THE UNIVERSITY OF
ALABAMA IN HUNTSVILLE

Large Language Models as Trust in Automation Analysis Tools

Derek Koehl and Lisa Vangsness

Challenge, Metacognition, and Perception Lab

Metacognition

Challenge

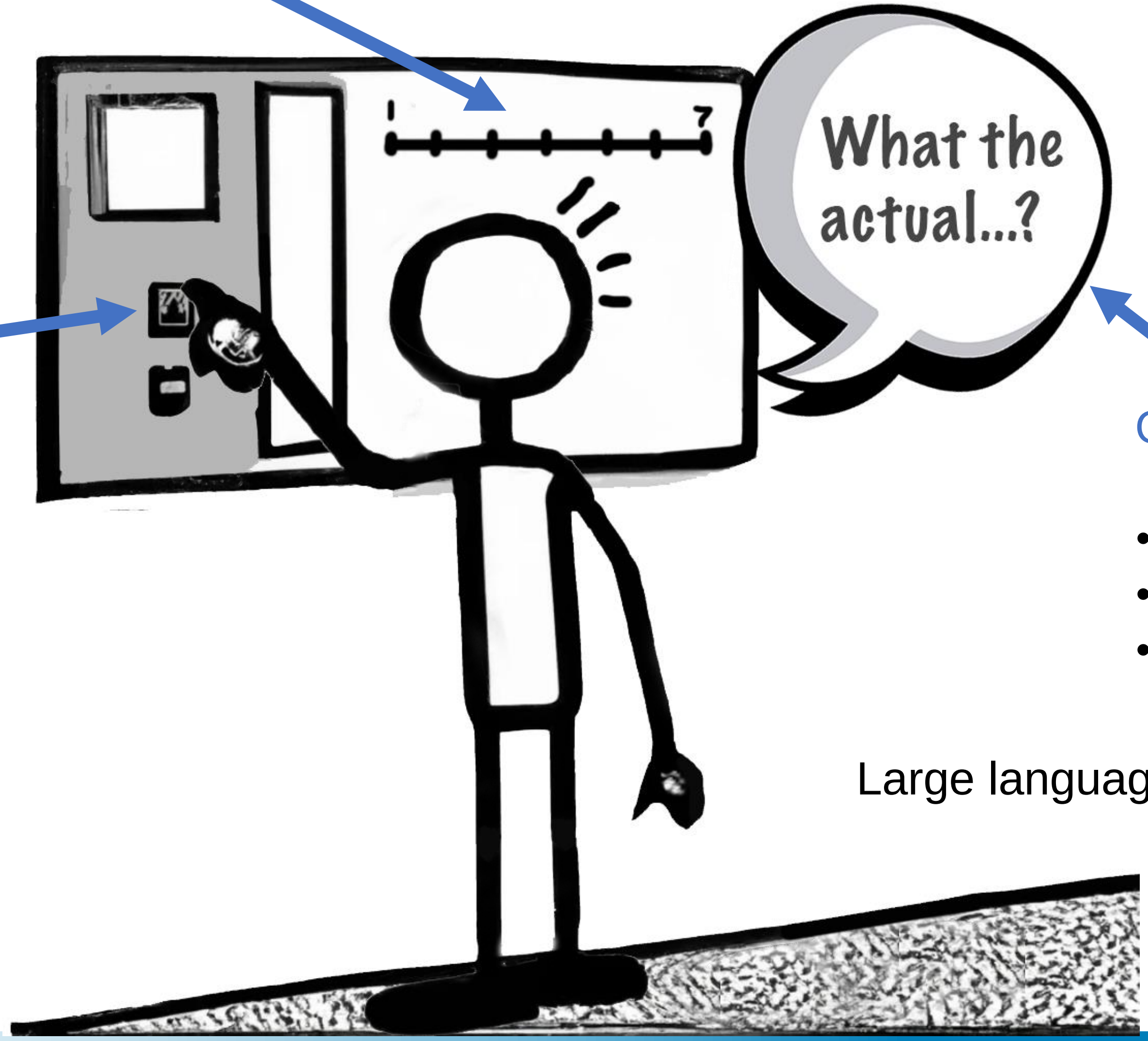
Perception

ChaMP Lab

The Why

Quantitative Self-Report Measures

Behavioral Measures



Qualitative Self-Reports

- Time-consuming analysis
- Subjective reliability
- Inter-rater reliability

Large language models as analysis tools?
(Lau et al., 2018)

