AI for assessing confidence in published findings

Open Science Bootcamp!
August 9th 2023

Sarah Rajtmajer
Assistant Professor
The Pennsylvania State University
smr48@psu.edu
www.rajtmajerlab.net

PennState

rockethics institute

Using ML and AI to understand science

ACM DIGITAL LIBRARY

≡ Article Navigation

CEUR Workshop Proceedings • Volume 2831 • 2021 • 2021 Workshop on Scientific Document Understanding, SDU 2021 • Virtual, Online • 9 February 2021 • Code 167870

Extraction and Evaluat
Information from Soci
Science

Understanding and predicting retractions of

WWW > Proceedings > WWW '22 > *A Study of Computational Reproducibility using URLs Linking to Open Access Datasets and Software*

SHORT-PAPER OPEN ACCESS

Sree Sai Teja Lank

ng URLs
e

PLOS ONE

RESEARCH ARTICLE

The evolution of scientific literature as metastable knowledge states

Sai Dileep Koneru[1]*, David Rench McCauley[2], Michael C. Smith[2], David Guarrera[2], Jenn Robinson[2], Sarah Rajtmajer[1]

1 The Pennsylvania State University, University Park, PA, United States of America, **2** Ernst & Young, McLean, VA, United States of America

* sdk96@psu.edu

Abstract

The problem of identifying common concepts in the sciences and deciding when new ideas have emerged is an open one. Metascience researchers have sought to formalize principles underlying stages in the life cycle of scientific research, understand how knowledge is transferred between scientists and stakeholders, and explain how new ideas are generated and

ducibility and Replicability of
ure Recognition Methods

Iasan Choudhury, Sarah M. Rajtmajer, and Jian Wu

l Dominion University, Norfolk VA 23529, USA
State University, University Park PA, USA
001,mchou001,j1wu}@odu.edu
[3] smr48@pdu.edu

out reproducibility in artificial intelligence (AI)
hers have reported unsuccessful attempts to di-
d findings in the field. Replicability, the ability
the same procedures on new data, has not been
r, we examine both reproducibility and replica-
apers on table structure recognition (TSR), an

Highlight: PSU effort for DARPA's SCORE program (September 2019 – May 2023)

DEFENSE ADVANCED RESEARCH PROJECTS AGENCY

EXPLORE BY TAG

ABOUT US / OUR RESEARCH / NEWS / EVENTS / WORK WITH US /

Defense Advanced Research Projects Agency > Our Research > Systematizing Confidence in Open Research and Evidence

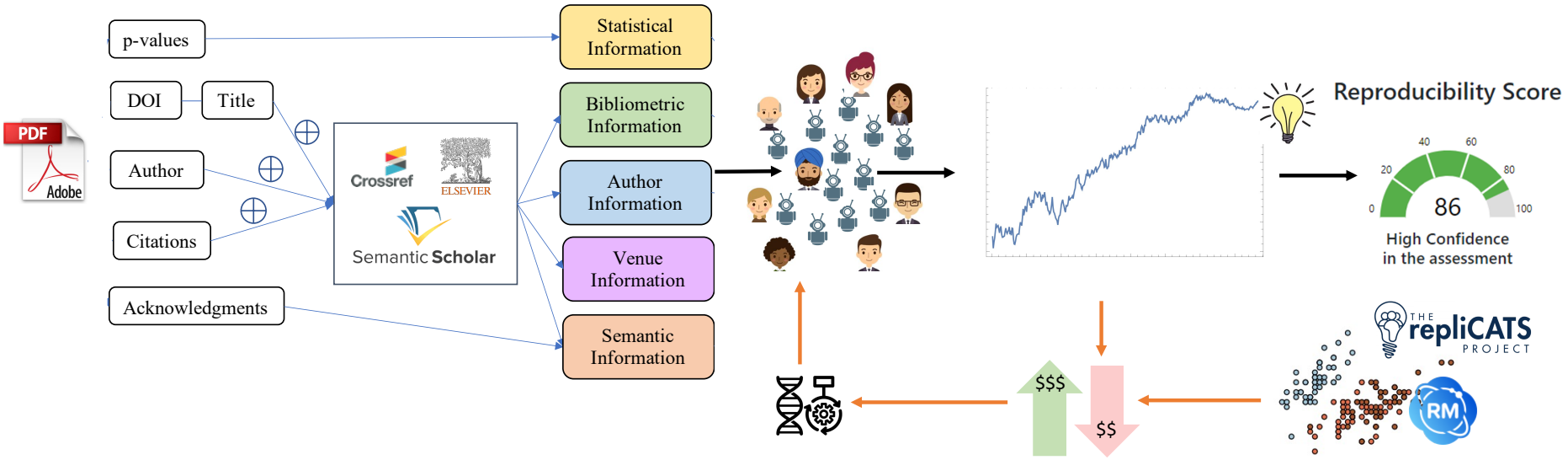
Systematizing Confidence in Open Research and Evidence (SCORE)

