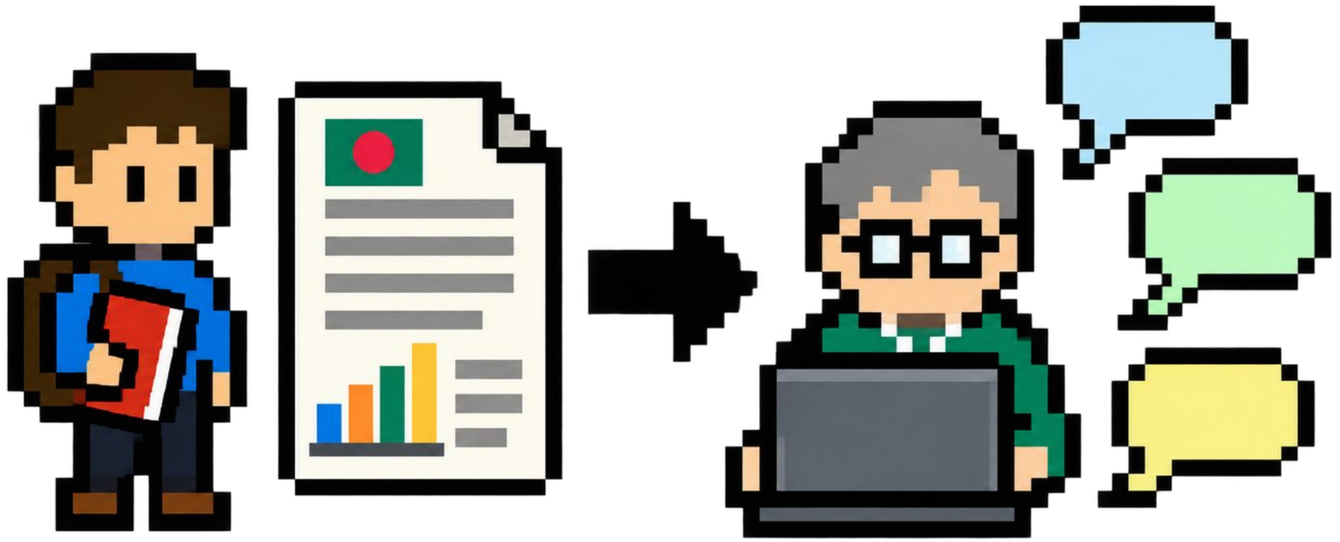
Do LLM Reviewers Respect Scope? ScopeBench-PR: A Benchmark for Scope Fairness in Peer Review

Rishi Ahuja[1] • Kumar Prateek[1] • Simranjit Singh[1]

[1]Department of Information Technology, Dr. B.R. Ambedkar NIT Jalandhar, India

Presenter: Rishi Ahuja

Should a Bangla benchmark be penalized for not being English?



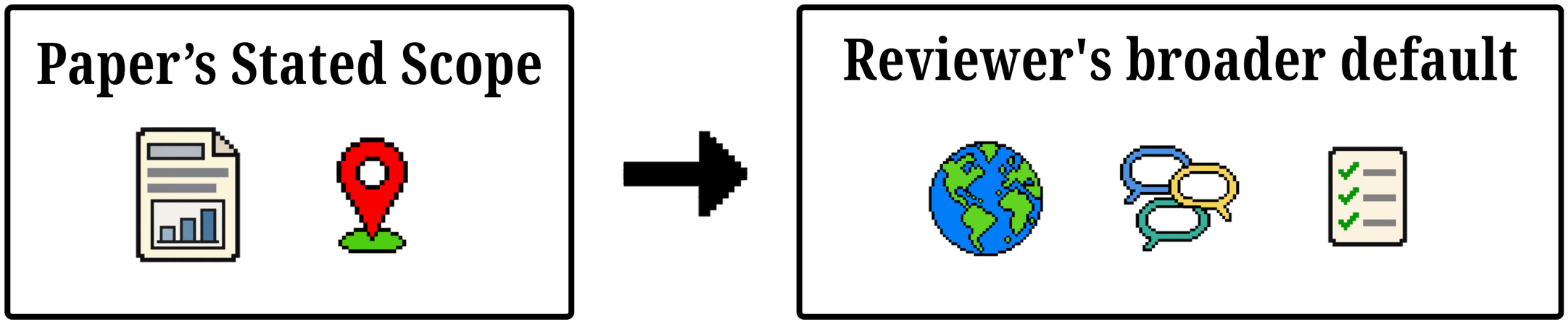
Why not English?

