Evaluating the Reliability of Large Language Models for Stress Detection

Visual Model

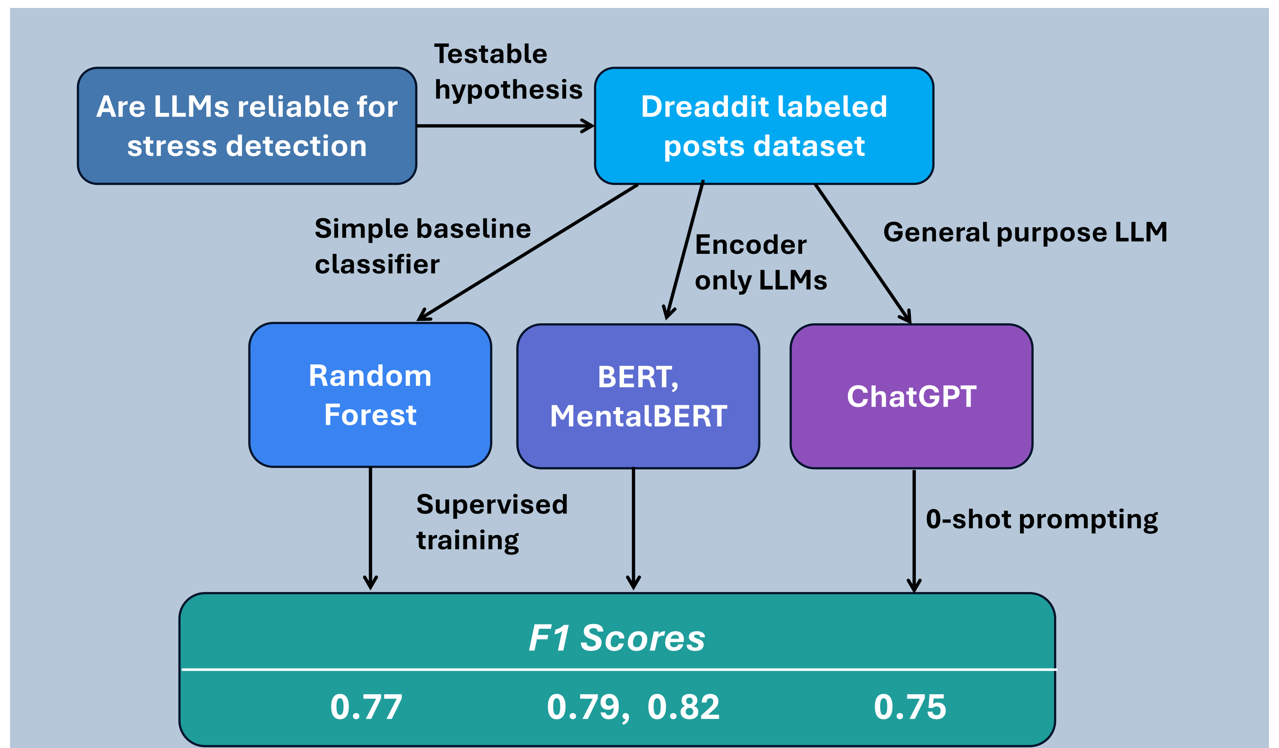


Figure created by finalist in PowerPoint, 2025.

Background

Stress

- Defined by APA as *"reaction to current or future pressures"*
- Severe stress → anxiety, depression, suicide risk
- Early detection is key for support

