# STRICTA: Structured Reasoning In Critical Text Assessment for Peer Review and Beyond


Nils Dycke, Matej Zečević, Ilia Kuznetsov, Beatrix Suess, Kristian Kersting, Iryna Gurevych

UKP Lab, AIML Lab, Synthetic Biology Lab

at Technical University of Darmstadt


## 1 Contributions

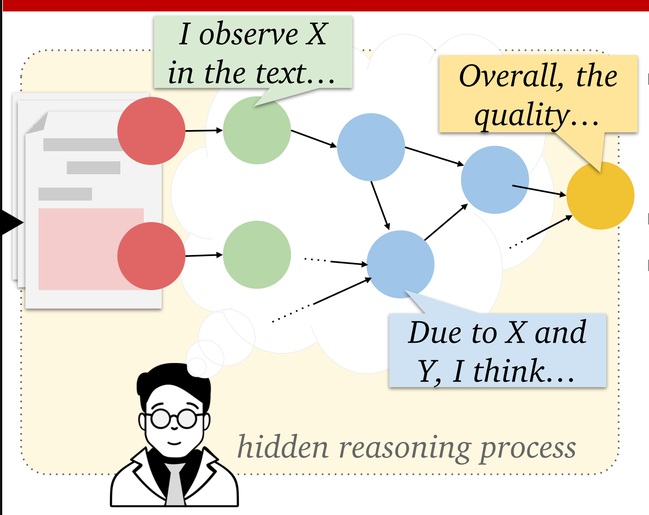
**3-step framework** *for modeling structured reasoning with uncertainty*

**Routed in causality** *using Pearl's SCMs*

**Expert multi-modal reasoning dataset** *with >4k reasoning steps on >20 biomedical papers*

**Insights in human and LLM assessment behavior** *e.g. background knowledge is a major source of variance*

## 2 Motivation

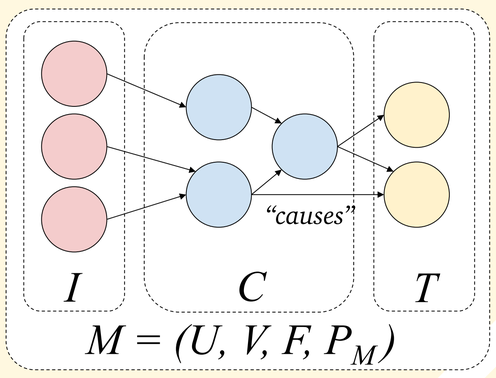



- **Critical text assessment is ubiquitous:** *peer review, essay grading, fact checking*
- **Unveiling the black box**
- **Structured reasoning** *because…*
  - ⊕ *fine-granular evaluation*
  - ⊕ *human behavior analysis*
  - ⊕ *human-LLM collaboration*

