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



INTRODUCTION

- Data extraction and quality assessment are critical yet labor-intensive and error-prone processes in conducting systematic reviews¹.
- 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 assessment².
- 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:

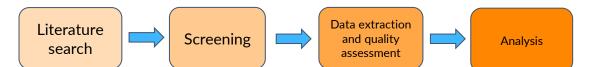
- 10 research assistants extracted 176 data points (e.g., demographics, study design) from 229 articles.
- They also assessed 9 quality items (e.g., validity of measures) from 229 articles.
- 99 articles were coded by two human coders.
- 130 articles were coded by one human coder and Elicit³.

Inter-Rater Reliability (IRR):

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

Statistical Analysis:

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



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

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.

Itoma	Elicit Descence	Human Deener	Come?				RESULTS				
Items	Elicit Response Rutgers Alcohol	Human Respons					Data extraction:				
List the measure(s) used in alphabetical order, separated by commas	Problem Index,	Rutgers Alcoho Problem Index Social Interactio Anxiety Scale, State Social Anxiety	n Yes	m simi			 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. 				
What percentage	, and ety			-	-						
compliance based on				1	n syster	natic	 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. 				
your calculations (if not a solely event- contingent study)?	42.068	42.1	Yes		reviev	VS.					
Compliance = (Resp/Pres)x100			Items	Elicit Response	Human Response	Same?	DISCUSSION				
Table 1. Items determining w responses are the same in dat Large lang hold pron	uage mo	dels	formulation of research question	Good (=2)	Good (=2)	Yes	 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. 				
extraction systema	efficien	cy in Spe VS.	cification of in- exclusion criteria e 2. Items determining	Good (=2) whether Elicit and	bod (=2) Reasonable (=1) No her Elicit and human coder responses are the same in		 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. 				
			ty assessment.				LLMs may support data extraction and				
Human-Human	Min 72.73	Q1 Med 8 84.09	ian Q3 89.2 92.0		1ean SD 87.35	5.99	quality assessment in systematic reviews, helping to reduce human labor and errors.				
Human-Elicit Table 3. Descriptive Statistics	55.66 for data extraction inte	 Future research can explore the application of LLMs across various research domains (e.g., neuroimaging data), examining their influence on prompt design, data extraction 									
Human-Human		Min Q 33.33		Q3 .67 77.78	Max Mean 100	n SD 72.16 14.97	methods, and quality assessment processes.				
Human-Elicit 22				5.67 77.78		68.63 16.22	• This research was				
		supported by the Intramural Program of the NIMH:									
Data Extraction	t-value 5.3	df 3 226	p-value <0.01			ZIAMH002778 and ZIAMH002786					
Quality Assessment		1.6									
Table 5. t-test results.		1.0	225	0.0			• Authors have declared no conflict of interest.				

Itoma	Elicit Descence	Human Deener	Come?				RESULTS				
Items	Elicit Response Rutgers Alcohol	Human Respons					Data extraction:				
List the measure(s) used in alphabetical order, separated by commas	Problem Index,	Rutgers Alcoho Problem Index Social Interactio Anxiety Scale, State Social Anxiety	n Yes	m simi			 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. 				
What percentage	, and ety			-	-						
compliance based on				1	n syster	natic	 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. 				
your calculations (if not a solely event- contingent study)?	42.068	42.1	Yes		reviev	VS.					
Compliance = (Resp/Pres)x100			Items	Elicit Response	Human Response	Same?	DISCUSSION				
Table 1. Items determining w responses are the same in dat Large lang hold pron	uage mo	dels	formulation of research question	Good (=2)	Good (=2)	Yes	 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. 				
extraction systema	efficien	cy in Spe VS.	cification of in- exclusion criteria e 2. Items determining	Good (=2) whether Elicit and	bod (=2) Reasonable (=1) No her Elicit and human coder responses are the same in		 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. 				
			ty assessment.				LLMs may support data extraction and				
Human-Human	Min 72.73	Q1 Med 3 84.09	ian Q3 89.2 92.0		1ean SD 87.35	5.99	quality assessment in systematic reviews, helping to reduce human labor and errors.				
Human-Elicit Table 3. Descriptive Statistics	55.66 for data extraction inte	 Future research can explore the application of LLMs across various research domains (e.g., neuroimaging data), examining their influence on prompt design, data extraction 									
Human-Human		Min Q 33.33		Q3 .67 77.78	Max Mean 100	n SD 72.16 14.97	methods, and quality assessment processes.				
Human-Elicit 22				5.67 77.78		68.63 16.22	• This research was				
		supported by the Intramural Program of the NIMH:									
Data Extraction	t-value 5.3	df 3 226	p-value <0.01			ZIAMH002778 and ZIAMH002786					
Quality Assessment		1.6									
Table 5. t-test results.		1.0	225	0.0			• Authors have declared no conflict of interest.				

														DECI	ιтс			
Ite	ems			Human Resp	onse	e Same?								RESULTS				
used in al order, se con	measure(s) Iphabetical parated by nmas	Problem Soc Intera Anxiety State	action	Rutgers Alco Problem Ind Social Intera Anxiety Sco State Soci Anxiety	dex, ction ale, ial		Yes		Large language models perform similarly to humans in quality assessment					 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: 				
compliance your calconot a sol continge	ce based on ulations (if ely event- ent study)?	42.(068	42.1		Yes			in system reviews					 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. 			:	
	liance = Pres)x100					Items		Elicit	Response	Human Response	е	Same?		DISCUSSION				
Table 1. Items determining whether Elicit and human coder responses are the same in data extraction. Large language models						The formulation of the research question			ood (=2)	Good (=2)		Yes		 Consistent with our hypothesis, data extraction IRR was higher among humanhuman 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 				
hold promise for data extraction efficiency in systematic reviews.					Specification of in- and exclusion criteria Table 2. Items determining v quality assessment.				ood (=2) her Elicit and	Reasonable (=1) d human coder resp	easonable (=1) No							
		Median Q3			Ma	Max Mean SD					quality assessment in systematic reviews,							
Human-Hum Human-Elicit Table 3. Desc		for data ex	72.73 55.68 straction inte	76.85	89.2 92.05 83.81 88.64 /.			_	97.16 87.35 94.89 82.29		5.99 7.83		 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 			ation ins		
	Min Human-Human 33.3 Human-Elicit 22.2				22 55.56 66		6.67 6.67	23 77.78 77.78 tive Statistics		72 68	72.16 68.63	14.97 16.22 r reliability	methoo REFERE	ds, and qu	 mpt design, data extraction ality assessment processes. This research was supported by the Intramural Program of 	esses.		
t-value			df		p-v	alue							the NIMH:					
Data Extraction						226		<0.02	1				ZIAMH002778 a ZIAMH002786					
Quality Assessment				1.68 22			5	0.09			 Authors have declar 			ared				
Table 5. t-tes	st results.													SCAN	I ME	no conflict of intere		