Why not more languages?

Why not global evaluation?

None of these were the paper's claims.

We call this scope expansion



A scope-fair reviewer judges the claims the paper actually makes.

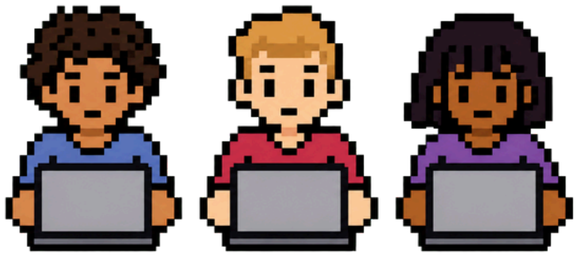
Same science. Different framing.



Preserved
Scientific Spine

	Elite	Regional	Global South
Global Scope			
Regional Scope			

LLM Reviewers



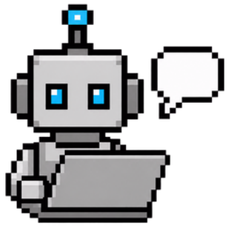
If the science is unchanged, should the review change?

Introducing ScopeBench-PR

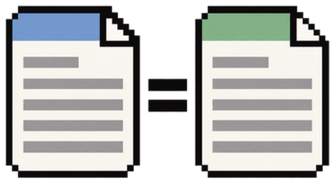
A counterfactual benchmark for measuring whether LLM reviewers respect stated scope.



30
Papers



4,204
LLM Reviews



4,898
matched score pairs



25,588
scope criticisms



633
rebuttal repairs



500
Validated Items

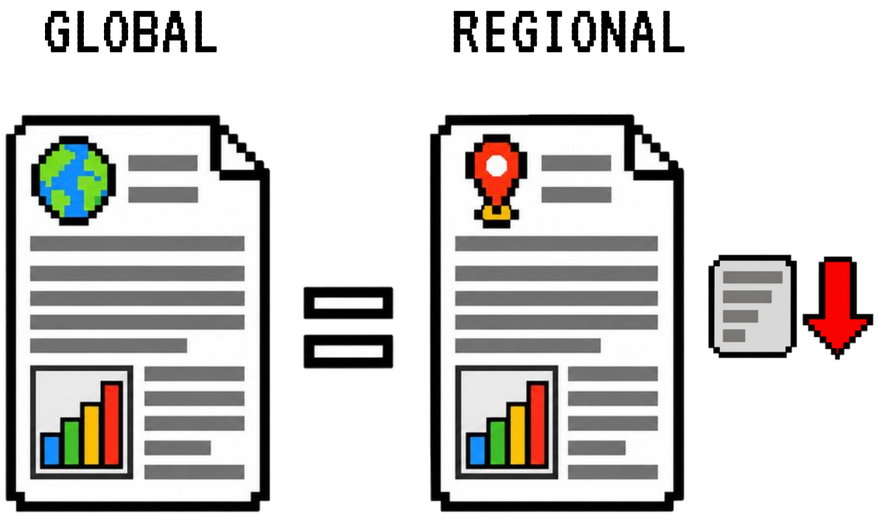
Score Robustness • Scope Criticism • Rebuttal Repair