Large Language Model

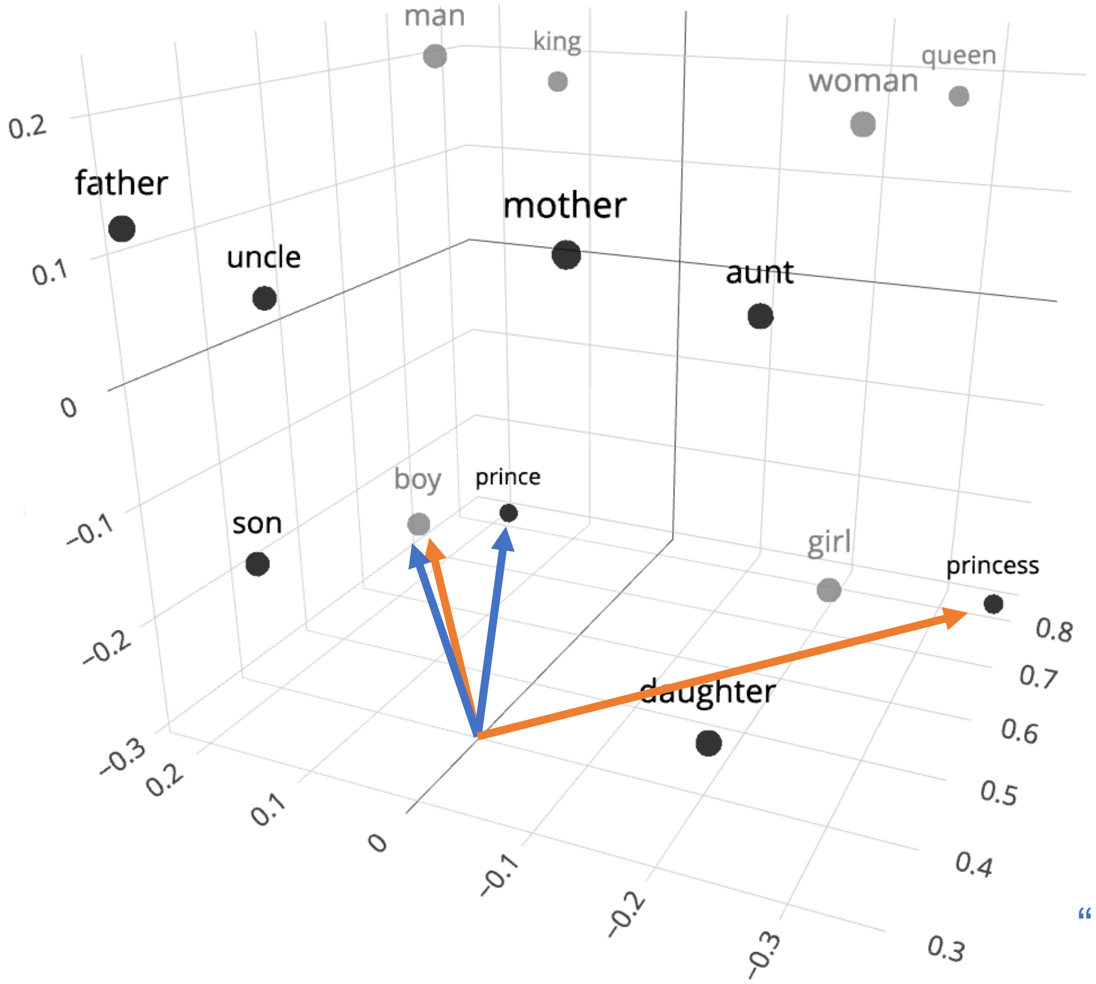
"man" is to "king" as "woman" is to "____________"?

Large language models contain latent semantic patterns.

$$\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman} \approx \overrightarrow{queen}$$

(Mikolov et al., 2013)

Large Language Model



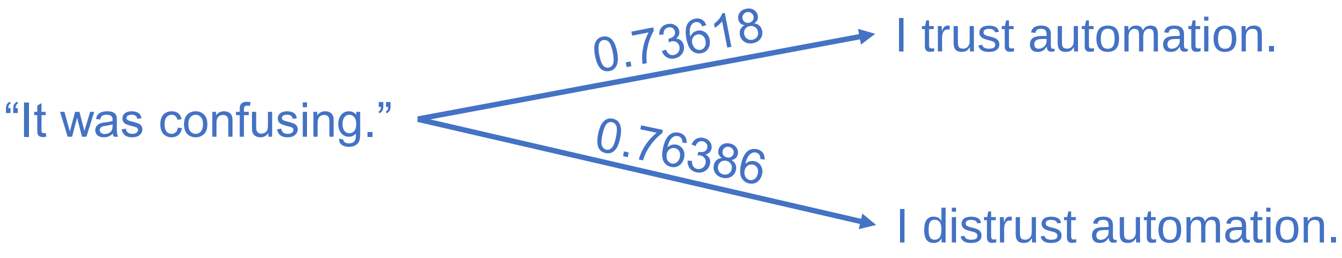
(Bandyopadhyay et al., 2022)

Unit of text (e.g., word, sentence) = a point in a high-dimensional space

low dimensional spaces: (x, y, z)
high dimensional: (p_1, p_2, p_3, …, p_{1536})

Cosine similarity

- measurement of angular distance
- high cosθ indicates similar semantic features



“It was confusing.”

0.73618

I trust automation.

