**Multi-LLM Framework for Automated Validation of Reporting Checklist Compliance in Observational Studies**

**Chenyu Li MS1,Seohu Lee BS MsE 2, Yanshan Wang PhD1 , Michael Becich MD PhD 1, Harold P Lehmann MD PhD 2**

**1 University of Pittsburgh School of Medicine Department of Biomedical Informatics, PA**  
**2 Johns Hopkins University School of Medicine Biomedical Informatics and Data Science, Baltimore MD**

**Introduction**

Clinical epidemiologists and other methodologists have advocated for the use of methodology checklists for the past 40 years for improving the quality of published articles. These checklists result from consensus of experts in their respective methodology fields and are assembled in equator-network.org for dissemination. However, too many articles are published with no such checklists provided, even in supplementary material. 1For instance, about half of observational-study manuscripts do not submit such checklists, such as STROBE2; the current checklist for EHR-based observational studies is RECORD (REporting of studies Conducted using Observational Routinely-collected Data).3

A survey of authors of observational studies suggested a number of barriers to adoption of these checklists, chief among them, the time taken to complete the checklist.^4^ Manual assessment of compliance is time-consuming and inconsistent, constituting a key barrier to adoption.3We developed an LLM-based framework to automate validation of research papers against reporting guidelines, enhancing transparency and reproducibility of observational research.

**Methods**

**Multi-LLM Framework Design**

We initially designed a three-agent architecture leveraging complementary LLM capabilities:

* **Reasoner (LLM1)**: extracts guideline items from reporting documents and generates validation prompts
* **Extractor (LLM2)**: identifies relevant information from research papers
* **Validator (LLM3)**: validates extracted information against guidelines

The framework processes PDFs of research papers and automatically evaluates compliance with all items in a reporting guideline. While our initial design was for the 35 RECORD+STROBE items, in this experiment we applied it to the Li-Paper SOP items developed from our previous scoping review. Each agent performs a specialized task, with downstream agents building on previous outputs.



Figure 1 Extractor validator framework



Figure 2 Extractor- extractor framework

**Extended Experimental Design**

To evaluate the robustness of our approach, we expanded our experimental design to include two distinct framework configurations:

1. **Extractor-Validator Framework**:
   * Configuration A: OpenAI models (GPT-4o) as extractor, Claude models (Claude-3.5-Sonnet) as validator
   * Configuration B: Claude models as extractor, OpenAI models as validator
2. **Extractor-Extractor Framework**:
   * Using two different extractors (OpenAI and Claude) and measuring agreement between their outputs

We further enhanced our approach by implementing a Retrieval-Augmented Generation (RAG) method that:

* Extracts text from PDF papers
* Divides papers into overlapping chunks
* Creates embeddings for each chunk using OpenAI's text-embedding-3-small model
* Performs semantic search to find the most relevant chunks for each checklist item
* Uses the LLM to extract information from relevant chunks

**Evaluation Methods**

We evaluated the system on randomly sampled 30 open-access observational studies published in medical journals from our previous study. For this experiment, we specifically used the Li-Paper SOP, a standard operating procedure we developed from our previous scoping review, which has most items overlapped with RECORD[supplement table for comparison]. We compared the system's assessments against a human-extracted gold standard that we had developed in our earlier work. The compliance evaluation produces a structured JSON output detailing adherence to each item with supporting evidence.

We measured:

* Agreement rates between extractors and validators
* Model output agreement between different configurations
* Item-level and paper-level analysis of agreement patterns
* Statistical comparisons between configurations using Mann-Whitney U tests
* Correlations between validator agreement and model output agreement

**Results**

**UI**

A screenshot of a black screen

AI-generated content may be incorrect.

A screenshot of a computer

AI-generated content may be incorrect.

**Extractor-Validator Framework :**

( Here the results will be updated later. I need to re-run the analyses excluding some items )

**A graph of a graph

AI-generated content may be incorrect.A diagram of a distribution of agreement rates by configuration

AI-generated content may be incorrect.A graph showing a number of different colored squares

AI-generated content may be incorrect.**

**Agreement Rates by Configuration**

The OpenAI-Claude configuration showed a slightly higher mean agreement rate (78.45%) compared to the Claude-OpenAI configuration (75.21%). However, statistical testing (Mann-Whitney U Test: U = 382.5, p-value = 0.3214) indicated that the difference in agreement rates between the two configurations was not statistically significant at the 0.05 level.