Dr. Greg Witkop

The Department of Defense (DoD) often leverages social and behavioral science (SBS) research to design plans, guide investments, assess outcomes, and build models of human social systems and behaviors as they relate to national security challenges in the human domain. However, a number of recent empirical st endently reproduced or repl on, DARPA's Systematizing Cor "confidence scores" to differen er of SBS research to unders explainable confidence scores nable DoD personnel to quick aim, and thereby increase the effective use of SBS literature and research to address important human domain challenges, such as enhancing deterrence, enabling stability, and reducing extremism.

Develop and deploy tools to assign *explainable* "confidence scores" to research results and claims

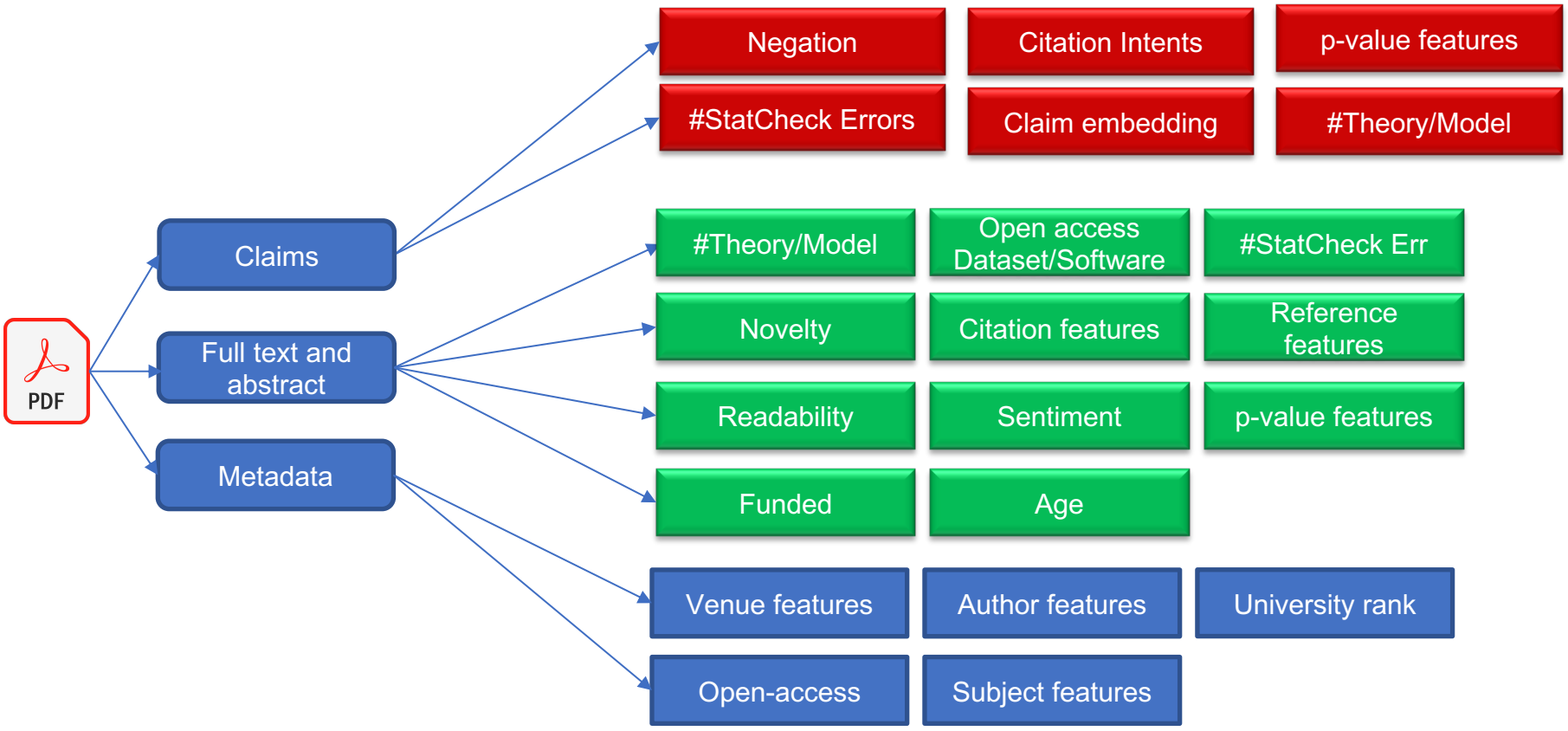
XAI (artificial prediction markets) *and* crowd+AI hybrid markets



