ISCTM
International Society for CNS Clinical Trials and Methodology

The application of Large Language Models to Psychiatric Measurement

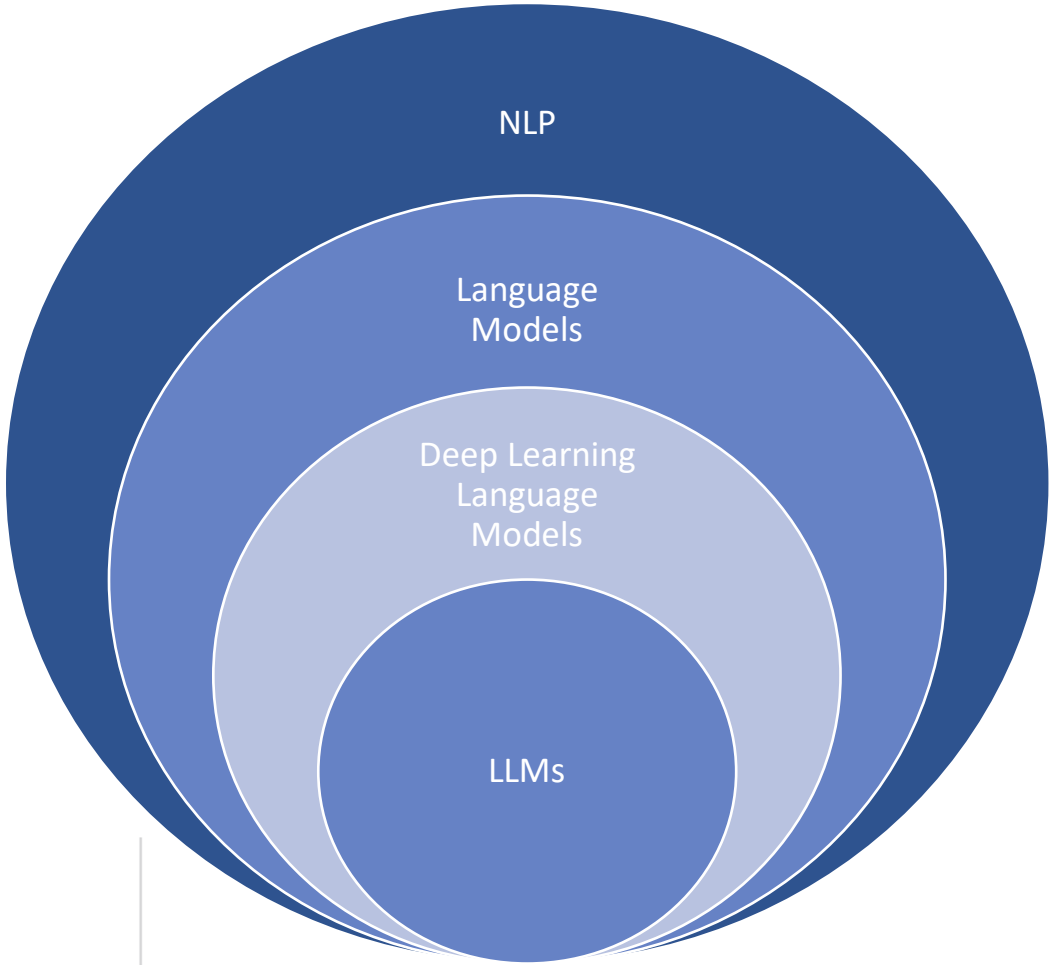
