Context-Aware AI for Constructive Peer Review

Measuring and reducing toxicity in academic peer review

Tara Dixit

Motivation

Peer review shapes what researchers learn next.

- Good reviews point authors toward improvement
- Harsh reviews can discourage collaboration and growth
- The same criticism can be rigorous without being personal, speculative, or emotionally charged

Research goal: build AI tools that understand academic feedback in context

Research Question

Can AI distinguish rigorous criticism from harmful tone?

In ≈1,000 ICLR peer reviews, I studied how negativity and toxicity appear in scholarly feedback, then tested whether LLMs can reduce harmful language while preserving reviewer intent.

Core tradeoff: reduce toxicity without rewriting valid scientific feedback

Dataset and Construct

Academic toxicity is not just internet toxicity.

≈1,000 peer reviews

ICLR reviews were scored across four dimensions designed for academic feedback rather than general online speech.

Emotive tone	frustration, condescension	**Constructiveness**	specific, actionable feedback
Personal attacks	comments aimed at authors	**Speculation**	unsupported assumptions

Method

One pipeline: measure, validate, and intervene.

1 Score reviews

LLM academic-toxicity dimensions

2 Compare tools

VADER sentiment and Perspective toxicity

3 Validate construct

Pearson and Spearman correlations

4 Rewrite selectively

Minimal edits to reduce harmful tone

The same construct guides both measurement and intervention.

What Academic Toxicity Looks Like

Often it is harshness, not obvious abuse.

Harsh but scientifically relevant:

"The writing is terrible."

Context-aware revision:

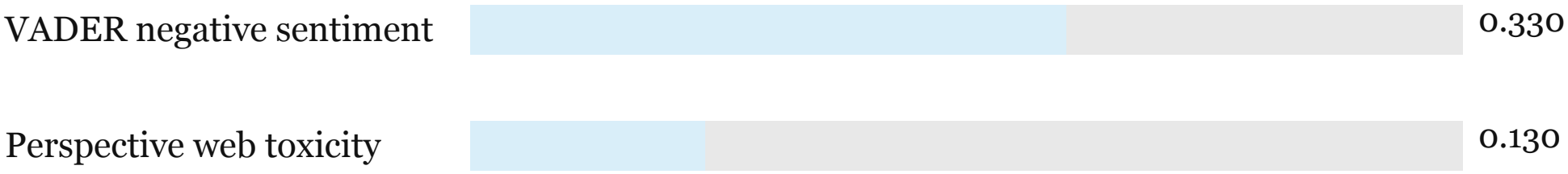
"The writing could benefit from further refinement."

Same criticism, but one is less harmful.

Validation Result

Academic toxicity aligned more with negative sentiment than web toxicity.

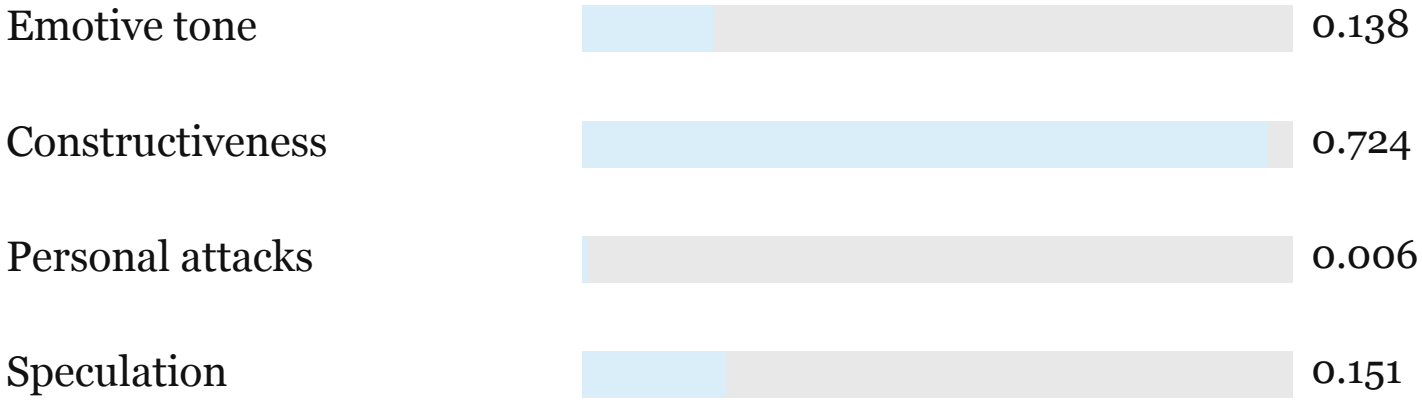
Average correlation with academic-toxicity dimensions



Interpretation: harmful academic feedback is often expressed through negative, discouraging tone rather than insults, threats, or other web-toxicity signals.

Corpus Findings

Most reviews were professional, but frustration still appeared.



Takeaway: personal attacks were rare, while emotive and speculative language created the clearest opportunity for intervention.