Artificial prediction markets — *populated by artificial agents (trader-bots)* — purchase assets representing "will replicate" and "will not replicate" outcomes of notional replications of claims appearing within research papers. Agent reasoning is based on ***human-interpretable signals***, including full text of scientific papers, metadata for specific papers, and metadata about the community and the field.

Hybrid scenario: SMEs engage alongside bot traders

Signals (features) extracted from full text and assembled from metadata



12 Claim level features

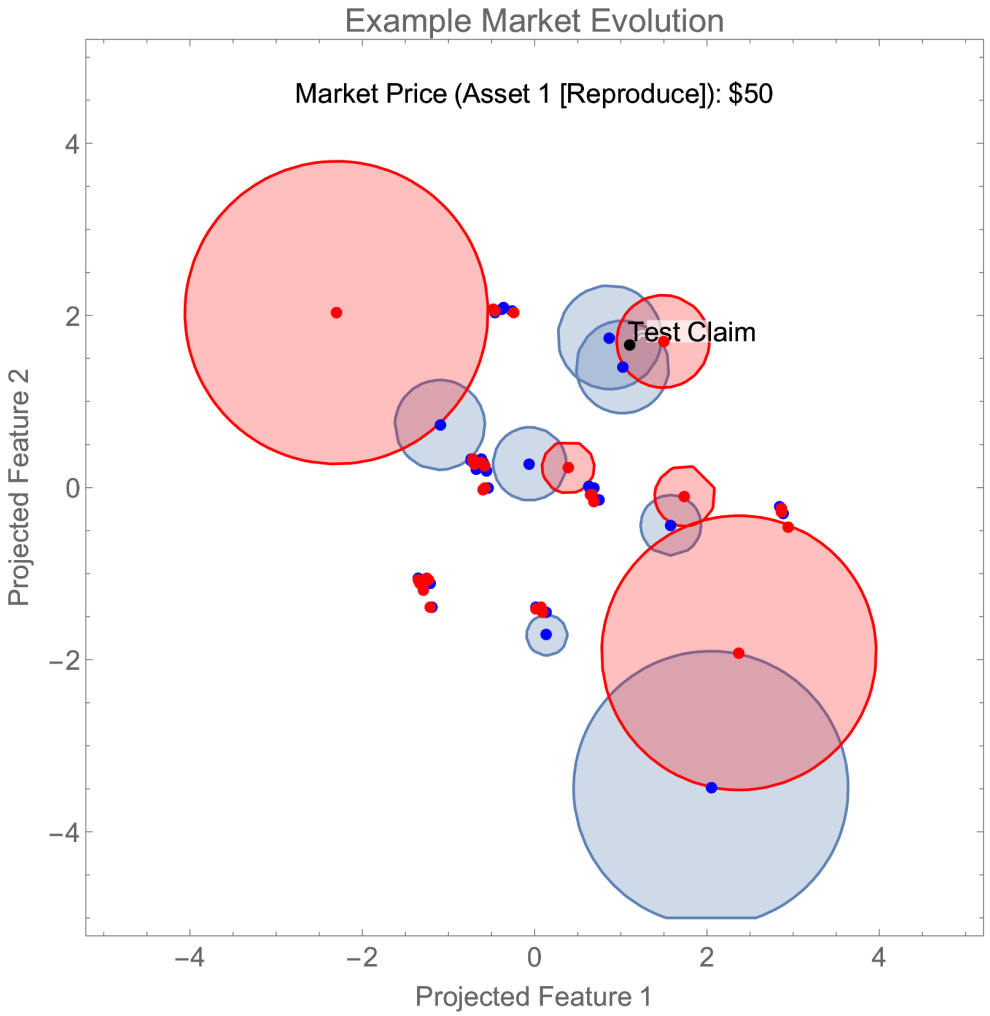
33 Citation features

15 Metadata features

Total:
up to 60 features

Artificial prediction markets

- Synthetic agents interact in a simple binary option market using a logarithmic market scoring rule.
- Agents in the market bid in geometric regions of feature space, shown as circles (for simplicity).
- The agents are sensitive to asset price, which causes their bid behavior to evolve in time.
- Convergence in the market is equivalent to a geometric equilibrium.

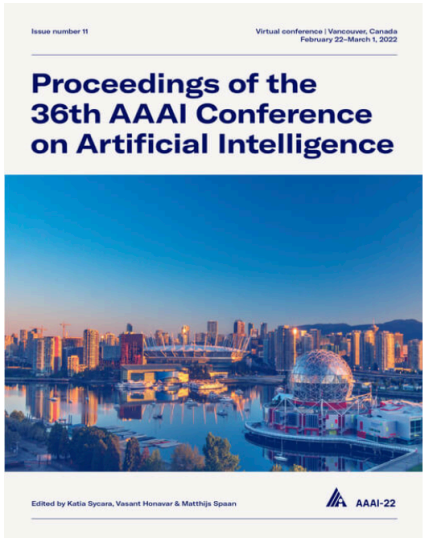


(above) A toy market with input data from RPP

Note 1: High dim feature space is projected down for visualization.
Note 2: We multiply the price by 100 and convert to dollars.)