Isaac R. Galatzer-Levy

NYU Grossman School of Medicine; Google

Disclosure

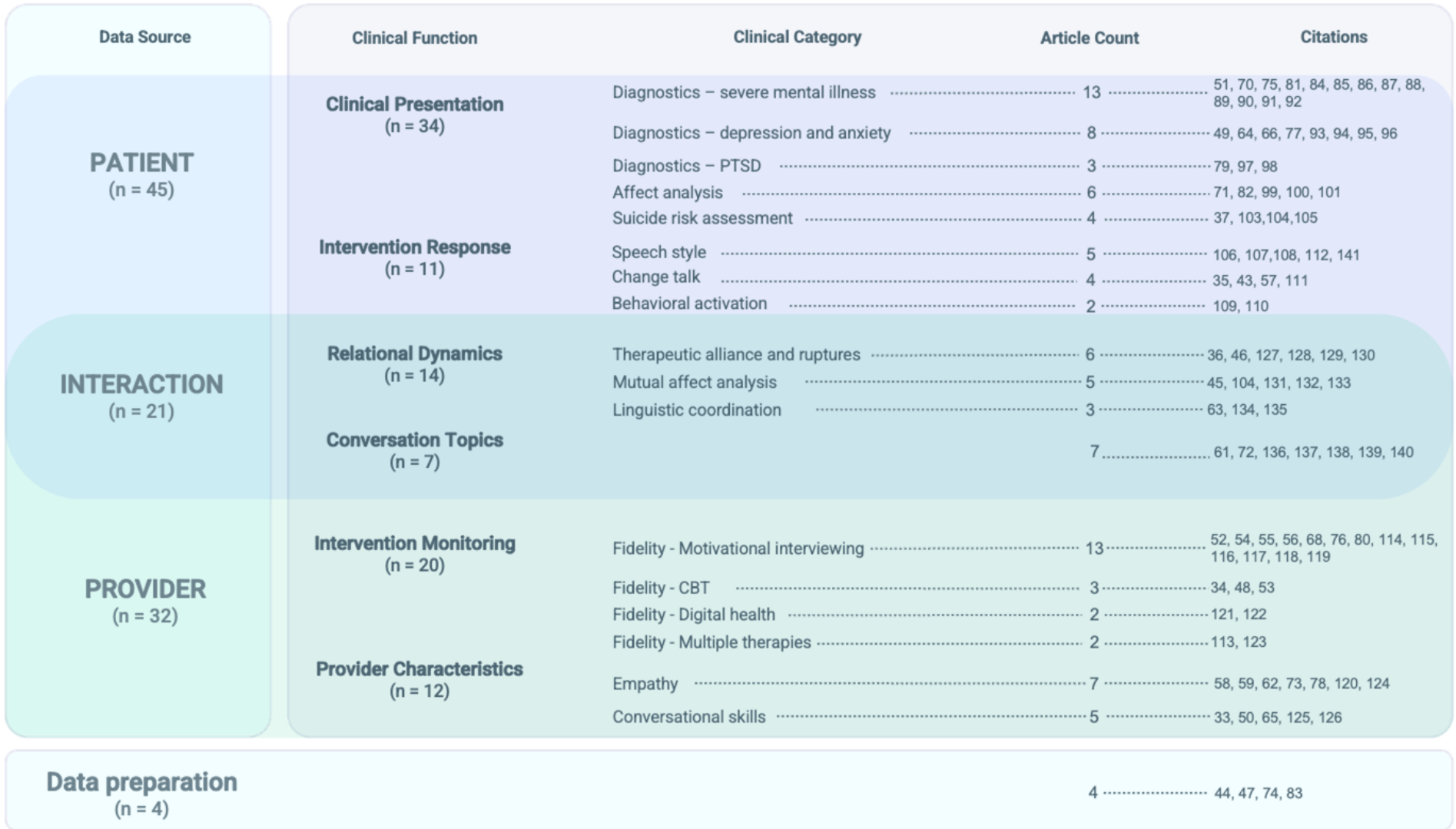
- I am employed by Google LLC.

