How effective is AI for evaluating open science practices?

25 March 2026

Dr. Yu Heng Daryl Lee (UCL)

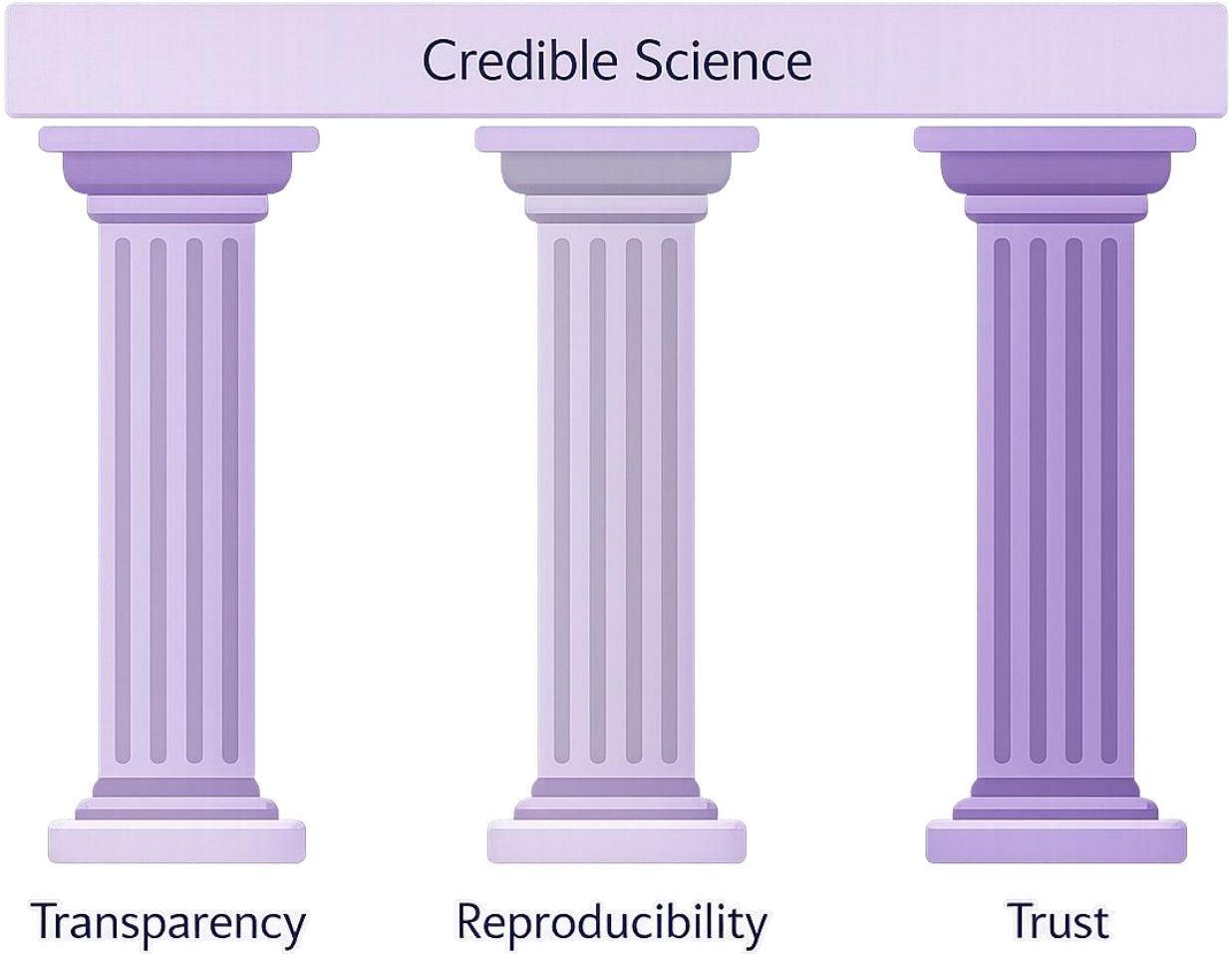
Please confirm your attendance today by following the QR code

UNIVERSITY OF WARWICK

Overview


Background
Methods
Results
Legal & Ethics
Future Directions

Why Open Science Matters?