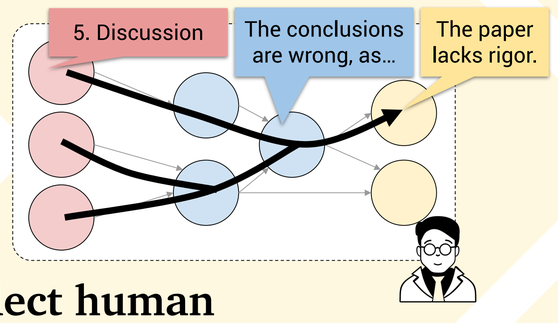
## 3 STRICTA

### I. Design Structured Causal Model

**Discover reasoning steps and links** *e.g. by expert interviews*

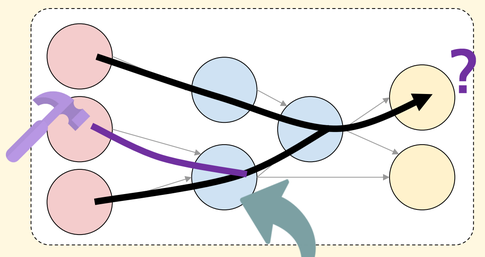



### II. Populate SCM




**Collect human reasoning** *as annotations*

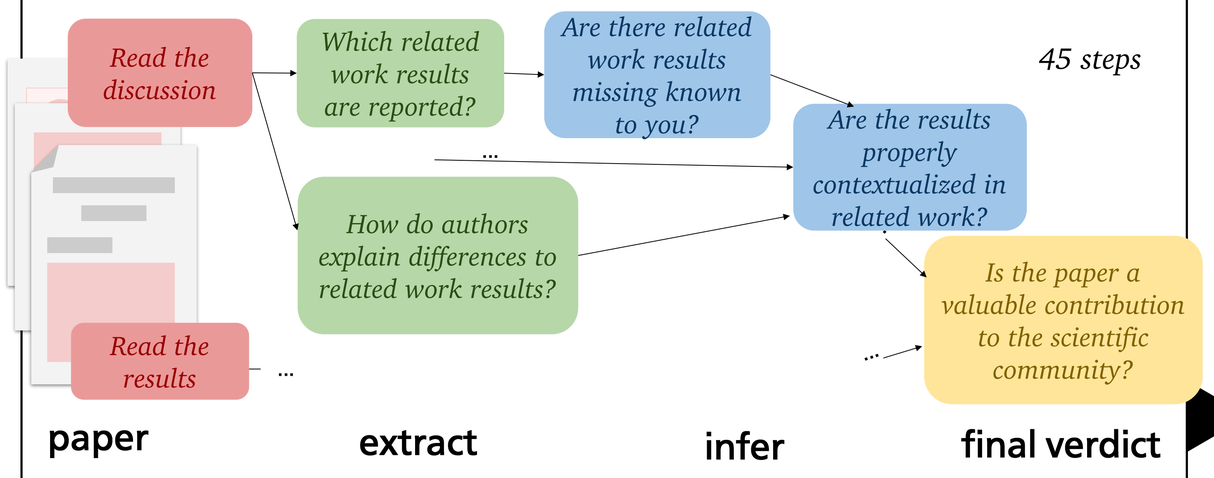
### III. Analysis + LLMs



**Study causal reasoning** **Test LLMs**

## 4 Biomedical Paper Assessment

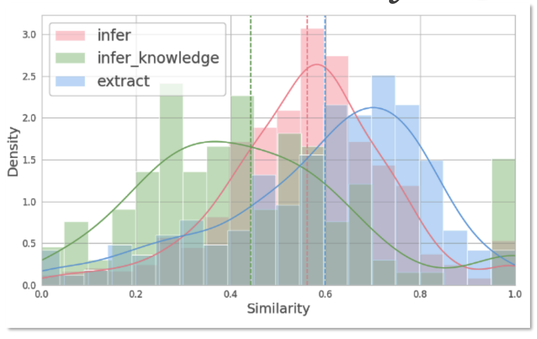
**Biomedical Assessment Workflow** *based on interviews*




### Expert-generated Dataset

| Dataset Statistics | |
|---|---|
| *# step answers* | *4371* |
| *# worklow answ.* | *93* |
| *# papers* | *22* |
| *Krippendorff's α* | *0.42* |

### Background knowledge *invokes dissimilarity*




## 5 Causal Analysis + LLM Experiments

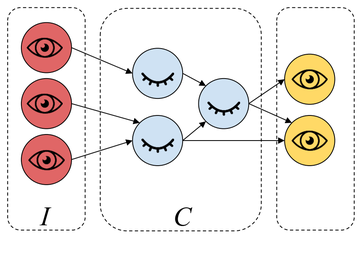
**Which factors shape the assessment most?**

| Steps' Average Causal Effect on Verdict | |
|---|---|
| *Consistency of conclusions to RQs* | *0.37* |
| *Relevance of conclusions to science* | *0.20* |
| *Clarity of the paper* | *0.20* |

› *Humans show a* **positive bias** *towards* **well-written papers.**

*… and a lot more analysis including counterfactuals etc.*

### Can LLMs explain human's final verdicts?




‹ *LLMs perform best on* **extraction** *and* **weight factors differently**

| | BERT-F1 ↑ | SummaC ↑ | TRUE ↑ | F1 ↑ |
|---|---|---|---|---|
| human*† | **0.799**$_{\pm.06}$ | −0.151$_{\pm.30}$ | 0.151$_{\pm.27}$ | **0.801** |
| Llama3 Prg† | 0.752$_{\pm.10}$ | −0.274$_{\pm.36}$ | 0.098$_{\pm.30}$ | 0.170 |
| Mixtral Prg† | 0.761$_{\pm.09}$ | **-0.149**$_{\pm.26}$ | 0.120$_{\pm.32}$ | 0.559 |
| GPT3.5t Prg† | 0.759$_{\pm.10}$ | −0.178$_{\pm.35}$ | **0.163**$_{\pm.37}$ | 0.531 |
| GPT4o Prg† | 0.780$_{\pm.07}$ | −0.186$_{\pm.30}$ | 0.139$_{\pm.35}$ | 0.720 |
| majority$^{io}$ | | | | 0.854 |
| human*$^{io}$ | 0.799$_{\pm.06}$ | −0.158$_{\pm.29}$ | 0.150$_{\pm.27}$ | 0.801 |
| Llama3$^{io}$ | 0.786$_{\pm.07}$ | −0.141$_{\pm.30}$ | 0.145$_{\pm.35}$ | 0.657 |
| Mixtral$^{io}$ | 0.794$_{\pm.07}$ | **-0.077**$_{\pm.27}$ | 0.161$_{\pm.37}$ | 0.822 |
| GPT3.5t$^{io}$ | **0.805**$_{\pm.07}$ | −0.125$_{\pm.34}$ | **0.214**$_{\pm.41}$ | 0.789 |
| GPT4o$^{io}$ | 0.795$_{\pm.07}$ | −0.094$_{\pm.28}$ | 0.194$_{\pm.40}$ | **0.876** |
| GPT4o§ | 0.776$_{\pm.07}$ | −0.154$_{\pm0.28}$ | 0.188$_{\pm.39}$ | 0.828 |


Funded by the European Union

hessian.AI

ATHENE Nationales Forschungszentrum für angewandte Cybersicherheit

Bundesministerium für Bildung und Forschung

ACL 2025 VIENNA