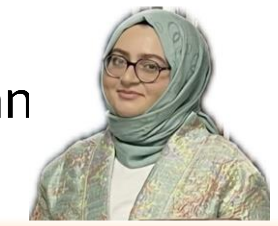
Leveraging Large Language Models (LLMs) for data extraction and quality assessment in psychiatry systematic reviews: A comparison of inter-rater reliability between Elicit and human coders.

Presenter: Cansu N. Erkan

Email: cansu.erkan@nih.gov



Erkan, C.N., Gu, G., Tandilashvili, E., Meigs, J.M., Lee, K., Metcalf, O., Livinski, A., Pine, D.S., Pereira, F., Brotman, M.A., & Henry, L.M.

INTRODUCTION

- Data extraction and quality assessment are critical yet labor-intensive and error-prone processes in conducting systematic reviews[1].
- Large language models (LLMs) have the potential to reduce human labor and enhance efficiency in this process.
- In existing research applying LLMs to systematic reviews, accuracy remains variable in data extraction and largely unexplored in quality assessment[2].
- In our systematic review, we utilized Elicit, a set of commercially available LLMs designed for systematic reviews, as a secondary coder for both data extraction and quality assessment.
- **Objective:** To evaluate Elicit's accuracy in data extraction and quality assessment compared to two human coders.
- **Hypothesis:** Inter-rater reliability between two human coders will be higher than between one human and one Elicit coder for both data extraction and quality assessment in our systematic review.

METHODS

Data Extraction & Quality Assessment:

- 10 research assistants extracted 176 data points (e.g., demographics, study design) from 229 articles.
- They also assessed 9 quality items (e.g., validity of measures) from 229 articles.
- 99 articles were coded by two human coders.
- 130 articles were coded by one human coder and Elicit[3].

Inter-Rater Reliability (IRR):

$$Data\ Extraction\ IRR = 100 - \frac{\#\text{ of discrepant data points}}{\text{total }\#\text{ of extracted data points}} * 100$$

$$Quality\ Assessment\ IRR = \frac{\#\text{ of items agreed upon}}{\text{total }\#\text{ of items}} * 100$$

Statistical Analysis:

- Independent samples t-tests were conducted using R 4.4.1 to compare reliability between groups.

Literature search → Screening → Data extraction and quality assessment → Analysis

Example Items for Data Extraction :

- General Study Characteristics (e.g., authors, journal, title)
- Sample Characteristics (e.g., sex, age, race, ethnicity)
- Clinical Characteristics (e.g., medication use, psychopathology, symptoms)
- EMA (e.g., compliance, drop-out, enrollment)

Example Items for Quality Assessment:

- Sample Description
- Study Procedure
- Formulation of Hypothesis
- Specification of inclusion/exclusion criteria

Items	Elicit Response	Human Response	Same?
List the measure(s) used in alphabetical order, separated by commas	Rutgers Alcohol Problem Index, Social Interaction Anxiety Scale, State Social Anxiety	Rutgers Alcohol Problem Index, Social Interaction Anxiety Scale, State Social Anxiety	**Yes**
What percentage compliance based on your calculations (if not a solely event-contingent study)? Compliance = (Resp/Pres)x100	42.068	42.1	**Yes**

Table 1. Items determining whether Elicit and human coder responses are the same in data extraction.

Large language models perform similarly to humans in quality assessment in systematic reviews.

Items	Elicit Response	Human Response	Same?
The formulation of the research question	Good (=2)	Good (=2)	**Yes**
Specification of in- and exclusion criteria	Good (=2)	Reasonable (=1)	**No**

Table 2. Items determining whether Elicit and human coder responses are the same in quality assessment.

Large language models hold promise for data extraction efficiency in systematic reviews.

	Min	Q1	Median	Q3	Max	Mean	SD
Human-Human	72.73	84.09	89.2	92.05	97.16	87.35	5.99
Human-Elicit	55.68	76.85	83.81	88.64	94.89	82.29	7.83

Table 3. Descriptive Statistics for data extraction inter-rater reliability.

	Min	Q1	Median	Q3	Max	Mean	SD
Human-Human	33.33	66.67	66.67	77.78	100	72.16	14.97
Human-Elicit	22.22	55.56	66.67	77.78	100	68.63	16.22

Table 4. Descriptive Statistics for quality assessment inter-rater reliability.

	t-value	df	p-value
Data Extraction	5.33	226	<0.01
Quality Assessment	1.68	225	0.09

Table 5. t-test results.

RESULTS

Data extraction:

- Human-human coders showed higher IRR (M=87.35, SD=5.97, range = 72.73 – 97.16) than human-Elicit coders (M=82.29, SD=7.83, range = 55.68 - 94.89), $t(226)=5.33$, $p<.001$.

Quality assessment:

- There was no difference between groups: human-human: M=72.17, SD=14.97, range = 33.33 – 100.00; human-Elicit: M=68.63, SD=16.22, range = 22.22 – 100.00, $t(225)=1.68$, p=0.094.

DISCUSSION

- Consistent with our hypothesis, data extraction IRR was higher among human-human coders than human-Elicit coders.
- Contrary to our hypothesis, quality assessment showed no significant difference between the groups, suggesting similar performance.
  - Given that quality assessment requires more analytical and subjective reasoning than data extraction, this finding was unexpected.
- While Elicit's data extraction performance has not yet reached the level of human coders, it shows promise for improving efficiency in evidence synthesis.
- LLMs may support data extraction and quality assessment in systematic reviews, helping to reduce human labor and errors.
- Future research can explore the application of LLMs across various research domains (e.g., neuroimaging data), examining their influence on prompt design, data extraction methods, and quality assessment processes.

REFERENCES

SCAN ME

- This research was supported by the Intramural Program of the NIMH: ZIAMH002778 and ZIAMH002786

- Authors have declared no conflict of interest.