Adapting a Large Language Model (LLM) to Assess Clinical Ratings of Thought Disorder in Psychosis

Ryan Partlan, Simran Bhola, Sandy Yin, Zak Singh*, Sunny Tang

Feinstein Institutes for Medical Research
Northwell Health

TANG LAB

Contact me!
rpartlan@northwell.edu

Data and Objectives

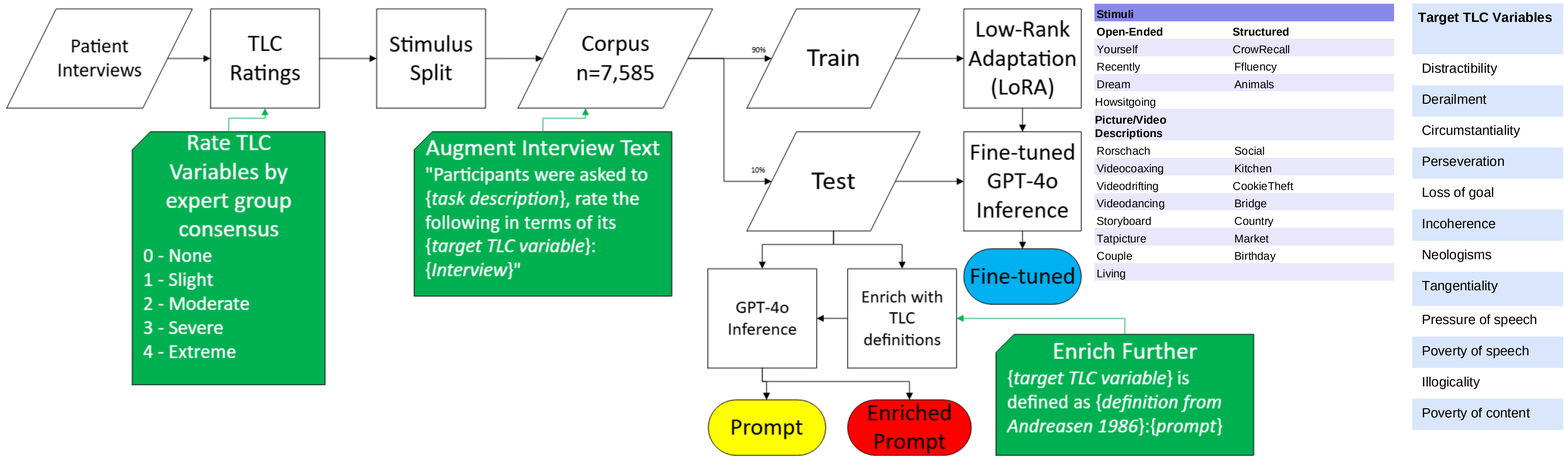
Expert annotators perform post-hoc analyses to assign ratings on the Scale for Assessment of Thought, Language, and Cognition (TLC) to transcripts from participants responding to language tasks.

We combined annotated transcripts from four studies:

	Healthy Volunteers	Any Psychiatric Disorder (PD)	
n=640	279 (43.6%)	361 (56.4%)	
Mean Age	26.3		29.9
Gender			
Man	111(17.3%)	214 (33.4%)	
Woman	147 (23%)	132 (20.6%)	
Non-binary	20 (3%)	10 (6.4%)	
Unknown	1 (0.16%)	5 (3.2%)	
Diagnosis			
Schizophrenia		0 158 (24.7%)	
Bipolar + Psychosis		0 48 (7.5%)	
Unspecified PD		0 48 (7.5%)	
Schizoaffective		0 38 (5.9%)	
Schizoaffective-BT		0 23 (3.6%)	
Schizoaffective-DT		0 16 (2.5%)	
Schizophreniform		0 14 (2.2%)	
MDD + Psychosis		0 14 (2.2%)	
Brief PD		0 1 (0.16%)	
Substance-induced PD		0 1 (0.16%)	
None	279 (43.6%)		0

- We investigate the capability of LLMs to assign TLC ratings
- We discern the linguistic tasks for which LLMs make the best symptom assessments.
- We discern which TLC variables LLMs can accurately evaluate.

Methods



Stimuli	
Open-Ended	**Structured**
Yourself	CrowRecall
Recently	Ffluency
Dream	Animals
Howsitgoing	
Picture/Video Descriptions	
Rorschach	Social
Videocoaxing	Kitchen
Videodrifting	CookieTheft
Videodancing	Bridge
Storyboard	Country
Tatpicture	Market
Couple	Birthday
Living	

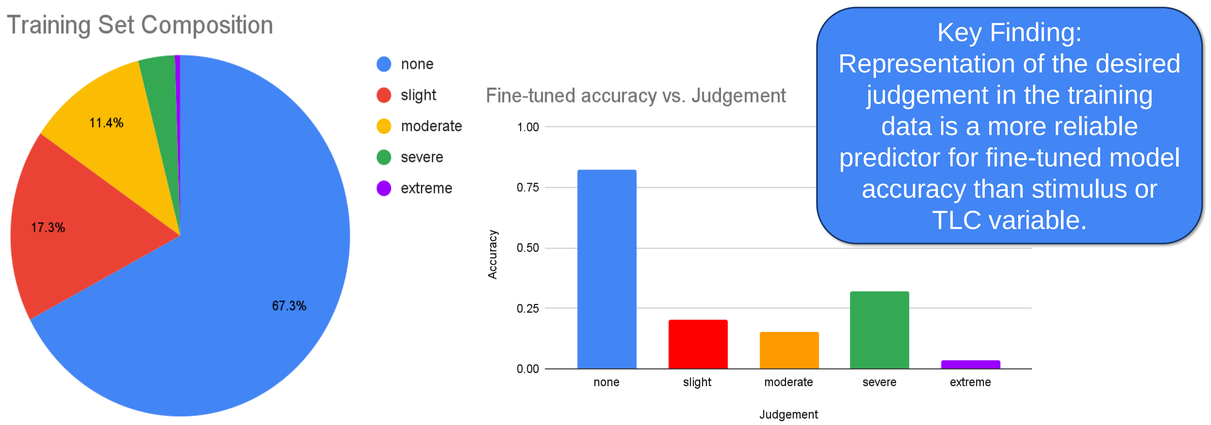
Target TLC Variables
Distractibility
Derailment
Circumstantiality
Perseveration
Loss of goal
Incoherence
Neologisms
Tangentiality
Pressure of speech
Poverty of speech
Illogicality
Poverty of content