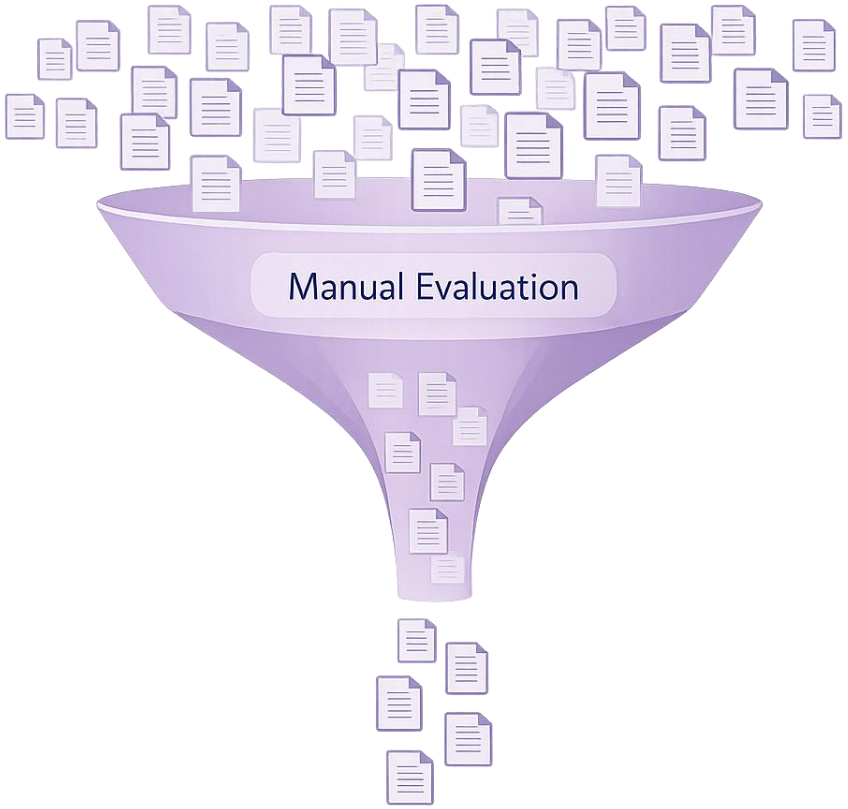


The Reality: Low Adoption

- **Preregistration: ~3% (2014 – 2017) → ~13% (2022)**
- **Public data: ~2% → ~18%**
- **Analysis code: ~1% → ~8%**

Hardwicke et al. (2022, 2024)

The Evaluation Bottleneck



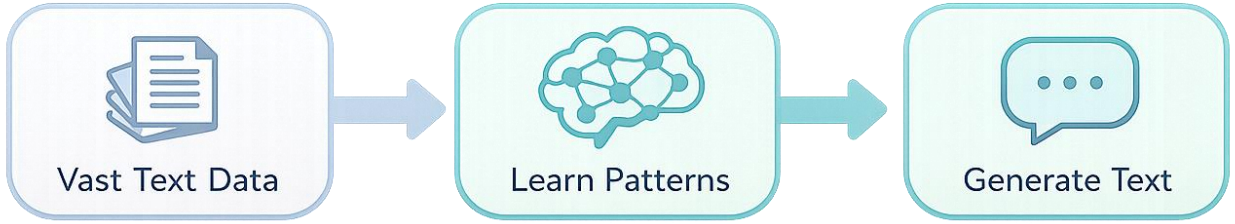
- **Manual coding is time-consuming and labour-intensive**
- **Requires substantial expertise**
- **Limits scale, frequency, and coverage of assessments**

Existing Approaches & Limitations

- Text-mining tools (e.g., Serghiou et al., 2021)
  - Require programming skills
- Machine learning approaches
  - Narrow scope (e.g., funding disclosures only)
  - Need large, manually labelled training datasets
- No accessible, comprehensive solution

LLMs: A Brief Primer

- AI ➔ LLMs (e.g., GPT / Claude) ➔ Chatbots (web interfaces)
- LLMs learn by predicting the next word across vast text data