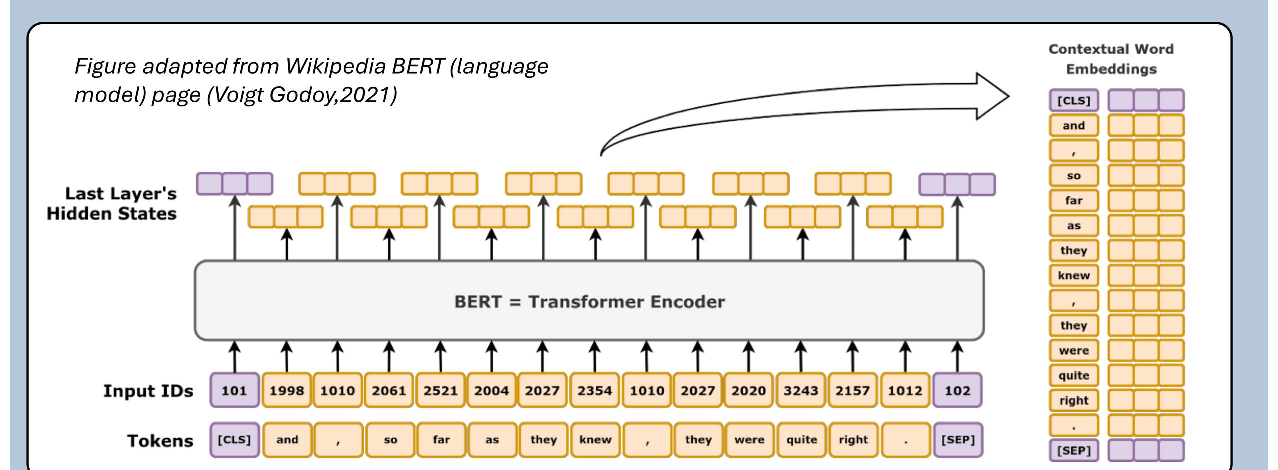
AI in Mental Health

- Generative AI & LLMs: accessible, affordable counseling tools
- *However,* their effectiveness is still unclear

Models Studied

- **BERT**: transformer-based model, captures word context
- **MentalBERT**: fine-tuned for mental health text
- **ChatGPT**: generative general model; no direct stress classifier

Figure adapted from Wikipedia BERT (language model) page (Voigt Godoy,2021)



Key Risks & Considerations

- Detecting mental health conditions is safety-critical
- LLMs, trained on massive data, risk bias and opacity
- Evaluation of accuracy and fairness is vital

Research Questions

1. Can LLMs reliably detect stress in mental health?
2. How do simpler, interpretable models compare with LLMs?

Approach

__Evaluated:__
ChatGPT-4o, BERT, and MentalBERT

__Baseline:__
Random Forest is a lightweight ensemble of decision trees

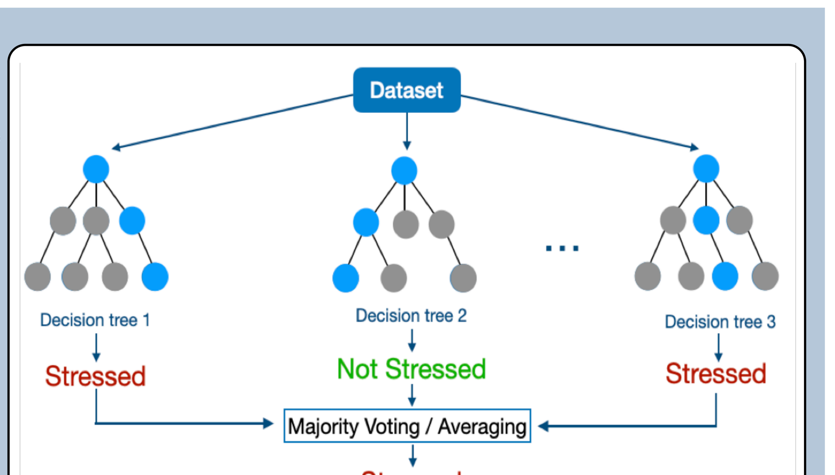


Figure created by finalist in PowerPoint, 2025.

Methodology

Data Collection

Dataset: Dreaddit contains **190K** Reddit posts across 5 domains: abuse, social, anxiety, PTSD, financial

- **3.5K** human-annotated segments (Amazon Mechanical Turk)
- Split into training, validation, and test sets

Dreaddit	Columns
Metadata	id, subreddit, post_id, social_timestamp, social_upvote
Post	text, label (stress or not), confidence (% of votes)
Syntax	Flesch-Kincaid grade, readability index
LIWC features	lex_liwc_*, **90+ total**
DAL features	lex_dal_*, **9 total**

Table made by finalist in Excel, 2025.