0.76386

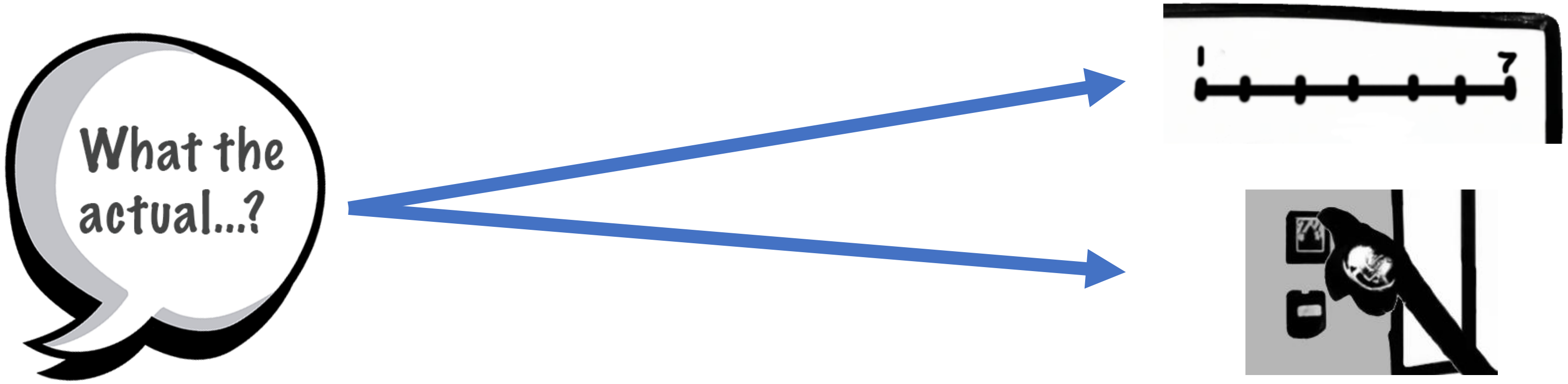
I distrust automation.

Exploratory Research Questions

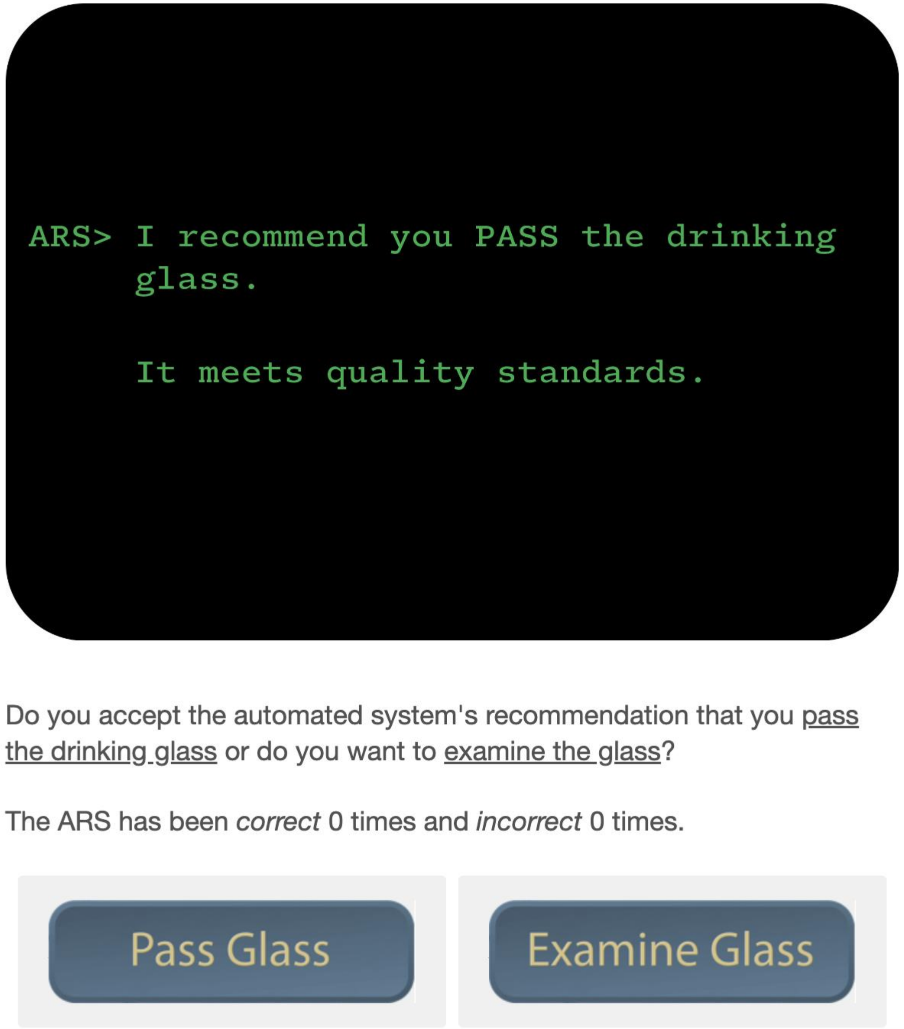
Can cosine similarities calculated against trust/distrust sentences predict a self-report Likert rating of trust?

Can cosine similarities calculated against trust/distrust sentences predict a behavioral measure of trust?

What sample size is necessary to achieve a well-trained model for prediction?