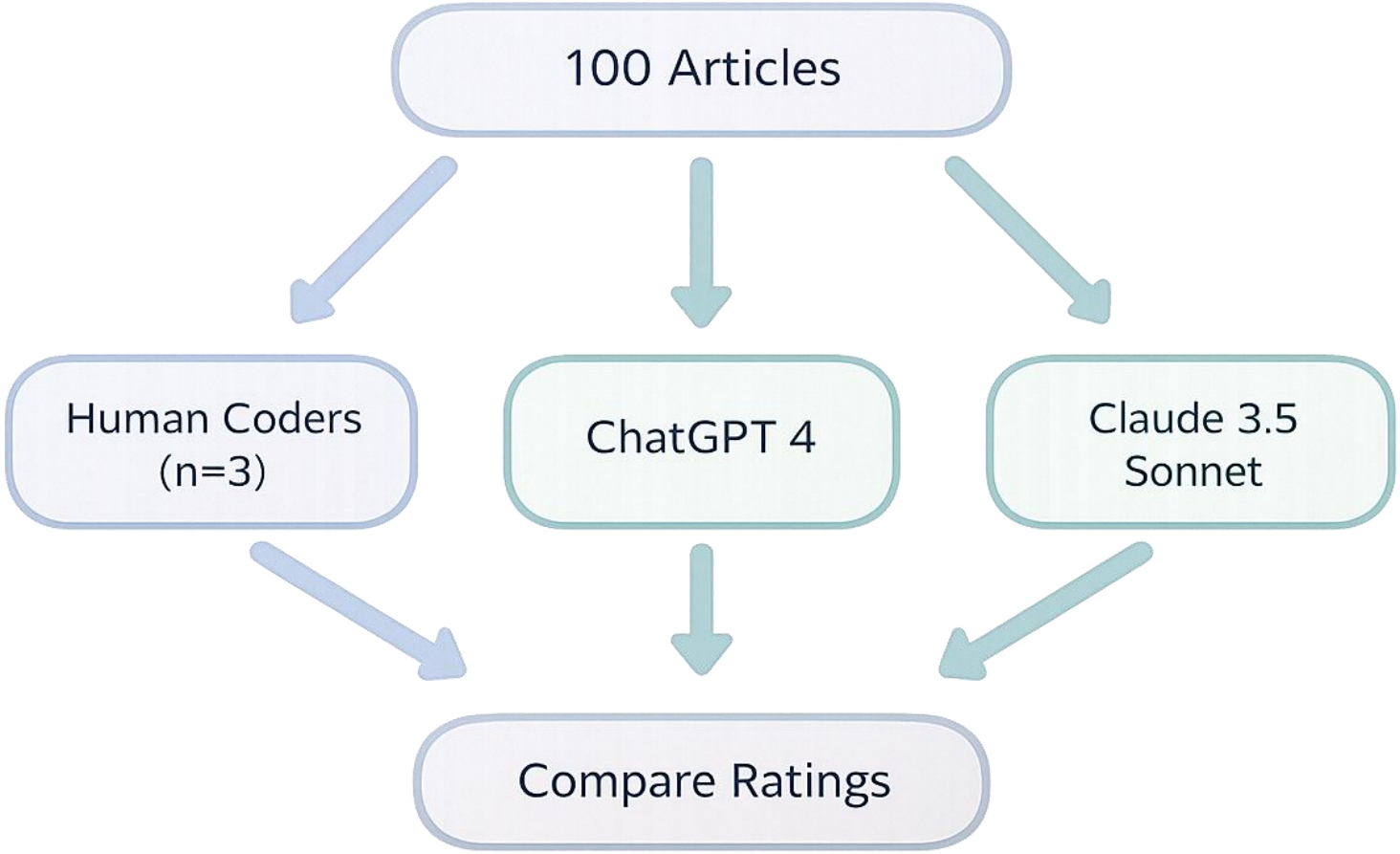
- Modern chatbots are multimodal: can process text, images, tables, and PDFs
- Accessible via web interfaces, no coding required

Our Research Question

Can AI chatbots evaluate open science practices as reliably as human experts?

- Comparing ChatGPT 4 and Claude 3.5 Sonnet against human coders
- Focus on brain science articles

Study Design Overview



Our Sample: REF 2021

- **Research Excellence Framework (REF) 2021**
  - **UK's national research quality assessment**
  - **Universities submit their best research outputs**

- **100 articles randomly selected from UCL Brain Sciences submission**
  - **Diverse disciplines: psychology, psychiatry, neuroscience, neurology, language sciences**
  - **Mix of study types: experimental, observational, meta-analyses, genomics**