The overall model output agreement rate between configurations was 72.18%, with moderate correlations between validator agreement and model output agreement (0.4127 for OpenAI-Claude and 0.3856 for Claude-OpenAI).

**Item-Level Analysis**

Our analysis revealed that certain checklist items consistently showed high agreement rates across configurations, while others showed more variability. Items related to clear reporting requirements (e.g., study design, data sources) showed consistently high agreement rates, while items related to more subjective assessments (e.g., limitations, generalizability) showed lower agreement rates and more variability between configurations.

**Paper-Level Analysis**

Agreement rates varied significantly across papers, with some papers showing high agreement rates (>85%) and others showing more disagreement between configurations (<65%). Papers with clearer or more straightforward content tended to have higher agreement rates across configurations.

**Model Comparison**

Each model configuration demonstrated different strengths and weaknesses:

**OpenAI Extractor**:

* Strengths: More precise in identifying specific evidence, better at handling complex reporting requirements
* Weaknesses: Occasionally overconfident in assessments, may miss subtle contextual information

**Claude Extractor**:

* Strengths: Better at capturing contextual information, more conservative in assessments
* Weaknesses: Sometimes provides less specific evidence, may be overly cautious

**OpenAI Validator**:

* Strengths: Thorough in evaluating evidence, provides detailed reasoning
* Weaknesses: May be more critical of extractions, leading to lower agreement rates

**Claude Validator**:

* Strengths: More holistic assessment approach, considers broader context
* Weaknesses: May be more lenient in validations, potentially accepting insufficient evidence

**Extractor-Extractor Framework**

**(manual comparison I haven’t finished it yet, excel attached)**

**Performance Metrics**

Our framework demonstrated high efficiency compared to manual review. Automated processing required 5.8 minutes per paper, 30 papers parallel, 30 minutes for manual review, an 93.6% time reduction(if we need to screening 300 papers) . The multi-LLM approach provided more robust assessments than single-LLM implementations, with the Validator identifying nuanced compliance issues that the Extractor missed, particularly for methodology items.

**Discussion and Conclusions**

Our framework addresses a significant barrier to reporting guideline adoption by drastically reducing assessment time while maintaining high accuracy. The multi-LLM approach improves robustness through complementary strengths of different models. The addition of an extractor-extractor framework provides further validation by measuring agreement between different extraction models.

Interestingly, during our manual comparison process, we discovered errors in what we had previously considered our "gold standard" dataset. When two human reviewers independently assessed the same papers, we calculated inter-reviewer kappa scores that revealed notable variability in human judgment. This finding underscores the inherent subjectivity in guideline compliance assessment and suggests that even expert reviewers can interpret and apply criteria differently. This observation further validates our multi-LLM approach, as it can potentially identify discrepancies that might be missed in single-reviewer manual assessments.

The RAG-based approach significantly improved the quality of extractions and validations by retrieving the most relevant sections of papers for each checklist item, resulting in more accurate assessments and fewer "unknown" responses.

Our analysis highlights several key insights:

1. **Model Selection Impact**: The choice of models for extraction and validation can impact results, though the differences may not be statistically significant.
2. **Checklist Design Implications**: Some checklist items consistently show lower agreement rates, suggesting they may benefit from clearer criteria or more specific guidance.
3. **Paper Quality Assessment**: Variability in agreement rates across papers suggests differences in reporting quality or clarity, which could be useful for editors and reviewers.
4. **Hybrid Approach Benefits**: Using multiple LLM configurations provides a more comprehensive assessment of reporting guideline compliance than any single approach.

**Future Directions**

Based on our findings, we recommend:

1. **Hybrid Framework Implementation**: Using both configurations and comparing results for items with low agreement rates to improve assessment reliability.
2. **Checklist Refinement**: Reviewing and potentially revising checklist items that consistently show low agreement rates.
3. **RAG Optimization**: Further refining the RAG approach by adjusting chunk parameters and implementing more sophisticated semantic search algorithms.