Introduction

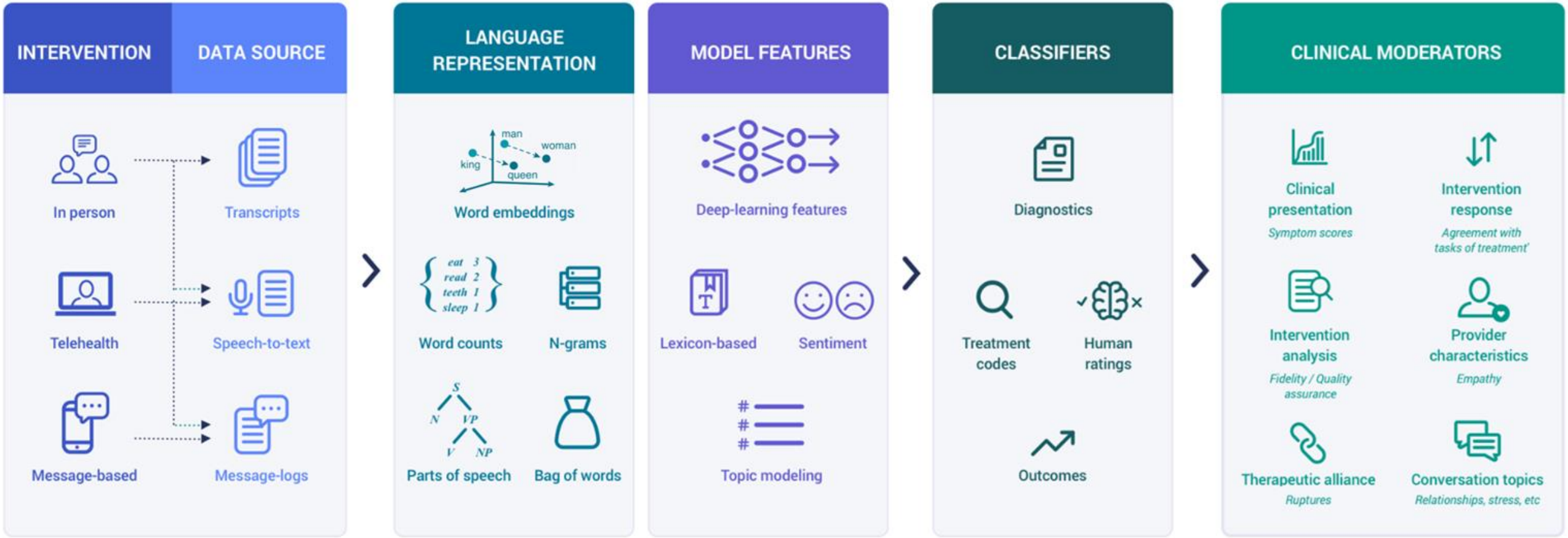


Meta AI

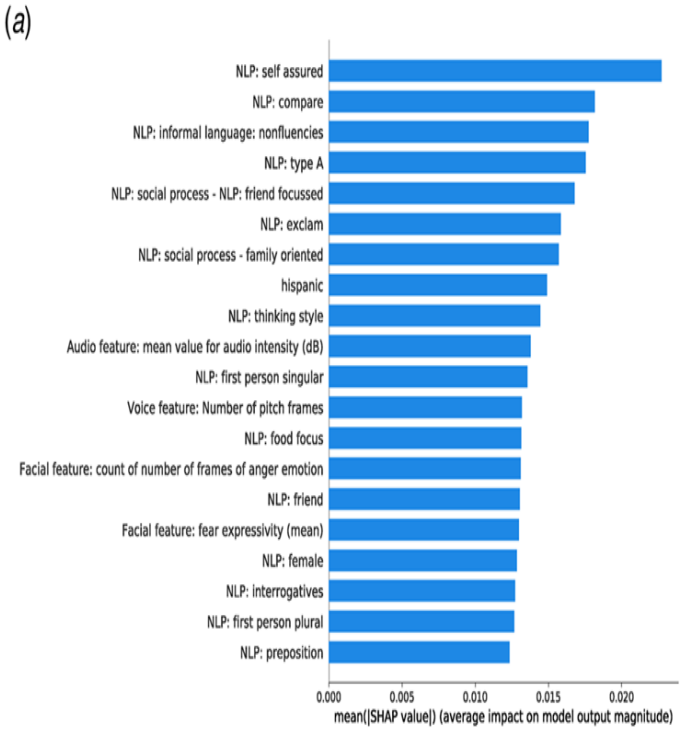
Bard

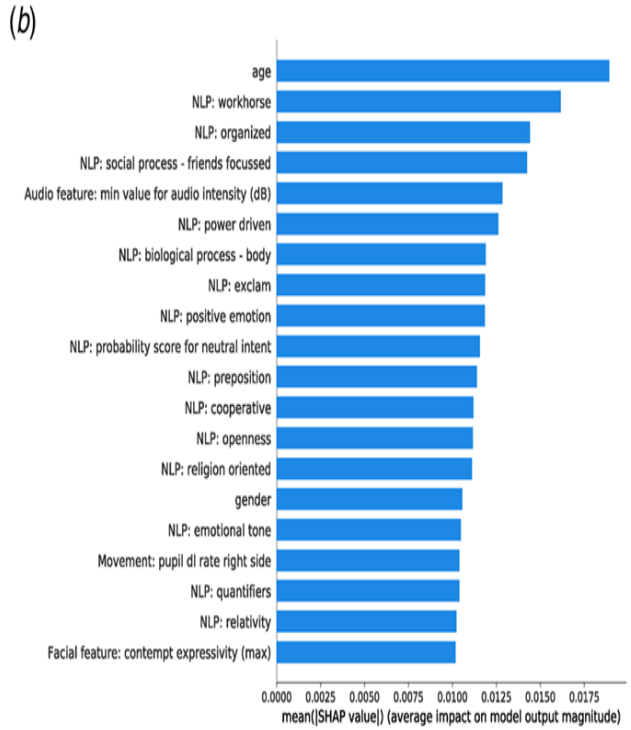


Malgaroli et al. (2023), Translational Psychiatry



Malgaroli et al. (2023), Translational Psychiatry





- Sample of 81 ED patients following a life-threatening event
- Text: transcripts from open-ended interviews
- The ”NLP” variables are dictionary features (e.g., positive emotion), based on lexicons by experts
- An ensemble of NLP, audio, and facial features allowed to diagnose PTSD and Depression with good accuracy