The 6-Item Transparency Checklist

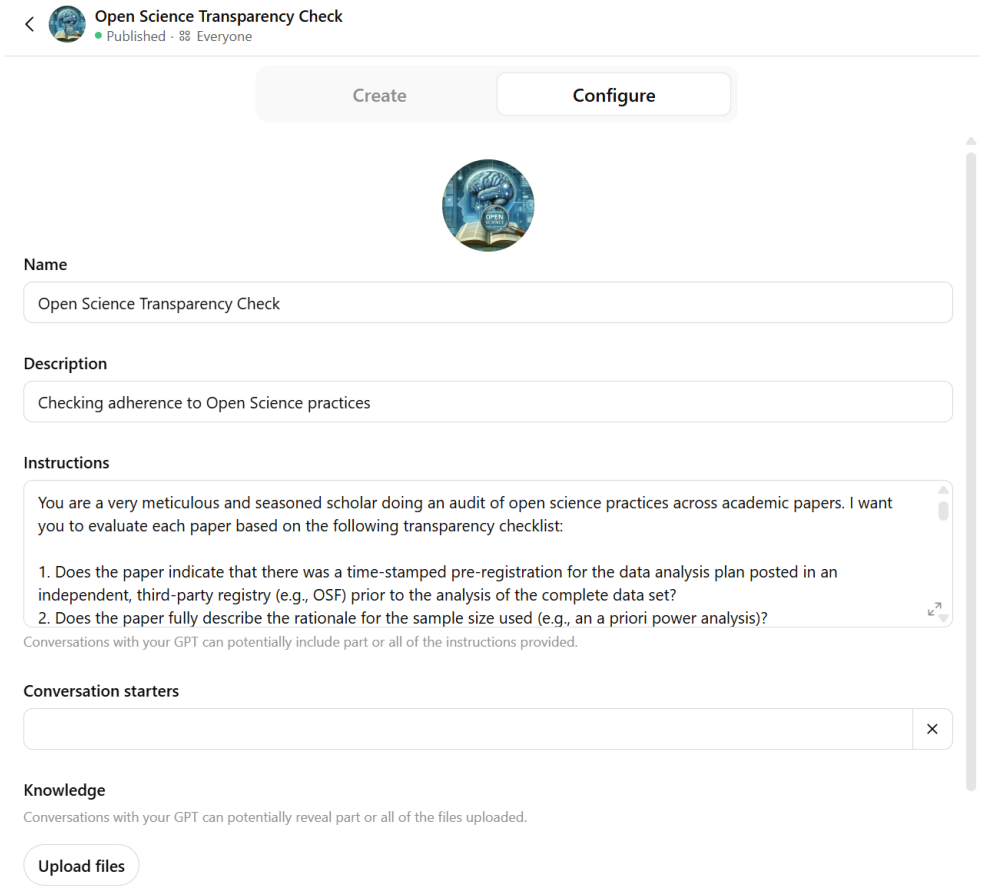
Study Planning	
PreReg	Preregistration statement
SampleSize	Sample size rationale

Research Outputs	
DataAvail	Data availability statement present
OpenData	Data publicly accessible
OpenCode	Code/software shared
OpenMaterials	Materials accessible

1 **Criterion met** **0** **Not met**

Adapted from Aczel et al. (2020) Transparency Checklist

AI Setup



- **ChatGPT 4** (January 2024)
  - Custom GPT: “Open Science Transparency Check”
  - Tailored instructions with evaluation criteria and examples
- **Claude 3.5 Sonnet** (September 2024)
  - Projects feature with identical instructions
  - 200,000 token context window (~500 pages)
- **Shared approach**
  - Full PDF upload for each article
  - Standard web interfaces only (no API, no coding)

Example Evaluation Output

7251.pdf
PDF

Please evaluate this paper.

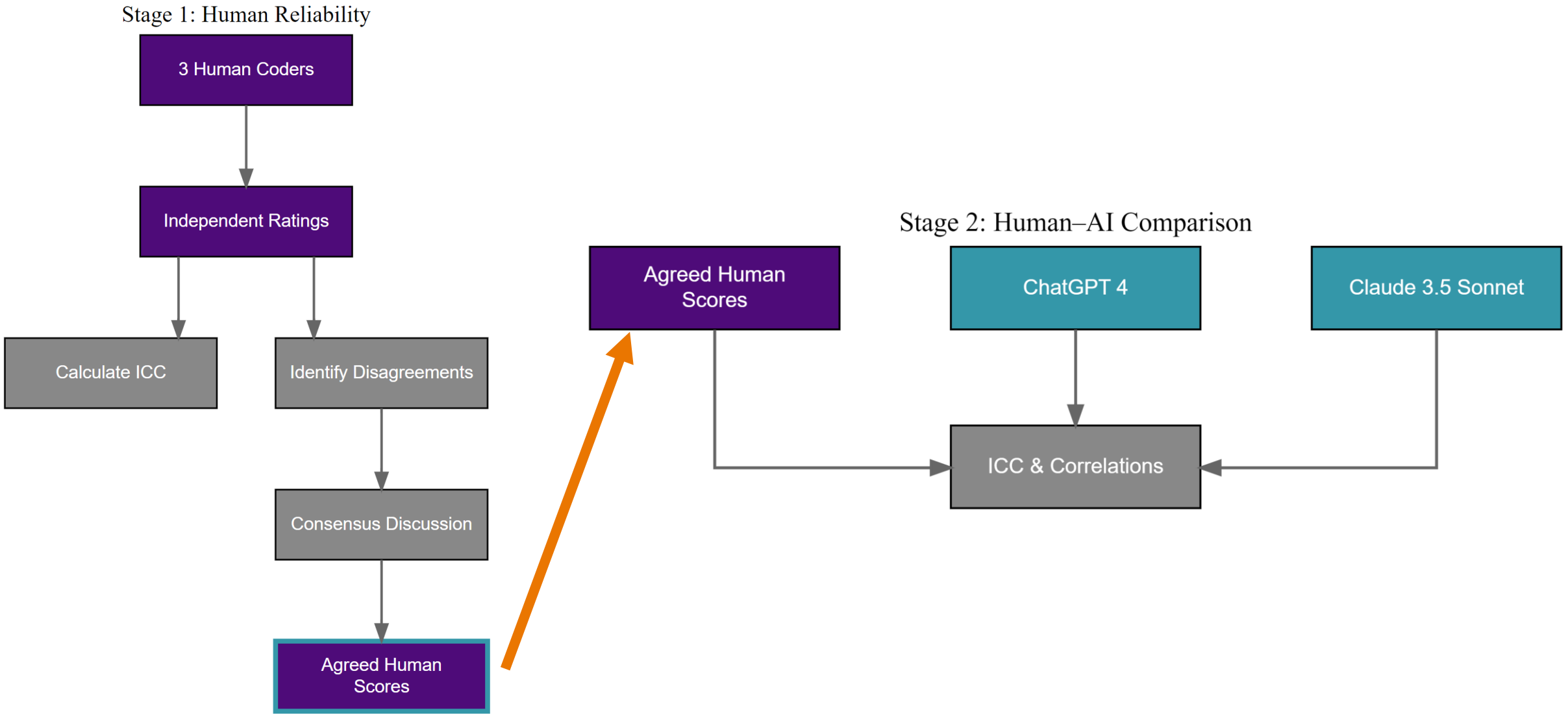
Transparency Checklist Evaluation:

