NeurIPS 2026 POSITION PAPER TRACK

Text-Based AI Tools for Research Integrity Must Be Audited on Linguistic Fairness Before Deployment

ARE WE DETECTING RESEARCH MISCONDUCT — OR WRITING STYLE?

Shuai Shao[1,2],* · Yongkang Wan[1],* · Daoyin Dang[1] · Lanyun Zhu[3]
Di Yang[1,2] · Yutong Bai[4] · Yan Wang[5],† · Jiangtao Wang[1,2],†

FULL AUTHOR AFFILIATIONS

1 School of Artificial Intelligence and Data Science, University of Science and Technology of China, China

2 Suzhou Institute for Advanced Research, University of Science and Technology of China, China

3 School of Computer Science and Technology, Tongji University, China

4 Beijing Tiantan Hospital, Capital Medical University

5 School of Automation and Intelligence, Beijing Jiaotong University

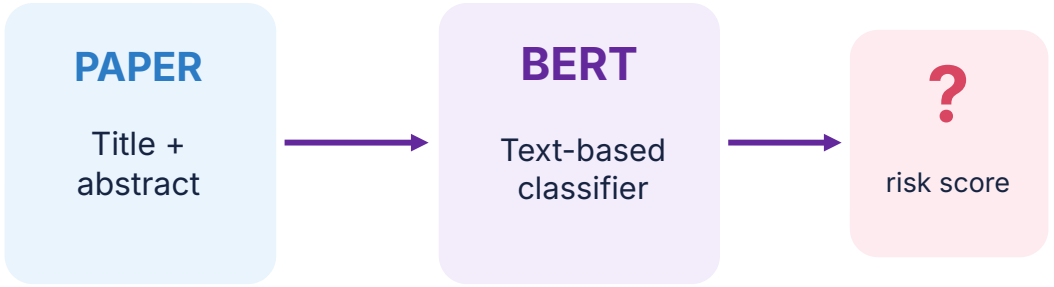
* Equal contribution † Corresponding authors

A position for safer, fairer use of AI-based research integrity screening

Why question AI integrity screening?

AI tools can support large-scale screening. But a false positive may harm a real researcher.

HOW THE DETECTOR WORKS



Key concern

A detector may use writing style as a shortcut for suspected misconduct.

False alarms can damage careers and research communities.

THE BMJ CASE · JAN 2026

Paper-mill detection in cancer research

- Based on paper titles and abstracts
- Original model and code were not publicly released
- We independently reconstructed the BERT pipeline