Psychological Medicine

cambridge.org/psm

Original Article

Cite this article: Schultebraucks K, Yadav V, Shalev AY, Bonanno GA, Galatzer-Levy IR (2020). Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood. *Psychological Medicine* 1–11. https://doi.org/10.1017/S0033291720002718

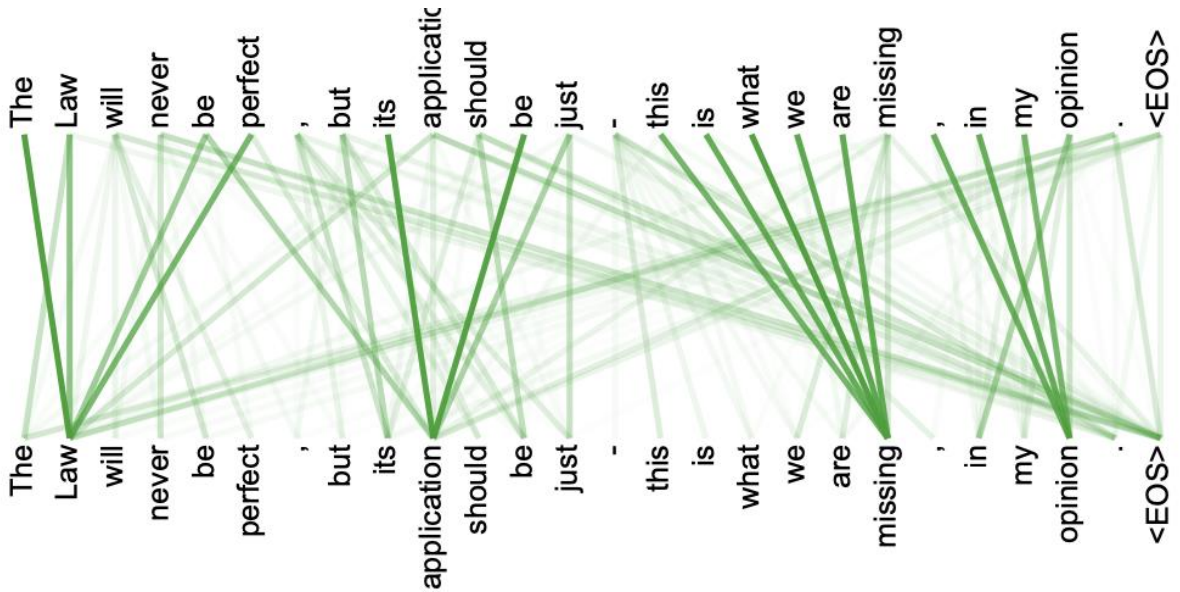
Received: 20 January 2020

Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood

Katharina Schultebraucks[1,2,3], Vijay Yadav[4], Arieh Y. Shalev[2], George A. Bonanno[5] and Isaac R. Galatzer-Levy[2,4]

[1]Department of Emergency Medicine, Vagelos School of Physicians and Surgeons, Columbia University Irving Medical Center, New York, New York, USA; [2]Department of Psychiatry, New York University Grossman School of Medicine, New York, New York, USA; [3]Data Science Institute, Columbia University, New York, New York, USA; [4]AiCure, New York, New York, USA and [5]Department of Counseling and Clinical Psychology, Teachers College, Columbia University, New York, New York, USA

Transformers: A **generic model architecture** with performance that scales well with parameters (model size) and training samples (data size).