Dreaddit adopts **LIWC** (Linguistic Inquiry and Word Count) and **DAL** (Dictionary of Affect in Language) features as **lookup-tables**

- Pronoun use (*"I," "we"*), social words
- Tone, clout, positive/negative emotion, anxiety terms
- Sentence length, syntactic complexity, readability
- Affective ratings: pleasantness, activation, imagery

Baseline Model: Random Forest

- Uses only structured linguistic features for detecting stress
- Interpretable, lightweight, and efficient
- Benchmarks whether costly LLM training yields real gains

Model Evaluation

BERT & MentalBERT: evaluated with Python's scikit-learn
ChatGPT-4o: tested in **zero-shot** setting

- Designed prompts for binary stress classification
- No fine-tuning, each post processed independently
- Prevents context bias; responses rely only on the input text

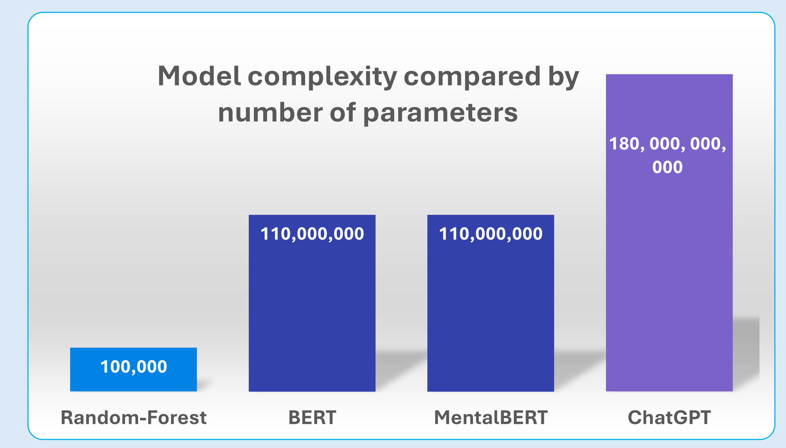
Example Prompt (Zero-Shot Classification)

"Long story short my family in NE Ohio is really abusive so I had to leave the state and stay with family down south. It isn't working out and they're sending me packing to Ohio because I guess I'm a financial problem even though I got a job here. I have nowhere I can stay. I'm even getting rid of my beloved cat so I can have options. I can't go back to my family in Ohio."

Consider the above post to answer the question: Is the poster likely to suffer from very severe stress?

Only return Yes or No. Give me your confidence in the answer on a scale from 0 to 1.

Random-Forest **1000x** less than BERT/MentalBERT which is **1000x** less than ChatGPT



__Graph is in log-scale__

Graph created by finalist in Excel, 2025.

Results & Analysis

To evaluate models, I use **Precision, Recall,** and **F1-score**

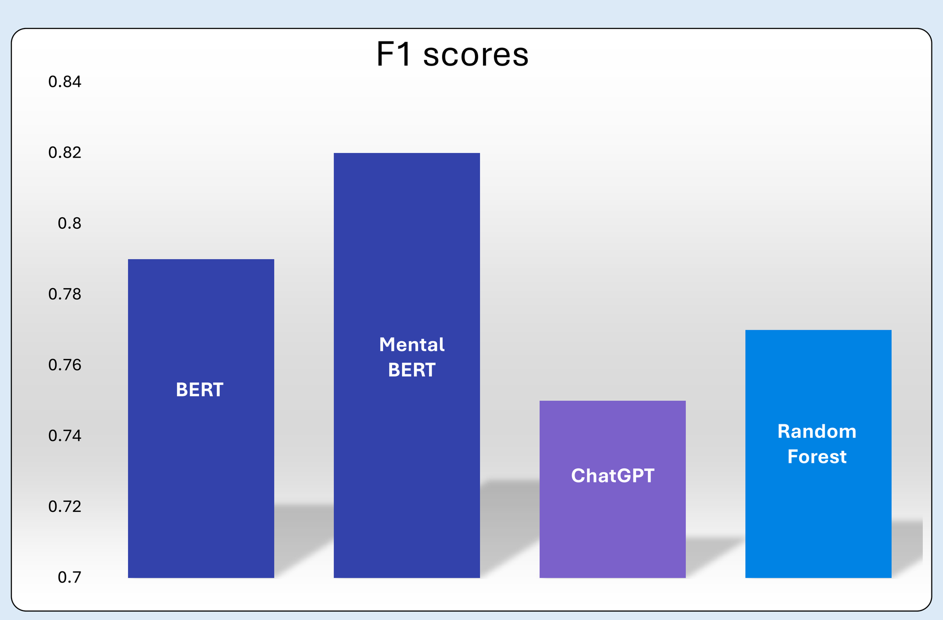
True Positives (TP): Correctly identified stress-related posts
False Positives (FP): Posts incorrectly classified as stress
False Negatives (FN): Stress posts missed by the model