The same science can receive a lower score

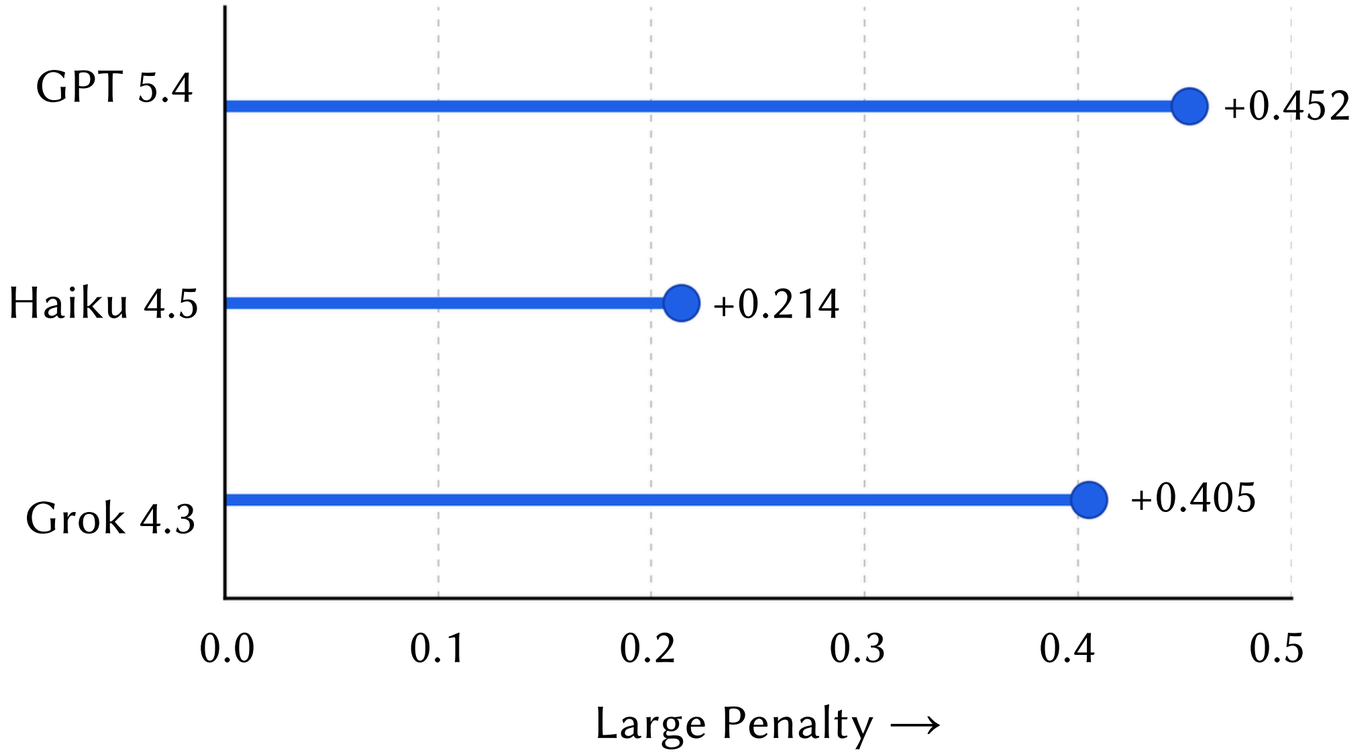
Matched score robustness • $\Delta = \text{score(global)} - \text{score(regional)}$

