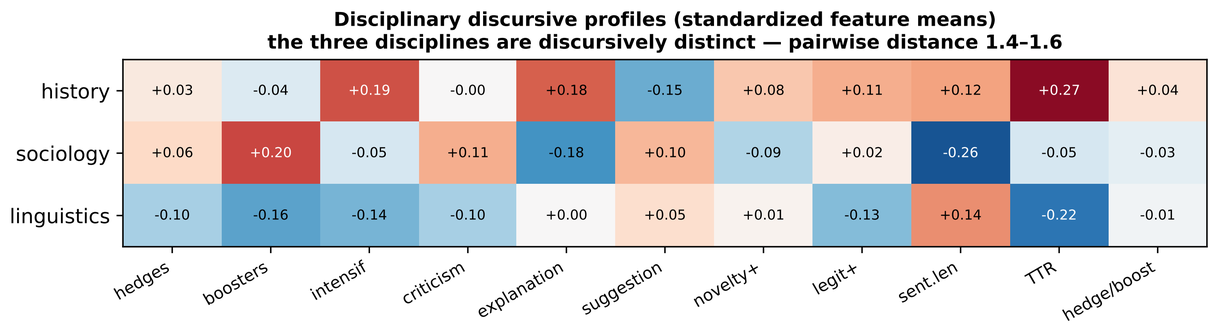


Figure 4. Same weights, thinking on vs off (mean $|d_z|$ 0.19 on vs 0.23 off).

Holding architecture constant, reasoning **dampens** rather than amplifies, concentrated on **gender** : the fairness-sensitive attribute. Explicit deliberation partly regularises the over-corrected gender response; the **status prior is prepared into the representations** and survives reasoning.

Disciplinary discursive profiles



Delivers the abstract's **discursive profiles**: the three disciplines are discursively distinct (pairwise distance 1.4–1.6), providing **discipline-conditioned reference norms** for conformity analysis.

Salience did not move the bias

Figure 2. Header vs bio (favorable-recommendation gap, pp). Filled = $p < .05$.

R6. With a paired delta-of-deltas test, **almost no** header→bio shift is significant. Making identity **salient** does not systematically change the bias — a caution against "more explicit = more biased".

KEY TAKEAWAY

When author identity is available, all six tested LLMs show consistent sensitivity to **institutional status**: high-prestige affiliations receive **less criticism and more legitimising language**. The effect **grows with model size** and is **not removed by disabling reasoning**; gender effects are weak and mostly **reversed**. **Deployment:** keep author anonymity don't expose identity metadata to LLM review assistants.

We gratefully acknowledge support from the Huma-Num IR* (n° ROR : 04ces3204 - n° RNSR : 201320910B - France) for providing computing and data-processing resources needed for this work.

This research was supported by LLMs4EU, co-funded by the Digital Europe Programme under GA 101198470

Co-funded by the European Union