Gamified Survey



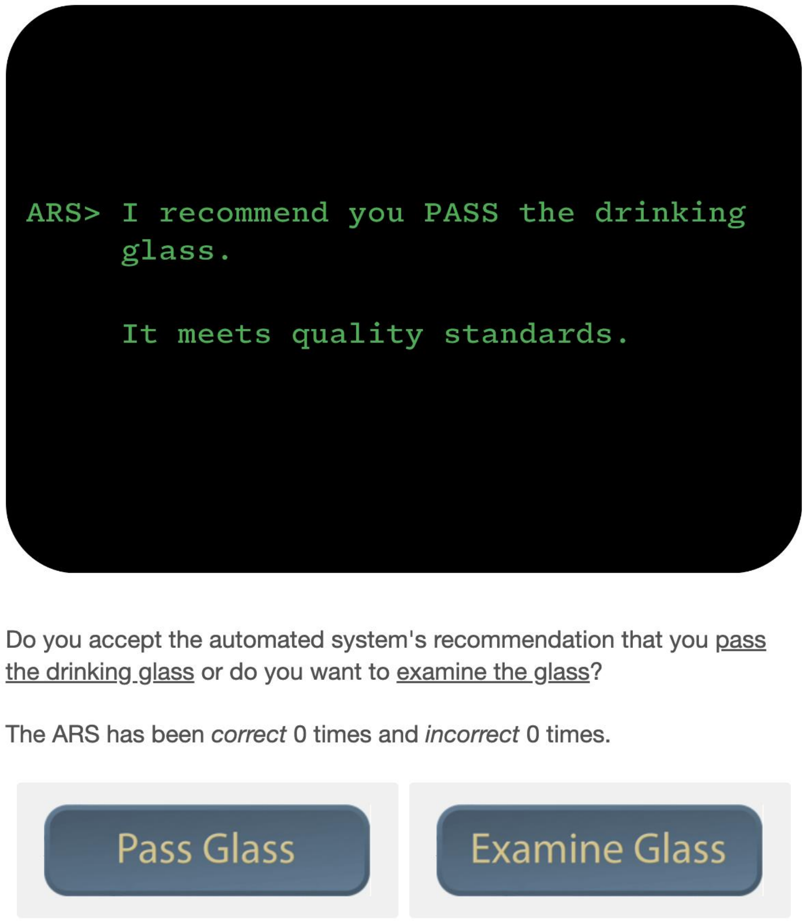
```
ARS> I recommend you PASS the drinking
     glass.

     It meets quality standards.
```

Do you accept the automated system's recommendation that you pass the drinking glass or do you want to examine the glass?

The ARS has been *correct* 0 times and *incorrect* 0 times.

(Yu et al., 2017)

Gamified Survey



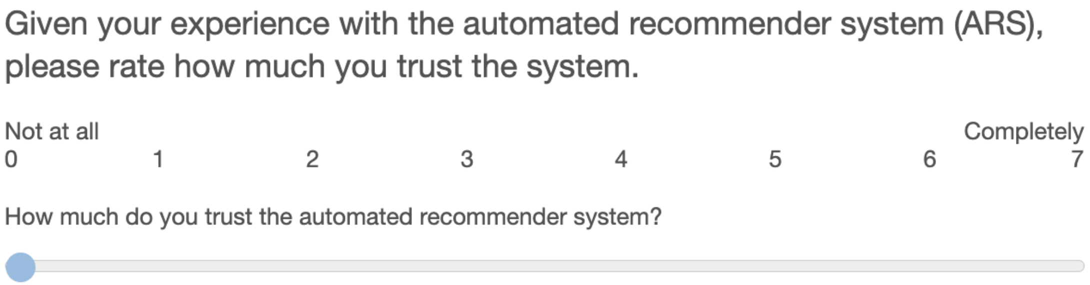
```
ARS> I recommend you PASS the drinking
     glass.

     It meets quality standards.
```

Do you accept the automated system's recommendation that you pass the drinking glass or do you want to examine the glass?

The ARS has been *correct* 0 times and *incorrect* 0 times.

(Yu et al., 2017)

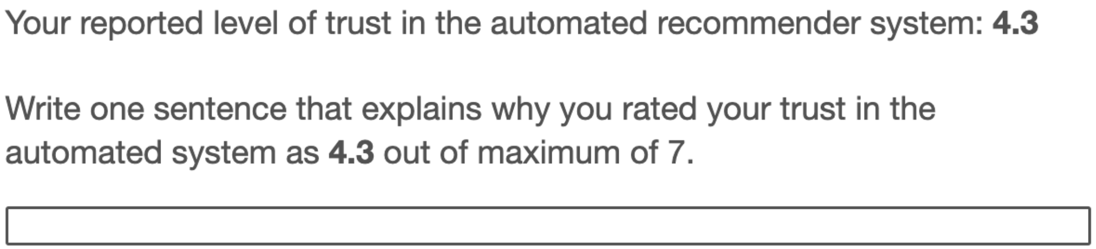
Behavioral measure



Given your experience with the automated recommender system (ARS), please rate how much you trust the system.

How much do you trust the automated recommender system?

Self-report rating



Your reported level of trust in the automated recommender system: **4.3**

Write one sentence that explains why you rated your trust in the automated system as **4.3** out of maximum of 7.