1. **Pre-registration:**
   There is no mention of a pre-registration for the data analysis plan. (0)
2. **Sample Size Rationale:**
   The paper does not provide a detailed rationale for the sample size used in the experiments. It mentions the number of participants recruited but does not discuss an a priori power analysis or any other justification for the chosen sample sizes. (0)
3. **Data Availability:**
   The paper indicates that anonymized participant data and model simulation data are available at a public repository: GitHub repository ↗. (1)
4. **Code Availability:**
   The code used for all models and analyses is made publicly available at the same GitHub repository: GitHub repository ↗. (1)
5. **Materials Availability:**
   The paper does not specify the availability of all instructions, stimuli, and test materials used in the study. There is no mention of whether these materials have been made publicly accessible. (0)
6. **General Statement of Availability:**
   The paper includes a clear statement concerning the availability and location of the research items, including data and code, in the "Data availability" section. (1)

Total Score: 3/6

Note: The paper makes a positive contribution by sharing data and code openly, but it lacks transparency regarding pre-registration, sample size rationale, and materials availability.

Analysis Approach



Human Inter-Rater Reliability

Agreement Statistics

- ICC(1,1) (single rater) = .79 [.71, .86]
- ICC(1,2) (averaged) = .88 [.83, .92]

Pairwise Correlations

- DS–CN: r = .80
- CN–AP: r = .71
- DS–AP: r = 1.00

Mean Scores (out of 6)

- DS: 0.57, CN: 0.91, AP: 0.62 (*SD*s ~ 1.0)

Sources of Disagreement

- Information overlooked (e.g., buried in methods/supplements)
- External link verification differences