Precision = TP / (TP + FP) % of predicted stress cases correct
Recall = TP / (TP + FN) % of actual stress cases detected

F1-score = 2 × (Precision × Recall) / (Precision + Recall)
Balances precision and recall as their harmonic mean

	Precision	Recall	F1-Score
BERT	0.77	0.81	0.79
MentalBERT	0.79	0.85	0.82
Chat-GPT	0.71	0.81	0.75
Random Forest	0.74	0.80	0.77

Table created by finalist in Mac Numbers, 2025.



Graph created by finalist in Excel, 2025.

MentalBERT achieved the highest **F1-score (0.82)**, demonstrating the advantage of domain-specific fine-tuning for stress detection

For **ChatGPT**, I tested on 714 social media posts: TP=294, FP=118, FN=69, and **F1-score=0.75**

For comparison, a **flip-the-coin strategy** would get Precision=0.5, Recall=0.5, and **F1=0.5**

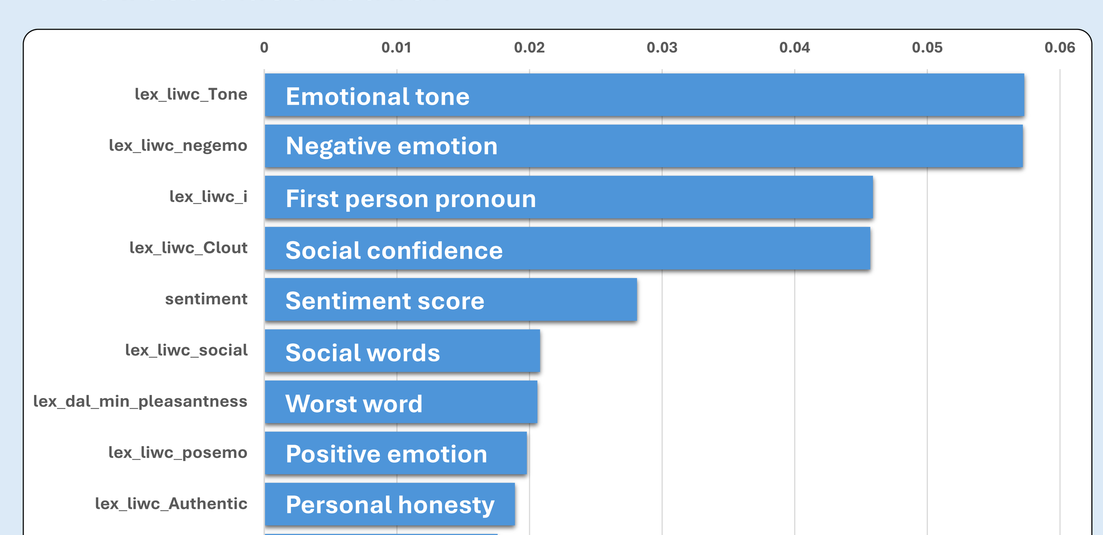
Despite its simplicity, **Random Forest** (using LIWC & DAL features) was **competitive with BERT** with **F1-score = 0.77**

- Outperformed ChatGPT-4o on stress detection
- Efficient, interpretable, practical

Top Features Driving Classification

Random Forest is **interpretable**:

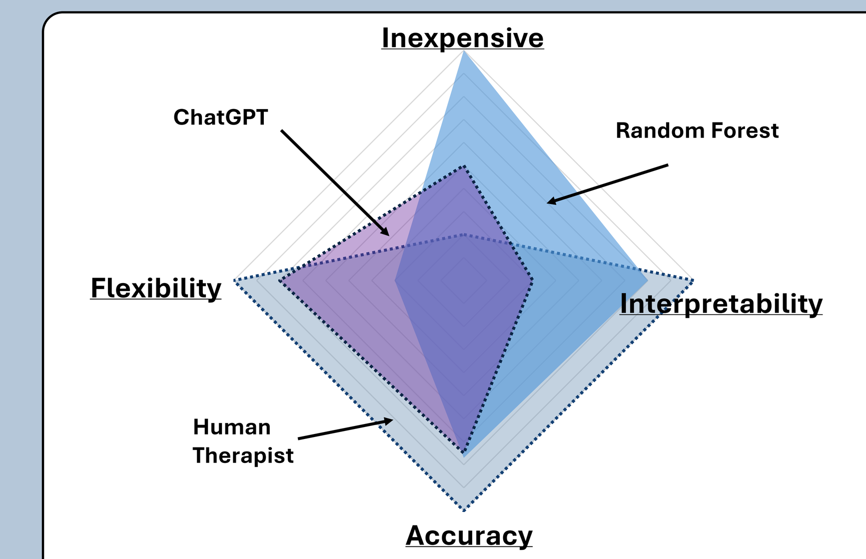
- Examined the weights it assigns to features
- Top-ranked linguistic and lexical cues show what drives a stress classification



Graph created by finalist in Excel, 2025.

Conclusions & Implications

- **LLMs are unreliable:** low precision & recall make them risky for stress detection in sensitive settings
- **Current LLMs cannot substitute for human therapists**
- **High cost, small gain:** LLMs require vast training and computation but offer only marginal improvement
- **Simple could be better:** Random Forest, with basic linguistic features, performed comparably to BERT and outperformed ChatGPT. **It is important to consider simpler, well-tuned models before defaulting to costly, heavyweight LLMs**



Graph created by finalist in Excel, 2025.

Stanford study (2025, June) also found LLMs unreliable:

- Moore et al. tested LLMs in simulated therapy settings
- Models showed stigma toward mental health conditions, often responded inappropriately to suicidal ideation or delusions
- ChatGPT appropriateness: **60–80%** vs. **93%** for human therapists

Future Work

Lightweight AI for triage	• Develop simple, interpretable models for early stress detection
Cross evaluation	• Test LLMs across demographics to uncover gender and minority bias
Improving reliability	• Boost LLM accuracy in safety-critical tasks • Transfer learning from mental health datasets • Study how state anxiety affects LLM performance
Safe & responsible AI:	• Create ethical guidelines for privacy and safety • Keep human oversight central, AI should augment, not replace, clinicians!

Key References

1. **Ben-Zion**, Z., Witte, K., Jagadish, A. K., Duek, O., Harpaz-Rotem, I., Khorsandian, M. C., ... & Spiller, T. R. (2025). Assessing and alleviating state anxiety in large language models. Digital Medicine, 8(1), 132.
2. **Devlin**, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of 2019 conference of the association for computational linguistics: human language technologies, vol 1 (pp. 4171-4186).
3. **Ji,** S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2021). Mentalbert: Publicly available pretrained language models for mental healthcare. arXiv preprint arXiv:2110.15621.
4. **Moore**, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025, June). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 599-627).
5. **Turcan**, E., & McKeown, K. (2019). Dreaddit: A reddit dataset for stress analysis in social media. arXiv preprint arXiv:1911.00133.
6. **Voigt Godoy,** D. (2021, June 6). BERT embeddings 01.png [PNG image]. Wikimedia Commons. https://commons.wikimedia.org/wiki/File:BERT_embeddings_01.png Wikimedia Commons
7. **Yang**, K., Ji, S., Zhang, T., Xie, Q., Kuang, Z., & Ananiadou, S. (2023). Towards interpretable mental health analysis with large language models. arXiv preprint arXiv:2304.03347.