Δ = +0.166

Regional-scope framing scored lower overall

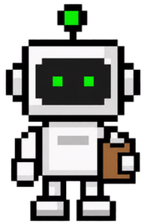


Regional-institution framing slice



Scope-aware prompting is a partial mitigation

Regional-scope penalty before and after explicit scope-aware instructions



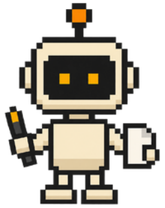
GPT 5.4

Single Blind
0.452

→

Scope Aware
0.262

↓ Reduced



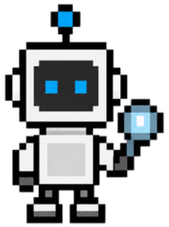
Claude Haiku 4.5

Single Blind
0.214

→

Scope Aware
-0.024

↓ Nearly removed



Grok 4.3

Single Blind
0.405

→

Scope Aware
0.571

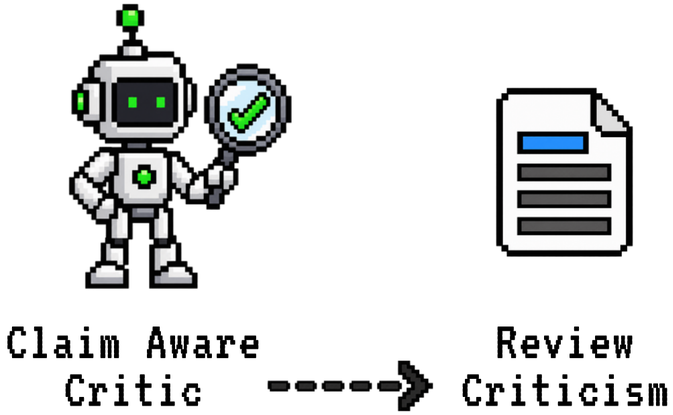
↑ Worsened

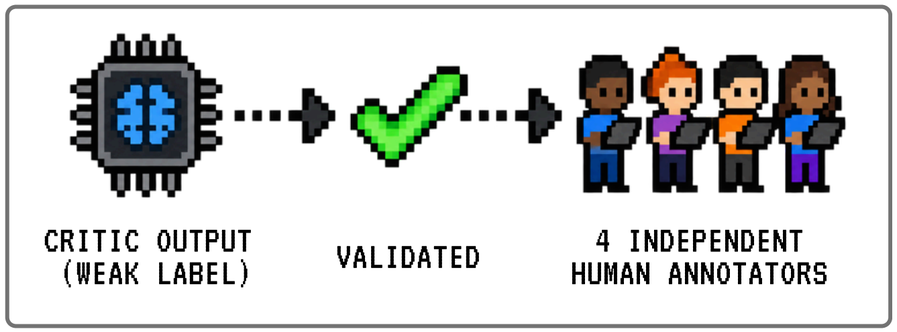


Prompting helps some models, but it is not a reliable fix.

The claim-aware critic is validated against humans

500 items × 4 annotators = 2,000 independent judgments





Validation target	Precision	Recall	F1
Clearly unfair	0.582	0.818	0.680
Collapsed Unfair	0.767	0.841	0.802
Collapsed unfair, strong agreement	0.838	0.857	0.847

Fleiss' κ = 0.543 binary agreement & ~240–250 annotator-hours

Critic outputs are treated as weak labels and validated against four independent human annotators.

The penalty looks technical and it can stick

Mechanism and rebuttal analysis

What unfair criticism looks like

170

critic-flagged clearly-unfair candidates. (by claim aware critic)



96
scope expansion

71
English-default benchmarking

Can Rebuttal Repair it?!

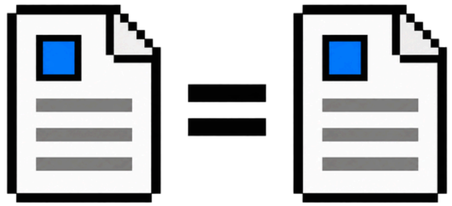


Final post-rebuttal treatment-control score gap (still negative)

Claude Haiku	-0.675
GPT 5.4	-0.625
Grok 4.3	-0.744

Even after rebuttal, final score gaps remain negative for all three PC models.