Overall Human-AI Agreement

Agreement Statistics

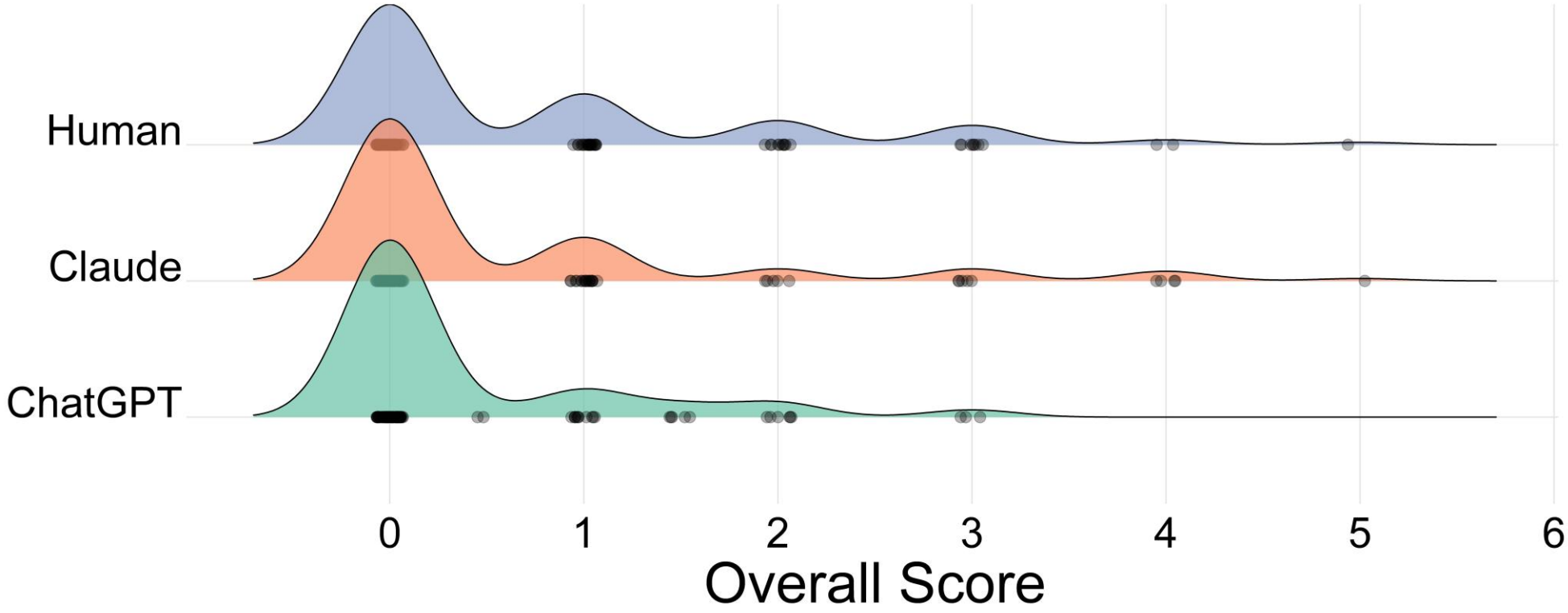
- ICC(2,1) (single rater) = .62 [.51, .72]
- ICC(2,3) (averaged) = .83 [.76, .88]

Pairwise Correlations

- Human–ChatGPT: *r* = .69
- Human–Claude: *r* = .62
- ChatGPT–Claude: *r* = .73

	Human (agreed)	ChatGPT 4	Claude 3.5 Sonnet
Mean Score	0.78	0.41	0.64
Divergent Articles		37/100	33/100

