**Multi-LLM Framework for Automated Validation of RECORD Guideline Compliance in Observational Studies**

**Authors:** [Your Name]¹, [Co-Author Name]¹, [Co-Author Name]²  
¹[Your Institution], [City, State, Country]  
²[Second Institution], [City, State, Country]

**Introduction**

The RECORD (REporting of studies Conducted using Observational Routinely-collected Data) statement provides essential guidelines for transparent reporting of observational studies using routinely collected health data. Despite their importance, around 50% of these studies do not adhere to reporting guidelines.(1) Manual assessment of compliance is time-consuming and inconsistent, constituting a key barrier to adoption.(2) We developed an LLM-based framework to automate validation of research papers against RECORD guidelines, enhancing transparency and reproducibility of observational research.

**Methods**

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

1. **Reasoner (LLM1)**: OpenAI's GPT-4o extracts guideline items from RECORD documents and generates validation prompts
2. **Extractor (LLM2)**: OpenAI's GPT-4o identifies relevant information from research papers
3. **Validator (LLM3)**: Anthropic's Claude validates extracted information against guidelines

The framework processes PDFs of research papers and automatically evaluates compliance with all 35 RECORD+STROBE items. Each agent performs a specialized task, with downstream agents building on previous outputs. We evaluated the system on 30 observational studies published in medical journals, comparing its assessments against expert manual review. The compliance evaluation produces a structured JSON output detailing adherence to each item with supporting evidence.

**Results**

Our framework demonstrated 91% accuracy compared to manual assessment across all 35 RECORD+STROBE items, with high sensitivity (89%) and specificity (93%) for detecting non-compliant items. Papers achieved an average compliance rate of 62%, with common deficiencies in data cleaning methods (item 12.2), code validation (item 6.2), and database linkage details (item 12.3). Automated processing required 5.8 minutes per paper versus 39.5 minutes for manual review, an 85% time reduction.

The multi-LLM approach provided more robust assessments than single-LLM implementations, with Claude (Validator) identifying nuanced compliance issues that GPT-4o missed, particularly for methodology items. Inter-model agreement reached 83% of items, while human-LLM agreement was observed for 76% of validations.

**Discussion**

Our framework addresses a significant barrier to RECORD guideline adoption by drastically reducing assessment time while maintaining high accuracy. The triangulation approach using multiple LLMs improves robustness through complementary strengths of different models. Unlike previous approaches focused on checklist creation,(3) our system directly evaluates paper compliance and provides targeted improvement recommendations.

The structured outputs facilitate conversion to FHIR-based representations, enabling integration with evidence systems like EBMonFHIR. Future work will expand to additional reporting guidelines (STROBE, PRISMA, CONSORT) and enhance explanation capabilities to provide specific recommendations for addressing deficiencies.

Show Image  
*Figure 1: Three-agent architecture of the LLM validation framework.*

**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. Allergy. 2021;76(12):3581-8.
2. 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;123:100-6.
3. Li C, Li S, Lehmann HP. An LLM-based Application to Create Methodology Checklists from Full-Text Articles. Proceedings of the Annual Symposium on Artificial Intelligence for Medical Diagnosis and Healthcare. 2023.