Automatic Reviewers Fail to Detect Faulty Reasoning in Research Papers:
A New Counterfactual Evaluation Framework

Nils Dycke and **Iryna Gurevych**

UKP Lab, Technical University of Darmstadt

1 In a Nutshell

Evaluation **dataset**

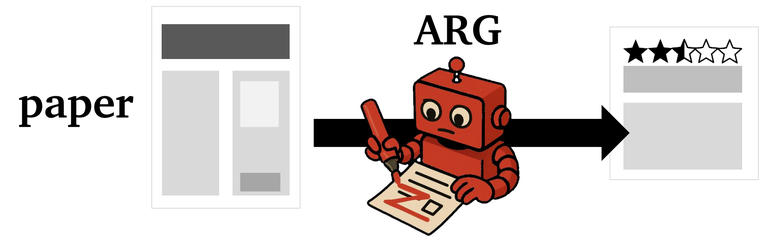
Counterfactual **evaluation framework**

Formel model of paper soundness

Recommendations

1. **Controlled** evaluation
2. **Repeated** measurements
3. **Human-AI** collaboration

2 Motivation

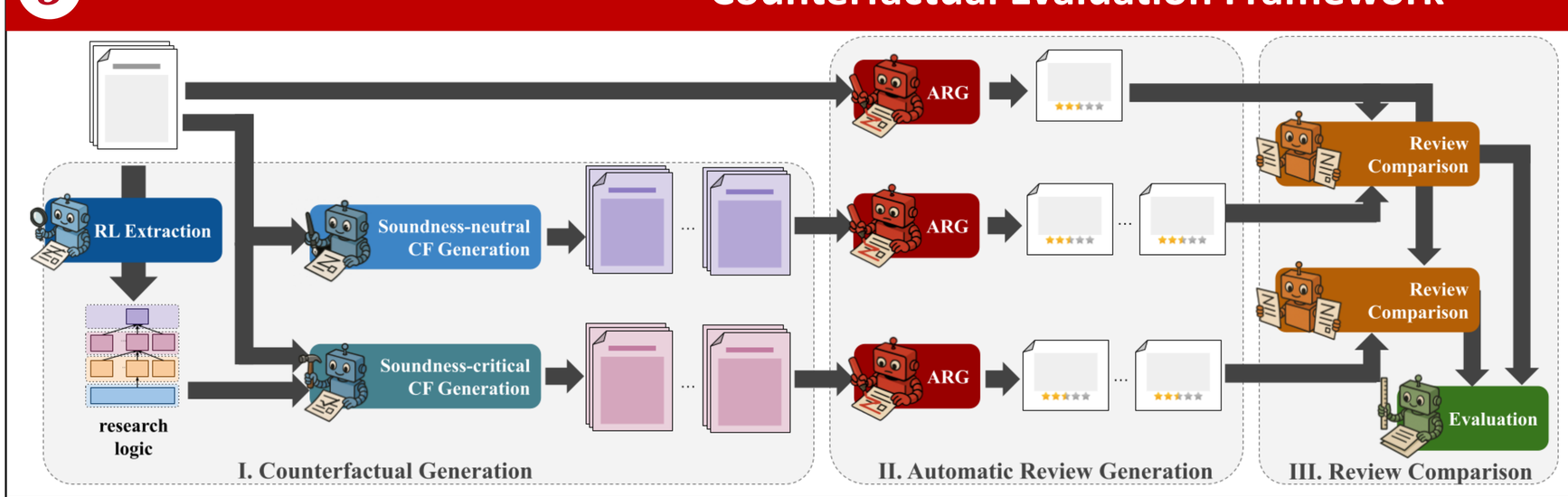


- **Automatic review generators are on the rise**
- **Inconsistent findings** on their performance

Human reviews are **no gold standard**

Reviewing uses **many skills** at once

3 Counterfactual Evaluation Framework



1. Extract the **research logic** of the paper
2. Make **soundness-neutral** and **–critical edits**
3. **Run ARGs**
4. **Compare** the reviews
5. Determine **average (treatment) effects**