Self-report sentence

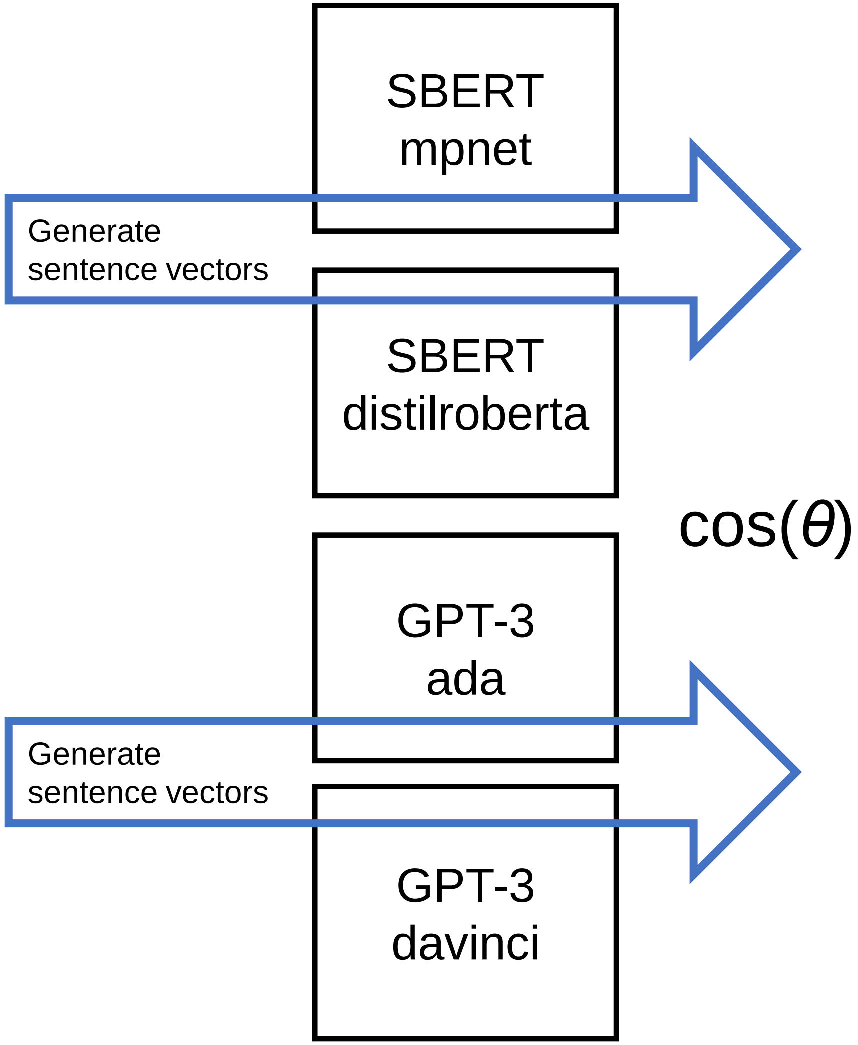
Analysis

N = 80

IS: n = 60

10 Sentence Sets
n = 15
n = 20
n = 25
n = 30
…
n = 60

8 Reference
I trust automation. I distrust automation.
…
I trust automated systems completely. I distrust automated systems completely.



Generate sentence vectors

SBERT mpnet

SBERT distilroberta

Generate sentence vectors

GPT-3 ada

GPT-3 davinci

cos(θ)

8 Similarity Scores for each sentence
0.829466819772240
0.837044959128676
…
0.808963624141162
0.817807812100523

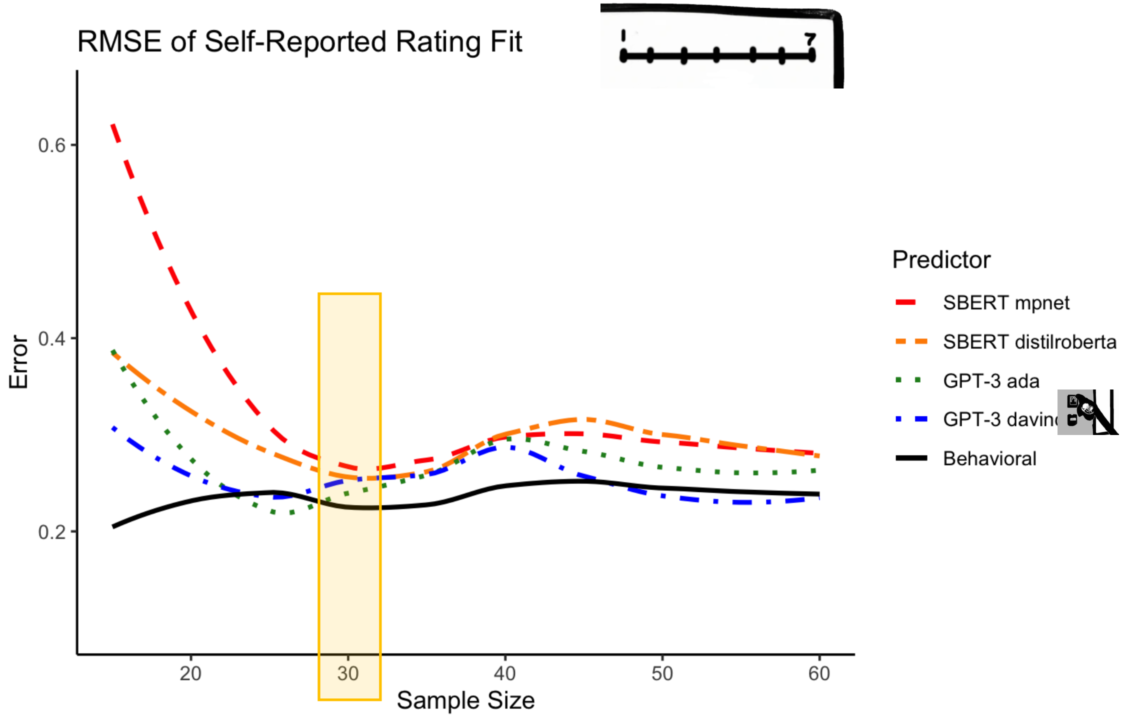
OOS: n = 20

Self-report Rating Linear Regression Models
Criterion
Self-report rating
Predictors
Similarity scores [or] Behavioral measure