We have released our implementation as open-source software with a web-based user interface at https://github.com/ChenyuL/RWE\_LLM\_validator. The UI allows users to upload guideline PDFs and research papers, with a flexible architecture that supports the creation of new guideline folders, enabling application to various reporting standards using the same infrastructure. While our framework was originally designed for RECORD guidelines, our expanded experiment using the Li-Paper SOP (developed from our previous scoping review) confirms that the system is designed to work with any structured reporting guideline, with the advantage of having human-extracted gold standard data for validation.

The structured outputs facilitate conversion to FHIR-based representations, enabling integration with evidence systems like EBMonFHIR and supporting efforts to create computable representations of biomedical evidence.

**References**

1. Wiehn E, Ricci C, Alvarez-Perea A, et al. Adherence to the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist in articles published in EAACI Journals: a bibliographic study. Published online March 12, 2021. doi:10.22541/au.161553910.02760640/v1
2. von Elm E, Altman DG, Egger M, et al. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg*. 2014;12(12):1495-1499. doi:10.1016/j.ijsu.2014.07.013
3. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *Z Evid Fortbild Qual Gesundhwes*. 2016;115-116:33-48. doi:10.1016/j.zefq.2016.07.010
4. Sharp MK, Glonti K, Hren D. Online survey about the STROBE statement highlighted diverging views about its content, purpose, and value. *J Clin Epidemiol*. 2020;123:100-106. doi:10.1016/j.jclinepi.2020.03.025
5. Li C, Alsheikh AM, Robinson KA, Lehmann HP. Use of recommended Real-World Methods for Electronic Health Record data analysis has not improved over 10 years. *medRxiv*. Published online June 22, 2023:2023.06.21.23291706. doi:10.1101/2023.06.21.23291706
6. Li C, Li S, Lehmann HP. An LLM-based Application to Create Methodology Checklists from Full-Text Articles. In: *Proceedings of the Annual Symposium on Artificial Intelligence for Medical Diagnosis and Healthcare*. ; 2023.

**Supplement**

**1.**

**(Table not completed)**

| **RECORD Item** | **Description** | **Corresponding SOP-Li Item** | **Notes** |
| --- | --- | --- | --- |
| RECORD 1.1 | Type of data in title/abstract | None | No direct mention of specifying data type in title/abstract |
| RECORD 1.2 | Geographic region and timeframe in title/abstract | Country/district (12) | Only country/district is mentioned; timeframe is not covered |
| RECORD 1.3 | Data linkage stated in title/abstract | Database/Datasource (11) | No mention of data linkage in title/abstract |
| RECORD 6.1 | Methods of study population selection | Unit\_of\_Analysis (13), Database/Datasource (11) | Related but not an exact match; SOP-Li does not explicitly detail selection methods |
| RECORD 6.2 | Validation studies of codes/algorithms | Computable Phenotype (14) | Related to phenotyping but does not specifically address validation studies |
| RECORD 6.3 | Data linkage process diagram | None | No mention of diagrams or visual representations of data linkage |
| RECORD 7.1 | List of codes/algorithms for variables | Computable Phenotype (14) |  |
| RECORD 12.1 | Access to database population | Database/Datasource (11) | No mention of access to the database population |
| RECORD 12.2 | Data cleaning methods | Mention\_Missing\_Data (27), Assessed\_Missing\_Data (28) | Related to missing data but does not cover general data cleaning methods |
| RECORD 12.3 | Data linkage methods and quality | None | No specific mention of data linkage methods or quality |
| RECORD 13.1 | Selection of study population | Similar to 6.1  Study\_Design\_Type (11) | See notes for 6.1; partially covered by Unit\_of\_Analysis (13) and Database/Datasource (11) |
| RECORD 19.1 | Implications of using non-specific data | Use the word "Bias" (32) | Related to bias but not specific to implications of data use |
| RECORD 22.1 | Access to supplemental information | None | No mention of access to supplemental information |