Nakshatri et al. (2021) Design and analysis of a synthetic prediction market using convex sets. Results in Control and Optimization. https://www.sciencedirect.com/science/article/pii/S2666720721000308#

System evaluation --> real replication data



The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)

A Synthetic Prediction Market for Estimating Confidence in Published Work

Sarah Rajtmajer,[1] Christopher Griffin,[1] Jian Wu,[2] Robert Fraleigh,[1] Laxmaan Balaji,[1] Anna Squicciarini,[1] Anthony Kwasnica,[1] David Pennock,[3] Michael McLaughlin,[1] Timothy Fritton,[1] Nishanth Nakshatri,[1] Arjun Menon,[1] Sai Ajay Modukuri,[1] Rajal Nivargi,[1] Xin Wei,[2] C. Lee Giles[1]

[1]The Pennsylvania State University
[2]Old Dominion University
[3]Rutgers University
{smr48,cxg286,rdf5090,lpb5347,acs20,amk17,mvm7085,tjf115,nzn5185,amm8987,svm6277,rfn5089,clg20}@psu.edu

Abstract

Explainably estimating confidence in published sc
work offers opportunity for faster and more robus
tific progress. We develop a synthetic prediction m
assess the credibility of published claims in the social
havioral sciences literature. We demonstrate our syst
detail our findings using a collection of known rep
projects. We suggest that this work lays the founda
a research agenda that creatively uses AI for peer revi