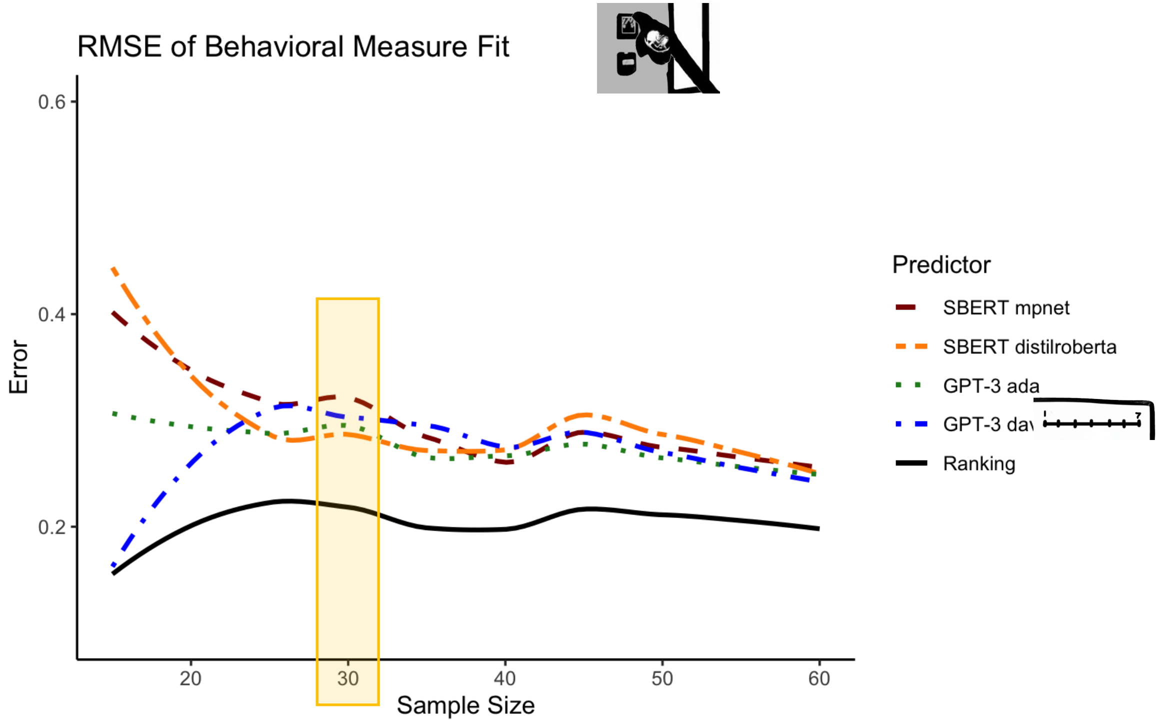
Behavioral Measure Linear Regression Models
Criterion
Behavioral measure
Predictors
Similarity scores [or] Self-report rating

Results

What size sample is necessary to achieve a well-trained model for prediction?