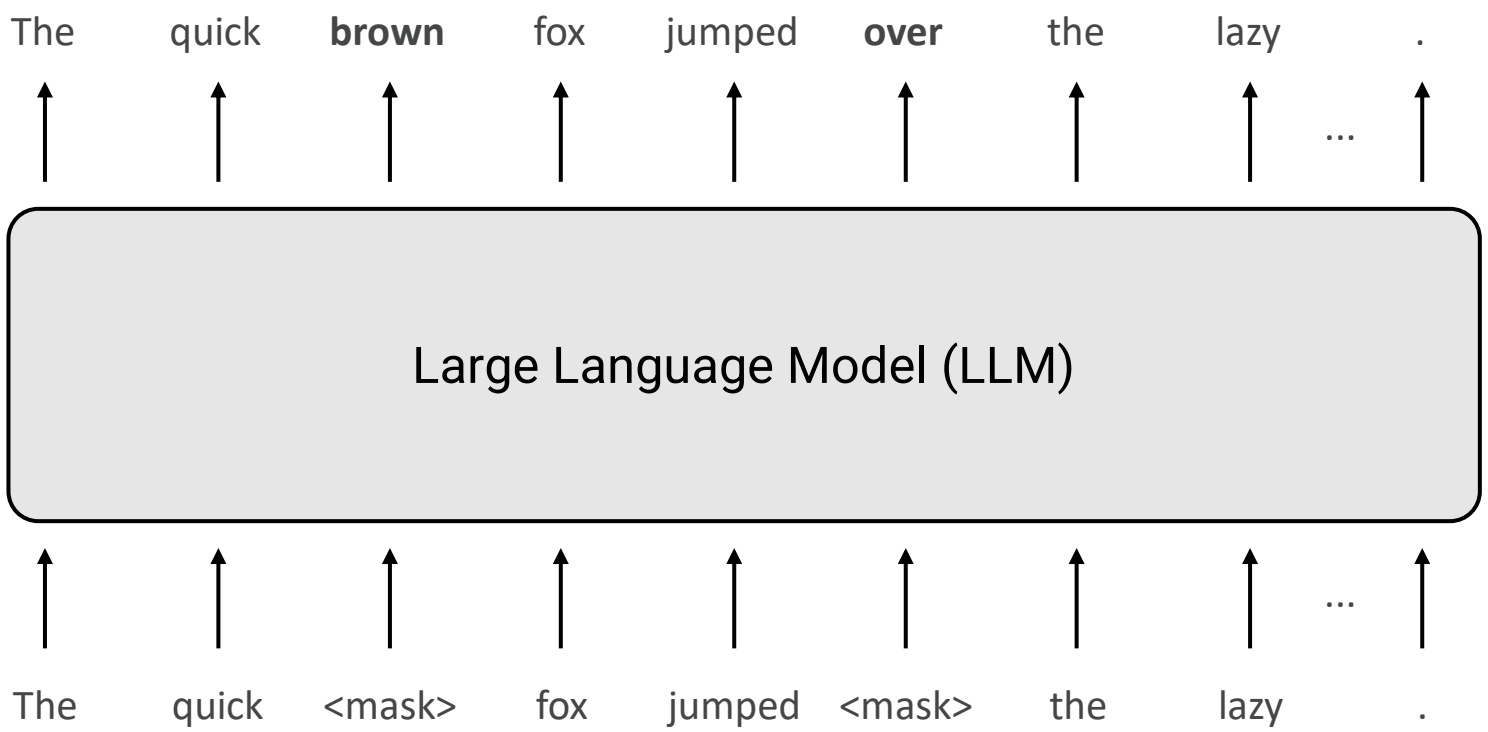
Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com
Noam Shazeer* Google Brain noam@google.com
Niki Parmar* Google Research nikip@google.com
Jakob Uszkoreit* Google Research usz@google.com
Llion Jones* Google Research llion@google.com
Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu
Łukasz Kaiser* Google Brain lukaszkaiser@google.com
Illia Polosukhin* ‡ illia.polosukhin@gmail.com

Abstract

The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.



Models are trained with “simple” pre-text tasks, such as predicting the next word, or filling in the blanks.

Surprisingly, this leads to model with the ability to generate complex language.

Current investigation

We investigated the ability of an LLM (Med-PaLM 2) to perform assessment of psychiatric functioning.

Our dataset included depression (n = 145) and PTSD assessments (n =115) and clinical case studies (n = 46) across high prevalence/high comorbidity disorders.