THREE-STAGE STUDY

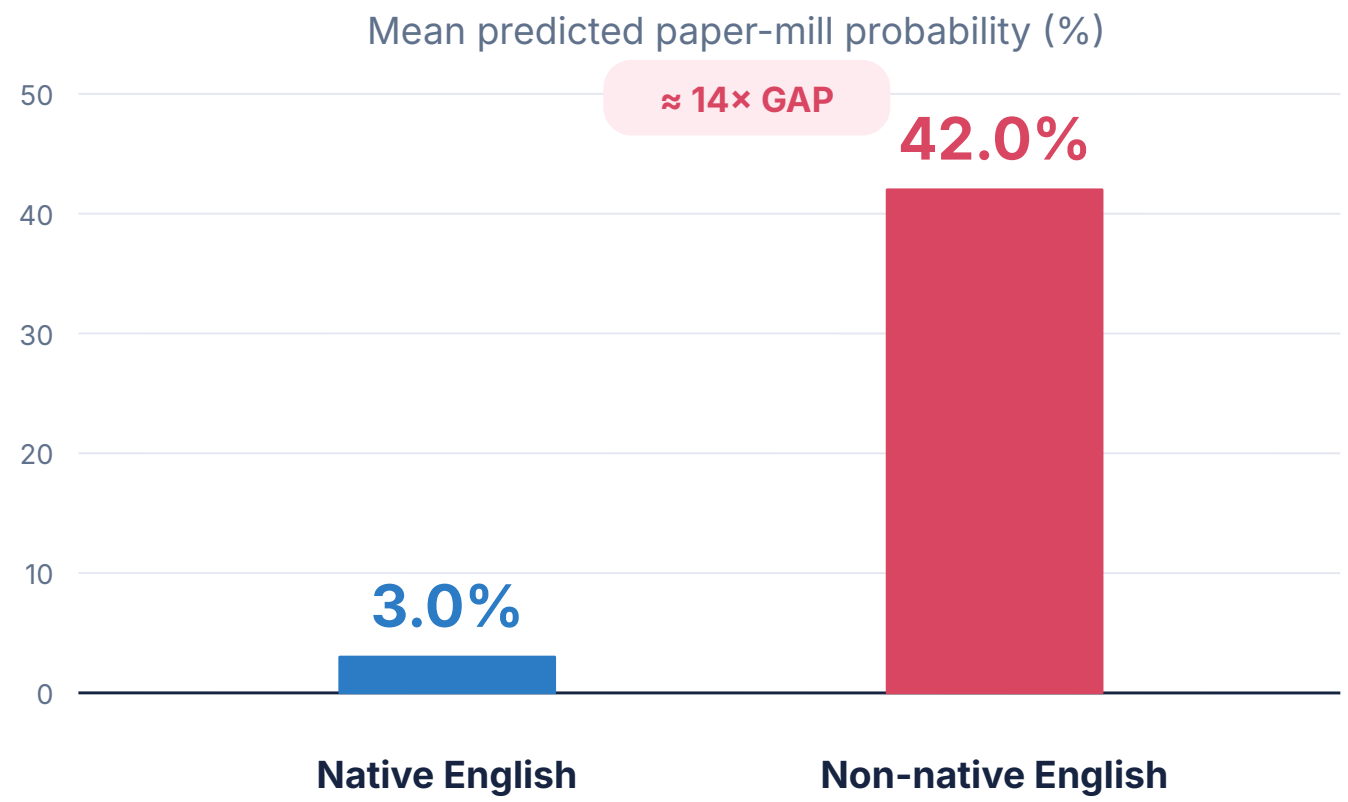
Reproduce | Compare | Control

Source: Shao et al., NeurIPS 2026 Position Paper

Scancar et al., The BMJ (2026)

A large gap in mean model scores

5,000 papers per group · High-impact journals (impact factor > 10)



42% vs 3%

The detector assigns much higher average paper-mill scores to the non-native English group.

Important caveat

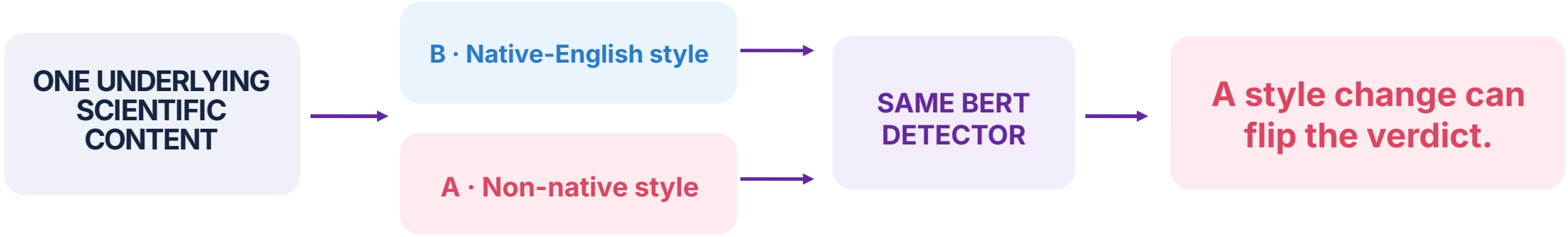
Chinese-authored papers proxy the non-native English group. These are distinct real papers, not matched rewrites.

Source: Shao et al., NeurIPS 2026 Position Paper

Paper Tables 2–3

Same content. Different writing styles.

Matched scientific content helps isolate the effect of linguistic style on the model's decisions.



PAIRED PREDICTION FLIPS

Experiment / matched content	A+ / B−	A− / B+
Exp 2 · Shared content skeletons	1,107	0
Exp 3 · Rewriting false positives	82	0
Exp 4 · Rewriting true negatives	98	0

A+ / B− = positive only for non-native style · A− / B+ = positive only for native style

Source: Shao et al., NeurIPS 2026 Position Paper

Paper Table 5

Why might this risk extend beyond one model?

Our paper argues that two structural factors make linguistic bias a concern for text-based tools.

FACTOR 01

Geographic bias in observed records

Retraction and misconduct databases reflect where cases are detected, reported, and recorded.

FACTOR 02

Language models encode writing style

Vocabulary, grammar, and syntax can signal language background, even when they do not signal misconduct.

If labels correlate with style, a model can learn linguistic shortcuts.

This is a general risk argument. The direct experiments in our paper are a case study of one detector.

Source: Shao et al., NeurIPS 2026 Position Paper

Paper Sections 4–5

Mandatory linguistic fairness auditing

Before text-based integrity tools are deployed for editorial screening or institutional decisions.

AUDIT LINGUISTIC FAIRNESS BEFORE DEPLOYMENT

01 Open source & reproducibility

Release code, models, and training procedures.

02 Training data transparency

Disclose data sources, labels, language-group composition, and limitations.

03 Independent fairness evaluation

Report sensitivity, specificity, and false-positive rates by language group.

04 Interpretability verification

Check whether model decisions rely on genuine misconduct signals or style proxies.

Our goal is not to stop useful detection tools, but to make them trustworthy and fair.

The takeaway

Detect misconduct. Not writing style.

3.0% vs 42.0%

Mean model score across real-paper groups

1,107 vs 0

One-way paired prediction flips (Exp 2)

Balance detection effectiveness with linguistic equity.

Thank you!

github.com/yongkangwan/Discussion-on-Fairness-in-AI-Tools