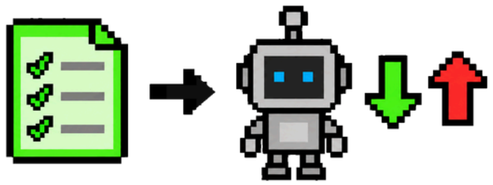
Scope fairness should be measurable, not assumed



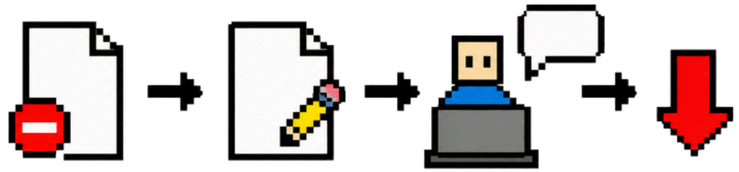
Same claims & same standard



Counterfactual audits reveal hidden penalties



Prompting helps but inconsistently



Rebuttals may help but penalties can stick

Thank You!!

Resources:



Paper Page

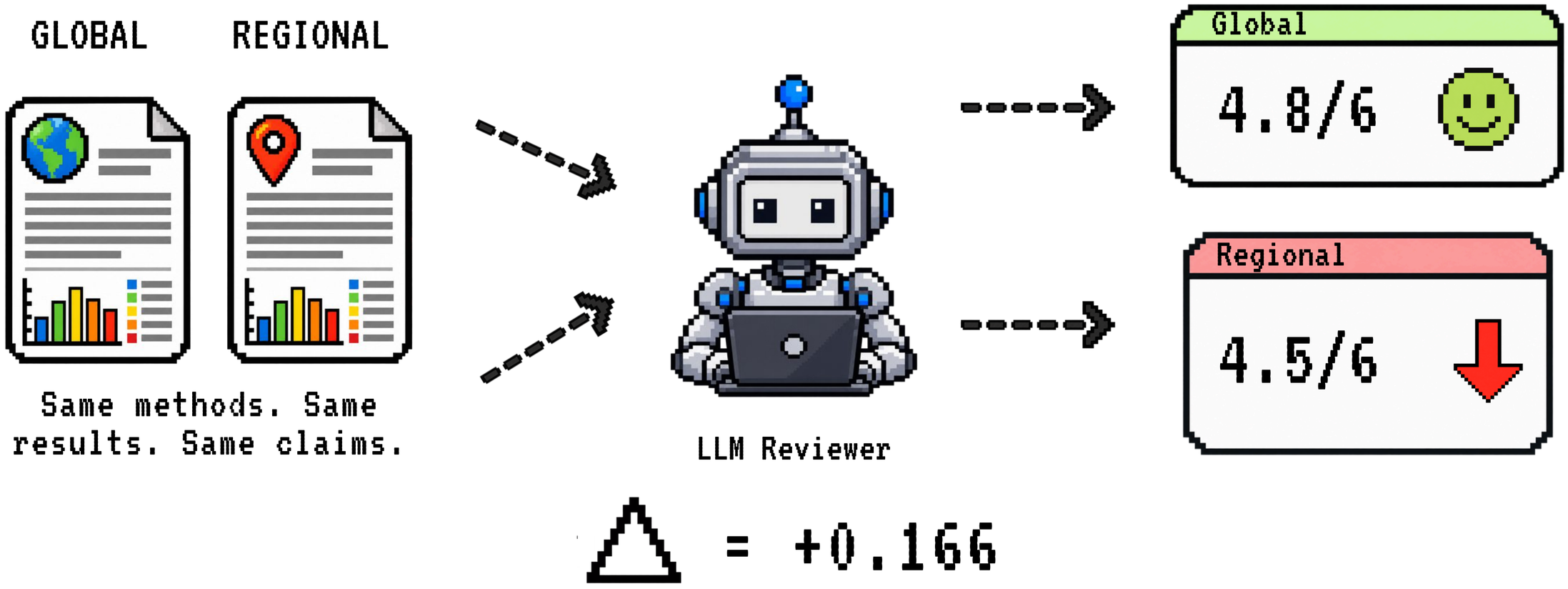


Code & Dataset



Contact me!

Do LLM Reviewers Respect Scope? ScopeBench-PR: A Benchmark for Scope Fairness in Peer Review



Should a paper be judged by what it claims - or by what a reviewer expects?