Example Prompt:

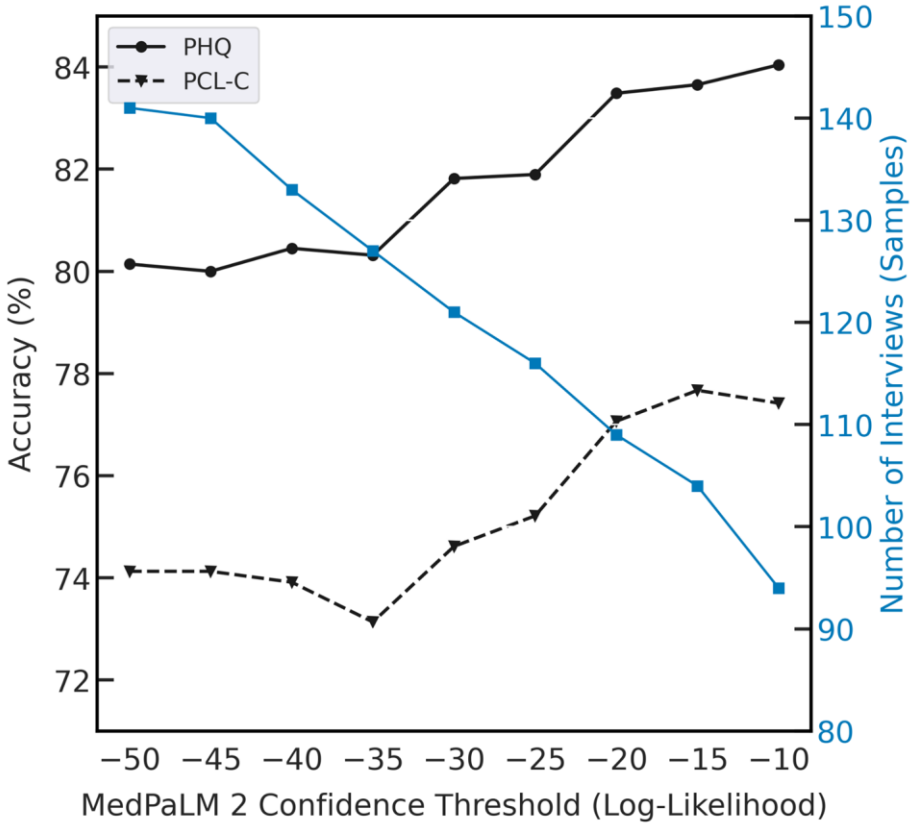
"Are you familiar with the [PHQ-8/PCL-C]? Based on the following clinical interview, what do you estimate the participants [PHQ-8/PCL-C] score is?"

… a 23-year-old woman who presented for an outpatient psychiatric evaluation 2 weeks after giving birth to her second child. She was referred by her breast-feeding nurse, who was concerned about the patient's depressed mood, flat affect, and fatigue …
… She was fully oriented and could register three objects but only recalled one after 5 minutes. Her intelligence was average. Her insight and judgment were fair to good.

LLM comparison to clinical evaluation

	Med-PaLM 2 PCL-C	Med-PaLM 2 PHQ-8
Accuracy	0.74	0.80
F1 Score	0.64	0.77
Precision	0.88	0.65
Sensitivity	0.30	0.75
Specificity	0.98	0.82
MAE	9.07	2.33
RMSE	11.2	3.93
Kappa with Clinical Ratings	0.33	0.55
Pearson r (p-value)	0.41 ($p < 0.01$)	0.55 ($p < 0.01$)

Human raters compared to Med-PaLM 2
$t(1,144) = 1.20$; $p = 0.23$; $r = 0.55$; $p < .01$



Predicting PTSD and Depression Scores

		Actual		
		MDD Term	PTSD Term	Total
Predicted	MDD Assessment	118	8	126
	PTSD Assessment	26	101	127
	Total	144	109	