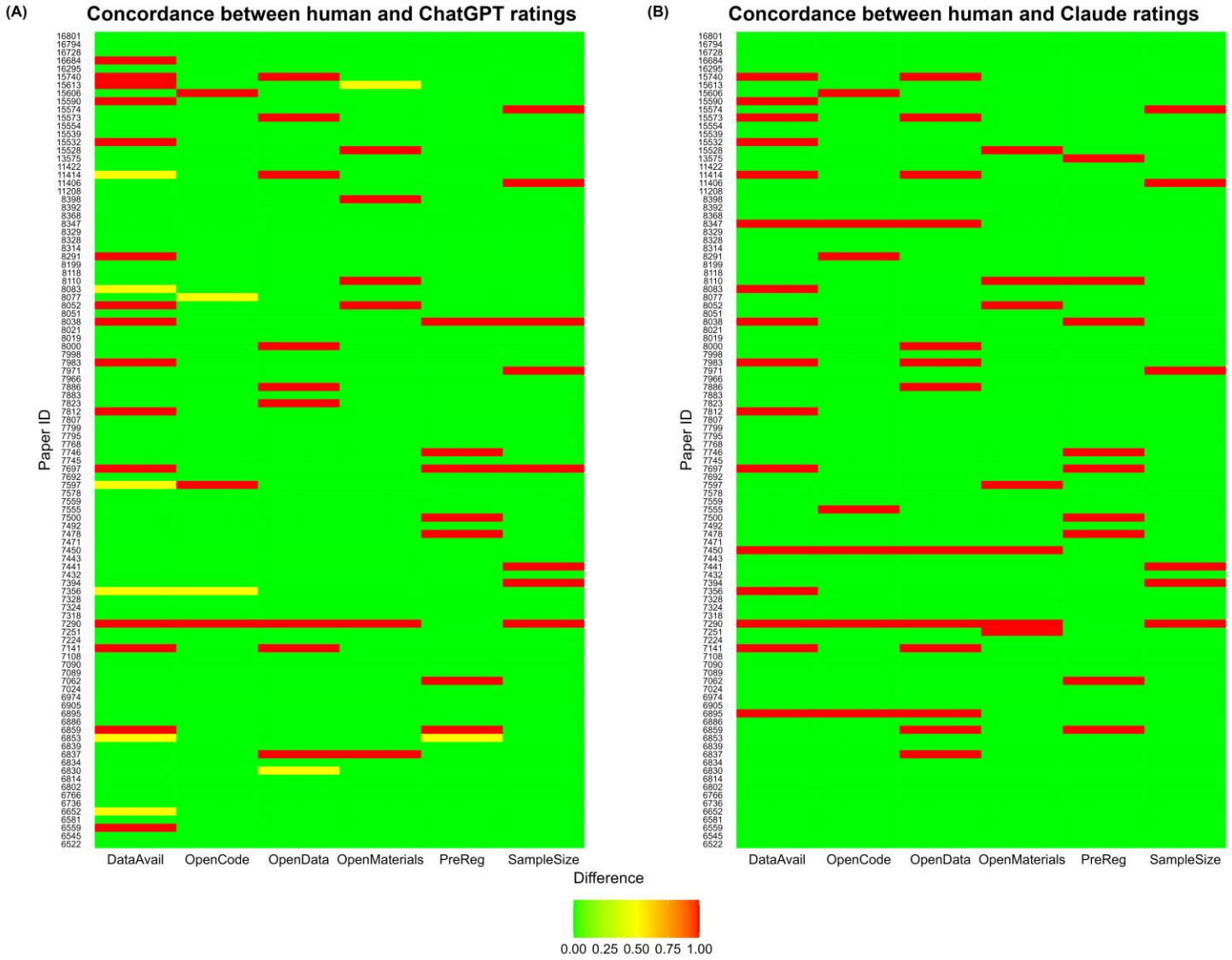
Score Distributions & the Floor Effect



- **Most articles scored 0 out of 6**
- **Floor effect: high agreement may partly reflect shared zeros**

Item-Level Agreement

- **High concordance: OpenCode, OpenMaterials, PreReg, SampleSize**
- **Moderate concordance: DataAvail, OpenData**



Formal Item-Level Agreement & Divergences

Table 1

Gwet's AC1 Coefficients for Agreement between Human and Chatbot Item-Level Ratings

Item	ChatGPT AC1 (95% CI)	Claude AC1 (95% CI)
DataAvail	.79 [.67, .90]	.77 [.66, .89]
OpenCode	.95 [.90, 1.00]	.91 [.85, .98]
OpenData	.87 [.78, .95]	.82 [.71, .92]
OpenMaterials	.93 [.88, .99]	.92 [.86, .98]
PreReg	.91 [.84, .97]	.90 [.83, .97]
SampleSize	.90 [.83, .97]	.92 [.86, .99]

Note. AC1 coefficients index agreement between the human consensus ratings and each chatbot (ChatGPT 4 and Claude 3.5 Sonnet) for each item.

Table 2

Divergences between Human and Chatbot Ratings of Open Science Practices

Item	Chatbot	Divergences	Human Higher	Chatbot Higher
DataAvail	ChatGPT	21	20	1
	Claude	16	12	4
OpenCode	ChatGPT	5	4	1
	Claude	7	2	5
OpenData	ChatGPT	10	4	6
	Claude	13	3	10
OpenMaterials	ChatGPT	7	6	1
	Claude	7	4	3
PreReg	ChatGPT	8	7	1
	Claude	9	9	0
SampleSize	ChatGPT	8	8	0
	Claude	6	6	0

Note. $N = 100$ articles. Divergences indicate the number of articles where chatbot and human ratings differed. 'Human Higher' means the human raters scored the item as 1 and the chatbot scored it as 0, and *vice versa* for 'Chatbot Higher'.

Further Analyses

1. **Systematic differences (Friedman test)**
    - Significant but small effect (Kendall's W = .12)
    - Humans scored slightly higher than ChatGPT
    - Median difference = 0 for all pairs
2. **Subgroup analysis (articles with non-zero scores, *n* = 48)**
    - ICC (single rater) = .42; ICC (averaged) = .68
    - Agreement extends beyond simply detecting absence
3. **Model robustness (ChatGPT 5 Pro) (September 2025)**
    - ICC (single rater) = .65; ICC (averaged) = .88
    - Strong correlation with human ratings (*r* = .67)
    - Core findings hold across model iterations

Illustrative Examples

"Custom computer code used to generate the findings of this study will be made available upon request to the Lead Contact. Simulation code can be found in..."

Open Code		
Human	ChatGPT	Claude
0	0.5	0

"The data that support the findings of this study and the custom code for data analysis are available upon reasonable request. The source data underlying Figs. 2–5 and Supplementary Figures 1, 3–7 are provided as a Source Data file."

Data Availability Statement		
Human	ChatGPT	Claude
0	0	1

"Our sample size was devised to allow us at least an 85% power to detect a difference in the prevalence of SDQ caseness from 10 to 20% between the paternal PTSD case group and the paternal non-PTSD case group."