Key Findings

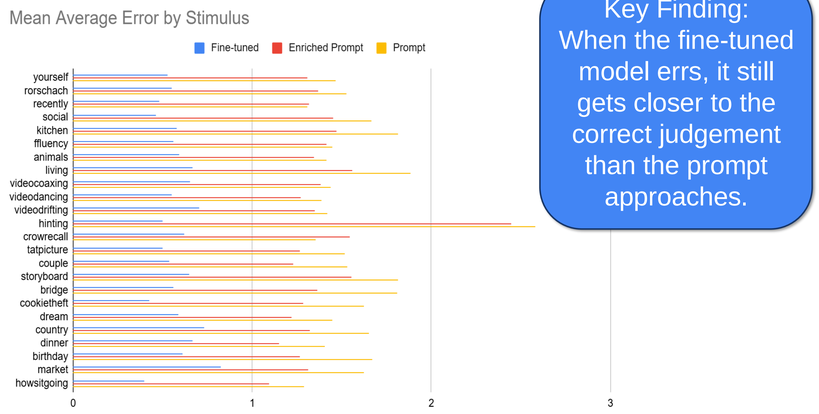
- Fine-tuned model is most effective and performs consistently across stimuli.
- All approaches have difficulty predicting scores for variables that rely on relevance and persistence of themes and topics, i.e tangentiality and circumstantiality
- Prompt enrichment renders negligent performance increase.
- Mean average error indicates that the fine-tuned model reasons beyond the training distribution.

Approach	Overall Accuracy	Mean Average Error
Prompt	0.10	1.6
Enriched Prompt	0.12	1.4
Fine-tuned	0.62	0.55

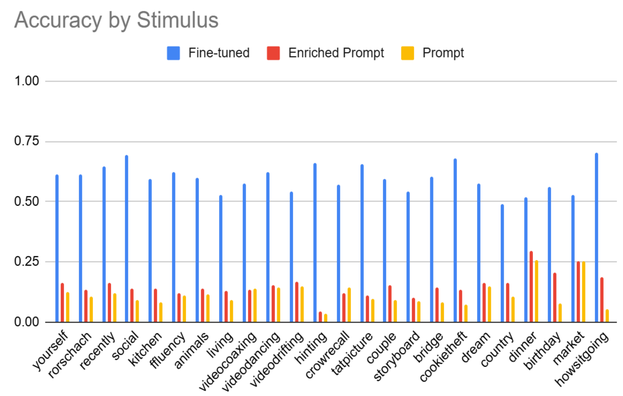
Results

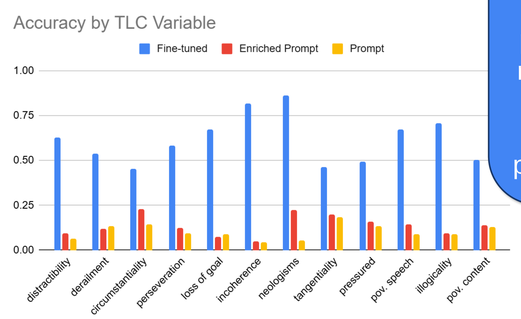


Key Finding: Representation of the desired judgement in the training data is a more reliable predictor for fine-tuned model accuracy than stimulus or TLC variable.

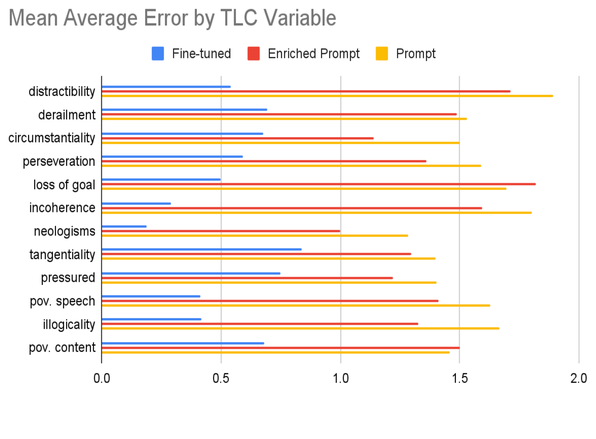


Key Finding: When the fine-tuned model errs, it still gets closer to the correct judgement than the prompt approaches.





Key Finding: Prompt enrichment increases performance across most categories, but not reliably. Fine-tuned model performs best overall



References

[1]Nancy C. Andreasen, Scale for the Assessment of Thought, Language, and Communication (TLC), Schizophrenia Bulletin, Volume 12, Issue 3, 1986, Pages 473–482, https://doi.org/10.1093/schbul/12.3.473

Institutions/Disclosures

Author Zak Singh is at Cambridge University, all other authors are at the Feinstein Institutes.

SXT received research funding from Winterlight Labs and holds equity with North Shore Therapeutics. She is also a consultant for both entities. She is on the advisory board for Psyrin, and serves as a consultant for Catholic Charities Neighborhood Services and LB Pharmaceuticals.