74% Classifying symptoms of PTSD based on PCL

80% Classifying symptoms of depression based on PHQ-9

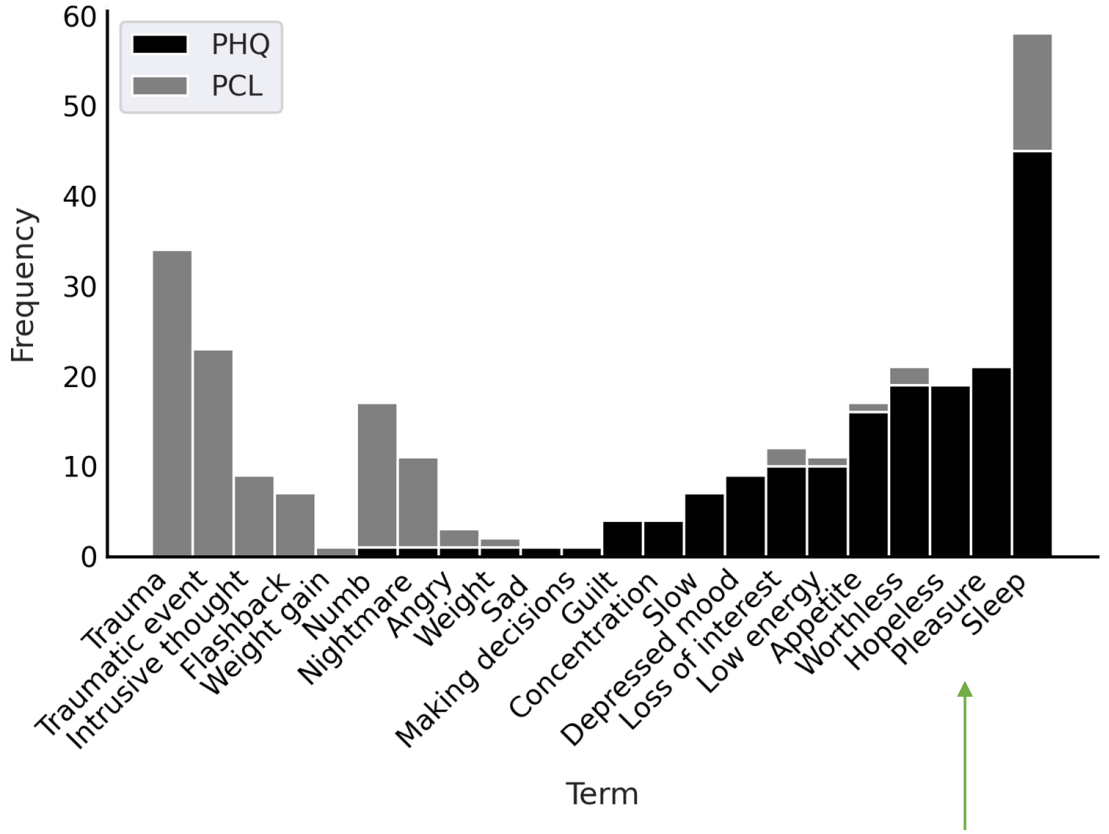
The LLM had **zero-shot performance** that was comparable to other models trained on in-context samples.

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., & Malgaroli, M. (2023). The Capability of Large Language Models to Measure Psychiatric Functioning. arXiv preprint arXiv:2308.01834.

Analyses of word frequencies show that Med-Palm 2 produces content-specific summarization.

However, it is difficult to assess how an LLMs reported reasoning corresponds to its classification of a given case.

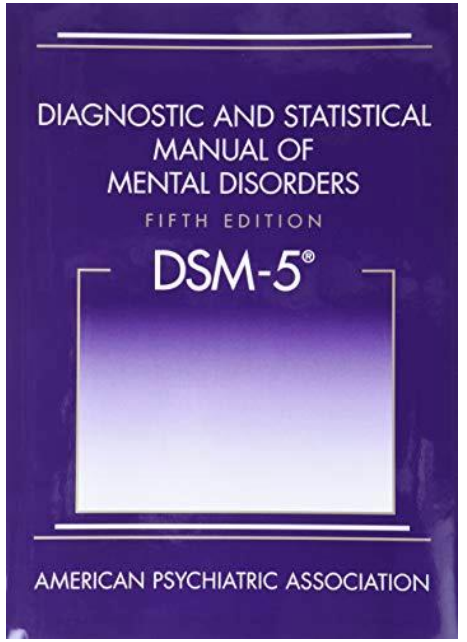
Transparency and interpretability are important properties.

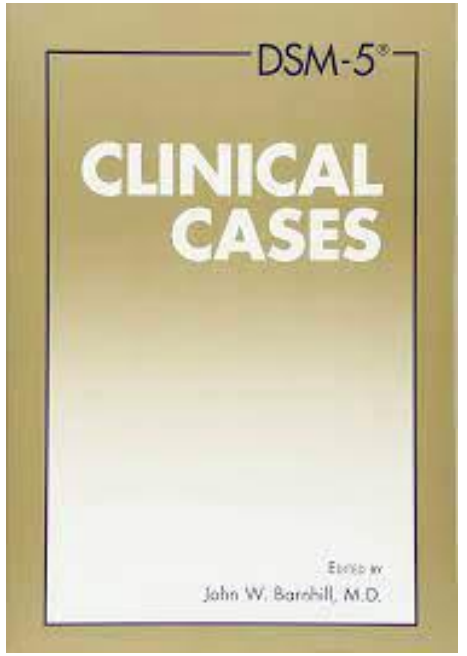


Frequency of words and phrases associated with Major Depressive Disorder (MDD) and Posttraumatic Stress Disorder (PTSD) associated with MDD and PTSD assessments.