Introduction

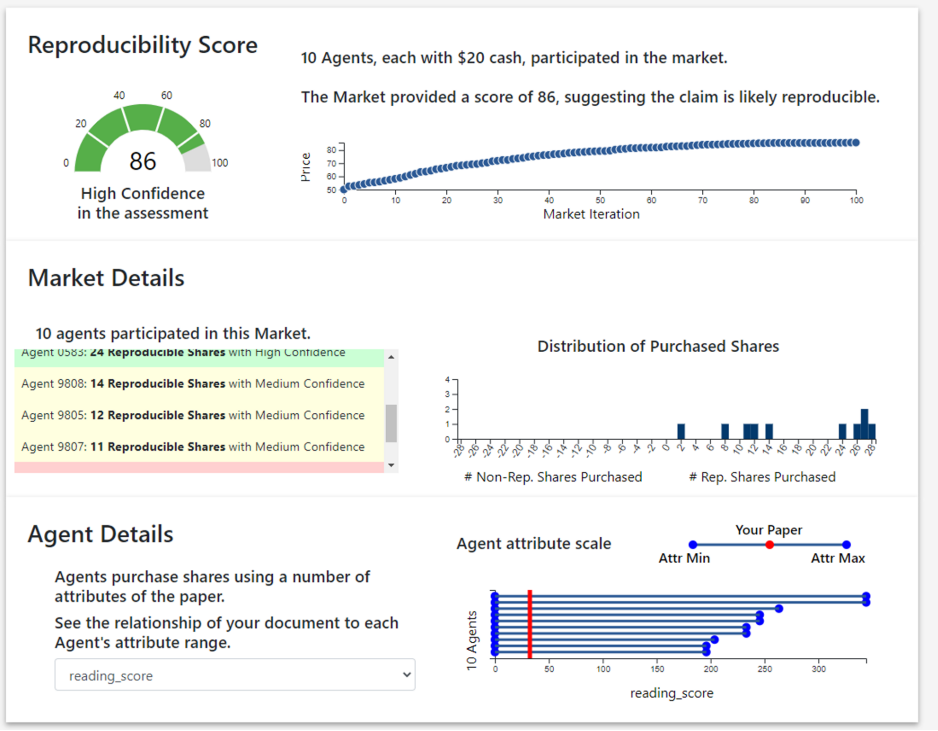
Concerns about the replicability, robustness ar
ducibility of findings in scientific literature hav

Results on scored papers. Our system provides a confidence score for 68 of 192 (35%) of the papers in our set. On the set of scored papers, accuracy is 0.894, precision is 0.917, recall is 0.903, and **F1** is **0.903** (macro averages). A sizeable un-scored subset of data (65%) is the trade-off for high accuracy on the scored subset of the data. A test point is un-scored when the system has determined it has insufficient information to evaluate it.

System non-scoring. Unlike most other machine learning algorithms, the synthetic market does not provide an evaluation for every input. Like its human-populated counterparts, the market is vulnerable to lack of participation (Arrow et al. 2008; Tetlock 2008; Rothschild and Pennock 2014). Agents will not participate if they have not seen a sufficiently similar training point (paper). This is more common when the training dataset is small; in experiments with larger datasets, we have observed participation increases. Meaningful ways to increase agent participation, including hybrid settings with human participants, are being explored.

System evaluation --> RAND grad students

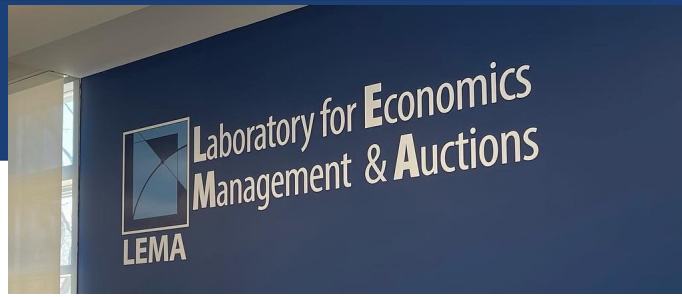
- **Claim submission**: User submits a paper (PDF) for evaluation.
- **Feature Extraction**: Extraction tools stage, followed by pass through feature extractor modules generate paper feature vector.
- **Evaluation through multiple prediction markets**: The feature vector is passed through multiple markets and results from each are collected.
- **SCORE and interpretability**: Results from the prediction markets are collated and a response containing the SCORE, interpretability and confidence is returned.



Explanations:

- **Level One:** Confidence in the claim's reproducibility through market score
- **Level Two:** Aggregated details related to agent participation in the system
- **Level Three:** Which agents participated + their confidence
- **Level Four:** Features corresponding to nearest training data points