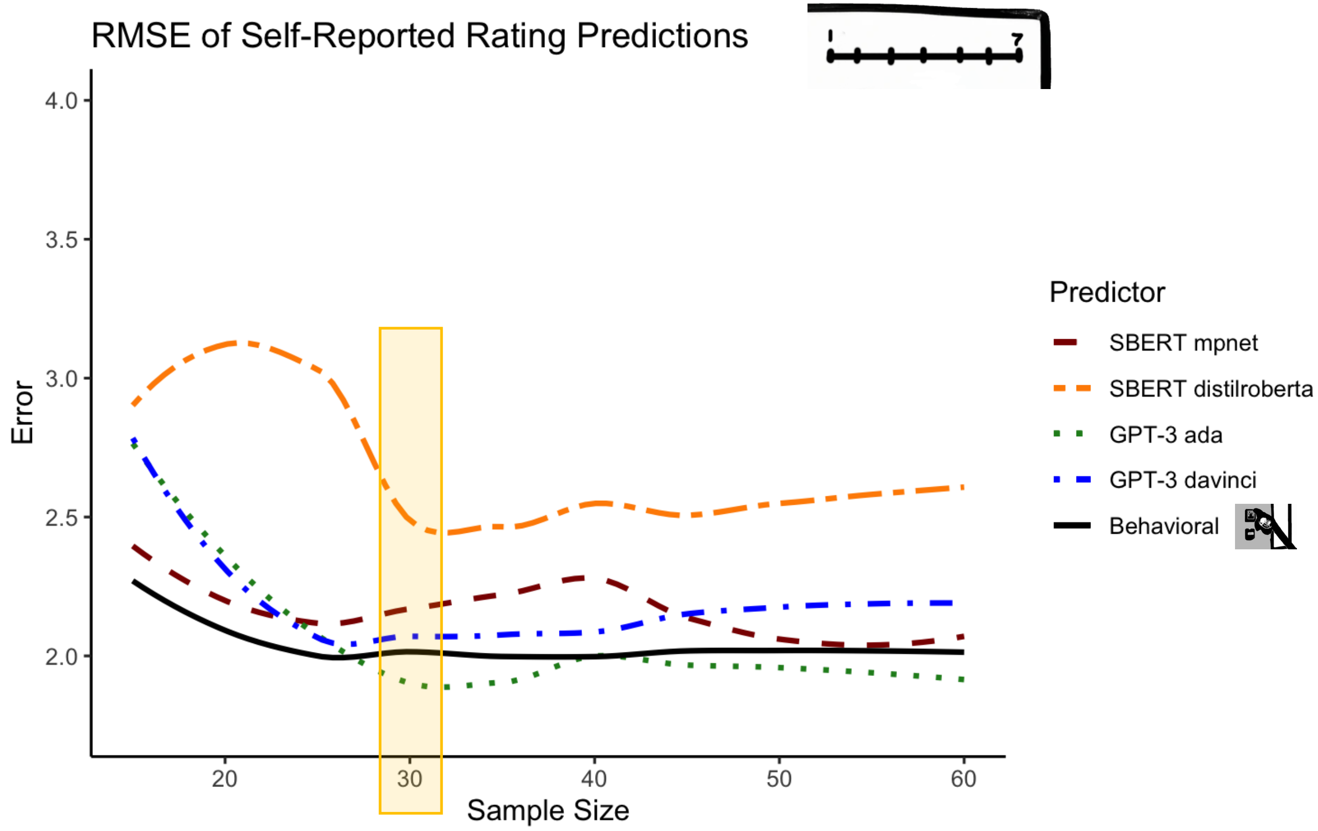


Note: Self-report rating criterion variable normalized for cross-chart comparisons



Results

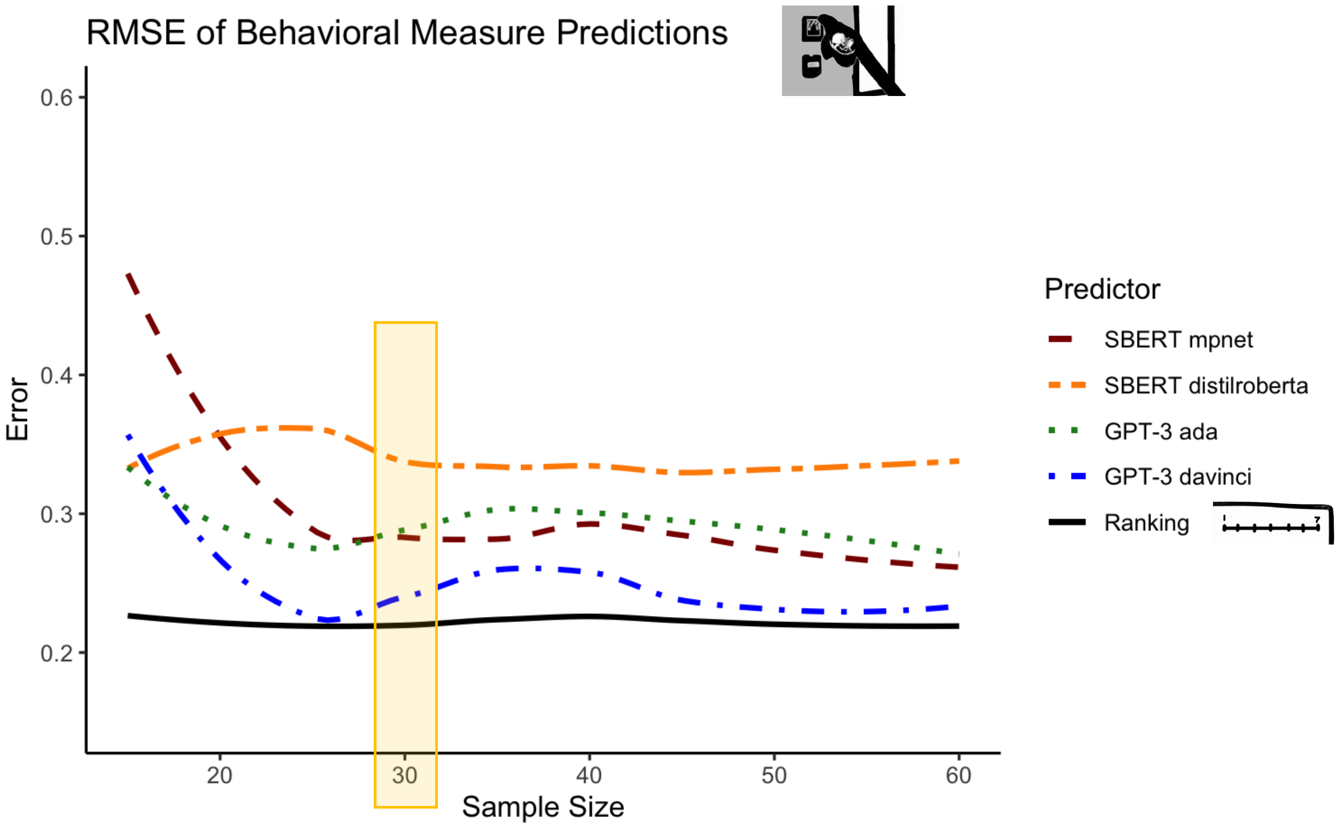
Can cosine similarities calculated against trust/distrust sentences predict a self-report Likert rating of trust?



	Sample Size					
	30		45		60	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE
GPT-3 ada	1.77	1.49	2.03	1.59	1.91	1.52
SBERT mpnet	1.98	1.71	2.33	1.87	2.09	1.72
GPT-3 davinci	2.19	1.73	2.30	1.89	2.19	1.81
SBERT distilroberta	2.22	1.71	2.55	2.02	2.61	2.03
Behavioral measure	2.04	1.69	2.03	1.61	2.02	1.53

Results

Can cosine similarities calculated against trust/distrust sentences predict a behavioral measure of trust?