Galatzer-Levy, I. R., McDuff, D., Natarajan, V., Karthikesalingam, A., & Malgaroli, M. (2023). The Capability of Large Language Models to Measure Psychiatric Functioning. arXiv preprint arXiv:2308.01834.

Clinical Case Evaluation





***n* = 46 clinical case studies:** *depressive disorders* (e.g. dysthymia, MDD, premenstrual dysphoric disorder; *n* = 12), *anxiety* (e.g. specific phobias, Generalized Anxiety Disorder; *n* = 6), *posttraumatic* (e.g. PTSD, acute stress disorder; *n* = 8), *substance and addiction related* (e.g. cocaine dependance; gambling disorder; *n* = 7), and *psychotic disorders* (schizophrenia, schizoaffective disorder; *n* = 7)

Example case input

Wyatt *was a 12-year-old-boy referred by his psychiatrist to an adolescent partial hospitalization program because of repeated conflicts that have frightened both classmates and family members. According to his parents, Wyatt was generally moody and irritable, with frequent episodes of being "a raging monster." It had become almost impossible to set limits. Most recently, Wyatt had smashed a closet door to gain access to a video game that had been withheld to encourage him to do homework. At school, Wyatt was noted to have a hair-trigger temper, and he had recently been suspended for punching another boy in the face after losing a chess match. Wyatt had been an extremely active young boy, running "all the time." He was also a "sensitive kid" who constantly worried that things might go wrong. His tolerance for frustration had been less than that of his peers, and his parents quit taking him shopping because he would predictably become distraught whenever they did not buy him whatever toys he wanted. Grade school reports indicated fidgetiness, wandering attention, and impulsivity. When Wyatt was 10 years old, a child psychiatrist diagnosed him as having attention-deficit/hyperactivity disorder (ADHD), combined type. Wyatt was referred to a behavioral therapist and started taking methylphenidate, with an improvement in symptoms. By fourth grade, his moodiness became more pronounced and persistent. He was generally surly, complaining that life was "unfair." Wyatt and his parents began their daily limit-setting battles at breakfast while he delayed getting ready for school, and then—by evening—continued their arguments about homework, video games, and bedtime. These arguments often included Wyatt screaming and throwing nearby objects. By the time he reached sixth grade, his parents were tired and his siblings avoided him. According to Wyatt's parents, he had no problems with appetite, and although they fought about when he would go to bed, he did not appear to have a sleep disturbance. He appeared to find pleasure in his usual activities, maintained good energy, and had no history of elation, grandiosity, or decreased need for sleep lasting more than a day. Although they described him as "moody, isolated, and lonely," his parents did not see him as depressed. They denied any history of hallucinations, abuse, trauma, suicidality, homicidality, a wish to self-harm, or any premeditated wish to harm others. He and his parents denied he had ever used alcohol or drugs. His medical history was unremarkable. His family history was notable for anxiety and depression in the father, alcoholism in the paternal grandparents, and possible untreated ADHD in the mother. On interview, Wyatt was mildly anxious yet easy to engage. His body twisted back and forth as he sat in the chair. In reviewing his temper outbursts and physical aggression, Wyatt said, "It's like I can't help myself. I don't mean to do these things. But when I get mad, I don't think about any of that. It's like my mind goes blank." When asked how he felt about his outbursts, Wyatt looked very sad and said earnestly, "I hate when I'm that way." If he could change three things in his life, Wyatt replied, "I would have more friends, I would do better in school, and I would stop getting mad so much."*

Example case output: ***Dx:*** **Disruptive mood dysregulation disorder Attention-deficit/hyperactivity disorder, combined presentation**

			Fisher's Exact Test	
	Category	Diagnosis	Phi(p - value)	Odds Ratio
Depression	1.00	0.83	-0.27(= 0.09)	0.32
Anxiety	1.00	0.83	-0.02(>0,99)	0.09
Psychosis	0.86	0.71	-0.04(>0,99)	0.23
Trauma & Stress	0.80	0.60	0.14(= 0.58)	0.16
Addictive disorder	1.00	1.00	-0.16 (= 0.57)	0.25
All	0.94	0.71		

Comparison to selecting diagnoses at chance

	Correct Diagnosis	Chance	χ2	*p*<=
Compared to All Diagnoses	29/40	1/297	120.41	0.0001
Compared to common diagnoses	29/40	1/40	19.60	0.001

Key considerations

- Underlying training data
- Robustness of results
- Prompt engineering vs. model tuning

Future directions

Evaluate real clinical notes

Compared LLMs and model fine tuning

Evaluate conversational agents to perform assessments from natural language

Thank you