Hybrid prediction markets

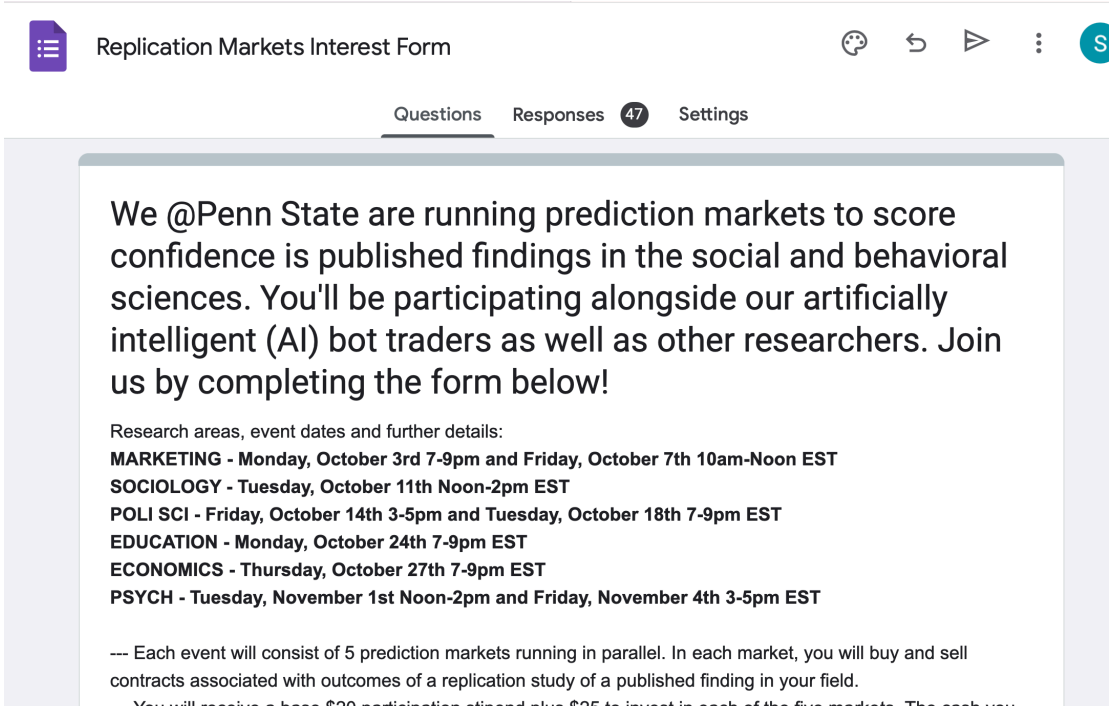


Virtual 2-hour long market events
April and May 2023

- 100+ participants
- Currently analyzing results, and conducting interviews with participants

Initial takeaways:

→ Major improvement on agent participation!
→ Change in individuals' evals before/after market based on surveys
→ Need more work to understand the right "balance" of bots and SMEs



Next steps… Of course… LLMs ☺

Can Large Language Models Discern Evidence for Scientific Hypotheses? Case Studies in the Social Sciences

Abstract: Although studies have shown that increases in the frequency of social media use may be associated with increases in depressive symptoms of individuals with depression, the current study aimed to identify specific social media behaviors related to major depressive disorder (MDD). Millennials (N = 504) who actively use Facebook, Twitter, Instagram, and/or Snapchat participated in an online survey assessing major depression and specific social media behaviors. Univariate and multivariate analyses were conducted to identify specific social media behaviors associated with the presence of MDD. The results identified five key social media factors associated with MDD. Individuals who were more likely to compare themselves to others better off than they were (p = 0.005), those who indicated that they would be more bothered by being tagged in unflattering pictures (p = 0.011), and those less likely to post pictures of themselves along with other people (p = 0.015) were more likely to meet the criteria for MDD. Participants following 300 + Twitter accounts were less likely to have MDD (p = 0.041), and those with higher scores on the Social Media Addiction scale were significantly more likely to meet the criteria for MDD (p = 0.031). Participating in negative social media behaviors is associated with a higher likelihood of having MDD. Research and clinical implications are considered.

Hypothesis: Is there an association between social media use and bad mental health outcomes?

Label: Yes

model	train corpus	test corpus	Acc	F1
MT-DNN	declarative	declarative	67.97%	0.523
MT-DNN	questions	questions	65.62%	0.497
MT-DNN	snli	declarative	42.97%	0.342
MT-DNN	snli	questions	21.87%	0.228
ESIM	declarative	declarative	64.84%	0.489
ESIM	questions	questions	61.72%	0.359
ESIM	snli	declarative	39.84%	0.335
ESIM	snli	questions	39.10%	0.306
Chat-GPT	zero shot	declarative		
Chat-GPT	demonstration learning			
Chat-GPT	prompt ensembling			
PaLM 2	zero shot	declarative		
PaLM 2	demonstration learning			
PaLM 2	prompt ensembling			

smr48@psu.edu