	Sample Size					
	30		45		60	
Model	RMSE	MAE	RMSE	MAE	RMSE	MAE
GPT-3 davinci	0.23	0.19	0.25	0.20	0.23	0.19
SBERT mpnet	0.26	0.22	0.29	0.23	0.27	0.21
GPT-3 ada	0.29	0.24	0.29	0.23	0.27	0.21
SBERT distilroberta	0.32	0.25	0.33	0.26	0.34	0.26
Self-report ranking	0.22	0.16	0.22	0.18	0.22	0.17

Questions?

References

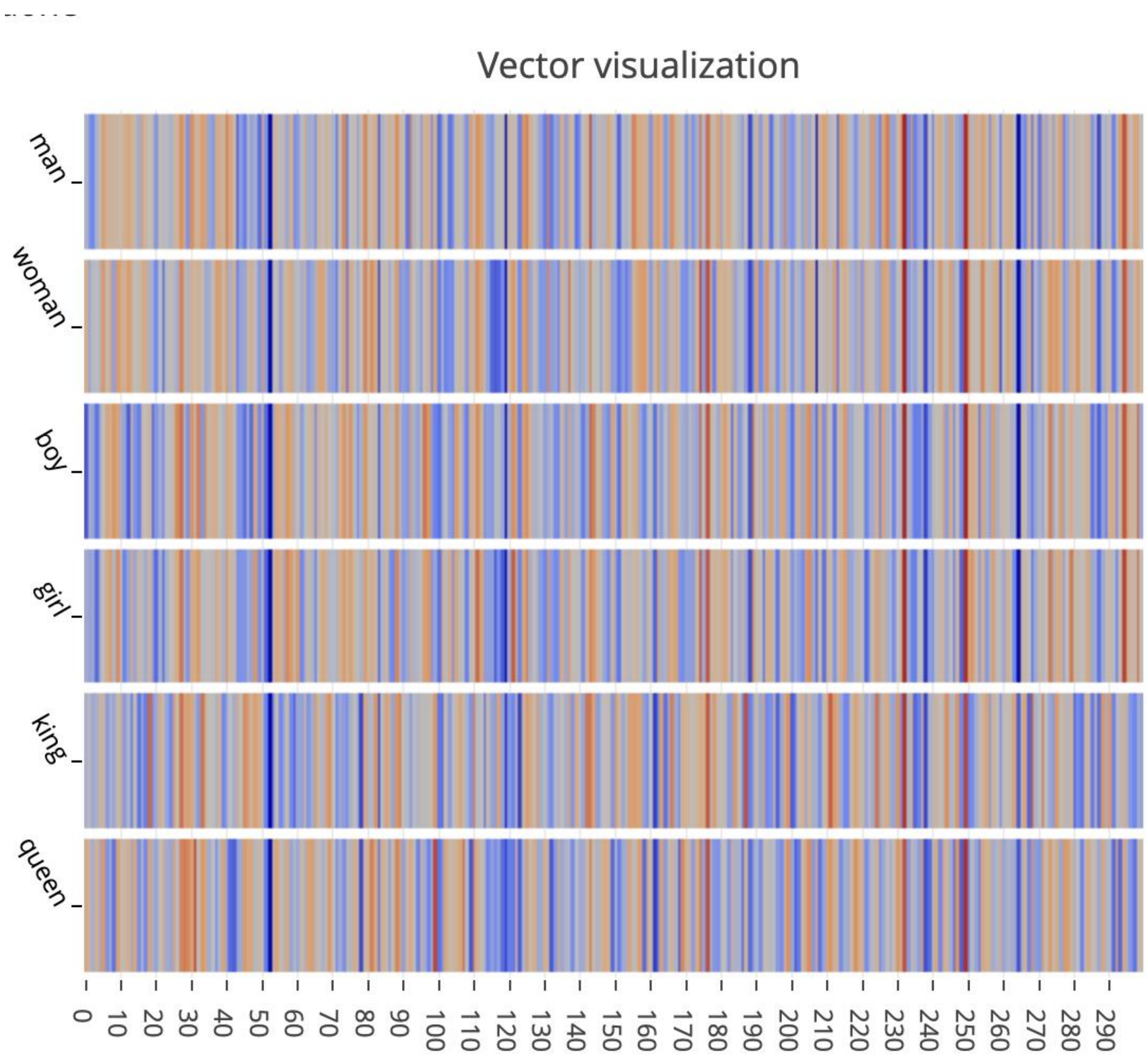
Bandyopadhyay, S., Xu, J., Pawar, N., & Touretzky, D. (2022). Interactive visualizations of word embeddings for K-12 students. *Proceedings of the AAAI Conference on Artificial Intelligence*, *36*(11), 12713–12720. https://doi.org/10.1609/aaai.v36i11.21548

Lau, N., Fridman, L., Borghetti, B. J., & Lee, J. D. (2018). Machine learning and human factors: Status, applications, and future directions. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *62*(1), 135–138. https://doi.org/10.1177/1541931218621031

Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User trust dynamics: An investigation driven by differences in system performance. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3025171.3025219

Latent Semantic Patterns



Cosine Similarity Function

$$\cos(\theta) = \frac{\vec{a} \cdot \vec{b}}{\parallel \vec{a} \parallel \parallel \vec{b} \parallel} = \frac{\sum_{i=1}^{n} a_i b_i}{\sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{a}_{\mathrm{i}}^{2}} \sqrt{\sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{b}_{\mathrm{i}}^{2}}}$$