From Detection to Intervention

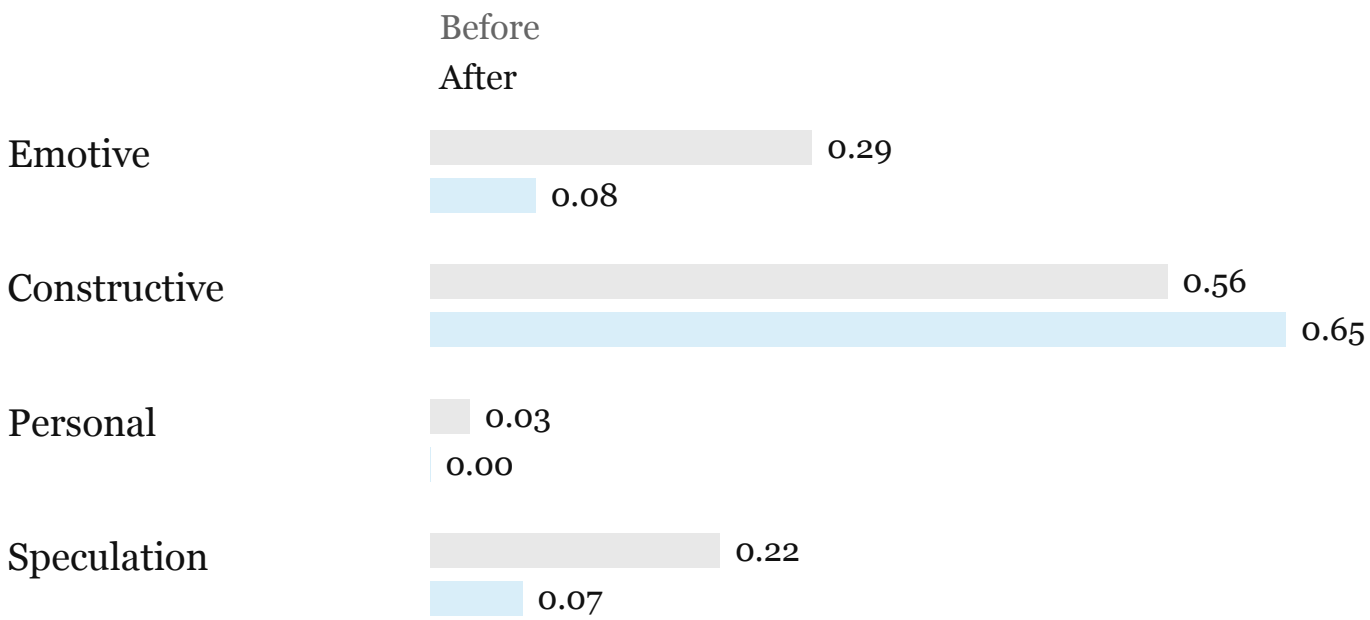
The model becomes an academic feedback editor.

- Identify the reviews with the highest toxicity scores
- Ask an LLM to make feedback more constructive, neutral, and professional
- Preserve the reviewer's original scientific judgment and recommendation

Goal: improve tone, not change the verdict

Intervention Result

Rewriting reduced harmful dimensions and increased constructiveness.



The strongest improvements appeared in emotive tone and speculation, while constructive feedback increased.

Preservation Challenge

A good intervention should not rewrite the reviewer.

- Full rewrites lowered toxicity but changed too much text
- Some reviews required substantial edits, but many only needed targeted phrase-level changes
- This motivated a stricter intervention objective

Optimization target: maximum toxicity reduction with minimum change ratio

Prompt Design

The optimized prompt made the smallest possible edits.

Only edit toxic sentences — leave all other text unchanged

Preserve technical content — do not change math, terminology, or claims

Avoid fluency edits — unless they directly reduce toxic tone

Target word-level changes — emotion, personal attacks, speculation

The prompt operationalizes alignment: helpful, harmless, and honest feedback

Example Revision

The technical criticism stays; the phrasing becomes more useful.

Original review phrase		Targeted revision
"The paper is not written well and has several errors."	→	*"The writing would benefit from revisions, as there are several typographical errors."*

The model changes tone while preserving the review's substantive concern

Contribution

A context-aware approach to feedback alignment.

- Defines academic toxicity as a distinct language construct
- Shows why web-toxicity tools miss important harms in peer review
- Tests LLM interventions that reduce harshness while preserving reviewer intent

A small editing task becomes a model-alignment problem

Why This Matters for AI

Peer review is a high-stakes test of context-aware language understanding.

Helpful	Harmless	Honest
Preserve actionable feedback	Reduce personal and emotional harm	Keep criticism rigorous

This is the same tension many frontier AI systems face: make language more useful and safer without distorting the underlying meaning.

Next Steps

Toward deployable, trustworthy review assistance.

- Validate edits with human reviewers and authors
- Compare multiple LLMs, including OpenAI and Anthropic models
- Measure bias, meaning preservation, and reviewer agency
- Build an interface that suggests edits rather than silently rewriting reviews

Long-term goal: AI that supports rigorous, constructive scientific communication

Questions or feedback?