Sample Size Rationale		
Human	ChatGPT	Claude
1	0	0

The Copyright Question

- **Uploading full articles to third-party AI services**
  - **Who owns the content?**
  - **Can it be used for model training?**
- **Our study: REF 2021 context**
  - **Open access policy required manuscript deposit**
  - **All embargoes expired by 2022 → green OA versions available**
- **Legal landscape evolving rapidly (two major US rulings in 2025)**

UK & EU Legal Landscape

UK

- Copyright, Designs and Patents Act 1988
- Section 29A: TDM exception for non-commercial research
- But copy must not be "transferred to any other person"

EU

- Directive on Copyright in the Digital Single Market
- Article 3: Research organisation exception
- Article 4: General TDM (subject to publisher opt-outs)

Key precedent: *Infopaq International A/S v Danske Dagblades Forening* (2009)

- Even 11-word extracts can constitute infringement

US Fair Use Doctrine

- 17 U.S.C. § 107: Four-factor balancing test
  - Purpose and character of use
  - Nature of the original work
  - Amount used
  - Effect on market
- Key question: Is the use “transformative”?

- Established precedents:
  - *Authors Guild v. Google* (2015)
  - *Authors Guild v. HathiTrust* (2014)
- Recent AI-specific rulings:
  - *Bartz v. Anthropic* (2025) – training on copyrighted works can be transformative
  - *Kadrey v. Meta* (2025) – no market harm should be demonstrated

Safeguards & Best Practices

Practical safeguards

- Prefer open access sources
  - OA articles, author-accepted manuscripts, preprints
- Prevent use in model training
  - Consumer versions: manually opt out in settings
  - Institutional tiers: excluded by default

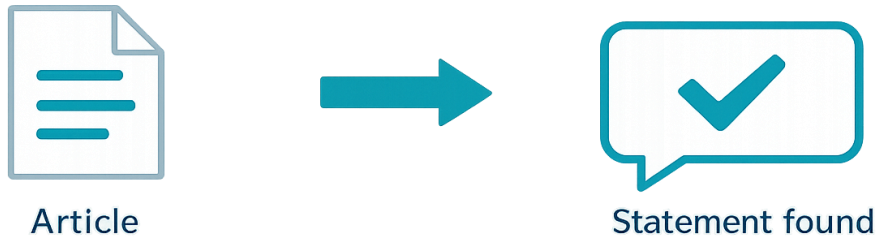
Ethical principles

- Transparency
  - Document methodology and safeguards applied
- Accountability
  - Align with institutional AI policies
- Data governance
  - Never upload sensitive data
- EU ALTAI; Lin (2025)

Practical Recommendations

- AI works well for (preliminary screening)
  - Code sharing statements
  - Materials availability
  - Preregistration mentions
  - Sample size rationale
- Human verification recommended for
  - Data availability statements (highest divergence)
  - Any item where AI scores “1” (verify links actually work)
- Suggested workflow
  - AI for initial large-scale screening
  - Human review for flagged or ambiguous cases
  - Final human sign-off for consequential decisions

Beyond Statements: The Next Frontier



- **Current AI limitation**
  - Can only assess what is *stated* in the article
  - Cannot verify external links or repository contents
- **The verification gap**
  - Does the link actually work?
  - Is the shared content complete and appropriate?
  - Can the analysis be reproduced?
- **Next frontier: computational verification**
  - AI agents that can browse, download, and execute code
  - Early efforts: CORE-Bench (Siegel et al., 2025); REPRO-Bench (Huang et al., 2025)
  - Moving from transparency *claims* to transparency / replicability *reality*

Agentic AI for Reproducibility

Terminal-based coding agents	Development environments
Claude Code (Anthropic)	VS Code (+ extensions)
Codex CLI (OpenAI)	Cursor
Gemini CLI (Google)	Antigravity

Enabling technologies

- Model Context Protocol (MCP): universal connector for AI agents
- Agent Skills: domain-specific expertise packages

Application to reproducibility (Current research direction)

- Clone repositories, download data, execute analysis code
- Verify whether reported results can be reproduced

Take-Home Message

Embrace AI for evaluating open science practices

- AI chatbots show moderate-to-good agreement with experts
- Legal and ethical concerns are addressable
- Human oversight remains essential
- Agentic tools offer rapidly maturing capabilities

The Research Team

DR

Yu Heng Daryl Lee

DR

Adam Parker

PROF

Courtenay Norbury

PROF

David Shanks

Funding: ESRC grants ES/S014616/1 and ES/Y002482/1, UCL Research Culture grant

Resources

Custom GPT

RSOS (2026)

Thank you!

Please confirm your attendance today by following the QR code

We value your feedback:

UNIVERSITY OF WARWICK