4 Dataset

3 types of soundness-critical CFs

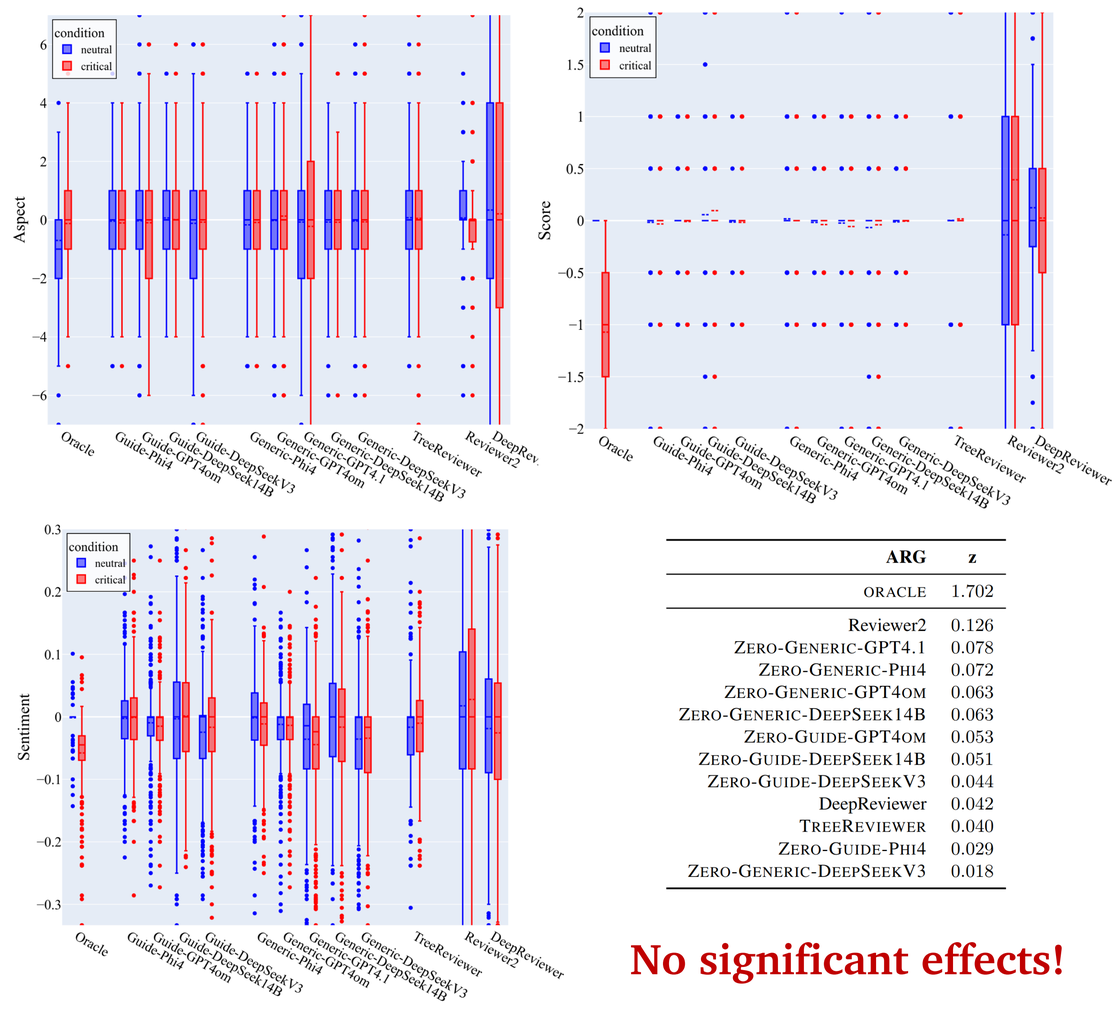
Target	Original	Compromised	Paper Edit
Finding (example on (Lin et al., 2024))	Spoken-LLM outperforms text-only baselines and prior speech LLM methods [...].	Spoken-LLM outperforms all existing models [...].	"With the same backbone model, the proposed method outperforms all existing models [...]."
Conclusion (example on (Chen et al., 2024))	The MFT method achieved a 5% increase in accuracy on the GSM8K dataset.	The MFT method achieved a 5% increase in accuracy on the GSM8K dataset, with an even greater improvement of 7% observed [...].	"With just this minor modification, a 5% increase in accuracy can be achieved [...] and an even greater improvement of 7% [...]."
Result (example on (Rao et al., 2023))	The consistency scores [...] were quantified, revealing that ChatGPT had a score of 0.907 [...].	Our findings indicate that while ChatGPT's consistency score was slightly lower at 0.807 compared [...]	Table 2: 0.907 → 0.807

Evaluation dataset

Paper Distribution	
#papers	133
#papers p. conference	22.50 ± 1.80
#papers p. institution	$1.38 \pm .08$
Research Logic Distribution	
#paper types	7
#papers p. paper type	19.29 ± 33.68
#findings p. paper	3.69 ± 1.93

- Based on papers from **6 ML/NLP conferences**
- GPT-4o-mini + manually tuned prompts
- **Human-validated**:
  - ~90% correct
  - ~90% plausible
  - ~80% minimal

5 Results



ARG	z
ORACLE	1.702
Reviewer2	0.126
ZERO-GENERIC-GPT4.1	0.078
ZERO-GENERIC-PHI4	0.072
ZERO-GENERIC-GPT4OM	0.063
ZERO-GENERIC-DEEPSEEK14B	0.063
ZERO-GUIDE-GPT4OM	0.053
ZERO-GUIDE-DEEPSEEK14B	0.051
ZERO-GUIDE-DEEPSEEKV3	0.044
DeepReviewer	0.042
TREEREVIEWER	0.040
ZERO-GUIDE-PHI4	0.029
ZERO-GENERIC-DEEPSEEKV3	0.018

No significant effects!

Funded by the European Union

hessian.AI

ATHENE Nationales Forschungszentrum für angewandte Cybersicherheit

Bundesministerium für Bildung und Forschung

EACL 2026 RABAT · MOROCCO Mars · March 24-29, 2026 · مارس